УДК 004.4
ББК 32.973.26-018.1
Х17

А.А. Халафян
STATISTICA 6. Статистический анализ данных. 3-е изд. Учебник — М.: ООО «Бином-Пресс», 2007 г. — 512 с.: ил.

В книге освещены основные разделы шестой версии программы STATISTICA. На примерах, большинство которых — из встроенной в программу библиотеки Examples, дано подробное описание технологии работы с модулями программы. Уделено внимание постановочной части статистических методов и интерпретации результатов анализа. Рассмотрены процедуры управления данными, графические возможности программы, разведочный анализ данных, одномерные и многомерные статистические методы, углубленные методы анализа, временные ряды и прогнозирование, формирование отчетов, составление макросов.

Книга написана на основе курсов, читаемых в Кубанском государственном университете. Простая и доступная для широкого круга читателей форма изложения делает возможным самостоятельное изучение программы STATISTICA.

Адресована студентам, изучающим экономические и математические дисциплины, а также аспирантам, преподавателям вузов, научным работникам различных направлений, специалистам в области аналитики и логистики, т.е. будет полезна всем, занимающимся обработкой статистических данных и использующих современные компьютерные технологии.

ISBN 978-5-9518-0215-6

Содержание

Лучше быть приблизительно правым,
чем абсолютно точно ложным.
Джон Мэйнард Кейнс

Предисловие

Статистика возникла в XVI в. в Италии и ограничивалась сбором данных о состоянии государства [1]. Сегодня трудно назвать область человеческих знаний, которая обходилась бы без сбора и анализа данных.

Статистика проникла практически во все сферы человеческой деятельности — технику, экономику, экологию, медицину, естественные науки, военное дело, социологию, политологию и т.д. Это наука, которая позволяет увидеть закономерности в хаосе случайных данных, выделить устойчивые связи в них, определить наши действия с тем, чтобы увеличить долю правильно принятых решений среди всех принимаемых нами [2].

Математическая статистика разрабатывает математический аппарат установления статистических закономерностей и получения научно обоснованных выводов о массовых явлениях из данных наблюдений или экспериментов.

Многие методы статистического анализа вышли за рамки классической математической статистики (например, кластерный анализ, многомерное шкалирование, моделирование структурными уравнениями). Поэтому вполне обосновано в [3] предлагается отличать прикладную статистику от математической статистики.

Прикладная статистика определяется как самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов.

Под методами математической статистики предлагается понимать те методы статистической обработки исходных данных, разработка и использование которых апеллируют к вероятностной природе этих данных. Эти методы предусматривают возможность вероятностной интерпретации обрабатываемых данных и полученных в результате обработки статистических выводов.

Условно выделены три центральные проблемы прикладной статистики [3]:

- статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными;
- разработка статистических методов классификации объектов и признаков;
- снижение размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных.

Вычислительные процедуры прикладного статистического анализа являются достаточно трудоемкими при их реализации. Особенно актуальной проблема трудоемкости становится при многомерном анализе данных. Поэтому эффективная, грамотная, всесторонняя статистическая обработка данных даже небольшого объема практически невозможна без использования вычислительной техники. В настоящее время новый импульс развития и использования получили компьютерные технологии обработки и анализа данных. Разработка собственных компьютерных программ не всегда целесообразна, так как большой сегмент рынка прикладных программ занимают пакеты по статистической обработке данных. Это профессиональные пакеты (*SAS, BMDP*), универсальные пакеты (*STADIA, OLIMP, STATGRAPHICS, SPSS, STATISTICA*, ...), специализированные (*BIOSTAT, MESOSAUR, DATASCOPE*, ...) [4]. Благодаря деятельности корпорации Софтлайн, являющейся дилером компании производителя *StatSoft* (*USA*), одним из наиболее известных в России пакетов для прикладного статистического анализа данных является пакет *STATISTICA*.

STATISTICA — это универсальная интегрированная система, предназначенная для статистического анализа и обработки данных [2]. Содержит многофункциональную систему для работы с данными, широкий набор статистических модулей, в которых собраны группы логически связанных между собой статистических процедур, специальный инструментарий для подготовки отчетов, мощную графическую систему для визуализации данных, систему обмена данными с другими Windows-приложениями.

С помощью реализованных в системе *STATISTICA* языков программирования (*SQL, STATISTICA BASIC*), снабженных специальными средствами поддержки, легко создаются законченные пользовательские решения и встраиваются в различные другие приложения или вычислительные среды.

Система *STATISTICA* производится фирмой *StatSoft Inc.* (США), основанной в 1984 г. в городе Тулса (США). Первые программные продукты фирмы (*PsyhoStat-2,3*) были предназначены для обработки социологических данных. В 1985 г. *StatSoft* выпускает первую систему статистического анализа для компьютеров *Apple Macintosh* (*StatFast*) и статистический пакет для *IBM PC* (*STATS+*). В 1986 г. начинается работа по созданию интегрированных статистических пакетов комплексной обработки данных.

В 1991 г. выходит первая версия системы *STATISTICA/DOS*. Эта программа представляла собой новое направление развития статистического программного обеспечения, так как в ней реализован графически ориентированный подход к анализу данных. Программа обладала рядом существенных преимуществ перед другими статистическими программами. Например, за счет оптимизации удалось добиться повышения скорости обработки более чем в 10 раз, программа могла анализировать фактически неограниченный объем данных. В 1992 г. вышла версия *STATISTICA* для *Macintosh*. В 1994 г. выходит версия *STATISTICA 4.5* для *Windows*, которая сразу же занимает лидирующее положение среди статистических пакетов. В результате сравнительного тестирования с профессиональными системами *BMDP 1.0, SPSS 6.1, STATGRAPHICS 1.0, SYSTAT 5.01* она получает первое место в некоторых ведущих научных и компьютерных изданиях.

В конце 1995 г. вышла версия *STATISTICA 5.0* с более удобным пользовательским интерфейсом и полной совместимостью с *Windows 95*. В этой версии реализованы новые мощные возможности численного и графического анализа данных. *STATISTICA 5.0* полностью удовлетворяет основным стандартам среды *Windows*. Это стандарты пользовательского интерфейса (*MDI*); использование технологий *DDE* (динамического обмена данными из других приложений); *OLE* (связывания и внедрения объектов, поддержка основных операций с буфером обмена) и др. В отличие от предыдущих версий в *STATISTICA 5.0* включен внутренний язык программирования *STATISTICA BASIC*, который позволит пользователю расширять возможности системы. Например, нарастить систему по своему усмотрению, добавив собственную панель инструментов с тем или иным методом статистического анализа.

В 1996–1998 гг. появились новые выпуски программы – *STATISTICA* 5.1, 5.1–97 и 5.1–98, в которые были добавлены новые специализированные модули, учтены все новые форматы *Windows и MS Office*, дополнены и улучшены существующие процедуры.

Система *STATISTICA* имеет более полумиллиона зарегистрированных пользователей во всем мире. Пользователями системы являются крупнейшие университеты, исследовательские центры, компании, банки всего мира, государственные учреждения. Имеются версии системы на немецком, французском, японском, испанском, польском и других языках. В 1999 г. состоялся выпуск русской версии *STATISTICA 5.1*. Корпорацией Софтлайн во главе с В.П. Боровиковым издано большое число книг с подробным описанием системы *STATISTICA 5.0*. Технология работы с основными процедурами проиллюстрирована на большом количестве примеров. На сайте *www.statsoft.ru* можно найти всю необходимую информацию о пакете *STATISTICA*.

Появление операционной системы *Windows XP* привело к необходимости структурных изменений программы *STATISTICA* и созданию новой версии *STATISTICA 6.0*. В ней существенно изменены структуры интерфейса, диалоговых окон. Некоторые модули исключены, так как в новой версии за счет расширения возможностей они потеряли актуальность.

Предлагаемый учебник посвящен описанию новой версии пакета – STATISTICA *6.0*. При рассмотрении примеров в основном использованы легко доступные пользователям файлы данных из встроенной в программу *STATISTICA* библиотеки *Examples*. Учебник написан по материалам лекционных курсов и семинарских занятий, проводимых на факультете прикладной математики Кубанского государственного университета по дисциплинам: теория вероятностей и математическая статистика (ЕН.Ф.00) – для специальности 080801 Прикладная информатика (в экономике); 010501 Прикладная математика и информатика; математическая статистика (ЕН.Ф.00); эконометрика (ОПД.Ф.00) многомерные статистические методы (ОПД.Р.00) – для специальности 080116 Математические методы в экономике; методы социально-экономического прогнозирования (ДС) для специальности 080801 Прикладная информатика (в экономике).

В 3-е издание книги вошли новые разделы с описанием модулей: анализ главных компонент и классификация, деревья классификации, анализ соответствий, многомерное шкалирование, моделирование структурными уравнениями, анализ выживаемости, общие модели дискриминантного анализа. Рассмотрены способы создания макросов.

Так как в книге не излагаются основы теории вероятностей и математической статистики, читателям желательно иметь определенную математическую подготовку и опыт работы на компьютере. При этом нет необходимости знать сложные методы в деталях. С помощью программы *STATISTICA* можно научиться использовать их мощные возможности для анализа и интерпретации данных [2].

Учебник предназначен для самого широкого круга читателей: студентов, изучающих экономические и математические дисциплины, аспирантов, преподавателей вузов, научных работников различных направлений, специалистов в области аналитики и логистики. Будет полезна всем, занимающимся обработкой статистических данных и использующих современные компьютерные технологии.

Автор благодарит за экспертизу рукописи *Учебно-методическое объединение по образованию в области статистки и антикризисного управления при Московском государственном университете экономики, статистики и информатики, Научно методический совет по математике Министерства образования и науки РФ*; также рецензентов – доктора физико-математических наук, лауреата Государственной премии, профессора *О.Д. Пряхину*, доктора физико-математических наук, профессора *Е.А. Семенчина* за замечания, способствующие улучшению содержания 3-го издания книги; декана факультета прикладной математики *Ю.В. Кольцова* за создание условий, благоприятных для написания книги родных и близких за терпение и поддержку.

Опытом каждый называет свои ошибки.
Оскар Уайльд

Введение

Объектом исследования в прикладной статистике являются статистические данные, полученные в результате наблюдений или экспериментов. Статистические данные — это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих. Например, объекты исследования — страны мира и признаки, — географические и экономические показатели их характеризующие: континент; высота местности над уровнем моря; среднегодовая температура; место страны в списке по качеству жизни, доли ВВП на душу населения; расходы общества на здравоохранение, образование, армию; средняя продолжительность жизни; доля безработицы, безграмотных; индекс качества жизни и т.д.

Переменные — это величины, которые в результате измерения могут принимать различные значения.

Независимые переменные — это переменные, значения которых в процессе эксперимента можно изменять, а зависимые переменные — это переменные, значения которых можно только измерять.

Переменные могут быть измерены в различных шкалах. Различие шкал определяется их информативностью. Рассматривают следующие типы шкал, представленные в порядке возрастания их информативности: номинальная, порядковая, интервальная, шкала отношений, абсолютная. Эти шкалы отличаются друг от друга также и количеством допустимых математических действий. Самая «бедная» шкала — номинальная, так как не определена ни одна арифметическая операция, самая «богатая» — абсолютная.

Измерение в номинальной (классификационной) шкале означает определение принадлежности объекта (наблюдения) к тому или иному классу. Например: пол, род войск, профессия, континент и т.д. Часто номинальные переменные называют категориальными, или группирующими, так как они позволяют произвести разделение объектов исследования на подгруппы (подклассы). В этой шкале можно лишь посчитать количество объектов в классах — частоту и относительную частоту.

Измерение в порядковой (ранговой) шкале, помимо определения класса принадлежности, позволяет упорядочить наблюдения, сравнив их между собой в каком-то отношении. Однако эта шкала не определяет дистанцию между классами, а только то, какое из двух наблюдений предпочтительнее. Поэтому порядковые экспериментальные данные, даже если они изображены цифрами, нельзя рассматривать как числа и выполнять над ними арифметические операции [5]. В этой шкале дополнительно к подсчету частоты объекта можно вычислить ранг объекта. Примеры переменных, измеренных в порядковой шкале: бальные оценки учащихся, призовые места на соревнованиях, воинские звания, место страны в списке по качеству жизни и т.д.

При измерении в интервальной шкале упорядочивание наблюдений можно выполнить настолько точно, что известны расстояния между любыми двумя их них. Шкала интервалов единственна с точностью до линейных преобразований ($y = ax + b$). Это означает, что шкала имеет произвольную точку отсчета — условный нуль. Примеры переменных, измеренных в интервальной шкале: температура, время, высота местности над уровнем моря. Над переменными в данной шкале можно выполнять операцию определения расстояния между наблюдениями. Расстояния являются полноправными числами и над ними можно выполнять любые арифметические операции.

Шкала отношений похожа на интервальную шкалу, но она единственна с точностью до преобразования вида $y = ax$. Это означает, что шкала имеет фиксированную точку отсчета — абсолютный нуль, но произвольный масштаб измерения. Примеры переменных, измеренных в шкале отношений: длина, вес, сила тока, количество денег, расходы общества на здравоохранение, образование, армию, средняя продолжительность жизни и т.д. Измерения в этой шкале — полноправные числа и над ними можно выполнять любые арифметические действия.

Абсолютная шкала имеет и абсолютный ноль, и абсолютную единицу измерения (масштаб). Примером абсолютной шкалы является числовая прямая. Эта шкала безразмерна, поэтому измерения в ней могут быть использованы в качестве показателя степени или основания логарифма. Примеры измерений в абсолютной шкале: доля безработицы; доля безграмотных, индекс качества жизни и т.д.

Каждая измерительная шкала имеет соответствующую ей оценку среднего и разброса случайной величины. Так, например, в качестве оценки среднего для шкалы наименований целесообразно использовать моду — значение случайной величины, имеющее наибольшую частоту; для порядковой шкалы целесообразно использовать медиану — значение случайной величины, находящейся

в середине несгруппированного вариационного ряда; для более сильных шкал — среднее арифметическое.

Вообще говоря, конечная цель всякого исследования или научного анализа состоит в нахождении связей (зависимостей) между переменными. Философия науки учит, что не существует иного способа представления знания, кроме как в терминах зависимостей между количествами или качествами, выраженными какими-либо переменными. Таким образом, развитие науки всегда заключается в нахождении новых связей между переменными [6].

Одномерный статистический анализ совокупности данных, состоящих из наблюдений и характеризующих их переменных, заключается в рассмотрении каждой отдельной переменной и исследовании их попарной взаимосвязи. Естественно, такой подход весьма ограничен, так как закономерности и взаимосвязи, присущие всей совокупности, не возможно выявить, исследуя каждую переменную в отдельности. Поэтому наиболее интересным, с точки зрения прикладных исследований, разделом математической статистики является многомерный статистический анализ данных.

Многомерный статистический анализ — это раздел математической статистики, посвященный математическим методам построения оптимальных планов сбора, систематизации и обработки многомерных статистических данных, направленных на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака и предназначенных для получения научных и практических выводов. Под многомерным признаком понимается *p-мерный* вектор $X = (x_1, x_2, ..., x_p)$ показателей (признаков, переменных) $x_1, x_2, ..., x_p$, среди которых могут быть количественные, т.е. скалярно измеряющие в определенной шкале степень проявления изучаемого свойства объекта; порядковые (или ординальные), т.е. позволяющие упорядочить анализируемые объекты по степени проявления в них изучаемого свойства; и классификационные (или номинальные), т.е. позволяющие разбивать исследуемую совокупность объектов на не подающиеся упорядочиванию однородные (по анализируемым свойствам) классы [7].

Многомерный статистический анализ дает возможность получить общие выводы относительно всей совокупности данных [8]. Учитывая также и то, что анализируемые данные являются стохастическими, т.е. ограниченными и неполными, использование методов многомерного анализа является не только оправданным, но и существенно необходимым [9].

Изложению теоретических основ статистического анализа данных и разработанного математического аппарата посвящено большое количество прекрасных изданий. В то же время при пакетной реализации статистических методов особую актуальность приобрели их прикладные аспекты. Поэтому в учебнике уделено внимание не описанию математических методов статистического анализа, а рассмотрению постановочной части решаемых задач и интерпретации результатов статистического исследования — таблиц, графиков, сообщений пакета *STATISTICA*.

В гл. 1 изложены основные принципы работы с данными. Рассмотрены основные способы ввода данных в электронную таблицу *STATISTICA*, подготовленных в каком-либо другом приложении. Гл. 2 посвящена описанию основных принципов создания отчетов. *Отчет* — это тип документов *STATISTICA*, куда может быть выведена любая графическая и текстовая информация в формате *RTF* (*Rich Text Format — расширенный текстовый формат*).

В пакете *STATISTICA* представлено большое количество различных графических представлений данных. Это различные типы линейных графиков, диаграмм рассеяния, диаграмм размаха, круговые диаграммы частот и т.д. В гл. 3 показаны некоторые возможности пакета для построения двух- и трехмерных гафиков.

Вычислению описательных статистик, корреляционным матрицам, процедурам *t-критерия* сравнения средних, однофакторному дисперсионному анализу посвящена гл. 4. В гл. 5 содержится описание различных способов частотного анализа, который позволяет выявить взаимосвязь, установить характер этой взаимосвязи для двух переменных, измеренных в номинальной или порядковой шкале.

Большинство рассмотренных в книге модулей относится к методам параметрической статистики, в основе которых лежит предположение, что случайный вектор переменных образует некоторое многомерное распределение, как правило, нормальное или преобразуется к нормальному распределению. Если это предположение не находит подтверждения, следует воспользоваться непараметрическими методами математической статистики. В гл. 6 включены некоторые методы непараметрической статистики — сравнение средних, корреляционный анализ.

Возможности вероятностного калькулятора, способы генерации случайных чисел, проверка соответствия законов распределения известным законам, примеры решения задач по терии вероятностей и математической статистике изложены в гл. 7.

Только математическими методами можно установить тесноту и характер взаимосвязей различных переменных и степень их воздействия на интересующий исследователя результат. В таких исследованиях широко используются процедуры множественной регрессии. Регрессионный анализ тесно связан с другими статистическими методами — методами множественного корреляционного и дисперсионного анализа. В отличие от корреляционного анализа, исследующего направление и силу статистической связи переменных, регрессионный исследует вид зависимости переменных, т.е. математические модели зависимости количественной или качественной переменной от одной или нескольких других переменных. В дисперсионном анализе исследуется зависимость количественной переменной от одной или нескольких качественных переменных. В гл. 8–10 рассмотрены методы многомерного регрессионного и дисперсионного анализа. Описаны основные процедуры модулей «Дисперсионный анализ», «Множественная регрессия», «Нелинейная регрессия», «Нелинейное оценивание».

Канонический анализ является обобщением множественного корреляционного анализа как меры взаимосвязи одной переменной с множеством других переменных. Канонический анализ необходим, если имеются две совокупности

переменных и необходимо определить взаимосвязь между ними. В гл. 11 рассмотрен модуль «Канонический анализ».

В методах классификационного анализа с обучением и без обучения исследуется взаимосвязь между одной качественной переменной и совокупностью количественных переменных. Дискриминантный анализ и деревья классификации — это методы, позволяющие предсказывать принадлежность объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких независимых переменных. Кластерный анализ позволяет произвести разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. В гл. 12,13 рассмотрены методы классификационного анализа. Описаны модули «Дискриминантный анализ», «Общие модели дискриминантного анализа», «Кластерный анализ», «Деревья классификации».

Главными целями методов факторного анализа и анализа главных компонент и классификации являются сокращение числа переменных и определение структуры взаимосвязей между ними. Сокращение достигается посредством выделения скрытых общих факторов, объясняющих связи между наблюдаемыми признаками объекта, т.е. вместо исходного набора переменных анализируются данные по выделенным факторам, число которых значительно меньше исходного числа признаков. В гл. 14 описаны модули «Факторный анализ», «Анализ главных компонент и классификация».

Многомерное шкалирование можно рассматривать в качестве альтернативы факторного анализа. Основное предположение многомерного шкалирования заключается в том, что есть некоторое метрическое пространство существенных базовых характеристик и объекты можно представить как точки в этом пространстве. Предполагают, что более близким (по исходной матрице) объектам соответствуют меньшие расстояния в пространстве базовых характеристик. Следовательно, многомерное шкалирование — это совокупность методов, с помощью которых определяется размерность пространства базовых характеристик объектов и конструируется конфигурация объектов в этом пространстве. Это пространство (многомерная шкала) аналогично обычно используемым шкалам в том смысле, что значениям базовых характеристик объектов соответствуют определенные значения на осях пространства.

Анализ соответствий содержит описательные и разведочные методы анализа двухвходовых и многовходовых таблиц. Эти методы по своей природе похожи на методы факторного анализа и позволяют исследовать структуру группирующих переменных, включенных в таблицу частот сопряженности. Одна из целей анализа соответствий — представление содержимого таблицы относительных частот в виде расстояний между отдельными строками и/или столбцами таблицы в пространстве возможно более низкой размерности. Гл. 15 посвящена описанию модулей «Многомерное шкалирование» и «Анализ соответствий».

Объектом моделирования структурных уравнений являются сложные системы, внутренняя структура которых не известна. Исследуя параметры системы при помощи методов причинного моделирования, можно изучить ее структуру,

установить причинно-следственные взаимосвязи между элементами системы. В гл. 16 рассмотрены основные идеи причинного моделирования и описан модуль «Моделирование структурными уравнениями».

Информация, когда нет данных о наступлении интересующего нас события, называется неполной. Если есть данные о наступлении интересующего нас события, то информация называется полной. Наблюдения, которые содержат неполную информацию, называются цензурированными наблюдениями. Цензурированные наблюдения типичны, когда наблюдаемая величина представляет время до наступления некоторого критического события, а продолжительность наблюдения ограничена по времени. Использование цензурированных наблюдений составляет специфику статистического метода – анализа выживаемости, в котором исследуются вероятностные характеристики интервалов времени между последовательным возникновением критических событий. Такого рода исследования называются *анализ длительностей до момента прекращения*. Их можно определить как интервалы времени между началом наблюдения за объектом и моментом прекращения, при котором объект перестает отвечать заданным для наблюдения свойствам. Цель исследований – определение условных вероятностей, связанных с длительностями до момента прекращения. В гл. 17 рассмотрен метод анализа выживаемости и описан модуль «Анализ выживаемости».

Методы прогнозирования временных рядов являются важным инструментом в процессе принятия решений. Такие прогнозы можно применять при принятии тактических и стратегических решений. Прогнозировать можно при помощи регрессионных моделей, описанных в гл. 9, 10. Такие приемы приемлемы при рассмотрении причинно-следственной зависимости между переменными 1. Однако существуют и альтернативные методы прогнозирования, которые используют приемы анализа временных рядов. В гл. 18 описаны основные процедуры модуля «Временные ряды и прогнозирование».

Очень часто при статистической обработке однотипных наборов данных приходится периодически многократно выполнять одну и ту же серию операций. Создав макрос, можно автоматизировать статистический анализ данных и соответственно избавить пользователя от трудоемкой и зачастую рутинной работы. Особенно использование макросов актуально при реализации модулей многомерного статистического анализа. В гл. 19 излагаются основные приемы создания макросов.

Замечательно, что науке, начинавшейся с рассмотрения
азартных игр, суждено было стать важнейшим
объектом человеческого знания...
Пьер Симон Лаплас

Глава 1
Работа с данными

1.1. Инструменты для работы с данными

Данные в *STATISTICA* организованы в виде электронной таблицы. Таблица с исходными данными (таблицы хранятся в файлах с расширением **.sta*) является одним из типов документа в системе *STATISTICA* (другие типы документов — электронная таблица с результатами анализа, график, отчет). Каждый тип документа выводится в своем окне в рабочей области системы. Как только это окно становится активным, изменяется панель инструментов и меню. В нем появляются команды, доступные для этого типа документов.

Для работы с электронными таблицами исходных данных существует большое количество операций, которые доступны при помощи выпадающих и контекстных меню и из панели инструментов. Перечислим основные из них [2].

Операции, которые изменяют структуру электронной таблицы. Это операции добавления, удаления, копирования и перемещения переменных и случаев из электронной таблицы.

Операции по заданию спецификаций (имен, форматов и т.д.) для переменных и случаев.

Большое количество операций с выделенным блоком значений. Эти операции не меняют структуру файла, а изменяют только значения данных в таблице. Они включают стандартные операции с Буфером обмена, например, операции вырезать, копировать, вставить, очистить и др. Часть операций с блоком значений ориентирована на специфику статистической обработки, например: транспонирование

блока, заполнение случайными значениями, стандартизация значений в блоках, вычисление основных статистических характеристик блока значений, визуализация значений, блока значений и др.

Операции, реализованные при помощи метода **Drag and Drop** (перетащить — отпустить), включая операции по копированию, перемещению и автозаполнению блока и др.

Операции перекодировки и ранжирования переменных.

Программа поддерживает большое количество методов обмена с данными из других приложений. При этом реализованы способы ввода данных с использованием:

- буфера обмена;
- механизма динамического обмена данными *DDE* — динамического обмена данными *Windows*;
- средств импорта данных, которые позволяют импортировать данные практически из любой базы данных.

Язык программирования *STATISTICA Visual Basic* позволяет создавать дополнительные приложения, реализующие как простые преобразования данных, так и сложные вычислительные процедуры.

В *STATISTICA* можно записать макросы, которые автоматизируют повторяющиеся шаги или используются для автоматического создания программ.

В *STATISTICA* используется подмножество языка программирования *SQL* (язык запросов к базам данных) для задания критериев импорта записей из баз данных.

1.2. Структура электронной таблицы

Исходные данные организованы в виде таблицы. Электронная таблица состоит из строк и столбцов. В отличие от обычных электронных таблиц, в которых строки и столбцы равноправны, в *STATISTICA* они имеют разные смысловые значения. При этом столбцы таблицы называются *Variables* (переменные), а строки — *Cases* (случаи, наблюдения).

Каждая переменная имеет свое имя, формат и другие атрибуты (которые называются спецификацией переменной), задаваемые пользователем. Переменная представляет собой наблюдаемую величину. Результаты наблюдений записываются в строках таблицы — наблюдениях. Нулевой столбец, в котором по умолчанию указаны номера наблюдений, при необходимости может содержать имена случаев. Ими могут быть либо даты наблюдений, либо какие-то другие имена, обычно естественно возникающие в конкретной задаче, например, имена опрашиваемых, при сборе данных социологического исследования, поэтому в качестве имен случаев *STATISTICA* позволяет использовать либо число, либо текстовое значение или значение даты. Электронная таблица с исходными данными в *STATISTICA* называется *Spreadsheet*. Электронные таблицы с исходными данными хранятся в файлах с расширением **.sta*. В дополнение к значениям переменных *STATISTICA* может хранить в файле с исходными

данными и дополнительную информацию как об индивидуальных переменных, так и обо всей таблице в целом. В электронной таблице *Spreadsheet* пользователь может задать спецификации переменных:

- формат отображения (например, число десятичных знаков или формат значений даты или времени);
- определенные значения, которые нужно пропускать при расчетах (т.е. коды пропущенных данных);
- длинные имена переменных и комментарии;
- длинные метки и комментарии для отдельных значений (см. ранее);
- формулы, которые можно использовать для задания, перекодирования или преобразования каждой переменной;
- динамические связи между файлом данных *STATISTICA* и другим *Windows*-совместимым файлом с использованием механизма *DDE*.

Настройки внешнего вида файла данных (высота и ширина столбцов, цвета и шрифт) также хранятся вместе с данными и могут быть использованы для упрощения идентификации отдельных файлов или наборов данных из разных проектов. Окно спецификаций переменной можно вызывать двойным щелчком на имени переменной в таблице исходных данных.

Для удобной работы с переменными, принимающими текстовые значения, реализован так называемый механизм двойной записи. Согласно этому соглашению каждому текстовому значению переменной ставится в соответствие некоторое число. Таким образом, устанавливается соответствие вида *число = текстовое значение*. Оно может быть установлено автоматически (самой системой при вводе данных) или определено пользователем. При работе с данными всегда можно переключиться с текстовой на числовую форму просмотра исходных данных. Наличие описанного механизма двойной записи позволяет удобно вводить текстовые значения, выполнять необходимые преобразования и, кроме того, любой статистический анализ над текстовыми переменными так, как если бы они принимали числовые значения.

Поясним принцип двойной записи на основе данных из таблицы на рис. 1.1. В нем приведены некоторые экономические показатели 10 крупнейших стран мира по численности городского населения: общее число жителей (млн чел.) на 1990, 1995, 2000 гг.; доля (%) городского населения на 1995 г.; наличие крупных запасов нефти и газа (более 1,5 млрд т); структура ВВП (%) в промышленности, сельском хозяйстве, сфере услуг. Перечисленным показателям в файле данных соответственно присвоены имена: *Нас.90, Нас.95, Нас.00, Нас. гор., Нефть, Газ, Пром., С/х, Услуги*. Две переменные *Нефть* и *Газ* содержат текстовые значения. Щелкните 2 раза левой кнопкой мыши на имени переменной, например, *Газ*. Предположим, в окне спецификаций переменных, нажав на кнопку **Text Labels** (текстовые ярлыки) и открыв окно **Text Labels Editor** (редактор текста ярлыков), сделаны следующие присвоения: *1 = есть, 0 = нет*. Тогда для переключения отображения с числовых значений на текстовые и наоборот надо нажать на панели инструментов на кнопку **Show/Hide Text Labels**.

	Крупнейшие страны мира по численности населения								
	1 Нас.00	2 Нас.95	3 Нас.00	4 Нас.гор.	5 Нефть	6 Газ	7 Пром.	8 С/х	9 Услуги
Китай	1120	1121	1275	30,3	есть	есть	48	21	31
Индия	830	935	1010	26,8	нет	нет	30	29	41
США	250	263	250	76,2	есть	есть	26	2	72
Бразилия	150	162	170	78,2	нет	нет	37	14	49
Россия	289	149	146	73	есть	есть	38	7	55
Япония	124	125	126	77,6	нет	нет	38	2	60
ФРГ	80	82	82	86,5	нет	нет	38	2	60
Индонезия	180	198	215	35,4	есть	есть	42	17	41
Великобритания	57	57	69	89,5	нет	нет	32	2	66
Франция	56	58	59	72,8	нет	нет	27	2	71

Рис. 1.1

Наличие механизма двойной записи существенно упрощает работу с переменными, принимающими текстовые значения. Например, при вводе данных, вместо того чтобы последовательно вводить текстовые значения, можно сначала ввести числовые значения, а потом приписать им текстовые эквиваленты. Метки значений — это комментарии или описания (до 40 символов), которые можно присвоить определенным текстовым (числовым) значениям наборов данных. Каждое значение переменной может иметь присвоенную ему метку. Их можно отображать и изменять при помощи **Text Labels Editor** (редактора текста ярлыков), который также доступен через кнопку на панели инструментов или через меню **Data**. Рассмотрим вновь данные из примера. Мы можем приписать метки для значений *есть* — запасы газа более 1,5 млрд т и *нет* — запасы газа не более 1,5 млрд т. В меню **Data** выберите команду **Text Labels Editor**. Откроется соответствующее окно (рис. 1.2), из которого можно извлечь необходимую информацию или произвести в нем необходимые отображения и редактирование. Опишем кратко приемы редактирования:

- для перемещения между полями можно воспользоваться стрелками перемещения курсора;
- для редактирования содержимого какого-либо поля нужно дважды щелкнуть на нем;
- для удаления (вставки) новых строк можно использовать мышь: левее поля **Text Labels** щелкните на уровне соответствующей строки левой кнопкой мыши, далее, переместив на поле **Text Labels**, щелкните правой кнопкой мыши и произведите необходимые действия — **cut** (вырезать), **paste** (вставить).

Для того чтобы открыть это окно через панель инструментов, надо вывести на панель инструментов кнопку **Text Labels Editor**, которая имеет вид, идентичный кнопке **Show/Hide Text Labels**. Выберите пункт меню **View/Toolbars/Customize** (вид/панели инструментов/настройка). В открывшемся окне надо выбрать элемент

Data (данные) в списке **Categories** (категории), затем выделить **Text Labels Editor** в списке **Commands** (команды) и поместить кнопку на панель инструментов электронной таблицы. Еще более простой способ настроить панель инструментов для пользователя — щелкнуть правой кнопкой мыши на панели инструментов и вывести на нее дополнительные опции.

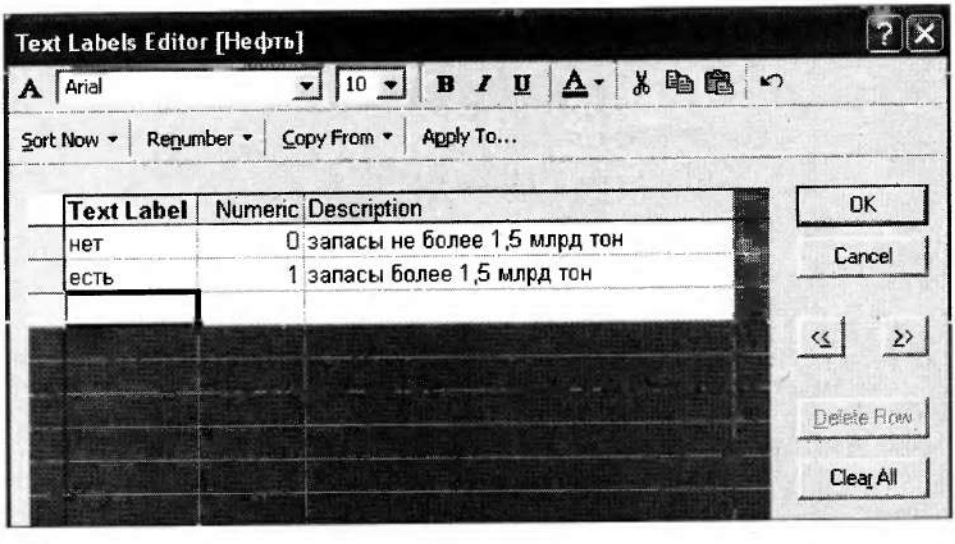


Рис. 1.2

Значения дат в *STATISTICA* хранятся в юлианском формате, как целые числа, представляющие число дней, прошедших с 1 января 1900 г. Например, дата, отображаемая как 1/21/1968, в юлианском формате представляет число 24858; при этом десятичные знаки интерпретируются как время. Хранящиеся таким образом значения дат можно использовать в любой процедуре анализа. В то же время в отчетах и на графиках можно отображать даты в общепринятом формате (например, для отметок на шкале). Юлианские значения дат в таблице исходных данных можно отображать как в числовом (юлианском) формате, так и в одном из заранее заданных форматов отображения дат. Чтобы изменить формат отображения даты, выберите тип *Date* (дата) в диалоговом окне текущих спецификаций или из выпадающего меню *Format Cells* (формат ячеек) и укажите один из предлагаемых в списке форматов отображения (рис. 1.3).

При вводе даты в новую переменную сначала необходимо изменить в диалоговом окне текущих спецификаций формат отображения переменной с типа *Number* (установленного по умолчанию) на тип *Date*, а затем выбрать нужный формат отображения. Данные можно вводить в любом из заданных форматов. Не обязательно в том формате, который выбран, можно просто вводить двухзначные числа через пробел, программа распознает форматы, преобразует в нужный и сохранит введенные значения.

При работе с реальными данными часто приходится иметь дело с ситуациями, когда часть данных не была по каким-либо причинам измерена.

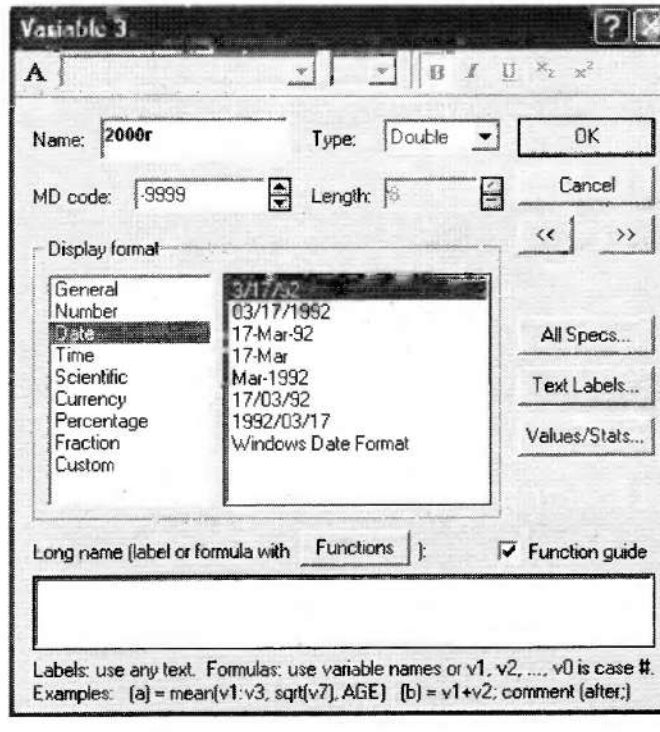


Рис. 1.3

В этом случае в соответствующую ячейку электронной таблицы не заносится никакое значение. Ячейка остается пустой. Однако при внутреннем хранении данных *STATISTICA* приписывает всем пустым ячейкам — пропущенным наблюдениям данных, некоторый специальный код *Missing Data Code* (код пропущенных данных). Код пропущенных значений устанавливается в спецификации переменной. Значение этого кода по умолчанию равно –9999. Пользователь всегда имеет возможность установить другое значение этого кода для каждой конкретной переменной. Способ, которым пропущенные данные обрабатываются при статистическом анализе, может корректироваться индивидуально для каждого вида анализа. Обычно он может быть установлен из стартовой панели конкретного статистического модуля. Пользователь имеет возможность устранить данные из вычислений, заменить их средним значением или интерполировать их. Возможны и другие способы обработки пропущенных наблюдений. Имеется возможность заменить в исходном файле данных все пропущенные значения переменной на среднее значение. Для этого в меню **Data** выберите команду **Replase Missing Data by Means** (замена пропущенных значений на среднее).

Создание нового файла с данными в системе *STATISTICA* может быть осуществлено при помощи меню **File** (файл) или из выпадающего меню на панели инструментов. Выберите команду **New** из меню **File**. В появившемся диалоговом окне (рис. 1.4) выделите вкладку **Spreadsheet** (таблица) и укажите *Namber of variables* (число переменных), и *Namber of cases* (число случаев). Нажмите **OK**.

Программа автоматически откроет пустую электронную таблицу **Spreadsheet** соответствующего размера. Переменные по умолчанию имеют имена *Var1, Var2, ...*, и заданное число пронумерованных случаев, которые не имеют имен. Для сохранения файла выберите команду **Save** (сохранить) из меню **File**. В появившемся диалоговом окне наберите имя файла. Нажмите **OK**. В заголовке окна с электронной таблицей автоматически отобразится имя файла с расширением *sta* и его размер. Если выделить вкладки **Report, Macro, Workbook,** то можно создать соответственно *отчет, рабочую книгу, макрос.*

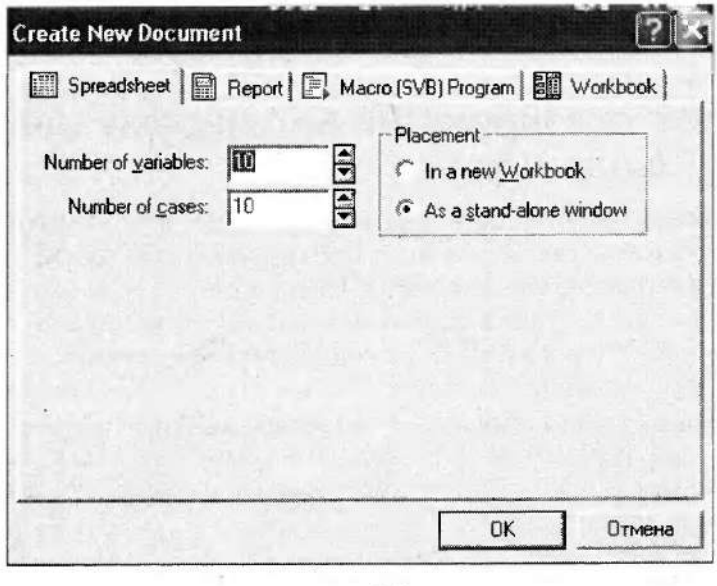


Рис. 1.4

Задать имена случаев можно, дважды щелкнув левой кнопкой мыши на поле имени (где указан номер случая). Для перехода от имени случаев к номерам можно воспользоваться кнопкой **Show/Hide Case Names.** Согласно стандартным соглашениям об электронных таблицах, для выделения всей таблицы исходных данных нужно щелкнуть на ее левом верхнем углу. Это может оказаться полезным, например, для копирования всего файла в буфер обмена.

Для ввода данных в таблицу надо установить указатель мыши на ячейку, в которую необходимо ввести данные. Щелкнуть мышью, для того чтобы сделать эту ячейку активной, и ввести необходимое значение с клавиатуры. Перейти к другой ячейке можно одним из следующих способов:

- нажать на клавишу **Enter**. После этого курсор переместится либо вправо, либо вверх, в зависимости от тех установок, которые определены в постоянных параметрах конфигурации системы;
- воспользоваться стрелками перемещения курсора;
- щелкнуть мышью на любой другой ячейке.

Текстовые значения можно непосредственно ввести в таблицу или ввести сна чала соответствующие им числовые значения.

Для исправления данных в ячейке надо установить указатель мыши в ячейк в которой необходимо исправить значение, щелкнуть мышью и ввести новое значе ние. Чтобы не удалять содержимое ячейки при вводе нового значения, нужно дваж щелкнуть на ней до начала ввода. При этом будет включен режим редактировани а курсор установлен внутри ячейки. Для написания заголовка к именам случаев (на пример, Страны мира) надо щелкнуть левой клавишей мыши два раза в верхнем ле вом поле таблицы и ввести соответствующее имя. Аналогичным способом можно за писать дополнительную информацию о файле (например, Крупнейшие страны ми по численности городского населения) в верхнем свободном поле таблицы (рис.1.1).

1.3. Основные операции над переменными и наблюдениями

Во всех командах, которые будут описаны далее, используется соглашение о том, что для задания имен переменных в диалоговых окнах можно дважды щелкнуть на поле ввода имени переменной и выбрать нужную переменную из появляющегося списка имен. Операции над переменными доступны либо через меню **Data** (данные), либо при помощи кнопки на панели инструментов **Vars**, либо через контекстное меню, щелкнув правой кнопкой мыши на имени переменной.

При помощи команды **Add** (добавить) можно добавить переменные (пустые столбцы) в электронную таблицу, при этом размер таблицы увеличивается. В диалоговом окне, которое появится после выбора этой команды, необходимо задать следующие параметры:

- *How many* (сколько). Позволяет задать число добавляемых переменных. Для электронной таблицы это число не ограничено (естественное ограничение — размер жесткого диска на вашем компьютере).
- *After* (после). Здесь необходимо задать имя переменной, после которой предполагается вставить новые переменные.
- *Name* (имя). Можно указать имена вставляемых переменных.

Команда **Move** (переместить) позволяет переместить переменные (как одну, так и несколько). При этом перемешаются непосредственно столбцы электронной таблицы. В диалоговом окне команды необходимо задать диапазон перемещаемых переменных и номер переменной, после которой необходимо их вставить.

Команда **Copy** (копировать) предназначена для копирования на указанное место столбцов с их содержимым. В диалоговом окне команды необходимо задать параметры: с какой переменной; по какую переменную; вставить после какой переменной. При этом вместе с переменными будут скопированы формат, длинное имя, формулы и т.д.

При помощи команды **Delete** (удалить) можно удалить столбцы. В диалоговом окне команды надо указать имена переменных начала и конца диапазона удаления.

Для преобразования данных в одной строке и перекодирования отдельных переменных можно воспользоваться формулами в таблице исходных данных. Двойной щелчок на имени преобразуемой переменной открывает диалоговое окно спецификаций переменной, в котором формулу преобразования или перекодировки можно ввести непосредственно в поле **Long name (Label or formula with Functions)** (длинное имя (метка или формула с функцией)).

По соглашениям об использовании формул в электронных таблицах *Windows* (например, *MS Excel*) формулы должны начинаться с символа «=». В противном случае программа не определит, что введенный текст является формулой. Переменные вызываются по именам или по номерам, например, *v1, v2,* ... Для выражений, содержащих условия преобразования, можно использовать логический оператор.

Чтобы пересчитать значения переменной согласно введенной формуле, надо нажать на **OK**. Откроется окно, в котором будет предложено подтвердить команду **Recalculate the variable now** (пересчитать переменную сейчас, если формула записана верно).

Команда **Date Operations** (действия с датами) позволяет провести ряд полезных операций над значением дат. Например, создать новую дату из двух или трех переменных, в которых хранятся значения дня, месяца или года. Либо разбить уже существующую дату на три переменные — день, месяц и год. Имеется возможность преобразования значения в формате даты в текстовое значение. Для создания даты из переменных со значениями дня, месяца и года надо, используя переключатель **Create Date from 2 or 3 Variables** (создать дату из 2 или 3 переменных), указать в рамке **Source Variables** (исходные переменные) имена переменных, из которых необходимо взять значения дня, месяца и года. В рамке **Destination Variable** (создаваемая переменная) задать имя переменной и формат даты. По аналогии с предыдущим, используя переключатель **Split Date into 1 to 3 Variables** (разделить дату на 2 или 3 переменные), можно разделить дату на 2 или 3 переменные.

Команда **Recalculate** (пересчет) предназначена для пересчета значений переменных, которые связаны при помощи формул. Имеются возможности установить опцию автоматического пересчета значений переменной при изменении данных в электронной таблице. Можно пересчитывать не все значения переменной, а лишь некоторое подмножество случаев. Для этого в рамке **Subset** (подмножество) необходимо указать диапазон случаев. Команда также доступна при помощи кнопки на панели инструментов **X = ?**

Команда **Shift (Lag)** (сдвинуть (задержка)) предназначена для сдвига значений переменной на несколько случаев вниз или вверх. Это число называется лагом. Сдвиг может быть вперед (вниз) или назад (вверх) по отношению к текущему состоянию. В диалоговом окне **Shift (Lag)** необходимо задать имя переменой, лаг, направление сдвига (вправо или влево).

При помощи команды **Recode Variables** (перекодировать переменные) можно перекодировать значения переменной, при этом исходные значения переменной заменяются новыми значениями. Например, пусть имеется переменная *Нас.гор*, которая принимает значения 30,3, 26,8,..., 72,8. Нужно разбить эти значения на три

группы: *низкая*, если значение переменной меньше 40, *средняя*, если значение ременной больше либо равно 40 и меньше 80, и *высокая*, если значение перем ной больше либо равно 80. В открывшемся одноименном с командой окне, в р ке **Category 1**, выбрав опцию *Include if* (включить если), наберите *v4 < 40*, в р **New Value1** — *низкая*; аналогично в рамке **Category 2**, выбрав опцию *Include if*, берите *40 <= v4 and v4 < 80*, в рамке **New Value2** — *средняя;* в рамке **Categor** выбрав опцию *Include if*, наберите *v4 >= 80*, в рамке **New Value3** — *высокая*. жмите **OK**. В результате перекодировки числовые значения будут заменены категориальные (рис. 1.5). Всего предусмотрено 17 категорий.

	Крупнейшие страны мира по численности населения								
	1 Нас.90	2 Нас.95	3 Нас.00	4 Нас.гор.	5 Нефть	6 Газ	7 Пром.	8 С/х	9 Услуги
Китай	1120	1121	1275	низкая	есть	есть	48	21	31
Индия	830	935	1010	низкая	нет	нет	30	29	41
США	250	263	250	средняя	есть	есть	26	2	72
Бразилия	150	162	170	средняя	нет	нет	37	14	49
Россия	289	149	146	средняя	есть	есть	38	7	55
Япония	124	125	126	средняя	нет	нет	38	2	60
ФРГ	80	82	82	высокая	нет	нет	38	2	60
Индонезия	180	198	215	низкая	есть	есть	42	17	41
Великобритания	57	57	69	высокая	нет	нет	32	2	66
Франция	56	58	59	средняя	нет	нет	27	2	71

Рис. 1.5

Команда **Rank** (ранжировать) (рис. 1.6) позволяет ранжировать одну или б лее переменных. Содержимое столбца будет заменено рангами значений.

Для сохранения исходных значений столбца (переменной) надо сделать к пию переменной и произвести ее ранжирование.

Рассмотрим назначение функциональных кнопок диалогового окна:

- **Variables** (переменные) позволяет выбрать переменные для ранжирования;
- **Cases** (наблюдения) предназначена для выбора набора случаев, которы надо ранжировать;
- **Weight** (задать вес). Можно задать вес выделенных ранее переменных с по мощью другой переменной таблицы исходных данных.

Рассмотрим функциональное назначение полей выбора основных опций ран жирования:

- **Assign rank 1 to** (присвоить ранг 1). Значения можно ранжировать по воз растанию, т.е. *smallest value* (наименьшее значение) начинается с 1. Или по убыванию, т.е. *largest value* (наибольшее значение) начинается с 1;
- **Rank for ties** (совпадающие ранги). Опция *Mean* (средний) означает что рангам совпадающих значений присваивается средний из этих рангов. Опция *Sequential* (последовательный) означает, что каждое совпадающее значение ранжируется последовательно в порядке их появления в столбце.

Опции *Low* (низший), *High* (высший) означают, что каждому совпадающему значению присваивается соответственно наименьший или наивысший из рангов совпадающих значений;

- **Type of ranks** (типы рангов). Опция *Regular* (обычный) – диапазон ранжирования от 1 до *n* (*n* – число случаев в таблице). Опция *Fractional* (дробный) – диапазон ранжирования от 0 до 1. Опция *Fractional as %* (дробный в %) – ранги являются процентным соотношением, основанным на дробном ранжировании значений переменной.

Командой **Standardize** (стандартизация) все значения выбранных переменных заменяются на стандартизованные значения, вычисляемые следующим образом:

стандартное значение = (исходное значение – среднее)/стандартное отклонение.

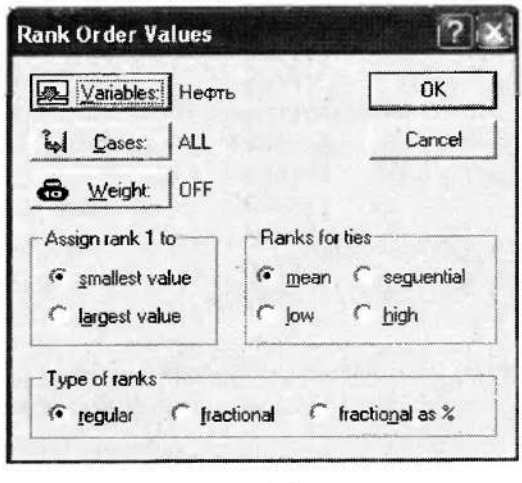


Рис.1.6

В открывшемся диалоговом окне надо выбрать переменные для стандартизации, а также подмножество наблюдений для стандартизации (по умолчанию выбираются все наблюдения) и задать веса для наблюдений. Задание весов эквивалентно тому, что каждое наблюдение используется при вычислении среднего и стандартного отклонения несколько раз, пропорционально весу.

Несколько подробнее рассмотрим команду **Variable Specs** (спецификации переменной), открывающую основное диалоговое окно (рис.1.4), в котором задаются все спецификации переменной. Альтернативный способ вызова этого диалогового окна – дважды щелкнуть на имени переменной в электронной таблице. Под спецификацией переменной в системе *STATISTICA* понимается имя переменной, ее формат, код для пропущенных значений в этой переменной, метка, формула или связь *DDE*. Например, в поле **Name** (имя) задается имя переменной. Для быстрого перехода к именам можно использовать кнопки быстрого перехода: к следующей переменной {>>} и предыдущей переменной {<<}. Значения переменной по умолчанию

отображаются в формате восьми значащих цифр и тремя разрядами после деся тичной точки. Число десятичных разрядов в представлении числа должно бы меньше ширины столбца и может быть задано либо при помощи опции *Decim* (десятичные знаки) в данном окне, либо при помощи кнопки **Increas Decim** на панели инструментов.

Формат отображения исходных данных задается в группе полей, объедине ных в рамке **Display Format** (формат отображения). Необходимо задать тип да ных (категорию) и выбрать для нее в рамке справа способ представления. Рас смотрим основные типы категорий.

Категория *Number* (число). Этот формат используется для представления чи словых или текстовых переменных.

Категория *Date* (дата). Даты в системе *STATISTICA* хранятся в юлианско формате. Если переменная имеет формат даты, то можно выбрать различные фо мы представления и коды, которые им соответствуют. Коды могут использовать ся для ссылки на соответствующий формат.

ФОРМАТ	*КОД*
3/16/89	*DATE1*
03/16/1989	*DATE2*
16 -Мар-89	*DATE3*
16-Мар	*DATE4*
Мар-1989	*DATE5*
16/03/89	*DATE6*
Windows Format	*DATE7*

Категория *Time* (время). Аналогично датам эти переменные хранятся в виде чисел – доля дня, прошедшего с полуночи. Возможны следующие способы представления:

ФОРМАТ	*КОД*
3/16/89 11:39 PM	*TIME1*
89/3/16 23:39	*TIME2*
11:39PM	*TIME3*
23:39	*TIME4*
11:39:01PM	*TIME5*
23:39:01	*TIME6*
Windows Format	*TIME7*

Категория *Scientific* (научный). Предоставляет возможность представления чисел в научной записи. Например, число 0,123 будет представлено как 1,23E–01.

Категория *Currency* (денежный формат). После выбора этого формата перед числовым значением или после него появится знак соответствующей денежной единицы, например *$, DM, etc.*

Категория *Percentage* (проценты). Задает формат представления переменной, которая принимает значения в виде процентов. При этом, например, число 0,1 отобразится как 10%, а 15,5 – как 1550%.

Кнопка **All Specs** осуществляет переход в меню для просмотра свойств всех переменных.

Кнопка **Text Labels editor** (редактор текста ярлыков) редактирует метки переменных. Можно добавлять, сортировать, перенумеровывать ярлыки.

Кнопка **Values/Stats...** вычисляет основные статистики переменной.

Для выполнения основных операций над наблюдениями надо либо нажать на кнопку **Cases** (наблюдения) на панели инструментов, либо в меню **Data** выбрать команду **Cases**, либо через контекстное меню, щелкнув правой кнопкой мыши на любом поле имен случаев. Команды **Add Cases** (добавить наблюдения), **Move Cases** (переместить наблюдения), **Delete Cases** (удалить наблюдения), **Copy Cases** (копировать наблюдения) выполняются как и соответствующие команды для переменных.

Команда **Case Names Manager** (диспетчер имен случаев) открывает диалоговое окно, в котором можно задать длину имен всех случаев, высоту строк, присвоить именам случаев значение какой-либо переменной (опция **From**) или наоборот переменной присвоить имена случаев (опция **To**).

Диалоговое окно **Case Selection Conditions** (условия выбора случаев) предназначено для выбора подмножества случаев для анализа. Оно доступно из меню **Tools** (инструменты), команды **Selection Conditions** (условия выбора) и команды **Edit** (правка), либо из стартовых окон модулей, в которых предусмотрен выбор подмножества случаев для анализа, нажатием кнопки **Cases**. Рассмотрим вкладку **Selection**.

Если установить флажок на *Enable Selection Conditions* (включение условий выбора), то будут открыты поля для задания условий выбора случаев. Необходимо выбрать опцию, которая указывает, используются ли наблюдения для анализа (*Include cases)* или же не используются *(Exclude cases)*. В поле **Expression** записываются условия выбора. Если проводится анализ по одной переменной, то условия выбора случаев могут быть произведены по другой переменной. Например, стандартизуется переменная *V1*, а условие выбора случаев имеет вид *V2 >= 5* или *v0 <= 8* (напомним, что *V0* означает номер наблюдения).

В поле **Or case number** можно просто перечислить номера случаев, которые либо включаются в анализ, либо не включаются. Например, 1:5 означает, что в анализ включены (не включены) наблюдения 1–5; а 1; 3; 5 означает, что в анализ включены (не включены) наблюдения 1, 3, 5. Номера можно перечислить просто через пробел.

Можно задавать сложные условия выбора при помощи логических операторов **AND** (и), **OR** (или), **NOT** (нет). При этом должны соблюдаться определенные правила. Можно ссылаться на переменную, используя либо ее номер (например, *v1, v5*), либо имя. Текстовые значения необходимо заключать в одиночные кавычки (например, *v5* — «есть»), а сложные условия — в скобки. Полный список доступных операторов следующий:

=	*(равно)*
#, <>, *X*	*(не равно)*
<	*(меньше)*
>	*(больше)*
<=	*(меньше или равно)*

>=	*(больше или равно)*
NOT, ~	*(логическое отрицание)*
AND, &	*(логическое И)*
OR	*(логическое ИЛИ).*

Кнопка **Open** (открыть) активизирует сохраненное ранее при помощи кнопки **Save As** (сохранить как) условие выбора случаев.

На вкладке **Display** можно использовать специальный формат редактирования условий выбора.

Вкладка **Subset/Random Sampling...** (подмножество/случайное формирование) соответствует одноименной команде из меню **Data** (данные).

1.4. Основные операции с таблицами данных

Команда **Input Spreadsheet** (ввод крупноформатной таблицы) предназначена для обозначения таблицы как крупноформатной. Анализ и построение графиков возможен только с крупноформатными таблицами.

Команда **Subset/Random Sampling** (подмножество/случайное формирование) предназначена для создания новой таблицы, состоящей из подмножества случаев выбранных переменных. Выделите вкладку и нажмите одноименную кнопку. Программа откроет диалоговое окно (рис. 1.7). Нажмите кнопку **Variables** и выберите переменные, например, *Пром., С/х, Услуги.* В рамке **Subset Selection Rules** (правила выбора подмножеств) надо выбрать правило формирования подмножеств. Выберите опцию *Use selection condition expression* (использовать выбранное выражение). Нажмите кнопку **Cases** и укажите условие выбора, например *V3 < 250.* Щелкните по **OK**, появится таблица (рис. 1.8) с именами случаев (названия стран) с численностью населения на 2000 г. меньшей, чем 250 млн чел.

Опция *Simple random Sampling* (простой случайный выбор) предусматривает два различных способа случайного выбора: *Percent of cases* — в поле задается процент наугад выбранных случаев; *Approximate number of cases* — в поле задается приближенное число выбранных случаев.

Если установить флажок на *With replacement,* осуществляется случайный выбор с возвращением, т.е. выбранный случайным образом случай возвращается в исходное множество и может быть вновь выбран.

Выделите опцию *Systematic random Sampling* (систематический случайный выбор). Укажите значение параметра k, который может принимать значения от 1 до n. Пусть, например, $k = 3$. Программа из первых трех случаев наугад выберет один случай, затем из следующих трех случаев выберет следующий и т.д., пока из оставшихся случаев не будет выбран последний. Нажмите **OK**. В таблице, изображенной на рис. 1.9, приведено сформированное подмножество.

При проведении исследований может появиться необходимость в объединении файлов различными способами. Команда **Merge** (слитие) предназначена для соединения данных из двух фалов. В команде **Merge** реализованы три способа объединения:

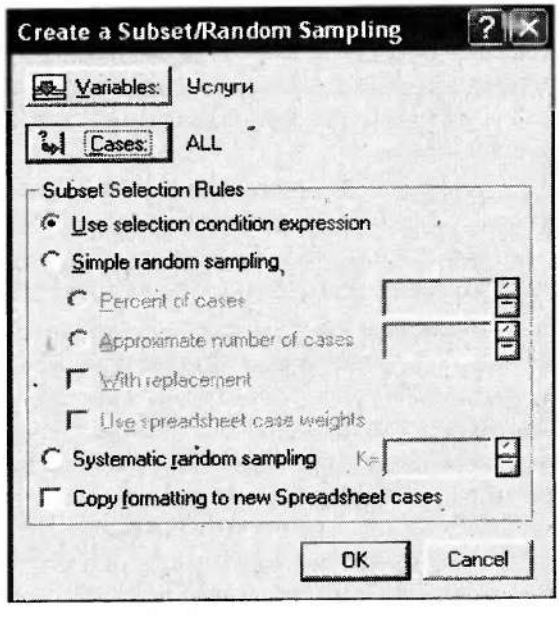


Рис. 1.7

	1 Пром.	2 С/х	3 Услуги
Бразилия	37	14	49
Россия	38	7	55
Япония	38	2	60
ФРГ	38	2	60
Индонезия	42	17	41
Великобритания	32	2	66
Франция	27	2	71

Рис. 1.8

	1 Пром.	2 С/х	3 Услуги
Индия	30	29	41
Россия	38	7	55
Индонезия	42	17	41

Рис. 1.9

- **Variables** (объединение переменных), если надо добавить данные другого файла, содержащего те же наблюдения справа от исходного;
- **Cases** (объединение наблюдений), если надо добавить данные другого файла, содержащего те же переменные вниз исходному, число переменых обоих файлов должно совпадать;
- **Text labels** (объединение текстовых значений), если надо объединить текстовые присвоения и метки из двух файлов. В программе не предусмотрено объединение файла с самим собой. Если же необходимо это сделать, надо при помощи команды **Save As** (сохранить как) сделать копию файла.

Рассмотрим более подробно наиболее сложную процедуру объединения первым способом. Если количество и порядок в двух файлах данных совпадают не полностью (например, во втором файле пропущено хотя бы одно наблюдение), то соответствие следующих за этим наблюдением данных будет нарушено. Чтобы избежать этой ситуации, надо выбрать один из трех предусмотренных режимов объединения.

Not Relational (нереляционный). При выборе этого режима переменные из второго файла будут просто добавлены к переменным первого файла. Этот режим устанавливается по умолчанию. Число случаев и переменных в файлах может быть различным.

Relational (реляционный). Для реляционного режима объединения требуется в каждом из наборов данных определить переменную — ключ. Программа будет для каждого наблюдения проверять значение ключа в обоих файлах и проводить объединение наблюдений только тогда, когда значения ключей совпадают. Файлы должны быть отсортированы по значениям ключей. Наблюдения из второго файла будут совмещены с соответствующими наблюдениями из первого файла в соответствии со значениями ключа. При этом объединяются последовательные записи с одинаковыми значениями ключа. Если в файлах имеется неравное количество записей с одинаковыми значениями переменной — ключа, к файлу с меньшим количеством записей в нужных местах добавляются коды пропущенных данных.

Relational Hierarchical (реляционный иерархический). Все делается аналогично предыдущему режиму за исключением последнего шага. Вместо недостающих записей вставляются данные из последней записи, имеющей те же значения переменной — ключа.

Предположим надо объединить файл, изображенный на рис. 1.10, с файлом, изображенным на рис. 1.11 (в файлах определены переменные — ключи). Выберите в команде **Merge** способ **Variables**. В открывшемся окне **Open File to Merge** укажите, с каким файлом надо соединить, появится диалоговое окно (рис. 1.12). В верхней части окна указаны названия исходного файла и файла для объединения.

В рамке **Mode** следует указать режим объединения. Выберите опцию *Not Relational* и нажмите кнопку **OK**. Произойдет простое объединение файлов. Переменные из второго файла приписаны справа к переменным первого файла (рис. 1.13). Выберите опцию *Relational* и нажмите **OK**. В новом файле (рис. 1.14) будут объединены записи с одинаковыми значениями ключа. Причем к файлу с меньшим числом одинаковых значений ключа приписаны пустые ячейки.

	1 Key1	2 Var1
1	1	10
2	2	5
3	2	5
4	2	7
5	3	3

Рис. 1.10

	1 Key2	2 Var2
1	1	5
2	1	7
3	2	3
4	3	2
5	3	4

Рис. 1.11

Выберите опцию *Relational hierarchical* и нажмите **OK**. В новом файле (рис. 1.15) также будут объединены записи с одинаковыми значениями ключа, но вместо недостающих записей будут вставлены данные из последней записи, имеющей те же значения переменной — ключа.

В рамке **Unmatched Cases** указывается способ обработки данных, если наблюдения не совместны. Несовместимость наблюдений может возникнуть из-за неодинакового количества наблюдений в объединяемых файлах или когда некоторые наблюдения не удовлетворяют критериям реляционного объединения.

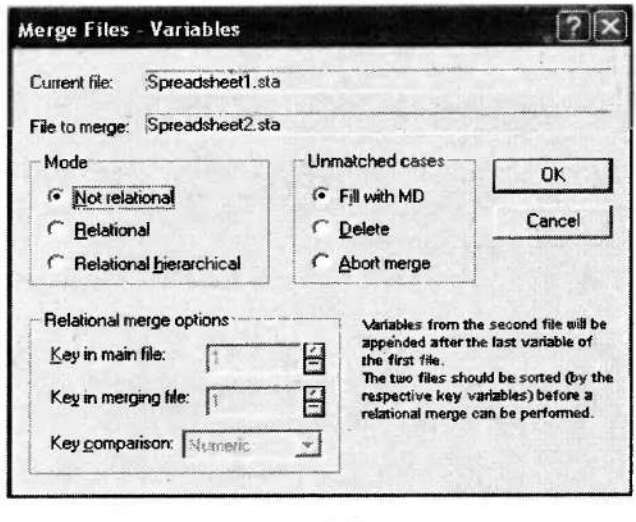


Рис. 1.12

	1 Key1	2 Var1	3 Key2	4 Var2
1	1	10	1	5
2	2	5	1	7
3	2	5	2	3
4	2	7	3	2
5	3	3	3	4

Рис. 1.13

	1 Key1	2 Var1	3 Key2	4 Var2
1	1	10	1	5
2			1	7
3	2	5	2	3
4	2	5		
5	2	7		
6	3	3	3	2
7	3	3	3	4

Рис. 1.14

	1 Key1	2 Var1	3 Key2	4 Var2
1	1	10	1	5
2	1	10	1	7
3	2	5	2	3
4	2	5	2	3
5	2	7	2	3
6	3	3	3	2
7	3	3	3	4

Рис. 1.15

Есть три способа обработки данных:

- **Fill with MD** (заполнить их кодом МД), при выборе этого варианта недостающие значения переменных для несовместимых наблюдений заполняются кодом пропущенных данных;
- **Delete** (удалить), при выборе этого варианта несовместимые наблюдения из результирующего файла будут удалены;
- **Abort Merge** (прервать объединение), при наличии несовместимых наблюдений в файлах на экран выдается сообщение об ошибке и выполнение процедуры объединения прерывается.

В рамке **Relational merge options** указываются номера столбцов ключей и сравнение по ключу (текстовое или числовое).

Команда **Sort** (сортировка) предназначена для сортировки данных таблицы. Сортировка (перестановка целиком всей строки таблицы) может осуществляться по значениям переменных (ключей) или по именам наблюдений, если они есть. В программе реализован принцип иерархической сортировки, т.е. строки сначала сортируются по значениям первого ключа; если у случаев ключи совпадают, то они будут отсортированы по второму ключу и т.д. Изначально доступно три ключа, для того чтобы получить доступ к семи, надо щелкнуть кнопкой **More keys**. Если значения переменной – ключа числовые или текстовые, надо указать соответственно – *Numeric* или *Text*.

При сортировке по именам наблюдений надо выделить опцию *Case Name* и указать в поле формата *Text*. При этом сортировка будет осуществлена сначала

по первым буквам имен, при совпадении первых букв — по вторым и т.д. Если имена совпадают, то будет использован следующий ключ — переменная. Опция *Case Name* может быть выделена один раз в диалоговом окне и порядок перехода программы к сортировке по имени наблюдений будет зависеть от номера ключа. *Ascending* (возрастание) и *Descending* (убывание) определяют тип сортировки.

Команда **Transpose — Block** (транспонирование — блок) транспонирует предварительно выделенный квадратный блок электронной таблицы. Команда **Transpose — File** (транспонирование — файл) транспонирует весь файл данных. Переменные становятся наблюдениями, а наблюдения — переменными. Имена случаев становятся именами переменных и наоборот.

Команда **Verify Data** (проверка данных) предназначена для проверки введенных данных. После того как данные введены или импортированы в файл, может возникнуть необходимость проверить их на целостность (логическую непротиворечивость) и полноту. Так, можно определить, являются ли значения переменной допустимыми (например, переменная *Нефть* может принимать только два значения 0 и 1) или находятся ли они в пределах допустимого диапазона значений (например, переменная *Пром.* должна быть больше 26 и меньше 48). Критерии проверки могут быть как простыми (состоять из одного условия), так и очень сложными (содержать множественные логические условия). После выбора в меню **Data** команды **Verify Data** откроется диалоговое окно (рис. 1.16).

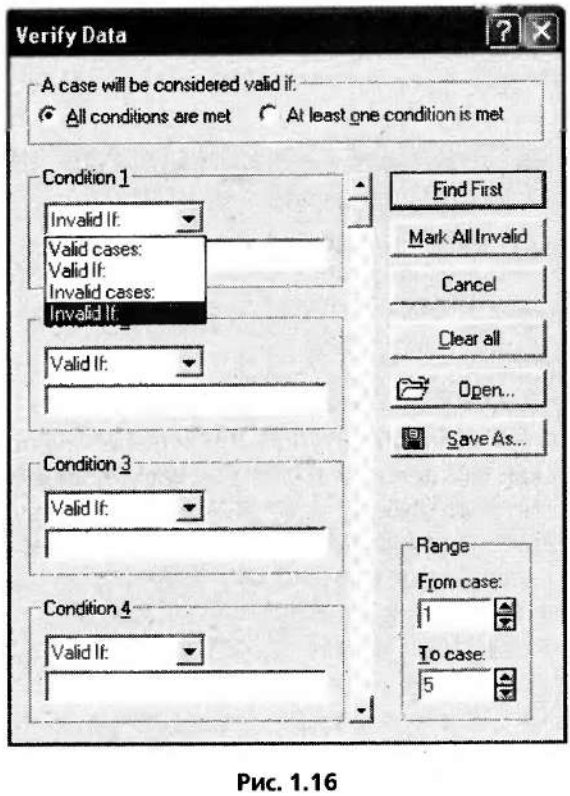


Рис. 1.16

Опция **All condition a met** означает: считать случай правильным, если он удовлетворяет всем условиям. Соответственно опция *At least condition is met* означает: считать случай правильным, если он удовлетворяет хотя бы одному условию. В рамке **Condition** надо указать тип условия:

- *Valid cases* (правильные наблюдения) — указываются номера правильных случаев;
- *Valid if* (правильное, если) — указывается формула, по которой определяется правильный случай;
- *Invalid cases* (неправильные наблюдения) — указываются номера неправильных случаев;
- *Invalid if* (неправильное, если) — указывается формула, по которой определяется неправильный случай.

Можно задать 16 условий. по которым будет осуществлена проверка.

В рамке **Range** указывается, с какого по какой случай производить поиск. Неправильные ячейки можно смотреть по очереди — *Find First*, либо выделить их все сразу — *Mark all invalid*. Программа начнет процедуру проверки набора данных или указанного диапазона. При обнаружении наблюдения, не удовлетворяющего условиям, программа выделит соответствующую строку данных и откроет диалоговое окно, в котором пользователь может пропустить указанное наблюдение или отредактировать его, а затем продолжить или остановить проверку. После окончания редактирования для продолжения процедуры проверки выделенного наблюдения надо нажать на кнопку **Continue**. Заданные условия можно сохранить — **Save as**, либо загрузить новые — **Open**.

1.5. Обмен данными с другими приложениями

Для ввода данных в электронную таблицу *STATISTICA*, подготовленных в каком-либо другом приложении, можно воспользоваться одним из следующих способов:

- буфером обмена;
- технологией динамического обмена данными;
- средствами импорта файлов.

Первый способ — самый быстрый и простой путь ввода данных из прикладных программ *Windows* [2]. Для реализации этого способа надо выполнить следующую последовательность шагов:

- в исходном материале выделить данные, которые необходимо скопировать;
- в меню **Edit** (правка) выбрать команду **Copy** (копировать); данные будут скопированы в буфер обмена;
- перейти в электронную таблицу *STATISTICA* и установить указатель там, куда следует скопировать данные, затем нажать кнопку мыши;
- в меню **Edit** выбрать команду **Past** (вставка), при этом данные будут скопированы в направлении вправо и вниз от места, обозначенного курсором.

Можно также воспользоваться одноименной кнопкой на панели инструментов. Содержимое буфера обмена может быть вставлено несколько раз.

Иногда необходимо установить связь между данными из какого-либо приложения (называют еще источником или сервером), например *Excel*, и таблицей *STATISTICA* (клиентский файл) таким образом, чтобы при изменении данных в сервере соответствующие изменения произошли в таблице *STATISTICA* — клиенте. Связи такого типа в *STATISTICA* устанавливаются при помощи процедуры динамического обмена данными (DDE).

В программе *STATISTICA* реализованы две возможности задания динамического обмена данными: при помощи команды **Past Special** (специальная вставка) и команды **DDE Links** (*DDE связи*) из меню **Edit**.

Для того чтобы установить динамический обмен данными при помощи команды **Past Special**, необходимо выполнить следующие действия:

- в исходном материале выделить данные, которые надо скопировать;
- в меню **Edit** выбрать команду **Copy**; данные будут скопированы в буфер обмена;
- перейти в электронную таблицу *STATISTICA* и установить указатель мыши в том месте, куда следует скопировать данные, и нажать кнопку мыши;
- выбрать команду **Past Special** в меню **Edit,** откроется окно *Специальная вставка*, в котором надо выделить нужный формат *(Text, Лист Microsoft Exel, либо HTML)* и выбрать опцию *вставить связь*, далее нажать **OK** (рис. 1.17). Данные из буфера обмена будут скопированы в указанное место таблицы *STATISTICA*.

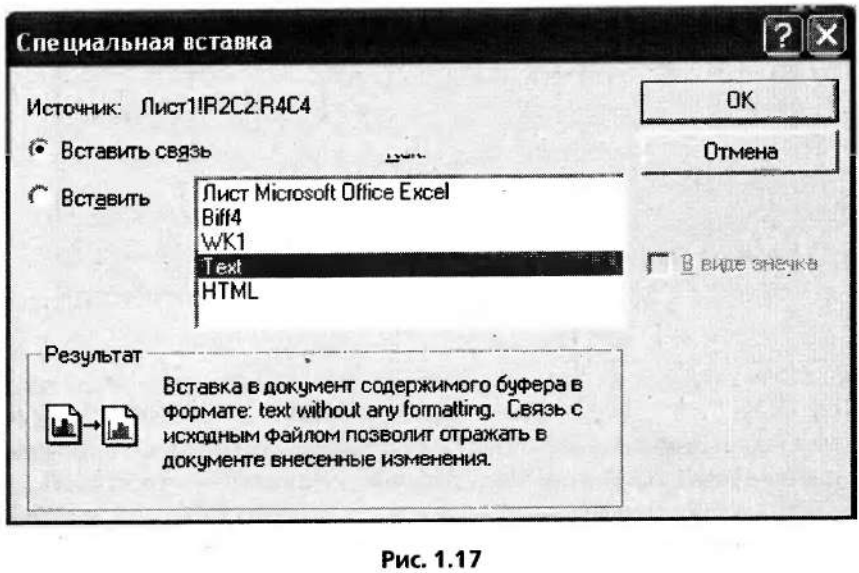


Рис. 1.17

Рассмотрим, как можно установить динамическую связь вторым способом — при помощи команды **DDE Links** из меню **Edit**.

Пусть необходимо установить **DDE** связь между источником *Книга 1*, который представляет собой таблицу чисел в *Excel* (рис. 1.18), и клиентским файлом в *STATISTICA*. С этой целью надо открыть файл *Книга 1*, перейти в программу *STATISTICA*, установить курсор на ячейку, начиная с которой будут отображаться данные из источника. Выберите команду **DDE Links** в меню **Edit**. Откроется диалоговое окно **Manage DDE Links** (рис. 1.19).

	A	B	C	D	E
1	1	2	3	4	5
2	5	4	3	2	1
3	2	3	4	5	1
4	3	4	5	1	2
5	4	5	1	2	3

Рис. 1.18

Окно диалога будет не заполнено, это означает, что текущих связей нет. В зависимости от ситуации отдельные кнопки могут быть неактивны.

Если бы были текущие связи, то они были бы прописаны в этом окне. Связи можно создать, редактировать, удалить, отключить, обновить.

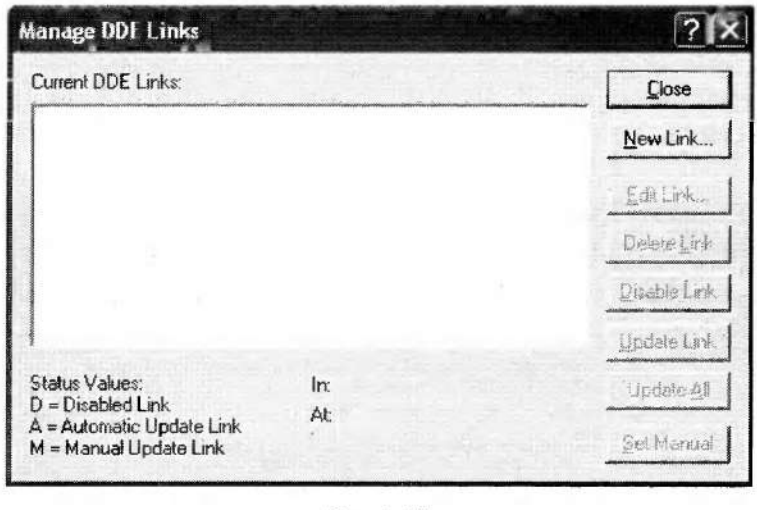


Рис. 1.19

Для создания связи щелкните по кнопке **New Link** (новая связь), откроется окно **New DDE Link**. Напечатайте инструкцию связи в поле **DDE Link** или, что значительно проще, используйте поля обзора **Service** (обслуживание), **Topics** (разделы), **Items** (элементы), чтобы формировать инструкцию блоками. На рис. 1.20 показано, что в соответствующих полях выбраны *Excel, Лист 1* из *Книги 1.xls* и в *Items* набраны *r1c1:r3c3*. При этом инструкция связи автоматически пропишется в поле **DDE Link**. Она связывает указанные ячейки в электронной таблице *Excel* с ячейками в электронной таблице *STATISTICA*. Несколько подробнее опишем поля обзора.

Обслуживание. Это поле обеспечивает список всех активных приложений-серверов. Например, если открыт *Microsoft Excel*, то *Excel* будет доступным ресурсом под обслуживанием.

Разделы. Это поле обеспечивает список всех доступных объектов (целей), которые связаны с обслуживанием. Например, если выбран *Excel* как обслуживание, поле *Разделы* перечислит все доступные электронные таблицы *Excel* в открытых рабочих книгах.

Элементы. Это поле обеспечивает список всех доступных элементов (ячеек), которые соответствуют обслуживанию и разделу, которые выбраны. Для определения диапазона ячеек необходимо использовать буквы *r [строка]* и *c [столбец]*. При этом первая пара символов обозначает левую верхнюю ячейку, а вторая пара — правую нижнюю ячейку.

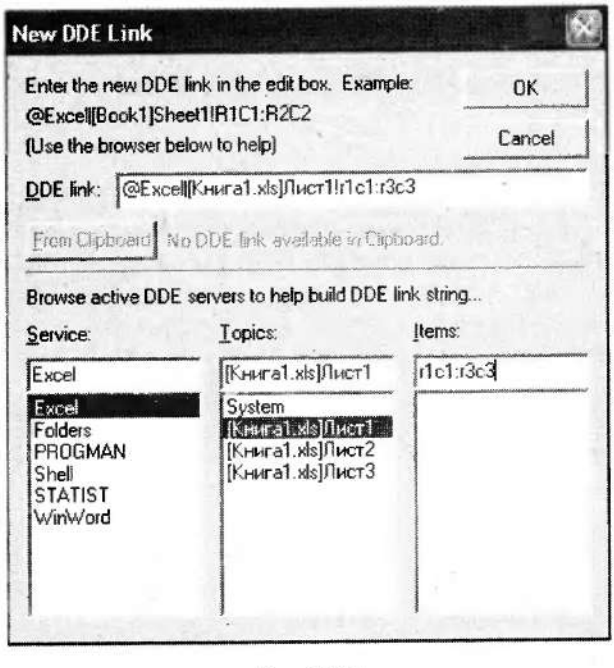


Рис. 1.20

Если в таблице *Excel* выделены данные и в меню **Edit** выбрана команда **Copy**, то при нажатии на кнопку **New Link**, связь будет автоматически создана и прописана в поле **DDE Link**.

Нажмите кнопку **OK**, окно **Manage DDE Link** примет вид, изображенный на рис. 1.21. Прописана инструкция связи и активизированы все кнопки.

В таблице *STATISTICA* (клиенте) появятся элементы из соответствующего диапазона таблицы *Excel* (сервера) (рис.1.22).

Рассмотрим работу кнопок окна **Manage DDE Link**.

Edit Link (редактор связи). После выбора существующей связи нажмите кнопку **Edit Link**, откроется окно **New DDE Link**, в котором можно редактировать связи.

Можете использовать инструкцию связи в поле редактирования или поля: *Обслуживание, Разделы, Элементы,* чтобы сформировать новую инструкцию связи. Например, в поле *Элементы* укажите диапазоны ячеек — *r1c1:r5c5*. Нажмите кнопку **OK**. В окне **Manage DDE Link** появится новая отредактированная связь, а в таблице *STATISTICA* — новые элементы (рис.1.23).

Заметим, что после редактирования связи исходный файл (сервер) копируется на то же место, которое было указано ранее курсором при создании связи.

Чтобы изменить место копирования, необходимо воспользоваться кнопкой **New Link**.

Delete Link (удалить связь). При помощи этой кнопки можно удалить выделенную связь.

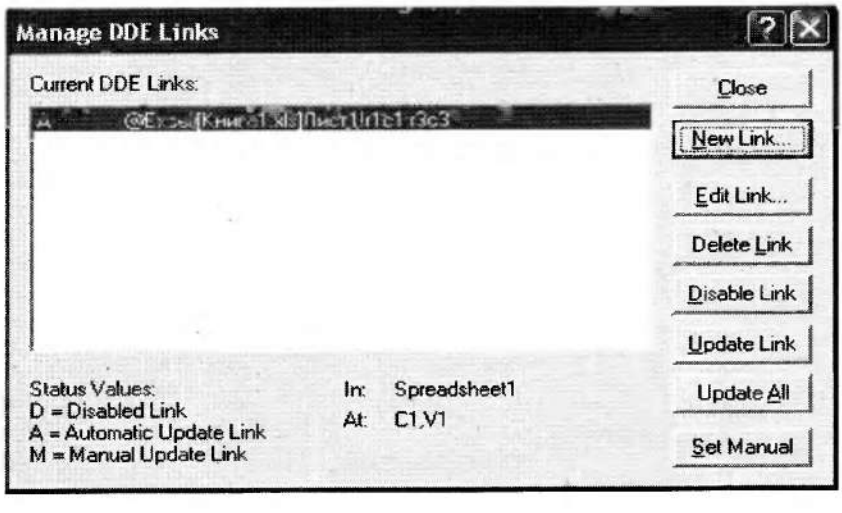


Рис. 1.21

Disable Link (отключить связь). Эта кнопка позволяет временно отключить динамические связи. Рекомендуется использовать при реализации процедур программы *STATISTICA*. В начале инструкции связи появится буква *D*. Для включения связи надо нажать на кнопку **Update Link**.

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1	1	2	3		
2	5	4	3		
3	2	3	4		
4					
5					

Рис. 1.22

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1	1	2	3	4	5
2	5	4	3	2	1
3	2	3	4	5	1
4	3	4	5	1	2
5	4	5	1	2	3

Рис. 1.23

Update Link (обновить связи). Произойдет подключение выделенной связи. В начале инструкции связи пропадет буква *D*.

Update All (обновить все). Будут подключены все связи.

Команда **Update** предназначена не только для подключения связи. Если при помощи кнопки **Set Manual** (ручная установка) выбран ручной режим обновления связи (в начале инструкции связи появится буква *M*), то после изменения данных в сервере автоматически не произойдут изменения в клиенте. Для того чтобы они произошли, необходимо нажать на кнопку **Update Link**. Для перехода к режиму автоматического обновления связи надо еще раз нажать на кнопку **Set Manual**, буква *M* заменится на букву *A*, что означает автоматический режим обновления связи. При каждом нажатии кнопка меняет свое название с **Set Manual** на **Set Auto** и наоборот.

Close (закрыть). Эта кнопка предназначена для выхода из диалога.

Третий способ ввода данных из других приложений — импорт файлов. Он реализован при помощи команды **Get External Data** (получение внешних данных) из меню **Data.** Эта команда формирует запросы из других баз данных. Для загрузки ранее сохраненного запроса надо выделить команды **Get External Data — Open Query from File** (получение внешних данных из файла). Для составления нового запроса следует выделить команды **Get External Data — Create Query** (создать запрос).

Запросы **Statistica Query** используются для получения данных, хранящихся в базе данных (БД). Программа *STATISTICA* позволяет обращаться к наиболее распространенным БД: *Oracle, MS SQL Server, Sybase, MS Access, Fox Pro* и др. Для доступа к данным используется драйвер *ODBC* (*Open DataBase Connectivity* — совместимость открытых баз данных), который позволяет приложению обращаться к БД на языке *SQL*. Запросы дают возможность легко выбрать из таблиц БД необходимые для статистического анализа данные и сохранить их в программе *STATISTICA*.

Рассмотрим все необходимые действия для того, чтобы создать запрос и импортировать данные из БД. В качестве примера используем первую из БД, которые поставляются вместе с программой *STATISTICA* и хранятся в файлах *C:\Program Files\ StatSoft\STATISTICA 6\Examples\Database\baseball.mdb* или *C:\Program Files\StatSoft\ STATISTICA 6\Examples\Database\Screw95.mdb* (предполагается, что программа *STATISTICA 6* установлена на диске *C*, в противном случае нужно указать правильное обозначение диска).

Для того чтобы создать запрос, в главном меню выберите команды **Data – Get External Data – Create Query**. Появится окно **Statistica Query** и в нем еще одно (рис.1.24) **DataBase Connection** (подключение базы данных). Здесь можно выбрать уже имеющееся подключение к нужной базе данных или создать новое. Для этого надо нажать кнопку **New...**,

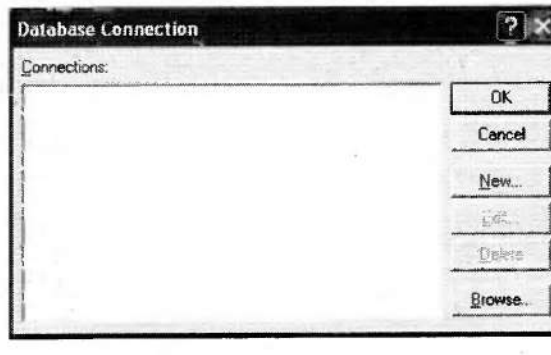


Рис. 1.24

в открывшемся окне **Data Link Properties** (поставщик данных, рис.1.25)

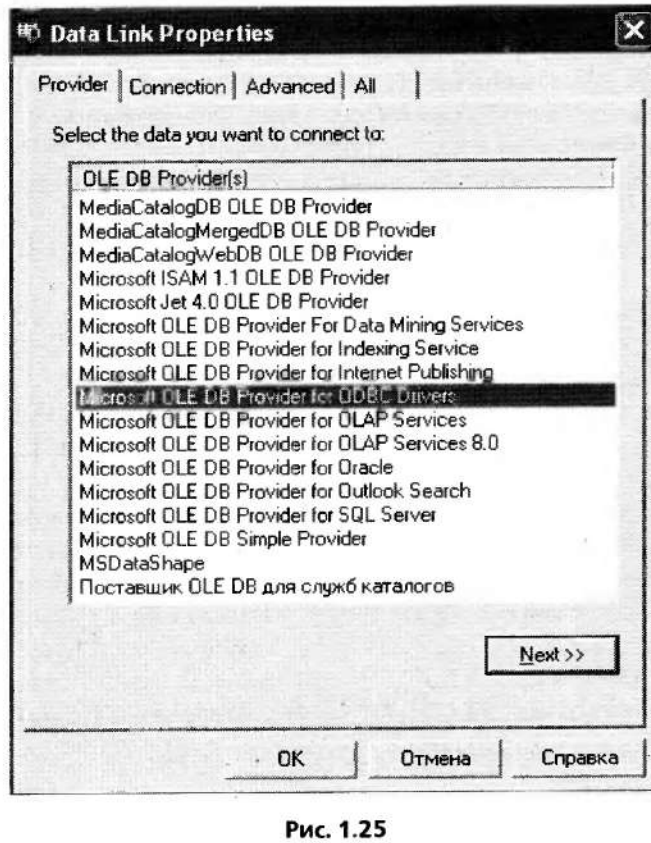


Рис. 1.25

на вкладке **Provider** (поставщик) выбрать драйвер, с помощью которого будет организован доступ к базе данных. Обычно используется *ODBC* драйвер, позволяющий работать с такими БД, как *MS Acceess, Visual FoxPro, dBase* и *Excel.* После того как драйвер выбран, переключитесь на следующую вкладку **Connection** (подключение) или нажмите кнопку **Next** (рис.1.26). Здесь для подключения данных **(Specify the following to connect to ODBC data)** необходимо в трех полях произвести определенные установки.

В поле **Specify the source of data:** (источник данных):

- либо, используя имя источника данных *(Use data source name)* из списка, указать, какая именно БД используется, т.е. задать специфику, например *MS Access* (*STATISTICA* импортирует данные практически из любых БД);
- либо, используя строку подключения *(use connection string)*, задать параметры подключения вручную, для чего необходимо будет указать драйвер, тип источника данных, создать имя для этого подключения и указать полный путь к файлам БД.

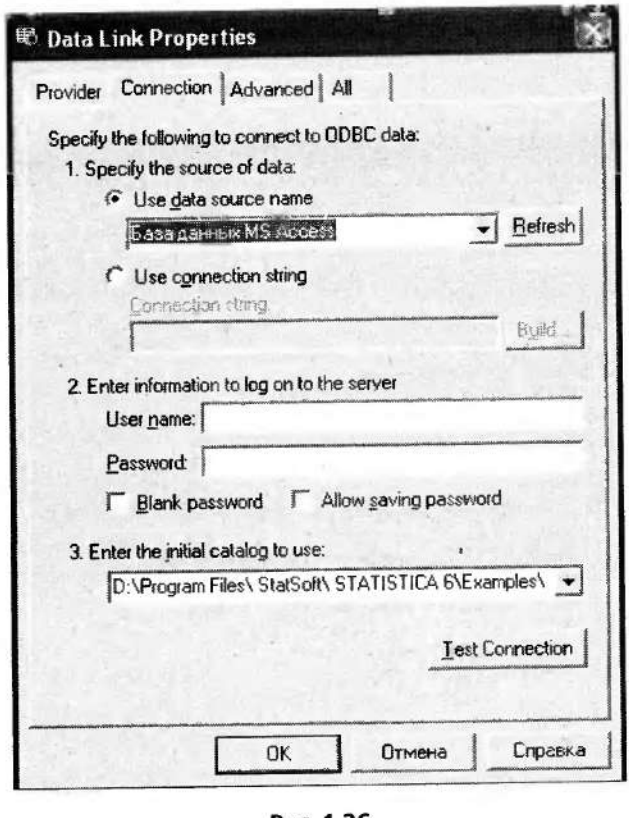


Рис. 1.26

В поле **Enter the User name and Password** (введите имя пользователя и пароль) надо указать имя пользователя и пароль, если они необходимы для подключения к БД.

В поле **Enter the initial catalog to use** (введите начальный каталог) следует указать полный путь к файлу БД, предварительно скопировав его из папки **Database.** В нашем случае это *C:\Program Files\StatSoft\STATISTICA 6\Examples\Database\baseball.mdb.*

После того как все параметры подключения указаны, можно протестировать его, нажав на кнопку **Test Connection** (проверить подключение). Если появилась надпись *Test connection succeeded,* то параметры указаны верно и можно нажать **OK.**

Вкладка **Advanced** предусмотрена для более сложного подключения к удаленным базам данных.

Последняя вкладка **ALL** позволяет изменять все произведенные настройки вручную.

После нажатия кнопки **OK** появится окно **Add a DataBase Connection** (добавить подключение базы данных) с параметрами подключения и останется лишь дать название созданному подключению, например: *Baseball* (рис.1.27). Нажмите кнопку **OK**, программа вернется в окно **DataBase Connection**, в котором будет прописано имя подключения *Baseball* и активизированы все кнопки. Нажмите **OK**, откроется главное окно модуля **Statistica Query** (рис.1.28). Слева в виде дерева отображаются таблицы и поля таблиц БД. Справа — рабочая область, с помощью мыши сюда можно перетащить таблицы.

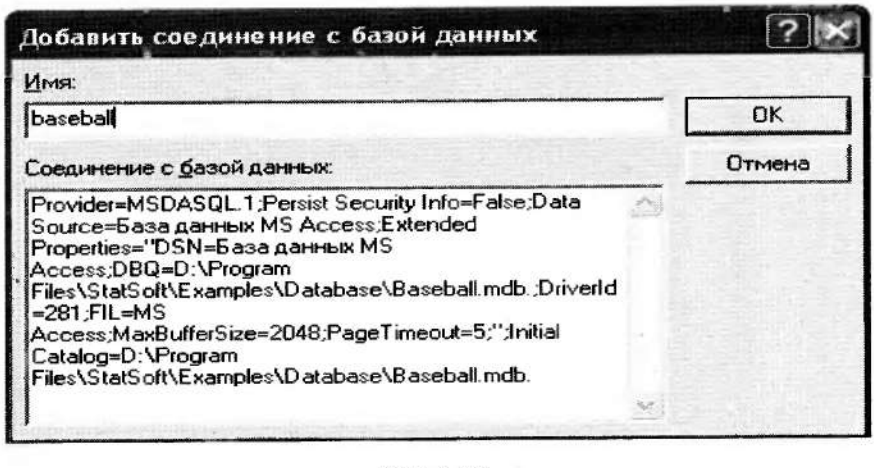


Рис. 1.27

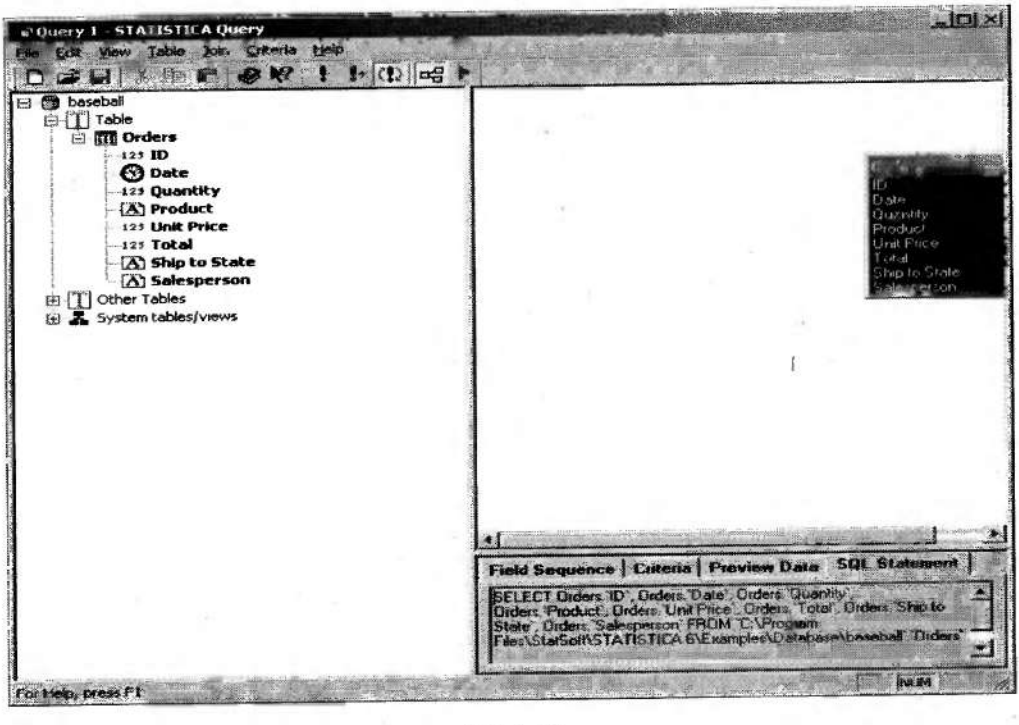


Рис. 1.28

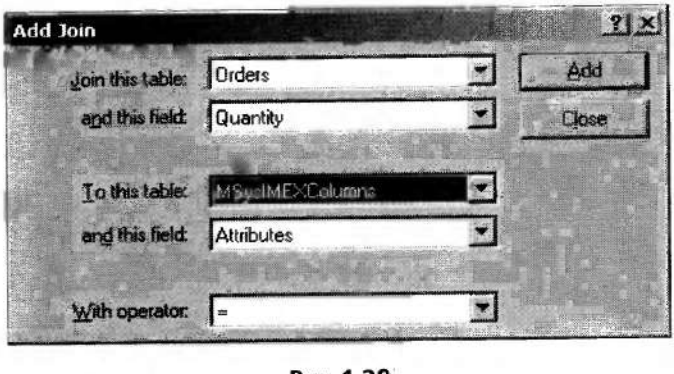


Рис. 1.29

С помощью команды **Join** (соединение) в главном меню можно добавлять, удалять и модифицировать связи между таблицами. После выбора этой команды откроется окно **Add Join** (добавить соединение), в котором надо указать названия таблиц и их полей, между которыми устанавливается связь. Так как БД содержит только одну таблицу *Orders* (заявки), на рис. 1.29 в качестве примера, сделаны установки для создания связи между полем *ID* таблицы *Orders* и полем *Attributes* одной из системных таблиц *MSyslMEXColumns*.

Внизу рабочей области диалога **Statistica Query** отображается вся информация о запросе. Первая вкладка **Field Sequence** отображает поля таблиц, из которых будут

выбираться данные по запросу. Чтобы добавить поле какой-либо таблицы к запросу, достаточно высветить его щелчком мыши в таблице.

На вкладке **Criteria** (критерий) можно задать достаточно сложные логические условия выбора данных.

Чтобы просмотреть полученные с помощью созданного запроса данные, прежде чем импортировать их в программу *STATISTICA*, необходимо открыть вкладку **Preview Data** (предварительный просмотр).

Вкладка **SQL Statement** предназначена для того, чтобы создавать *SQL* запросы вручную. Убедившись, что запрос правильно создан, можно импортировать данные в программу *STATISTICA*. Для этого нажмите кнопку на панели инструментов модуля **Statistica Query** или в главном меню **File** (файл) — **Return Data to Statistica**, программа вернется из **Statistica Query** в *STATISTICA* и откроется окно **Return External Data to Spreadsheet**. Прежде чем на мониторе отобразятся данные запроса, можно присвоить (изменить) ему имя, чтобы сохранить для дальнейшего использования. Кроме того, можно указать ячейку, начиная с которой следует вставить новые данные. Произведите при необходимости соответствующие установки в окне **Returning External Data to Spreadsheet** как это, например, сделано на рис.1.30. Нажмите кнопку **Run Now**, и данные запроса передадутся в таблицу *STATISTICA* вправо — вниз от ячейки (1, 1). Если в БД произойдут какие-либо изменения, то обновить данные запроса можно через главное меню **Data — Get External Data — Open Query from File.** Все запросы будут сохранены в файлах с расширением **. sqy*.

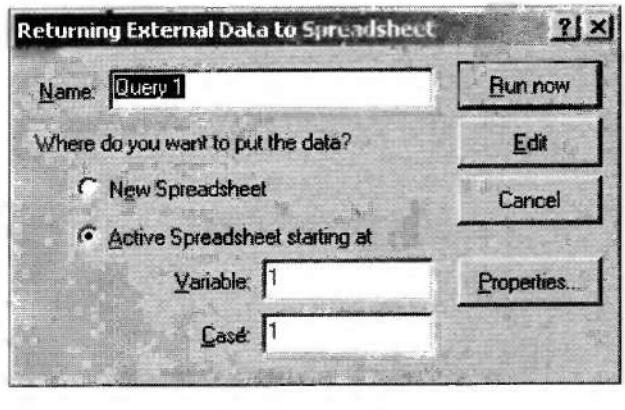


Рис. 1. 30

Создайте подключение к БД *Baseball*, в которой содержится информация о заявках на приобретение товаров, необходимых для игр: *Date* — дата заявки, *Product* — товар, *Quantity* — количество товара одного наименования, *Unit Price* — цена за единицу товара, *Total* — общая цена, *SalesPerson* — имя продавца, *ID* — номер заявки, *Ship to State* — название штата, из которого сделана заявка. Предположим, надо узнать, кто из продавцов оформил больше всего заявок на товары за определенный период (например, с февраля по май 1993 г.).

В диалоге **Statistica Query** курсором мыши перетащите таблицу *Orders* в рабочую область, выделите те столбцы таблицы (поля), данные из которых должны

быть использованы для анализа (рис.1.28). Выберите вкладку **Preview Data**, на которой можно просмотреть выбранные данные.

Чтобы задать условия отбора, воспользуйтесь вкладкой **Criteria**. Откроется окно **Add Criteria** (рис.1.31).

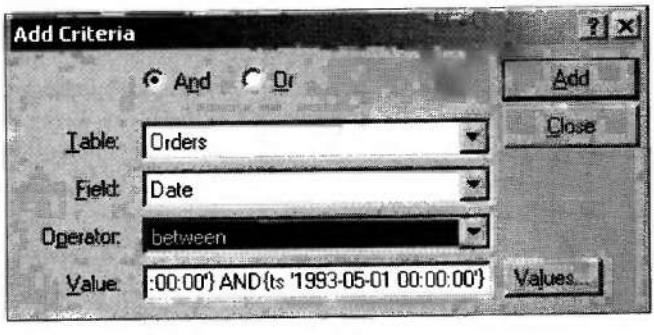


Рис. 1.31

В поле **Table** (таблица) выберите название таблицы *Orders*, в поле **Field** (поле) выберите название поля *Date* (дата) — столбец таблицы, из которого будут сравниваться данные. В поле **Operator** (оператор) надо выбрать правильный оператор сравнения, в нашем случае это будет оператор *between* (между). После того как задан оператор сравнения, необходимо указать граничные значения дат, соответствующие выбранному периоду времени. Нажмите кнопку **Values...** (значение...) и выберите левую границу интервала (рис.1.32).

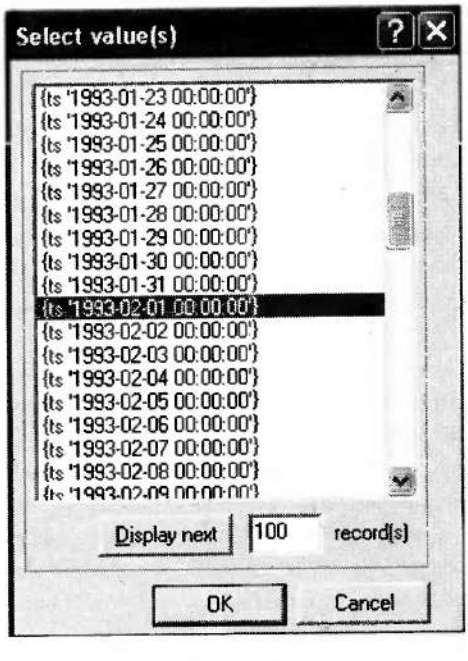


Рис. 1.32

В поле **Value** добавьте вторую границу самостоятельно, начав запись с оператора *AND*. В окне **Add Criteria** будет прописано условие выбора {ts '1993-02-01 00:00:00'}*AND{ts '1993-05-01 00:00:00'}* (рис. 1.31). Нажмите кнопку **Add**, в таблицу *STATISTICA* будут скопированы все выделенные поля БД (рис. 1.33).

	1 ID	2 Date	3 Quantit	4 Product	5 nit Pric	6 Total	7 Ship to State	8 Salesperson
61	92	02.04.1993	2	Jersey	30	60	AR	Paul
62	93	03.04.1993	3	Bat	45	135	KS	Paul
63	94	04.04.1993	5	Baseball	25	125	MN	Rob
64	95	05.04.1993	4	Baseball	25	100	MA	Paul
65	96	06.04.1993	1	Jersey	30	30	MN	Tom
66	97	07.04.1993	3	Jersey	30	90	MA	Rob
67	98	08.04.1993	3	Bat	45	135	DE	Tom
68	99	09.04.1993	5	Jersey	30	150	TX	Mark
69	100	10.04.1993	1	Glove	50	50	DE	Paul
70	101	11.04.1993	3	Baseball	25	75	PA	Tom
71	102	12.04.1993	3	Baseball	25	75	TX	Paul
72	103	13.04.1993	5	Jersey	30	150	MD	Rob
73	104	14.04.1993	5	Bat	45	225	DE	Tom

Рис. 1.33

С элементами данной таблицы можно проводить любые статистические исследования, доступные в программе *STATISTICA*. Например, можно построить гистограмму (рис. 1.34) и сделать вывод, что за период с февраля по май 1993 г. больше всех оформил заявок *Tom*, затем *Paul*, *Rob* и *Mark*.

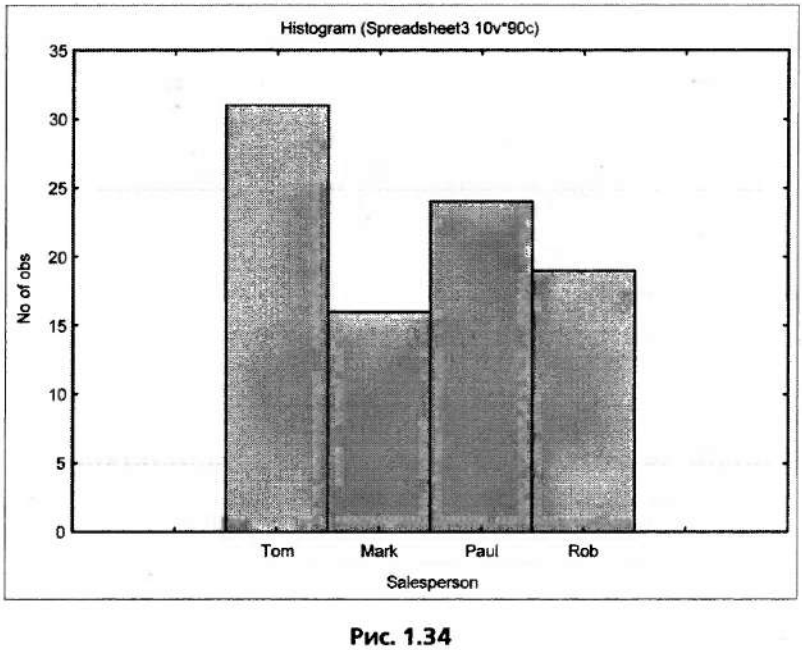


Рис. 1.34

Глава 2

Формирование *отчета* и *рабочей книги*

2.1. Назначение *отчета* и *рабочей книги*

При проведении статистического анализа важным моментом является удобный вывод результатов анализа. Кроме вывода результатов анализа в виде графиков и электронных таблиц в отдельных окнах на рабочем пространстве системы, в *STATISTICA* имеется возможность вывода этой информации и на другие каналы — принтер, в текстовый файл (для электронных таблиц) или в специальные файлы.

Файлы данных могут рассматриваться как *Workbook* (рабочая книга файлов), поскольку они сохраняют информацию обо всех дополнительных файлах, которые используются с текущим набором данных. *Рабочие книги* помогают организовать наборы дополнительных файлов (например, таблиц результатов, графиков, отчетов с текстом и графиками, программ пользователя, условий перекодирования и т.д.), которые были созданы или использованы во время анализа набора данных. Каждый раз, когда открывается или сохраняется файл, его имя автоматически добавляется в очередь файлов в поле рабочей книги. Когда очередь заполнена (по умолчанию длина очереди — 32),

то при добавлении новых файлов самые старые файлы удаляются из списка [6]. *Рабочие книги* могут создаваться автоматически и «вручную».

Любая графическая и текстовая информация в *STATISTICA* может быть выведена в файл в формате *RTF (Rich Text Format* — расширенный текстовый формат), который называется *отчетом* [2].

Отчет — это документ системы *STATISTICA* (файл в формате *RTF)*, в котором может сохраняться любая текстовая, численная и графическая информация. *Отчет* — один из типов документов в системе *STATISTICA* (другие типы документов — электронная таблица с данными, таблица *Scrollsheet* и график). В *STATISTICA* каждый тип документа выводится в собственном окне в рабочей области системы. Как только это окно становится активным, изменяется панель инструментов и меню. В нем появляются команды и инструменты, доступные для этого типа документа. *Отчеты* хранятся в файле с расширением **.rtf*.

Имеется возможность автоматического создания *отчета* — автоотчета. В этом случае любая таблица или график, которые выводятся на экран, будут автоматически выводиться в файл с *отчетом*.

Отчет служит для решения следующих задач:

1. Проведения статистического исследования. Для этого в него помещается достаточно большое количество промежуточной информации (графиков, таблиц и др.), чтобы в дальнейшем можно было «прокрутить анализ назад», сравнить его с другими *отчетами* или с результатами применения другого вида анализа к данным. В этом случае важно наличие в нем технической информации о том, когда он был создан, о файле данных и статистических процедурах, которые применялись для исследования и т.д.
2. Презентации и представления результатов исследования. С этой целью в него помещается небольшое количество графиков и таблиц. Основное внимание в *отчете* уделяется подаче материала, его размещению и форматированию. Он может быть частью другого, уже готового материала с *отчетом*.

Для решения этих задач в системе *STATISTICA* предусмотрены следующие подходы:

1. Направить всю необходимую информацию — таблицы, текст и графику — в специальное окно с *отчетом* и использовать все средства редактирования, доступные во встроенном текстовом редакторе системы *STATISTICA*. Если его возможностей недостаточно, то на заключительном этапе работы можно воспользоваться более мощным текстовым редактором, загрузив в него файл в формате *RTF*. Этот способ подготовки *отчета* — наиболее удобен для подготовки *отчета* первого типа (особенно при автоматическом создании *отчета*), а также для подготовки промежуточного варианта *отчета* второго типа.
2. Отдельные графики и таблицы можно скопировать или при помощи технологии *OLE* внедрить или связать с *отчетом*, который готовится в другом текстовом редакторе. Это удобно, если имеется небольшое количество документов *STATISTICA* для размещения в *отчете*. Такая возможность может быть с успехом использована по завершении статистического анализа,

после построения необходимых графиков и получения численных результатов. Для проведения статистического исследования этот способ подготовки *отчета* неудобен. Конечно, можно распечатать необходимые графики и таблицы непосредственно из *STATISTICA* и подложить их к *отчету*. Это наиболее простой способ получения высококачественных графиков и таблиц при печати.

2.2. Настройка программы для формирования *отчета* и *рабочей книги*

Для автоматического формирования *отчета* и *рабочей книги* нужно зайти в меню **File** (файл) и выбрать команду **Output Manager** (менеджер вывода) или в меню **Tools** (инструменты) выбрать команду **Options** и в появившемся диалоговом окне **Options** щелкнуть по вкладке **Output Manager** (рис.2.1).

Если в рамке **Place all results (Spreadsheets, Graphs) in** (поместить все результаты (таблицы, графики)) выбрать опции *Workbook* (книга), *Single Workbook* (одна книга) или *Existing Workbook* (существующая книга), то все результаты обработанных данных будут помещаться в отдельные окна, но все в одну и ту же книгу.

При выборе опции *Workbook containing the data file* (книга, содержащая файл данных) таблица с исходными данными будет помещена в отдельный файл.

Если установить флажок на *Also send to Report Window* (также послать окно *отчета*), *отчет* будет создаваться автоматически. После выбора этой опции можно указать, как будет формироваться *отчет*: *Multiplie Report* (много *отчетов*); *Single Report* (один *отчет*); *Existing Report* (существующий *отчет*).

- *Multiplie Report*. При выборе этой опции каждый анализ/график будет помещаться в отдельный *отчет*.
- *Single Report*. При выборе этой опции все анализ/графики будут помещаться в один *отчет*.
- *Existing Report*. При выборе этой опции все анализ/графики будут помещаться в уже существующий *отчет*, дополняя его. Полное имя *отчета* нужно указать, нажав на кнопку **Browse** (обзор).

При выборе опции *Individual windows* (индивидуальные окна) результаты анализа не будут автоматически вписываться в книгу.

Если установить флажок на *Display supplementary information* (показать дополнительную информацию), то можно выбрать степень полноты отображения дополнительной информации: *Brief* (кратко); *Medium* (средне); *Long* (длинно); *Comprehensive* (всесторонне).

- *Brief*. Будут выводиться лишь содержание выбранных таблиц и графики и более никакой дополнительной информации.
- *Medium*. Кроме таблиц и графиков будут также выводиться заголовок страницы (название модуля, дата, время, номер страницы), имя файла, условия выбора функции и другая дополнительная информация.

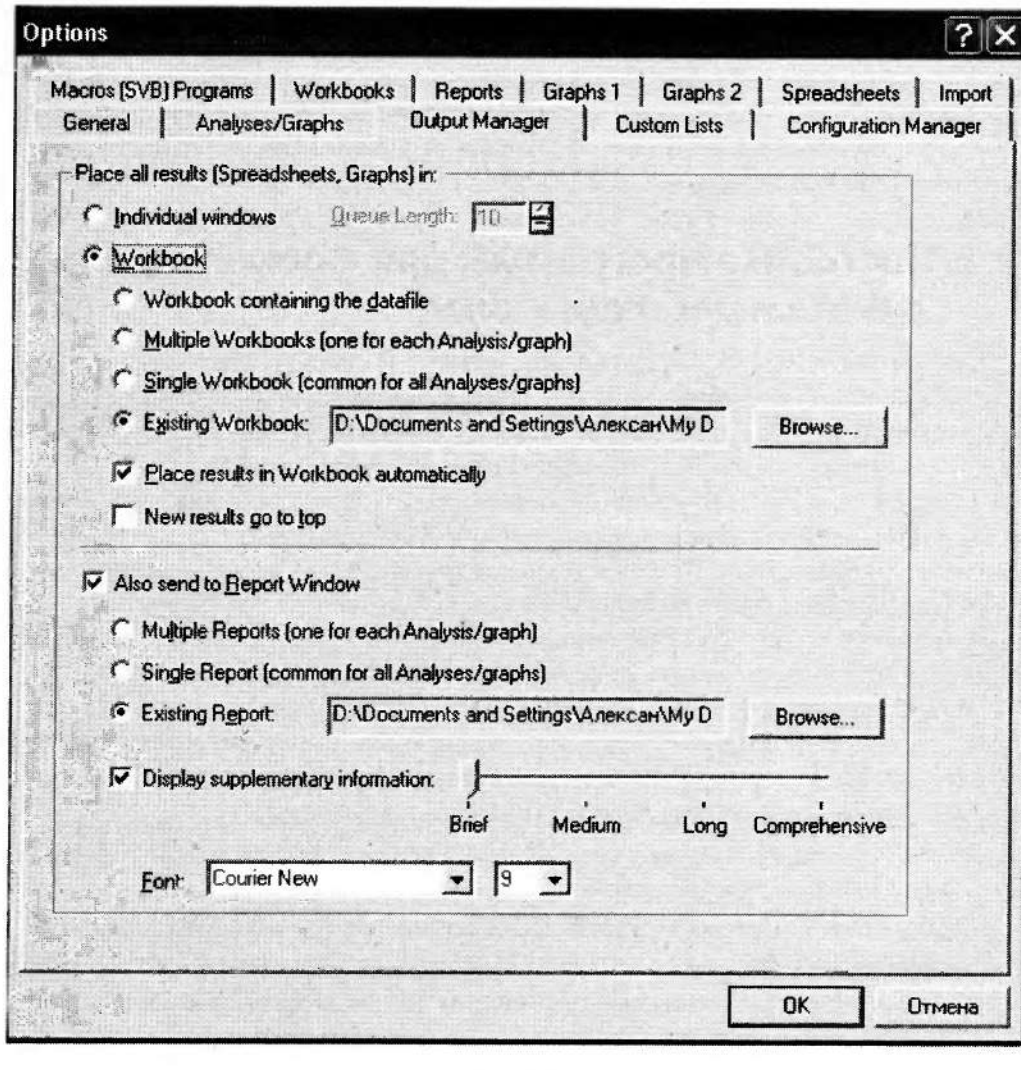


Рис. 2.1

- *Long*. Дополнительно к предыдущему включает длинные метки для переменных.
- *Comprehensive*. Дополнительно к предыдущему включает наиболее подробное описание данных (вся информация о двойной записи переменных и др.).

Также можно выбрать шрифт (*Font*) и размер шрифта дополнительной информации.

После установки нужных параметров в этом же диалоговом окне надо выделить вкладку **Reports** (рис. 2.2).

Если установить флажок на *Show object tree* (показывать дерево объектов), то в *отчете* будет отображена структура объектов. С помощью дерева объектов можно легко добавлять или удалять объекты в *отчет*.

Желательно установить флажок на *Save in RTF file format* (сохранять файл в формате *RTF*), так как формат **.rtf* поддерживают многие текстовые редакторы. В противном случае *отчет* будет сохраняться в формате **.str*.

Если выбрать опцию *Objects* (объекты), на просмотре и на печати таблицы будут представлены в виде объектов. При выборе опции *Full-sized Spreadsheets* (полноразмерные таблицы) таблицы на просмотре будут отражены в виде объектов, а на печати — в виде полных таблиц.

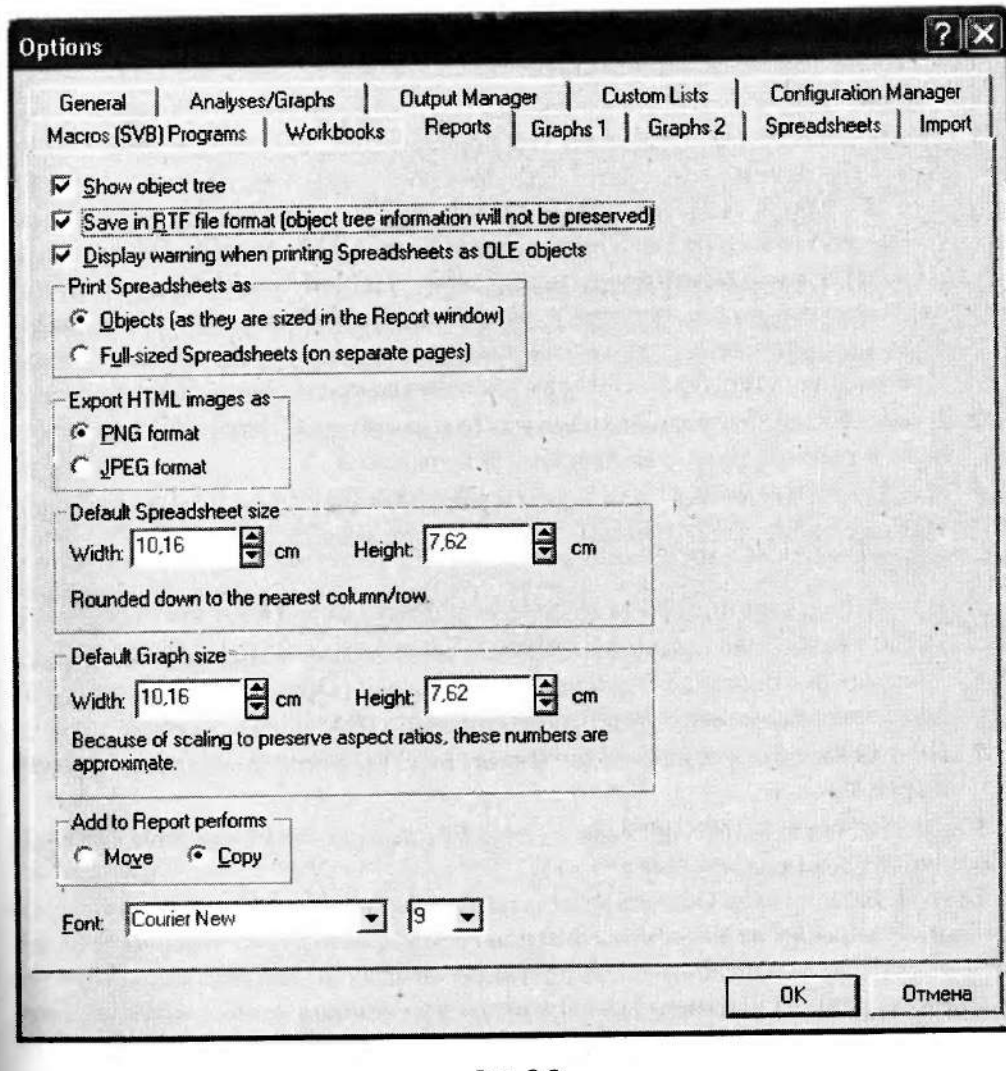


Рис. 2.2

В рамке **Add to Report performs** (выполняя добавление к *отчету*) можно выбрать способ заполнения *отчета*:

- *Move* (переместить) — после обработки таблиц с исходными данными все результаты анализа, графики и таблицы будут сразу помещены в окно *отчета*;
- *Copy* (копировать) — результаты будут выведены на экран в отдельных окнах, а в *отчет* помещены копии.

Также в этом окне можно задать параметры таблиц и графиков (высоту и ширину). Это удобно, когда нужно установить точные значения. К тому же можно выбрать шрифт, которым будет отображаться выводимая таблица.

Для автоматического формирования *отчета* или *рабочей книги* необходимо произвести следующую последовательность шагов:

1. Открыть файл исходных данных.
2. Создать файлы для *отчета* и *рабочей книги*. Для этого надо в меню **File** выбрать команду **Open** (открыть). В появившемся диалоговом окне **Create New Document** (см. рис. 1.4) выделите вкладку **Report** (*отчет*) или **Workbook** *(рабочая книга)*. Можно иначе — в меню **File** выбрать команду **Add to Report** (добавить в *отчет*) или **Add to Workbook** (добавить в *книгу*), а затем **New Report** (новый *отчет*) и **New Workbook** (новая *книга*). Откроются окна *отчета* и *книги*, например: **Report1** и **Workbook1**. Если предполагается результаты анализа внести в существующий *отчет* и *книгу*, надо открыть его, выбрав названия из предложенных перечней.
3. В меню **File** выбрать команду **Save as** (сохранить как) и сохранить файл *отчета* и *рабочей книги*, указав папку для хранения.
4. Выбрать в меню **File** команду **Output Manager**, в открывшемся окне **Options** на вкладке **Output Manager** произвести установки согласно рис. 2.1.
5. При выборе опции *Existing Workbook* откроется окно **Open**, в котором надо указать имя *книги*, например *Workbook1*, и нажать на кнопку *Открыть*.
6. При выборе опции *Existing Report* откроется окно **Open,** в котором надо указать имя *отчета*, например *Report1*, и нажать на кнопку *Открыть*.
7. При необходимости на вкладке **Report** надо произвести установки согласно рис. 2.2.

После составления *отчета* надо в меню **File** выбрать команду **Save** и сохранить отчет с расширением **.rtf (*.rts)*.

Если на рабочем окне **Options** на вкладке **Output Manager** (рис. 2.1) не выделена опция *Workbook* и не установлен флажок на *Also send to Report Window*, не будет происходить автоматическое формирование *отчета* и *рабочей книги*. *Отчет* и *рабочая книга* могут создаваться «вручную» при помощи команд **Add to Report** и **Add to Workbook** из верхнего меню или из меню **File**. Можно в автоматическом режиме создавать *отчет*, а «вручную» формировать *рабочую книгу* и наоборот.

2.3. Редактирование *отчета*

В пакете *STATISTICA* существуют различные способы редактирования *отчета*. При работе с *отчетом* можно редактировать его при помощи команд, находящихся в меню **Format**. Для этого надо открыть окно *отчета* и в меню **Format** выбрать нужную команду (рис. 2.3). Опишем эти команды.

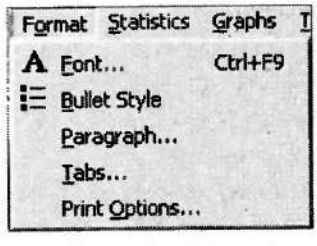


Рис. 2.3

Font (шрифт). В диалоговом окне команды можно выбрать шрифт, начертание, размер, атрибуты (зачеркнутый, подчеркнутый), цвет, набор символов.

Bullet Style (список). С помощью этой команды можно выбрать маркированный список.

Paragraph (абзац). В диалоговом окне команды можно выбрать отступы абзаца (слева, справа, отступ первой строки). Произвести выравнивание абзаца (по левому краю, по правому краю, по центру). Определить междустрочный интервал (одинарный, полуторный, двойной).

Tabs (табуляция). В диалоговом окне команды можно установить позиции табуляции.

Панель инструментов *отчета* напоминает панель инструментов *MS Word*. Она открывается автоматически, как только зашли в *отчет*, и позволяет редактировать *отчет*, не вызывая меню **Format**. На панели находятся основные команды форматирования текста.

Изменить ориентацию страницы (книжная, альбомная) можно в меню **File** командой **Print Setup**.

Глава 3
Графический анализ

3.1. Двухмерная графика

STATISTICA позволяет компактно описать данные, понять их структуру, провести классификацию, увидеть закономерности в хаосе случайных явлений. Вполне естественно, что многие явления, остающиеся за кадром, становятся отчетливыми, если для них найти соответствующее графическое представление, так как в некоторых случаях визуально оценить результат проще, чем численно.

Графические средства *STATISTICA* могут быть использованы в следующих целях [2]:

- для визуализации численных и текстовых значений непосредственно из электронных таблиц с исходными данными или таблиц с результатами анализа;
- для вывода результатов анализа в виде последовательностей графиков.

В программе реализовано большое многообразие графических представлений данных [10] и описать их в рамках данного пособия не представляется возможным. Поэтому коротко опишем основные из них.

В *STATISTICA* существуют различные способы доступа к графическим средствам:

- через верхнее меню, выбрав команду **Graphs** (графики);
- через контекстное меню, щелкнув правой кнопкой мыши на ячейке данных;
- при помощи панели инструментов **Graphs**, для ее вызова надо щелкнуть правой кнопкой мыши на панели инструментов и в появившемся контекстном меню выбрать пункт **Graphs** или выбрать меню **View** (вид) → **Toobars** (панели) → **Graphs**.

Прежде, чем ознакомиться с составом данной панели, отметим, что множество графиков в системе *STATISTICA* можно условно разделить на два класса:

- статистические графики;
- пользовательские (блоковые) графики.

Пользовательские графики дают возможность наглядно представить любые заданные пользователем комбинации значений из таблиц результатов или таблиц исходных данных (из строк, столбцов, из строк и столбцов, и (или) их частей). В отличие от пользовательских графиков, статистические графики — это заранее заданные представления данных.

На панели инструментов **Graphs** некоторые группы кнопок отделены друг от друга вертикальными вдавленными полосками. Это отделение *Stats 2D Graphs* (*статистических 2D графиков*), *Stats 3D Graphs* (*статистических 3D графиков*) и *Stats categorized Plots* (статистических категоризованных графиков) друг от друга.

Stats 2D Graphs — это визуальный анализ данных на плоскости, который осуществляется при помощи разнообразных гистограмм, диаграмм рассеяния, вероятностных графиков, линейных графиков, диаграмм диапазонов, диаграмм размахов, круговых диаграмм, столбчатых графиков, графиков последовательных значений и т.д.

Рассмотрим некоторые виды статистических *2D* графиков [6].

2D Histogramms являются графическими представлениями распределения частот выбранных переменных. Для каждого интервала (класса) рисуется столбец, высота которого пропорциональна частоте класса. Гистограмма наглядно показывает, какие значения или диапазоны значений исследуемой переменной являются наиболее частыми, насколько сильно они различаются, как сконцентрировано большинство наблюдений вокруг среднего, является распределение симметричным или нет, имеет ли оно моду или несколько мод. Различают несколько видов гистограмм.

2D Histogramms Regular (простые) представляет собой столбчатую диаграмму распределения частот для выбранной переменной (если выбрано более одной переменной, то для каждой из них будет построен отдельный график).

2D Histogramms Multiple (составные) изображают распределение частот для нескольких переменных на одном графике. Частоты для всех переменных откладываются по левой оси *Y*. Значения всех исследуемых переменных откладываются по одной оси *X*, что облегчает сравнение анализируемых переменных. Например, исследователя может заинтересовать динамика изменения веса студентов до и после сессии.

2D Histogramms Double-Y (с двойной осью *Y*). Гистограмму с двойной осью *Y* можно считать комбинацией двух по-разному масштабированных составных гистограмм. Для этой гистограммы можно выбрать две различные группы переменных. Для каждой из выбранных переменных будет изображено распределение частот, но частоты переменных из первого списка, называемого *Left Y* (левая ось *Y*), будут откладываться по левой оси *Y*, а частоты переменных из второго списка, называемого *Right Y* (правая ось *Y*), будут откладываться по правой оси *Y*. Имена всех переменных из двух списков будут внесены в условные обозначения и будут сопровождаться буквами *L* или *R*, обозначающими соответственно левую или правую ось *Y*. Этот график полезен для визуального сравнения распределений переменных с разными частотами.

2D Histogramms Hanging Bars (висячие столбцы). Гистограмма висячих столбцов является «наглядным критерием проверки на нормальность распределения», который помогает определить области распределения, где возникают расхождения между наблюдаемыми и ожидаемыми нормальными частотами. В то время как стандартным способом представления подогнанного к наблюдаемому распределению нормального распределения является наложение на гистограмму наиболее подходящей нормальной кривой, гистограмма висячих столбцов предлагает противоположный способ: столбцы, представляющие наблюдаемые частоты для последовательных диапазонов значений, «подвешиваются» к наиболее подходящей нормальной кривой. Если исследуемое распределение хорошо приближается к нормальной кривой, то нижние ребра всех столбцов должны образовать прямую горизонтальную линию.

Для построения гистограмм можно использовать кнопку со всплывающей подсказкой **2D Histogramms** на панели графиков или команды верхнего меню **Graphs → 2D Histogramms**. Откроется диалоговое окно **2D Histogramms** (рис. 3.1).

На вкладке **Advanced** (дополнительно) в поле **Graph type** (тип графика) указывается тип графика: *Regular; Multiple; Double-Y*.

В поле **Fit type** (тип подгонки) выбираются виды аппроксимирующих законов плотностей распределений: *Of* (выключить); *Normal; Beta; Exponential* и т.д.

В поле **Showing type** (тип показа) указываются форматы графиков: *Standard; Hanging Bars; Cumulative*. Последний формат дает графическое изображение накопленных частот.

В рамке **Intervals** (интервалы) производятся установки режимов категоризации.

В режиме *Integer mode* (целые числа), если не установлена галочка в поле **Auto**, программа округлит каждое значение выделенной переменной до целого числа и создаст одну категорию (или график в случае категоризованных графиков) для каждого целочисленного значения. При выборе этого метода кнопка **Change variable** (изменить переменную) позволит выбрать другую переменную. Если число целых категорий превзойдет 256, программа автоматически использует метод категоризации, включающий 16 категорий.

В поле ввода справа от режима *Categories* (категории) вводится необходимое число категорий. Программа разделит полный диапазон значений переменной на заданное число интервалов одинаковой длины (длина интервалов не будет целым числом).

После выбора опции *Boundaries* (границы) надо нажать кнопку **Specify Boundaries** (задать границы) и ввести список границ для выделенной переменной в появившемся диалоговом окне. Например, если ввести 1 3 4 9, то будут созданы 5 диапазонов значений выделенной переменной: $X <= 1$; $1 < X <= 3$; $3 < X <= 4$; $4 < X <= 9$; $X > 9$. Как видно из примера, интервалы могут иметь различную длину. Процедура работает, если в поле **Fit type** выбран режим *Off*.

Опцию *Codes* (коды) можно использовать, если переменная содержит коды, по которым нужно задать категории. После выбора этой опции надо нажать кнопку **Specify Codes** (задать коды) и ввести нужные коды в появившемся диалоговом окне. Процедура работает, если в поле **Fit type** выбран режим *Off*.

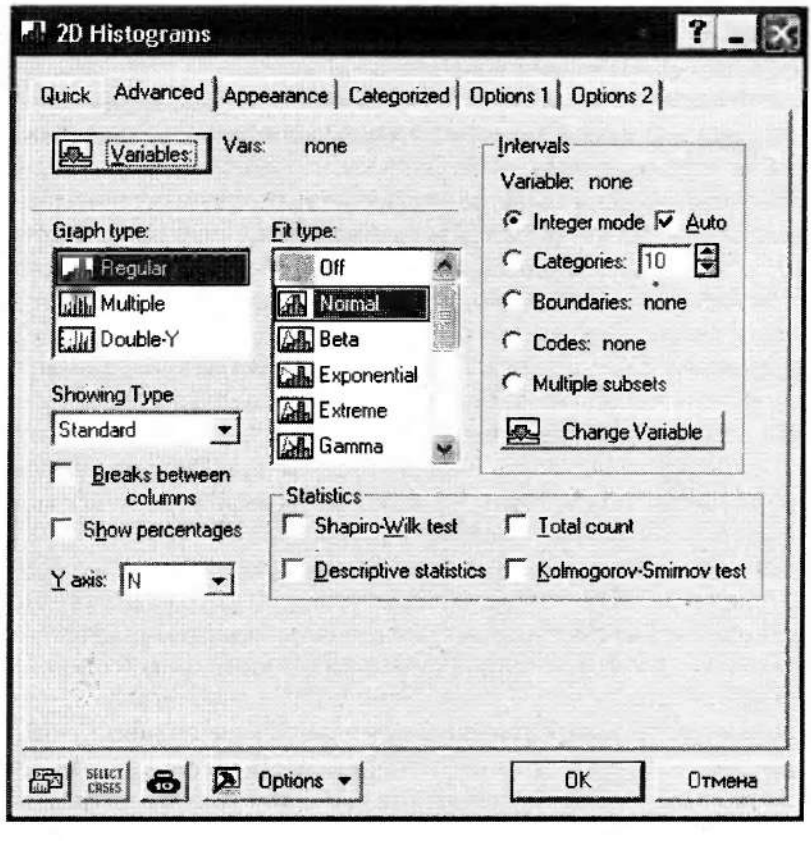


Рис. 3.1

После выбора метода **Multiple subsets** (сложные подгруппы) надо нажать кнопку **Specify subsets** (задание подгрупп) и в появившемся окне задать условия выбора. Этот метод позволяет использовать более одной переменной для определения групп.

В рамке **Statistics** выбираются критерии соответствия эмпирических распределений распределениям, приведенным в поле **Fit type**.

На вкладке **Appearance** (внешний вид) можно указать стиль графика и стиль документа.

В построенном программой графике можно изменить заголовок, название осей координат. Для этого надо выделить заголовок, щелкнуть на нем правой кнопкой мыши и в появившемся контекстном меню выбрать команду **Title Properties** (свойства заголовка). Аналогично щелчком правой кнопки мыши на любой из подписей осей координат можно вызвать их контекстное меню и произвести соответствующие изменения.

Есть и другой способ изменения обозначений графика. Двойным щелчком мыши в области графика откроется окно **All Options**, в котором можно изменить параметры графика. На вкладке **Graph Titles Text** можно изменить заголовок и формат заголовка диаграммы. Заголовки и формат заголовков осей можно изменить,

используя вкладку **Axis:Title**. По умолчанию выбрана ось *X*. В списке *Axis:* ее можно изменить на *Y left*, *Y right* или *Top*.

2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными *X* и *Y*. Данные изображаются точками в двумерном пространстве, где оси соответствуют переменным (*X* — горизонтальной, а *Y* — вертикальной оси). Если переменные сильно связаны, то множество точек данных принимает определенную форму. Подгонка функций к диаграммам рассеяния позволит увидеть зависимости между переменными. Если переменные не связаны, то точки образуют «облако рассеяния». В программе реализованы диаграммы рассеяния нескольких типов.

2D Scatterplots Regular (диаграммы рассеяния, простые) визуализируют зависимость между двумя переменными *X* и *Y*. Для уточнения типа зависимости можно поэкспериментировать с различными типами подгонки. Для этого нужно щелкнуть кнопкой **2D Scatterplots** внизу экрана. Открывшееся окно аналогично окну при построении нового графика. Но есть отличие. В первом случае свойства (тип подгонки, тип графика, переменные) будут применяться для уже построенного графика, а во втором — будет построен новый график.

2D Scatterplots Multiple (составные). В отличие от простой диаграммы рассеяния, на которой одна переменная представлена по горизонтальной, а вторая — по вертикальной оси, составная диаграмма рассеяния состоит из нескольких зависимостей и изображает несколько корреляций. Значения одной переменной (*X*) откладываются по горизонтальной оси, а по вертикальной — значения нескольких переменных (*Y*). Для каждой переменной *Y* используется разный цвет и вид точек, который указан в условных обозначениях, так что на графике можно отличить зависимости для различных переменных. Диаграмма рассеяния составного типа используется для сравнения структуры нескольких корреляционных зависимостей путем изображения их на одном графике, использующем один общий масштаб. Чтобы точки, соответствующие различным переменным по оси *Y*, не накладывались друг на друга, надо изменить вид маркеров (точек), соответствующих этим переменным. Для этого надо на любой из точек щелкнуть два раза левой кнопкой мыши и вызвать окно **General**, в котором нужно щелкнуть на кнопку **Markers** и в появившемся окне изменить размер, вид, цвет точек.

2D Scatterplots Double-Y (с двойной осью *Y*). Диаграмму рассеяния такого типа можно рассматривать как комбинацию двух составных диаграмм рассеяния для одной переменной *X* и двух различных наборов (списков) переменных *Y*. Для переменной *X* и каждой из переменных *Y* будет построена диаграмма рассеяния, но переменные из первого списка (*Left Y*) будут откладываться по левой оси *Y*, в то время как переменные из второго списка (*Right Y*), будут откладываться по правой оси *Y*. Имена всех переменных *Y* из двух списков будут включены в условные обозначения, сопровождаемые буквой *L* или *R*. Диаграммы рассеяния с двойной осью *Y* можно использовать для сравнения структуры нескольких корреляционных зависимостей путем изображения их на одном графике. При этом в силу независимости масштабов, используемых для двух списков переменных, этот график облегчает сравнение переменных, значения которых принадлежат разным диапазонам.

2D Scatterplots Frequency (частот). Программа подсчитывает частоты перекрывающихся точек. Размеры маркеров точек на графике соответствуют значениям частот. Имеет смысл использовать, когда хотя бы одна из переменных категориальная (измерена в номинальной шкале). Если переменные непрерывные и частоты равны 1, график совпадает с простой диаграммой рассеяния.

2D Scatterplots Bubble (пузырьков). Аналогична диаграмме частот, но должна быть назначена переменная весов.

2D Scatterplots Quartile (квантилей). На графиках квантилей изображается зависимость между квантилями двух переменных, позволяющая оценить сходство эмпирических распределений. Если точки данных попадают на линию регрессии, то можно сделать вывод, что две переменные имеют одинаковое распределение. По сути график квантилей — это изображение зависимости между функциями распределений переменных.

2D Scatterplots Voronoi (Вороного). Диаграмма этого вида является в большей степени аналитическим средством, чем средством графического представления данных. Программа разделяет пространство между точками данных, представленных координатами *X, Y* в двухмерном пространстве. Пространство между отдельными точками данных делится границами на такие области, каждая точка которых находится ближе к заключенной внутри точке данных, чем к любой другой соседней точке данных.

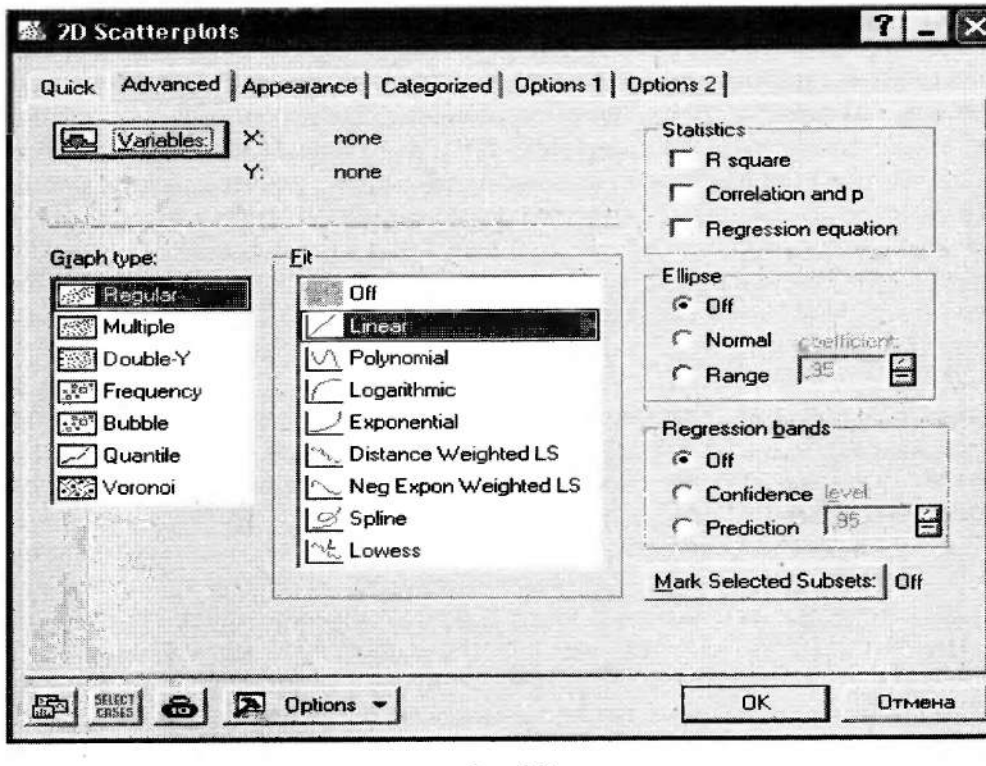


Рис. 3.2

Для построения диаграмм рассеяния можно использовать кнопку со всплывающей подсказкой **2D Scatterplots** на панели графиков или команды верхнего меню **Graphs → Graphs Scatterplots**. Откроется диалоговое окно.

2D Scatterplots (рис. 3.2). На вкладке **Advanced** в поле **Graph type** можно выбрать типы диаграмм: *Regular; Multiple; Double-Y; Frequency; Bubble; Quartile; Voronoi*. В поле **Fit type** можно осуществить подгонку функции, которая будет наложена на график. Возможен выбор следующих функций: *Linear; Polynomial; Logarithmic; Exponential; Distance Weighted LS; Neg Expon. Weighted LS; Spline; Lowess.*

В рамке **Elipse** опция *Normal* задает построение эллипса в предположении о нормальном распределении двумерной случайной величины (X,Y). Ориентация эллипса определяется знаком линейной корреляции между двумя переменными (более длинная ось эллипса накладывается на линию регрессии). Эллипс показывает прогнозируемый интервал для одного нового наблюдения при данных оценках параметров двумерного нормального распределения. Если число наблюдений мало, то эллипс может выйти за пределы области, показанной на графике.

Опция *Range* (размах) означает построение эллипса фиксированного размера. При этом длины его проекций на оси X и Y соответственно равны среднему (размах$*k$), где среднее и размах относятся к переменной X или Y, а k — текущее значение коэффициента, которое задается в поле **Coefficient**.

Опция *Regression bands* (границы регрессии) применяется для линейной или полиномиальной подгонки. Позволяет указать доверительные границы для выбранной линии регрессии. В поле **Level** (уровень) надо ввести значение вероятности того, что подогнанная линия попадет между доверительными границами.

В рамке **Statistics** выбираются статистические характеристики зависимости между переменными: *R square* (квадрат коэффициента корреляции); *Correlation and p* (коэффициент корреляции и уровень значимости p); *Regression equation* (уравнение регрессии).

2D Box Plots (графики ящика — диаграммы размаха). На диаграммах размаха диапазоны или характеристики распределения значений выбранной переменной (переменных) изображаются отдельно для групп наблюдений, заданных значениями категориальной (группирующей) переменной. Для каждой группы наблюдений вычисляется центральная тенденция (например, медиана или среднее) и вариационные статистики или статистики диапазона (например, квартили, стандартные ошибки или стандартные отклонения), и выбранные значения изображаются на диаграмме размаха выбранного типа. Программа вокруг средней точки рисует прямоугольник, представляющий выбранный диапазон разброса, и отрезок, также отражающий диапазон разброса, концы которого расположены вне прямоугольника. Для построения можно использовать кнопку со всплывающей подсказкой **2D Box Plots** на панели графиков или команды верхнего меню **Graphs → 2D Graphs → 2D Box Plots**. Откроется диалоговое окно модуля (рис. 3.3). В поле **Graph type** указывается тип графика диаграммы размаха: *Box Whiskers* (ящик с усами), *Whiskers* (усы), *Boxes* (ящики), *Columns* (столбцы), *High Low-Close* (верхние и нижние засечки). Можно выбрать один из двух форматов: *Regular* (простой), *Multiple* (составной).

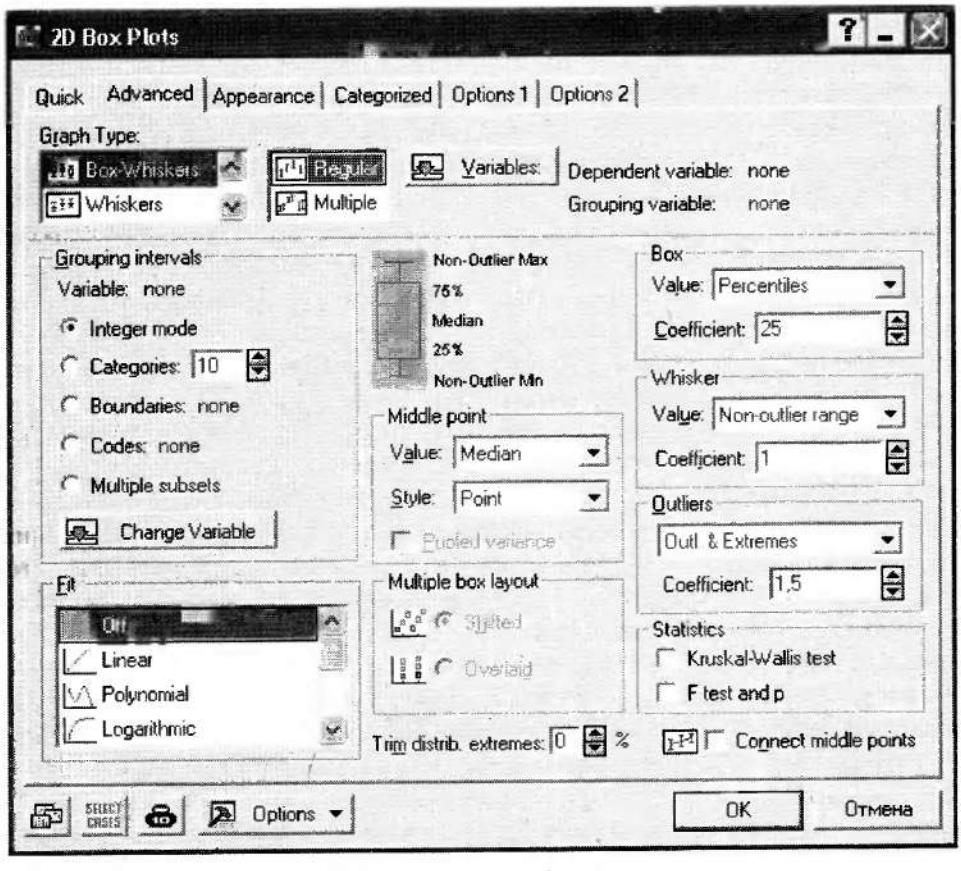


Рис. 3.3

Если нажать кнопку Variables, откроется диалоговое окно, в котором в правой части надо выбрать группирующую переменную (независимую), а в левой — зависимую. Если не предполагается при анализе использовать группирующую переменную, в правой части можно не указывать имя переменной.

В рамке **Grouping intervals** (группирование интервалов) указываются опции, часть которых устанавливается на вкладке **Categorized** (категоризация). Если выделить опцию *Codes*, кнопка **Change Variable** преобразуется в **Specify codes** (задать коды). Если выделить опцию *Multiple Subsets* (сложное подмножество), кнопка **Change Variable** преобразуется в **Specify Subsets** (задать подмножества). В рамке **Middle point** (средняя точка) выбирается статистика для оценки среднего: *Mean* (среднее) и *Median* (медиана) и стиль изображения среднего: *Point* (точка) и *Line* (линия).

В поле **Fit** можно осуществить подгонку функции к средним точкам диаграммы размаха путем выбора одной из заранее определенных функций или щелкнув мышью на кнопке **Custom Function** (пользовательский график) для самостоятельного задания функции, которая будет наложена на график. Возможен выбор следующих функций: *Linear; Polynomial; Logarithmic; Exponential; Distance Weighted LS; Neg Expon. Weighted LS; Spline; Lowess*. В рамках **Box** и **Whiskers** указываются статистики

для оценки разброса зависимой переменной: *Std.error* (стандартная ошибка), *Conf.Interval* (доверительный интервал), *Min–Max* (мини-макс), *Constant* (константа) для первого поля и *Std.dev.* (стандартное отклонение), *Std.error, Conf.Interval* для второго поля. Эти статистики соответствуют значению среднего — *Mean*. При изменении статистики оценки среднего на *Median* меняются статистики оценки разброса в указанных полях. В полях появятся соответственно оценки *Percentiles* (процентили) и *Non Outlier range* (без выбросов). Там же, в этих полях, указываются коэффициенты перед этими статистиками. В рамке **Outliers** (выбросы) задаются режимы обработки выбросов: *Outliers & Extremes* (выбросы и крайние точки); *Off* (выключить); *Outliers, Outl & Extremes*.

На *2D Line Plots* (линейных графиках) отдельные точки данных соединены линией. По оси *X* откладываются номера (имена) случаев, по оси *Y* — значения переменной. Это простой способ представления и исследования последовательностей значений. Выделяют несколько различных типов линейных графиков.

Regular (простые). Простые линейные графики используются для представления и исследования последовательностей значений (обычно, когда порядок значений является существенным). Кроме того, линейные графики применяются при построении графиков непрерывных функций, таких как функции подгонки или теоретические распределения. Пустая ячейка данных (т.е. пропущенные данные) «разрывает» линию.

Double-Y (с двойной осью *Y*). Линейный график с двойной осью *Y* можно рассматривать как комбинацию двух по-разному масштабированных составных линейных графиков. Для каждой выбранной переменной используется свой шаблон линии. Линейный график с двойной осью *Y* можно использовать для сравнения последовательностей значений нескольких переменных, накладывая их линейные представления на один график. В то же время в силу независимости шкал, используемых для двух осей, этот график может облегчить сопоставление переменных, трудно поддающихся сравнению (т.е. имеющих значения в разных диапазонах).

Multiple (составные). В отличие от простых линейных графиков, на которых представлена последовательность значений одной переменной, на составном линейном графике изображаются несколько последовательностей значений (переменных). Для каждой переменной используется и указывается в условных обозначениях свой шаблон и цвет линии. Этот тип линейных графиков используется для сравнения последовательностей значений нескольких переменных (или нескольких функций) путем изображения их на одном графике, использующем один общий масштаб (например, для сравнения нескольких одновременных экспериментальных процессов, социальных явлений, цен акций или товаров, форм кривых текущих характеристик и т.п.).

XY Trace (трассировочные). На трассировочных графиках сначала строится диаграмма рассеяния двух переменных, а затем отдельные точки данных соединяются линией (в порядке их считывания из файла данных). В этом смысле трассировочные графики визуализируют «путь» последовательного процесса (движение, изменение явления во времени и т.п.).

Aggregated (агрегированные). Агрегированными линейными графиками называются графики, которые изображают последовательность средних для последовательных подмножеств выбранной переменной. Можно выбрать число последовательных наблюдений, по которым будет вычислено среднее, а при необходимости диапазон значений в каждом подмножестве будет выделен значками типа отрезков. Агрегированные линейные графики используются для представления и исследования последовательностей большого числа значений.

Для построения линейных графиков надо выбрать кнопку со всплывающей подсказкой **2D Line Plots** или использовать команды **Graphs → 2D Graphs → Graphs Line Variables** (графики → *2D* графики → линейные графики переменных).

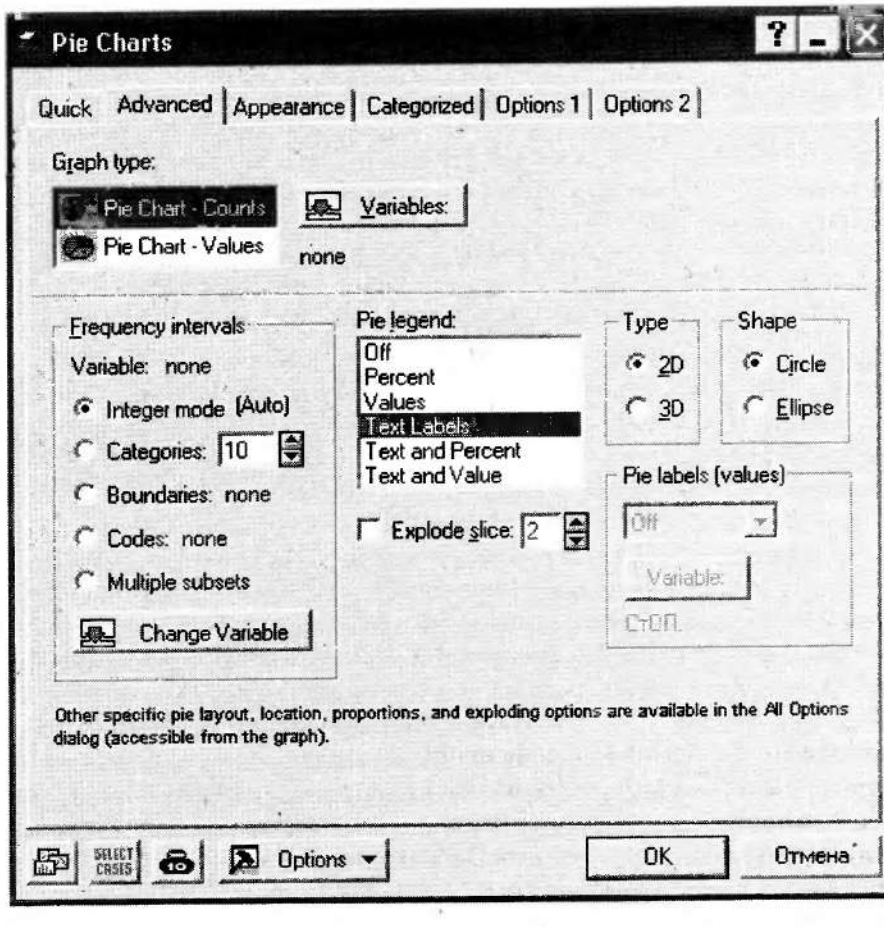


Рис. 3.4

2D Bar Column Plots (столбчатые диаграммы). На этой диаграмме последовательность значений представлена в виде столбцов (одному наблюдению соответствует один столбец). Если выбрано несколько переменных, то для каждой из них будет построен отдельный график. Можно построить составную диаграмму,

где все переменные будут отображены одновременно в виде групп столбцов (одна группа для каждого наблюдения). Для построения такой диаграммы надо нажать кнопку со всплывающей подсказкой **2D Bar/Column Plots** или воспользоваться командами **Graphs → 2D Graphs → 2D Bar/Column Plots (графики →** *2D* графики **→** графики строки/столбцы).

Pie Charts (круговая диаграмма). На графиках пропорции отдельных значений переменной *X* представлены в виде круговых секторов. Используются эти диаграммы для представления пропорций. Для построения диаграммы надо на панели графиков нажать кнопку со всплывающей подсказкой **Pie Charts** или в верхнем меню **Graphs** выбрать команды **Graphs** и **Pie Charts** (графики *2D* графики смешанные графики). Откроется диалоговое окно (рис. 3.4).

В поле **Graph type** указывается тип круговой диаграммы:

- *Pie Chart-Counts* (круговая диаграмма частоты), все значения выбранной переменной категоризуются по выбранному методу категоризации, а затем относительные частоты изображаются в виде круговых секторов пропорциональных размеров;
- *Pie Chart-Values* (круговая диаграмма значений), последовательность значений переменной изображается в виде последовательности круговых секторов, размер каждого сектора пропорционален значению переменной, одно наблюдение соответствует одному сектору.

В рамке **Frequency intervals** (частота интервалов) опции задаются аналогично ранее описанным в поле **Grouping intervals** диалогового окна **2D Box Plots** на рис. 3.3.

В рамке **Pie legend** (легенда круговой диаграммы) выбирается один из шести форматов: *Off* (нет); *Percent* (процент); *Values* (значение); *Text Labels* (текстовые метки); *Text and Percent*; *Text and Value*. Если задан тип графика *Pie Chart-Counts*, то при выборе формата *Text Labels* в условных обозначениях круговой диаграммы будут использованы описания категорий. Для формата *Values* используемые метки будут зависеть от выбора, сделанного в рамке **Pie Labels** (метки шрифта). Например, если в качестве меток выбраны имена наблюдений, то в качестве текстовых меток будут показаны реальные имена наблюдений, а не слова *Case 1*, *Case 2* и т.д. При одновременном выборе переменной в рамке **Pie Labels** и формата *Text Labels* программа будет использовать в условных обозначениях диаграммы текстовые значения выбранной переменной.

Меню команды **Graphs of Input Data** (графики входных данных) включает наиболее широко используемые типы статистических графиков. Оно не дает столько возможностей построения графика, сколько их существует в меню **2D Graphs**. Однако используя **Graphs of Input Data**, график можно построить гораздо быстрее, поскольку можно использовать контекстное меню; для построения нет необходимости выделять диапазон, так как график будет построен по столбцу, на одном из значений которого установлен курсор. Недостатком этих графиков является то, что далеко не все типы графиков, а только самые распространенные можно построить указанным способом. При построении графика таким способом следует учитывать и то, что бесполезно выделять определенные значения столбца или значения нескольких столбцов — график

все равно будет построен по всем значениям того столбца, со значений которого было начато выделение.

Построить же график выделенного диапазона (даже если выделены значения разных столбцов) можно, используя команду **Graphs of Block Data** (графики блока данных). Команды этого меню позволяют получить графическое представление значений из выделенной части (блока) таблицы результатов или таблицы исходных данных. Команды разделены горизонтальной чертой на две части. Графики нижней части называются *Custom Graphs* (пользовательские). Если в выделенном блоке содержатся несколько столбцов (переменных), то при использовании команд из верхней части графики для всех переменных будут расположены на одном листе; при выборе команд из нижней части графики для каждой переменной будут расположены на отдельных листах. Учитывая тот факт, что по типам графиков команды верхней и нижней частей идентичны, рассмотрим структуру команд пользовательских графиков.

Custom Graphs from Block by Column. При выделении значений из разных столбцов будет построено несколько диаграмм, каждая из которых станет соответствовать одному выделенному столбцу.

Custom Graphs from Block by Row. При выделении значений из разных строк будет построено несколько диаграмм, каждая из которых станет соответствовать одной строке.

Custom Graphs for Entire Column. График будет построен для целого столбца.

Custom Graphs from Entire Row. График будет построен для целой строки.

Заметим, что **Graphs of Block Data** и **Graphs of Input Data** доступны из верхнего меню **Graphs → Graphs of Input Data (Graphs of Block Data)** или через контекстное меню.

3.2. Средство «закрашивание»

Важным преимуществом диаграмм рассеяния является возможность находить «выбросы» (аномальные, нетипичные данные), которые влияют на значение коэффициента корреляции. Даже один выброс может значительно изменить коэффициент корреляции между двумя переменными. Средство **Brushing** (закрашивание) интерактивно удаляет выбросы, при этом можно непосредственно наблюдать за изменением аппроксимирующей функции или линии регрессии. На примере файла *Страны мира* рассмотрим основные принципы работы с закрашиванием.

Используя команды меню **Graphs 2D Graphs Scatterplots**, постройте диаграмму рассеивания для переменных *Нас.95*, *Нас.гор*. Из графика, приведенного на рис. 3.5, видно, что с увеличением численности населения доля городского населения убывает по зависимости, близкой к линейной (коэффициент корреляции равен — 0,81). Предположим, что точка, расположенная в нижней левой части плоскости, является выбросом.

Сначала определим, какой стране соответствует эта точка. Щелкните по кнопке с изображением прицела на панели инструментов (при подводе курсора мыши к ней высветится надпись **Brushing**). Откроется окно **Brushing** (рис. 3.6) на вкладке

Interactive (интерактивное). В рамке Action (операция) на вкладке **Normal** (нормальное) выделите опцию *Label* (метка). В рамке **Selection brush** (выбор закрашивания, чаще употребляется другое название инструмента «закрашивание» — кисть) выделите опцию *Point* (точка). Подведите «прицел» к точке и щелкните левой кнопкой мыши.

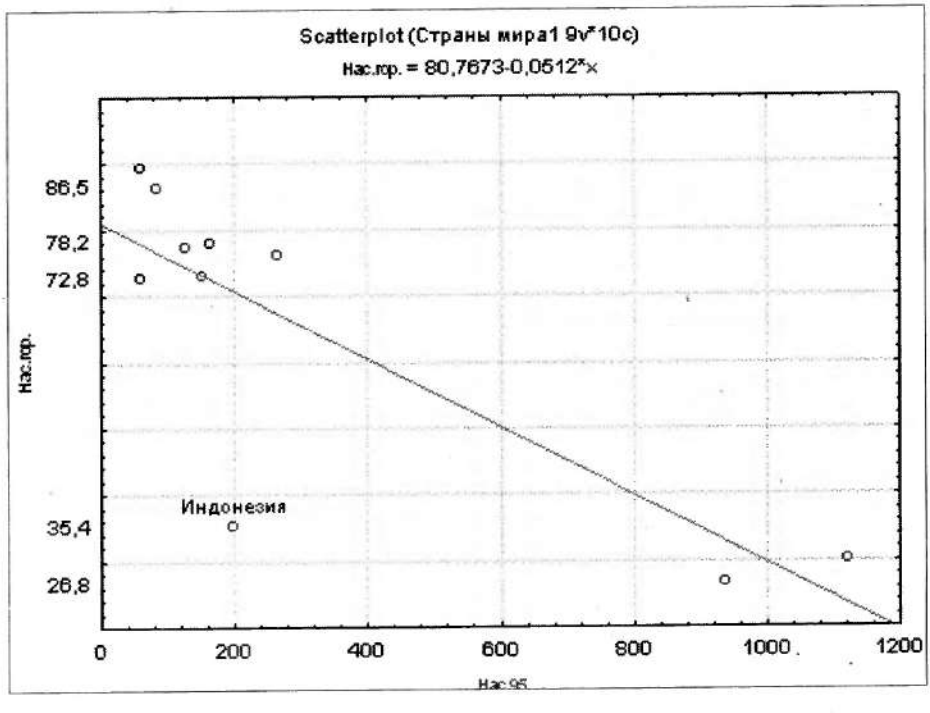


Рис. 3.5

Если выделен режим *Auto Update* (автообновление), то появится метка с обозначением имени страны *Индонезия*. Если режим *Auto Update* не выделен, то нажмите кнопку **Update** (обновить), которая расположена в верхней части окна. Если надо исключить эту точку из графика, на вкладке **Normal** выделите опцию *Turn OFF* (выключить), подведите «прицел» к точке, щелкните левой кнопкой мыши и нажмите кнопку **Update**. Точка исчезнет. Обратите внимание, что после этих действий появится новое уравнение регрессии над графиком, незначительно изменится положение прямой на плоскости, а корреляция существенно возрастет по абсолютной величине и составит — 0,95. Если произвести перечисленные действия, предварительно выделив опцию *Mark* (отметка), кружок, изображающий точку, будет закрашен.

При выборе кисти *Box* (блок), *Lasso* (лассо), *Slice X* (плоскость *X*), *Slice Y* операции, описанные ранее, можно одновременно произвести с группой точек.

Кистью *Box* точки выделяются на плоскости при помощи прямоугольного контура.

Кистью *Lasso* точки выделяются криволинейным контуром произвольной формы.

Кистью *Slice X* или *Slice Y* точки выделяются полосой определенной ширины и ориентации на плоскости. Если анализируется трехмерный график, становятся активными кисти *Cube* (куб) и *Slice Z*.

При выборе кисти *Box* или *Lasso* в нижней части окна активизируется опция *Draggable Brush* (подвижная кисть). Если ее выбрать, активными станут опции *Persist selection between brushes* (сохранение выбора кистей) и *Auto animate* (автоанимация) (рис. 3.7). Если выбрать опции *Slice X, Slice Y,* активными станут опции *Persist selection between brushes, Auto animate* и кнопка **Animate**. При помощи опции *Box* выделите на диаграмме рассеивания прямоугольную область и нажмите кнопку *Animate*, которая станет активной. Прямоугольный контур начнет перемещаться по плоскости, на которой изображена диаграмма рассеивания. Если была отмечена опция *Auto animate*, контур начнется перемещаться самостоятельно, сразу после появления на плоскости. Одновременно с процессом анимации появится окно **Animate** (рис. 3.8), на котором можно задать параметры анимации: направление и шаг движения по осям координат (*X Step, Y Step*); частоту перемещения (*Waiting time*); операции — остановить (*Pause*); сбросить (*Reset*); отменить анимацию (*Cancel*).

Если выбрана опция *Persist selection between brushes,* то при движении контура вид закрашенных точек не изменится. Если отметить опцию *Save Brushing settings*, то при закрытии окна **Brushing** и последующем открытии все сделанные установки будут сохранены.

Если выделить вкладку **Rev.** (обратное) (рис. 3.9), операции *Mark, Label, Turn OFF* поменяются на операции с противоположным смыслом действий — *Unmark* (исключить отметки), *Unlabel* (исключить метки), *Turn ON* (включить).

Если выделить вкладку **Toggle** (переключить), операции *Mark, Label, Turn OFF* поменяются на *Toggle mark* (переключить отметки), *Toggle label* (переключить метки), *Toggle ON/OFF* (включить/выключить) (рис. 3.10).

Кнопка **Quit** (выход) в верхней части окна предназначена для выхода из режима закрашивания кистью. Кнопка **Reset All** (отменить все) отменит все действия, произведенные с выделенными точками.

На вкладке **Extended** (другое) дополнительно к закрашиванию подмножества точек применяется их подсвечивание (рис. 3.11). При помощи опции *Selection by Sets* (выбор множеств) можно задать различные способы выбора множеств точек для подсвечивания:

- *Mark Points* (точки, предварительно отмеченные на вкладке **Interactive**);
- *Labeled Points* (точки с предварительно проставленными метками на вкладке **Interactive**);
- *Turned off Points* (точки, предварительно выключенные на вкладке **Interactive**);
- *Other Points* (другие точки, т.е. точки, не относящиеся ни к одному из перечисленных множеств).

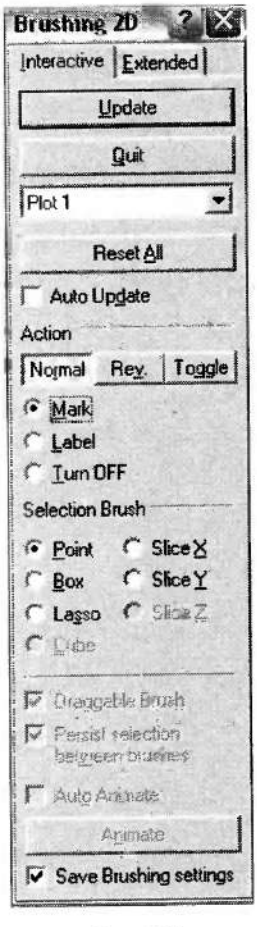


Рис. 3.6

Рис. 3.7

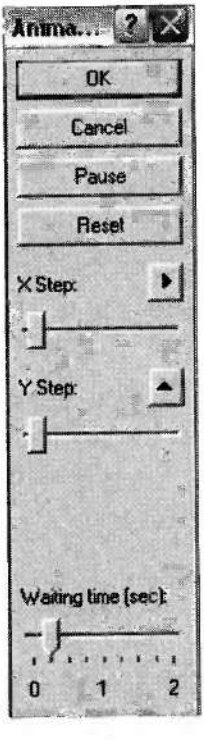


Рис. 3.8

Кнопка **Swap** (изменить) предназначена для изменения режима подсветки, т.е. если точки из выделенного множества подсвечены, то после нажатия на кнопку они станут не подсвеченными, и наоборот.

Опция *Selection by Ranges* (выбор диапазонов) осуществляет выбор множества точек заданием минимальных и максимальных значений координат точек по соответствующим осям.

Кнопка **Highlight** (подсветка) включает подсветку точек выделенного множества, кнопка **De-highlight** (без подсветки) выключает подсветку.

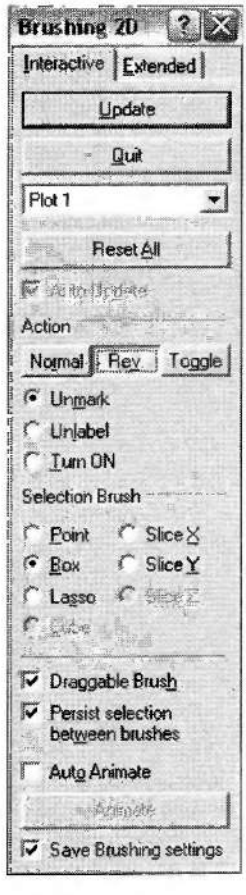


Рис. 3.9

Рис. 3.10

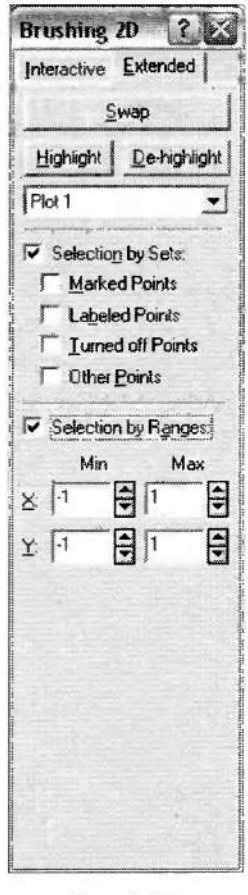


Рис. 3.11

3.3. Трехмерная графика

3D Graphs (трехмерные графики) позволяют анализировать данные в трехмерном пространстве. Например, строить трехмерное изображение последовательностей исходных данных для одной или нескольких выбранных переменных. Выбранные переменные можно представить по оси *Y*, последовательные наблюдения — по оси *X*, а значения переменных (для данного наблюдения) — по оси *Z*.

По своей идее **3D Graphs** сходны с составными линейными графиками, с тем лишь отличием, что для трехмерных графиков исходных данных ленты, линии, параллелепипеды и другие трехмерные представления значений каждой переменной не пересекаются (как в двухмерном графике), а «раздвигаются» в трехмерной перспективе.

3D Graphs применяются как для отображения данных, так и для аналитических исследований. Наиболее типичным приложением является наглядное представление имеющейся информации. Основное преимущество трехмерных представлений перед двухмерными составными линейными графиками заключается

в том, что для некоторых множеств данных при объемном изображении легче распознавать отдельные последовательности значений. При выборе подходящего угла зрения с помощью, например, интерактивного вращения линии графика не будут перекрываться или «попадать друг на друга», как часто бывает на составных линейных двухмерных графиках. Рассмотрим основные виды **3D Graphs**.

Raw Data Graphs (графики исходных данных) иллюстрирует соотношения между значениями переменных. Чтобы построить график, надо на панели инструментов **Graphs** выбрать кнопку со всплывающей подсказкой **3D Raw Data Plots.** Или через верхнее меню **Graphs → 3D Sequential Graphs → Raw Data Graphs** (графики → *3D* последовательные графики → графики исходных данных...). Откроется диалоговое окно **Raw Data Graphs**. На вкладке **Advanced** приведены возможные типы исходных графиков.

Columns (столбцы). Последовательный график представляет отдельные значения одной или нескольких переменных по оси *X* в виде серий трехмерных столбцов (параллелепипедов). Все серии отделены друг от друга промежутками вдоль оси *Y*. Высота каждого столбца по оси *Z* отвечает значению переменной, соответствующей данному наблюдению.

Blocks (блоки) представляет отдельные значения одной или нескольких переменных по оси *X* в виде серий «трехмерных блоков».

Ribbons (ленты) представляет отдельные значения одной или нескольких серий данных по оси *X* в виде серий «лент» в трехмерном пространстве.

Lines (линии) представляет отдельные значения одной или нескольких переменных по оси *X* в виде ряда непрерывных линий в трехмерном пространстве.

Spikes (всплески) представляет отдельные значения одной или нескольких переменных по оси *X* в виде серий «всплесков» (точек с перпендикулярами, опущенными на плоскость основания).

Contour/Discrete (линии уровня) представляет двухмерную проекцию *3D Rib-bons Graphs*. На этом графике каждая точка данных представлена в виде прямоугольной области. Значениям (или диапазону значений) точек данных соответствуют различные цвета или шаблоны (цветовые шаблоны описаны справа от графика). Значения из одной серии представлены по оси *X*, а сами серии откладываются по оси *Y*.

Surface (поверхность). На последовательном графике к точкам исходных данных подгоняется сглаженная сплайнами поверхность. Последовательные значения каждой серии откладываются по оси *X*, а сами последовательные серии представлены на оси *Y*.

Contours (контуры) представляет собой двухмерную проекцию сглаженной сплайнами поверхности, подогнанной к исходным данным. Последовательные значения каждой серии можно отложить по оси *X*, а сами последовательные серии представить на оси *Y*.

Часто может возникнуть проблема, когда одни части графика перекрывают другие. В некоторых случаях при очень большом числе наблюдений график почти невозможно понять, если смотреть на него под одним углом зрения. Поэтому при исследовании таких трехмерных графиков особенно полезно интерактивное вращение изображения на экране. Изменение угла зрения при отображении трех-

мерной диаграммы может оказаться эффективным средством для выявления некоторой структуры, которая видна только при определенном повороте графика. В трехмерной графике реализована возможность вращения графиков. Чтобы повернуть график, на панели инструментов **Graph Tools** надо нажать кнопку с изображением монитора. Появится окно. Перемещая ползунок в посолах прокрутки *Viewpoint distance/perspective (приблизить/удалить)*, *Vertical angle (вертикальный угол)* и *Horizontal angle (горизонтальный угол)*, можно выбрать наиболее удобный угол зрения.

Bivariate Histograms (гистограммы двух переменных) используются для визуализации табулированных значений двух переменных или для визуализации таблиц сопряженности двух переменных. Их можно рассматривать как сочетание двух простых гистограмм (т.е. гистограмм одной переменной), соединенных таким образом, чтобы можно было исследовать частоты совместного появления значений двух переменных. Распределение частот, изображенное на трехмерных гистограммах вызывает интерес по двум причинам:

- по форме распределения можно сделать вывод о природе исследуемой переменной, так если распределение бимодально, то можно предположить, что выборка не является однородной и состоит из наблюдений, принадлежащих двум совокупностям, которые возможно нормально распределены;
- многие статистики основаны на определенных предположениях о распределениях анализируемых переменных. Гистограммы двух переменных помогают проверить правильность этих предположений.

Чтобы построить такой график, на панели инструментов **Graphs** надо выбрать кнопку со всплывающей подсказкой **3D Bivariate Histograms.** Или можно открыть верхнее меню **Graphs**, в котором нужно выбрать команду **3D Sequential Graphs** (*3D* последовательные графики) и далее – команду **Bivariate Histograms**.

Процедуры сглаживания для **3D Bivariate Histograms** позволяют подгонять поверхности к трехмерным изображениям данных частот двух переменных. Так, например, каждая трехмерная гистограмма может быть превращена в сглаженную поверхность. Это представление нецелесообразно использовать для простых категоризованных данных. Для сглаживания надо щелкнуть дважды левой кнопкой мыши в области графика. Появится окно **All Options**, в котором можно изменить параметры графика. Например, выберите вкладку **Graph Layout**, в поле **Types** выделите **Surface** и нажмите **OK.**

3D Range Plots (*3D* диаграммы диапазонов) подобно статистическим *2D* диаграммам диапазонов, отображают диапазоны значений или столбцы ошибок, соответствующих определенным точкам данных. Диапазоны или столбцы ошибок не вычисляются по данным, а определяются исходными значениями выбранных переменных. Для каждого наблюдения строится один диапазон или столбец ошибок. Переменные диапазона можно понимать как абсолютные значения или как значения, отвечающие отклонениям от средней точки. На графике можно представить одну или несколько переменных. В основном диаграммы диапазонов используются для изображения:

- диапазонов значений для отдельных элементов анализа (наблюдений, выборок и т.д.);
- вариации значений в отдельных группах или выборках, когда величины вариации получены при независимых измерениях.

Иначе более целесообразно использовать *3D* диаграммы размаха, которые вычисляют вариацию для выборок, представленных на графике.

Основное различие между диаграммами диапазонов и диаграммами размаха состоит в том, что на диаграммах диапазонов все значения, определяющие диапазоны («средние точки», минимум и максимум), не вычисляются по данным, а являются исходными значениями переменных.

3D диаграммы диапазонов бывают различных типов: *Point Ranges* (точечные); *Border-style Ranges* (граничные), *Error Bar-style Ranges* (диапазоны ошибок); *Double Ribbon Ranges* (диапазоны двойных лент); *Flying Boxes* (летящие ящики); *Flying Blocks* (летящие блоки).

Проиллюстрируем построение точечной диаграммы диапазонов, используя файл, изображенный на рис. 3.12. В файле приведены значения температуры воздуха в различных городах: минимальная и максимальная летом, а также средняя в июле. На панели инструментов **Graphs** выберите кнопку со всплывающей подсказкой **3D Range Plots.** Или можно открыть верхнее меню: **Graphs** (графики) → **3D Sequential Graphs 3D** (последовательные графики) → **Range Plots** (диапазон графиков...). В открывшемся окне на вкладке **Quick** или на вкладке **Advanced** укажите тип диаграммы **Point Ranges и** нажмите кнопку **Variables.** В поле **Mode** (способ) выделите опцию *Absolute. В* списке **Mid-Point** выделите переменную *t в июле,* в списке **Max** — переменную *Макс. t летом,* а в списке **Мин** — переменную *Мин t летом.* Нажмите **OK.** Программа построит диаграмму, которая изображена на рис 3.13.

	1 Мин. t летом	2 t в июле	3 Макс. t летом
Краснодар	20	30	45
Москва	15	26	36
Воронеж	17	28	39
Волгоград	18	32	38
Сочи	25	40	47
Новосибирск	12	26	29

Рис. 3.12

Подобно статистическим *2D* диаграммам размаха на **3D Box Plot** (*3D* диаграммах размаха) диапазоны значений выбранной переменной строятся отдельно для групп наблюдений, определяемых значениями категориальной (группирующей) переменной. Центральная тенденция (например, медиана, среднее) и диапазон — вариационные статистики (например, квартили, стандартные ошибки, стандартные отклонения) вычисляются для каждой группы наблюдений. Стиль

изображения определяется типом графика: *Point Ranges* (точечные); *Border-style Ranges* (граничные), *Error Bar-style Ranges* (диапазоны ошибок); *Double Ribbon Ranges* (диапазоны двойных лент); *Flying Boxes* (летящие ящики); *Flying Blocks* (летящие блоки).

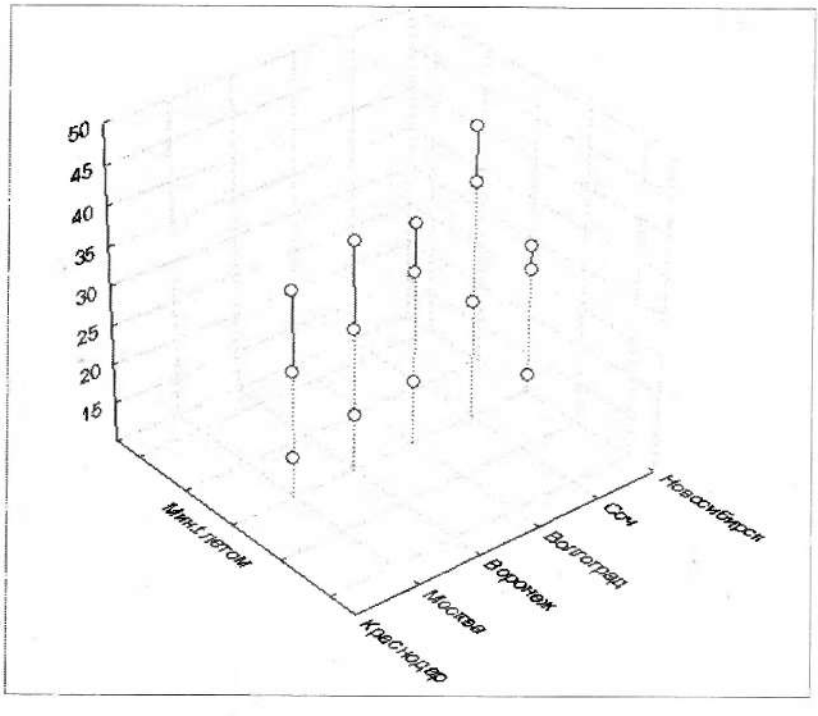


Рис. 3.13

Как правило, диаграммы размаха используются в двух случаях:

- для изображения диапазонов значений для отдельных наблюдений или выборок;
- для изображения вариации значений в отдельных группах или выборках. Например, диаграммы размаха, изображающие медиану или среднее для каждой выборки в виде точки внутри «летящего ящика», а также изображающие стандартные ошибки или квартильный размах в виде «летящих ящиков».

Когда на графике нужно представить только одну переменную, обычно достаточно воспользоваться *2D* диаграммой размаха. На этом графике также можно представить несколько переменных (путем сдвига изображений отдельных «ящиков» так, что для каждого наблюдения будет изображено последовательно столько «ящиков», сколько переменных используется для анализа). Тем не менее, для представления нескольких переменных на одном графике более подходящей является *3D* диаграмма размаха, так как она не «разбивает» строки пиктограмм для каждой переменной.

Для построения *3D* диаграмм размаха надо на панели инструментов **Graphs** щелкнуть кнопку со всплывающей подсказкой **3D Box Plots**. Можно также

использовать верхнее меню: **Graphs → 3D Sequential Graphs 3D → 3D Box Plots**. В открывшемся окне на вкладке **Quick** нажать кнопку **Variables.** В списке *Dependent Variable* выбрать зависимые переменные, в списке *Grouping Variable* обязательно указать группирующую переменную. Нажать на **OK**. В поле **Middle point** произвести соответствующие установки.

3D XYZ Graphs (трехмерные графики) представляют собой наиболее простой тип трехмерных зависимостей. Как правило, они используются для визуализации связей между непрерывными переменными. Хотя можно найти различные применения трехмерных графиков, тем не менее, их основное преимущество состоит в наглядном представлении сложных взаимосвязей между несколькими переменными. В программе реализованы различные типы трехмерных графиков.

Ограничимся описанием трехмерных диаграмм рассеяния (*scatter plots*) и категоризованных графиков (*catagorized 3D Plots*).

Scatterplot (диаграмма рассеяния) отражает взаимосвязь между тремя или более переменными в трехмерном пространстве, при этом каждой точке соответствует тройка координат *X*, *Y* и *Z*. Если выбрано более одной переменной *Z*, то будет построено несколько диаграмм рассеяния для различных наборов данных (соответствующих нескольким переменным *Z*), которые будут маркированы разными значками.

Space Plot (пространственный график) позволяет реализовать различные способы представления диаграмм рассеяния. Для этого предусмотрена возможность расположения плоскости *XY* на выбранном пользовательском уровне вертикальной оси *Z* (которая проходит через середину плоскости). Рекомендуется сопоставлять данные отдельным осям на графике таким образом, чтобы переменную, структуру связей которой необходимо выделить, обозначить как *Z*. Тогда, перемещая плоскость *XY* вдоль оси *Z* и интерактивно вращая изображение, можно попробовать найти такой уровень *Z*, на котором изменяется структура связей между *X* и *Y* (или *X*,*Y* и *Z*). Если ожидаемое изменение структуры слишком сложно для его исследования в одном «сечении», можно воспользоваться спектральным графиком, который позволит наблюдать несколько сечений.

Spectral Plot (спектральная диаграмма). Первоначально этот тип графиков применялся в спектральном анализе для исследования нестационарных временных рядов, например, речевых сигналов. На горизонтальных осях можно откладывать частоты спектра и последовательные временные интервалы, а на оси *Z* — спектральные плоскости для каждого интервала. Трехмерное пространство разделено на области, в которых данные «сжаты» в соответствующие спектральные плоскости. Для построения функциональных зависимостей (как в спектральном анализе) необходимо упорядочить данные таким образом, чтобы переменная *Y* содержала категоризующую информацию (т.е. была группирующей переменной). Спектральные диаграммы имеют явные преимущества перед обычными диаграммами рассеяния, когда необходимо исследовать, каким образом изменяется взаимосвязь между двумя переменными при различных значениях третьей переменной.

Значения переменных *X* и *Z* интерпретируются как координаты *X* и *Z* каждой точки, а значения переменной *Y* разделены на равноотстоящие группы, соответствующие положениям последовательных спектральных плоскостей.

Deviation Plot (диаграмма отклонений). На этом типе графиков точки данных (заданные координатами *X*, *Y* и *Z*) представлены в виде «отклонений» от определенного базового уровня на оси *Z*. Диаграммы отклонений похожи на пространственные графики. Однако на них в отличие от последних «плоскость отклонений» «невидима» и не обозначена положением плоскости *XY* (эти оси здесь всегда находятся в стандартном нижнем положении). С помощью диаграммы отклонений можно исследовать природу трехмерных наборов данных, изображая их в виде отклонений от произвольного (горизонтального) уровня. Как упоминалось ранее, такой метод «сечения» может выявить динамические связи между исследуемыми переменными.

Для построения трехмерной диаграммы рассеяния надо на панели инструментов **Graphs** щелкнуть кнопкой с всплывающей подсказкой *3D Scatterplots*. Можно также использовать верхнее меню. Для этого нужно выбрать пункт верхнего меню **Graphs → 3D XYZ Graphs → 3D Scatterplots**. В открывшемся окне на вкладке нажать кнопку **Variables**. В появившемся окне выбрать в списках *X*, *Y* по одной переменной, а в списке *Z* — одну или более переменных. В списке *Graph type* выбрать тип графика *(Scatterplot, Space Plot, Spectral Plot, Deviation Plot)*. Если выбрали тип *Spectral*, то в поле **Number of Planes** нужно указать число плоскостей. Если выбрали тип *Space Plot* или *Deviation Plot*, то в поле **Deviation level** необходимо установить свой пользовательский уровень плоскости *XY*, например, 100. Для определения коэффициента множественной корреляции нужно установить флажок в поле **Multiple r(z/xy), p** и нажать **OK**.

Categorized 3D Plots (категоризованные трехмерные графики) являются одним из наиболее мощных аналитических методов исследования. Основной идеей метода является разделение данных на группы для сравнения структуры получившихся подмножеств. Методы категоризации широко применяются как в разведочном анализе данных, так и при проверке гипотез и известны под разными названиями (классификация, группировка, категоризация, разбиение, расслоение и пр.). Для количественного описания различий между группами наблюдений разработаны многочисленные вычислительные методы, основанные на группировке данных (например, дисперсионный анализ). Однако графические средства (как категоризованные графики) дают особые преимущества и позволяют выявить закономерности, которые трудно поддаются количественному описанию и которые весьма сложно обнаружить с помощью вычислительных процедур (например, сложные взаимосвязи, исключения или аномалии). В этих случаях графические методы предоставляют уникальные возможности многомерного аналитического исследования. На категоризованном *3D графике* поверхности строятся сглаживанием или задаются пользовательским математическим выражением по категоризованным данным, т.е. по подмножествам данных. Причем все они изображаются в одном графическом окне, что дает возможность сравнивать эти подмножества (категории).

Для построения *Categorized 3D Plots* надо воспользоваться кнопкой с всплывающей подсказкой на панели инструментов **Graphs Categorized 3D Plots** или при помощи верхнего меню: **Graphs → 3D XYZ Graphs → Categorized XYZ Plots**.

Глава 4
Основные статистики

4.1. Описательные статистики

Модуль **Descriptive statistics** (описательные статистики) объединяет процедуры, наиболее часто используемые на начальном этапе обработки данных, когда выясняется структура, определяются зависимости между данными [2]. Статистики, используемые в данном модуле, в основном очень просты. Применение тех или иных статистик определяется использованием шкал, в которых произведено измерение признаков исследуемых объектов.

Номинальная шкала (шкала наименований) устанавливает принадлежность объекта измерения к некоторому классу.

Порядковая шкала осуществляет ранжирование объектов измерения, но не определяет расстояние между ними.

Интервальная шкала определяет расстояние между объектами, но начало отсчета и единица измерения выбираются произвольно исследователем.

Относительная шкала определяет расстояние между объектами при фиксированном начале отсчета, но произвольном масштабе измерения.

Абсолютная шкала определяет расстояние между объектами при фиксированном начале отсчета и фиксированном масштабе измерения.

Первые две шкалы применяются для измерения качественных переменных (признаков), остальные три — для измерения количественных переменных.

Mean (среднее арифметическое) — показывает центральное положение (центр) переменной и рассматривается совместно с доверительным интервалом.

Доверительный интервал представляет интервал значений вокруг оценки, где с данным уровнем доверия находится «истинное» (неизвестное) среднее генеральной совокупности. Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p = 0,95 равны соответственно 19 и 27, то можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее совокупности. Если установить больший уровень доверия, то интервал станет шире, поэтому возрастает вероятность, с которой он «накрывает» неизвестное среднее, и наоборот. Хорошо известно, например, что чем «неопределенней» прогноз погоды (т.е. шире доверительный интервал), тем вероятнее он будет точным. Заметим, что ширина доверительного интервала зависит от объема (*valid n*) выборки, а также от разброса (изменчивости) данных. Увеличение размера выборки делает оценку среднего более надежной. Увеличение разброса наблюдаемых значений уменьшает надежность оценки. Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Если это предположение не выполнено, то оценка может оказаться плохой, особенно для малых выборок. При увеличении объема выборки (скажем, до 100 или более) качество оценки улучшается и без предположения нормальности выборки.

Квантиль, соответствующая вероятности p, это значение переменной, ниже которой находится p-*я* часть (доля) выборки. Квантили, соответствующие вероятностям 0,25 и 0,75, называются соответственно *Lower & upper quartiles* (*нижней* и *верхней квартилью*; кварта — четверть).

Альтернативной оценкой среднего являются *median* (медиана) и *mode* (мода).

Медиана — это квантиль, соответствующая вероятности 0,5, т.е. это значение, которое разбивает выборку на две равные части по количеству элементов. Одна половина наблюдений лежит ниже медианы, вторая половина — выше. Если число наблюдений в выборке четно, то медиана вычисляется как среднее двух средних значений. Нижняя квартиль, медиана, верхняя квартиль делят выборку на 4 равные части. Как правило, используется для оценки среднего, если переменная измерена в порядковой шкале.

Мода — это значение переменной, соответствующее наибольшей частоте появления переменной в выборке. Как правило, используется для оценки среднего, если переменная измерена в номинальной или порядковой шкале.

Std.dev. (стандартное отклонение) — это корень квадратный из суммы квадратов отклонений значений переменной от среднего значения, деленное на $n - 1$.

Std.err.of mean (стандартная ошибка среднего) — это стандартное отклонение, деленное на корень квадратный из объема выборки.

Varience (коэффициент вариации) — это отношение стандартного отклонения к среднему.

Minimum (минимум) или *Maximum* (максимум) — это соответственно минимальное или максимальное значение выборки.

Range (размах) — это разность между максимальным и минимальным значениями выборки.

Quartiles range (квартильный размах) равен разности значений верхней и нижней квартилей, т.е. это интервал, содержащий медиану, в который попадает 50% выборки.

Skewness (асимметрия) — это мера симметричности распределения. Если распределение симметрично, то асимметрия равна нулю, если асимметрия существенно отличается от 0, то распределение несимметрично. Нормальное и равномерное распределения абсолютно симметричны. Асимметрия распределения с длинным правым хвостом положительна. Если распределение имеет длинный левый хвост, то его асимметрия отрицательна.

Kurtosis (эксцесс) — мера остроты пика распределения. Если распределение нормальное, то эксцесс равен 0. Если эксцесс положителен, то пик заострен, если отрицателен, то пик закруглен.

Для запуска программы в верхнем меню **Statistics** надо выбрать команду **Basic Statistic Tables** (основные статистики/таблицы). В появившемся меню надо выбрать команду **Descriptive statistics** (описательные статистики). Для выбора переменной, описательные статистики которой нас интересуют, надо нажать кнопку **Variables** и в открывшемся окне щелкнуть на имени переменной (переменных). Для просмотра результатов надо нажать кнопку **Summary**. **Descriptive statistics**. Откроется таблица с основными статистиками. Если нас интересуют другие статистики, необходимо указать их на вкладке **Advanced**, установив флажки напротив соответствующих статистик. При помощи кнопки **Select all stats** можно выбрать все статистики. Для анализа разброса данных предусмотрены графики *Box & Whisker* (ящики с усами), доступные на вкладке **Quick**.

Важным способом «описания» переменной является форма ее распределения, которая показывает, с какой частотой значения переменной попадают в определенные интервалы. Эти интервалы, называемые интервалами группировки, выбираются исследователем. Обычно исследователя интересует, насколько точно распределение можно аппроксимировать каким-либо стандартным распределением, например, нормальным. Простые описательные статистики дают об этом некоторую информацию. Более точную информацию о соответствии закона распределения нормальному можно получить с помощью критериев нормальности. Вкладка **Normality** предназначена для исследования возможности аппроксимации эмпирического закона распределения нормальным законом.

Если установить флажок на *Number of intervals* переменная воспринимается программой как непрерывная случайная величина и можно указать число интервалов разбиения диапазона ее изменения для построения гистограммы или *Frequency tables* (таблицы частот). При этом можно указать критерии соответствия эмпирического распределения нормальному закону *(Kolmogorof-Smirnof & Liliefors test for normality или Shapiro-Wilk's W test)*.

Если переменная является дискретной, то гистограмма визуализирует количественное соотношение различных значений переменной. Так, если установить флажок на *Integer intervals*, переменная воспринимается программой как дискретная случайная величина и число интервалов разбиения диапазона ее изменения определяется как число различных значений переменной.

При помощи кнопок в нижней части рабочего окна можно строить различные гистограммы.

Вкладка **Prob. & Scatterplots** предоставляет широкий набор способов графического исследования переменных.

Вкладка **Categ.plots** предназначена для графического исследования переменной с предварительным проведением категоризации (разбиения на различные подмножества).

Построим категоризованные гистограммы для данных из файла **Turtles** (черепахи) из библиотеки **Examples**, в котором приведены различные параметры измерений черепах: *LENGTH* (длина); *WIDTH* (ширина); *HEIGHT* (высота) и т.д. Построим гистограммы распределения частот длины черепах отдельно для особей мужского и женского пола. На вкладке **Categ.plots** нажмите кнопку **Variables** и выберите переменную *LENGTH*. Нажмите кнопку **Categorized histograms** и в открывшемся окне выделите группирующую переменную (можно выделять несколько группирующих переменных) *GENDER*. Нажмите **OK**. Из построенных гистограмм (рис. 4.1) видно, что для особей женского пола (*GENDER 2*) закон распределения длины тела более соответствует нормальному, чем для особей мужского пола (*GENDER 1*).

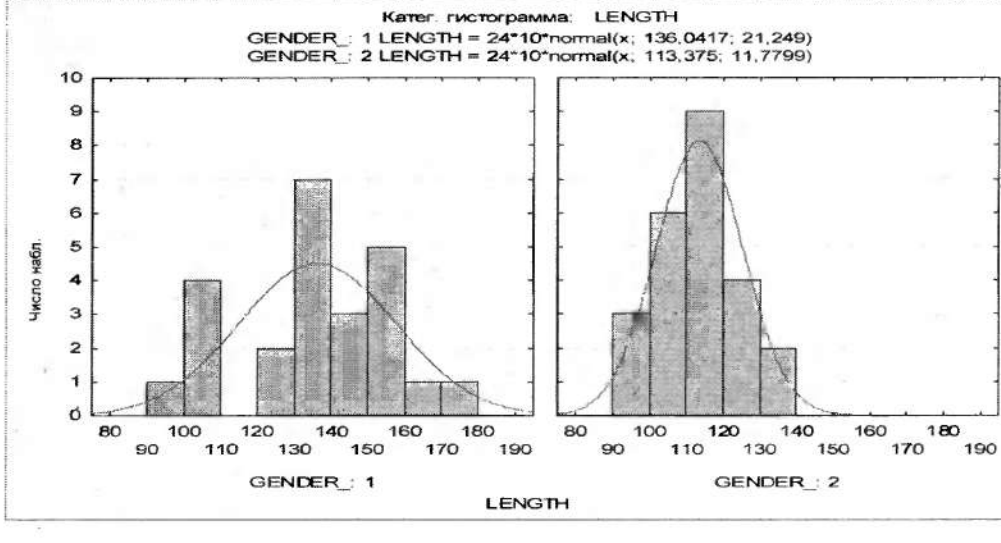


Рис. 4.1

Вкладка **Options** позволяет изменить некоторые параметры статистического исследования переменной.

4.2. Корреляционная матрица

Между переменными (случайными величинами) может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. Но между переменными может существовать и связь другого рода, проявляющаяся в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Такую связь называют стохастической. Она появляется в том случае, когда имеются общие случайные факторы, влияющие на обе переменные. В качестве меры зависимости между переменными используется коэффициент корреляции, который изменяется в пределах от −1 до +1. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения другой убывают. Если переменные независимы, то коэффициент корреляции равен 0 (обратное утверждение верно только для переменных, имеющих нормальное распределение). Но если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Чем ближе значение r к 1, тем зависимость сильнее. Коэффициент корреляции достигает своих предельных значений +1 или −1, тогда и только тогда, когда зависимость между переменными линейная. В модуле **Descriptive statistics** вычисляется коэффициент корреляции Пирсона, в предположении, что переменные измерены, как минимум, в интервальной шкале. Некоторые другие коэффициенты корреляции (например, корреляция Спирмена или тау Кендала) могут быть вычислены для более слабых шкал.

Принято считать, что при $|r| \le 0{,}25$ — корреляция слабая, $0{,}25 < |r| \le 0{,}75$ — умеренная, при $|r| > 0{,}75$ — сильная [12]. Сильная корреляция означает, что связь между переменными может быть близкой к линейной, но может быть явно нелинейной. В этом случае требуются дополнительные статистические исследования характера зависимости с применением процедур нелинейного оценивания, так как не имеется естественного обобщения коэффициента корреляции Пирсона на случай нелинейных зависимостей.

Для построения корреляционной матрицы в верхнем меню **Statistics** надо выбрать команду **Basic Statistic Tables,** откроется меню команды. После выбора команды **Correlation Matrices** откроется рабочее окно модуля. Имена переменных можно задать одним списком (кнопка **One variables list**) или двумя списками (кнопка **Two lists**). В первом случае будет построена квадратная корреляционная матрица, строки и столбцы которой представлены списком переменных. Элементы матрицы — коэффициенты корреляции между переменными, расположенными на пересечении строки и столбца. Во втором случае будет построена прямоугольная матрица, строки и столбцы которой представлены соответственно первым и вторым списком.

Кнопка **Scatterplot matrix for selected variables** позволяет построить графики функции рассеяния и гистограммы выбранных переменных. Вкладка **Advanced/plot** предоставляет расширенные услуги графической иллюстрации статистического анализа выделенных переменных. На вкладке **Options** можно изменить параметры корреляционного анализа. Если установить флажок на *Display*

simple matrix, то будет построена простая корреляционная матрица. Если установить флажок на *Display r, p-levels, and N's*, то в ячейках корреляционной матрицы наряду с коэффициентами корреляции будут даны соответствующие им значения уровней значимости. Если установить флажок на *Display detailed table of results*, то наряду с коэффициентами корреляции будут даны результаты статистического анализа переменных: средние; стандартные отклонения; значения *t-критерия* сравнения средних и др.

4.3. Критерий Стьюдента сравнения средних

Критерий Стьюдента (*t-критерий*) является наиболее часто используемым методом проверки статистической гипотезы о равенстве средних двух выборок.

Статистической гипотезой называется любое предположение о виде или параметрах некоторого закона распределения. Проверяемую гипотезу обычно называют нулевой и обозначают H_0 (например, H_0: между строками и столбцам таблицы нет зависимости; коэффициент корреляции равен 0; средние некоторого показателя в двух выборках равны, закон распределения признака соответствует нормальному закону и т.д.). Наряду с нулевой гипотезой рассматривают альтернативную, или конкурирующую гипотезу H_1, являющуюся логическим отрицанием H_0 (например, H_1: между строками и столбцам таблицы есть зависимость; коэффициент корреляции не равен 0; средние некоторого показателя в двух выборках не равны, закон распределения признака не соответствует нормальному закону и т.д.). Эти гипотезы представляют собой две возможности выбора в задачах статистической проверки гипотез. При этом возможны четыре случая, которые приведены в таблице, изображенной на рис. 4.2.

H_0	*Принимается*	*Отвергается*
Верна	*Правильное решение*	*Неправильное решение, ошибка 1-го рода*
Не верна	*Ошибка 2-го рода*	*Правильное решение*

Рис. 4.2.

Вероятность α совершить ошибку 1-го рода, т.е. отвергнуть гипотезу H_0, когда она верна, называется уровнем значимости критерия.

Вероятность $1 - \beta$ не допустить ошибку 2-го рода называется мощностью критерия. Вероятности α, β однозначно определяются выбором критической области. Очевидно, желательно сделать α, β, в как угодно малыми, однако это

противоречивые требования. Лишь при увеличении объема выборки возможно одновременное уменьшение вероятностей α, β [11].

Уровень значимости p — это максимально приемлемая для исследователя вероятность ошибочно отклонить нулевую гипотезу, когда на самом деле она верна, т.е. допускаемая вероятность ошибки первого рода. Величина уровня значимости устанавливается исследователем произвольно, однако обычно принимается равным 0,05, либо 0,01, либо 0,001. В программе *STATISTICA* приемлемой границей статистической значимости приняты значения p, меньшие либо равные 0,05. Если p меньше либо равно 0,05, то результат считается статистически значимым, если p меньше либо равно 0,01, то результат считается статистически высоко значимым.

Теоретически *t-критерий* может применяться, даже если размеры выборок очень небольшие, переменные нормально распределены (внутри групп), а дисперсии наблюдений в группах не слишком различны [2]. Предположение о нормальности можно проверить, исследуя распределение (например, визуально с помощью гистограммы) или применяя какой-либо критерий нормальности. Равенство дисперсий в двух группах можно проверить с помощью *F-критерия* или использовать более устойчивый критерий Левена. Если условия применимости *t-критерия* не выполнены, следует использовать непараметрические альтернативы *t-критерия* (см. § 6.2). Уровень значимости p *t-критерия* равен вероятности ошибочно отвергнуть гипотезу о равенстве средних двух выборок, когда в действительности эта гипотеза имеет место.

Чем более различны средние в группах, тем более сильная степень зависимости между независимой (группирующей) и зависимой переменными. Степень различия между средними в двух группах зависит от внутригрупповой вариации (дисперсии) переменных. Поэтому уменьшение внутригрупповой вариации увеличивает чувствительность критерия.

Например, это относится к экспериментам, в которых две сравниваемые группы основываются на одной и той же совокупности наблюдений (случаев), тестировавшихся дважды (например, больные до и после лечения). Такие группы данных (выборки) называются зависимыми. В подобных экспериментах значительная часть внутригрупповой изменчивости в обеих группах может быть объяснена индивидуальными различиями субъектов. Заметим, что такая ситуация не слишком отличается от той, когда сравниваемые группы совершенно независимы и индивидуальные отличия вносят вклад в дисперсию ошибки. Однако в случае независимых выборок ничего нельзя поделать с этим, так как невозможно определить (или удалить) часть вариации, связанную с индивидуальными различиями наблюдений.

Если та же самая выборка тестируется дважды, то можно легко исключить эту часть вариации. Вместо исследования каждой группы отдельно и анализа исходных значений можно рассматривать разности между двумя измерениями (например, до после лечения больного) для каждого наблюдения. Вычитая первые значения из вторых (для каждого наблюдения) и анализируя затем только эти «чистые» (парные) разности, исключим ту часть вариации, которая является результатом различия в исходных уровнях наблюдений. Именно так и проводятся вычисления в *t-критерии* для зависимых выборок. В сравнении с *t-критерием* для независимых

выборок такой подход дает всегда более точный результат (критерий становится более чувствительным). Теоретические предположения *t-критерия* для независимых выборок относятся также к критерию для зависимых выборок. Это означает, что парные разности должны быть нормально распределены. Если это не выполняется, то можно воспользоваться одним из альтернативных непараметрических критериев.

Для запуска программы в верхнем меню **Statistics** надо выбрать команду **Basic Statistic Tables** (основные статистики/таблицы). Откроется меню команды, в котором *t-критерий* представлен четырьмя процедурами:

- **t-test, independent, by variables** (*t-критерий* для независимых выборок) применяется, если надо сравнить средние случайных величин, полученных по двум разным (независимым) выборкам;
- **t-test, independent, by groups** (*t-критерий* для независимых выборок с группирующей переменной) используется, если надо сравнить средние случайных величин двух независимых групп, полученных из одной выборки при помощи группирующей переменной;
- **t-test, dependent samples** (*t-критерий* для зависимых выборок) применяется, если надо сравнить средние случайных величин двух зависимых групп;
- **t-test, single samples** (простые выборки).

В перечисленных процедурах в качестве нулевой гипотезы предполагается, что средние в группах равны.

Рассмотрим работу процедуры **t-test, dependent samples,** используя таблицу данных на рис. 4.3, в которой приведены данные пробега пятнадцати японских и европейских автомобилей. В столбце 2 указан тип топлива: P — бензин; $G + P$ — бензин и газ; D — дизель. В столбцах 3, 4, 5 приведены пробеги автомобилей до первой серьезной поломки, требующей ремонта на СТО — *Пробег1*; до капитального ремонта двигателя — *Пробег2*; после капитального ремонта — *Пробег3*.

	Пробег до кап.рем., после кап. рем.				
	1 Произв.	2 Тип.топл.	3 Пробег1	4 Пробег2	5 Пробег3
Opel Astra	Europe	P	65	240	230
Skoda Fabia 1,2	Europe	P	70	250	220
Mitsubishi Pinin	Japan	G+P	110	300	280
Skoda Ambiente 1,6	Europe	P	60	230	230
Nissan Almera 1,5	Japan	G+P	90	280	260
Nissan Maxima 2,0 QX	Japan	G+P	100	300	280
Audi A4 2.0 MultiTronic	Europe	P	80	250	230
Nissan Maxima 3,0 SE	Japan	P	110	310	310
Mitsubishi Pejero III	Japan	G+P	95	320	280
Toyota Corolla	Japan	G+P	100	300	300
Toyota Carina	Japan	D	110	310	300
VW Passat1,8 T	Europe	D	70	275	250
VW Bora 1,6	Europe	D	80	260	230
Subaru Legacy	Japan	D	105	315	350
VW Golf 1,6	Europe	D	75	250	240

Рис. 4.3

Предположим, что величины пробега автомобилей в столбцах таблицы имеют нормальное распределение. На основании этих данных нужно определить, существенно ли отличаются средние величины пробега автомобилей до капитального ремонта и после капитального ремонта двигателей; средние величины пробега в зависимости от типа используемого топлива и места производства. Так как сравниваемые группы основываются на одной и той же совокупности наблюдений (случаев), тестировавшихся три раза, необходимо использовать для анализа *t-критерии* для зависимых выборок.

После выбора команды **t-test, dependent samples** в открывшемся окне процедуры (рис. 4.4) нажмите кнопку **Variables**. Выберите две сравниваемые между собой переменные: *Пробег2*, *Пробег3*, нажмите **OK**. Программа вернется в диалоговое окно модуля, в котором надо нажать на кнопку **Summary: t-test**.

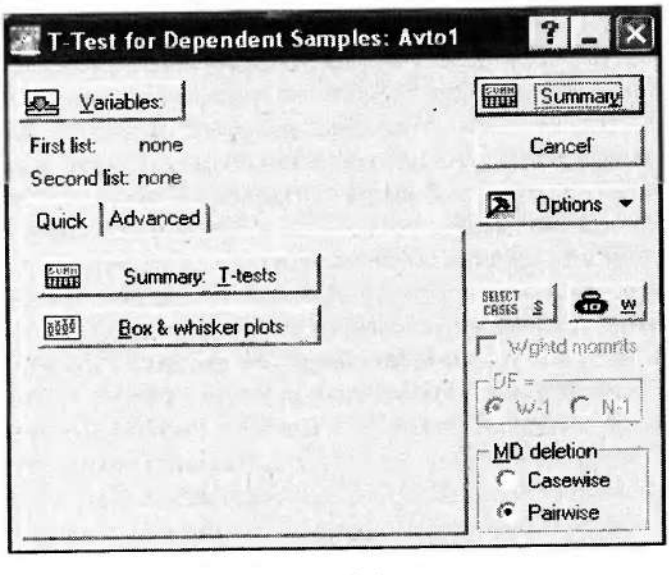


Рис. 4.4

Появится таблица с результатами анализа (рис. 4.5), в которой будут приведены значения следующих статистик:

- *Mean* (средние значения переменных);
- *Std. Dev.* (стандартные отклонения значений переменных);
- *N* (число наблюдений в группе);
- *Diff* (разница между средними);
- *t* (значение *t-критерия*);
- *df* (число степеней свободы);
- *p* (*p-уровень* значимости *t-критерия*).

Для нашего примера *p* меньше 0,05, поэтому гипотезу о равенстве средних отвергаем. Таким образом, средний пробег всех автомобилей до капитального ремонта значительно превышает средний пробег после капитального ремонта.

Variable	T-test for Dependent Samples (Auto) Marked differences are significant at p < ,05000 Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Пробег2	279,3300	30,52322						
Пробег3	262,6666	31,72801	15	16,66667	11,90238	5,423261	14	0,000090

Рис. 4.5

STATISTICA позволяет применять *t-критерий* для независимых выборок, используя одну независимую (группирующую) переменную (например, место производства) и одну зависимую переменную (например, пробег автомобиля). Обычно группирующая переменная является категориальной, т.е. содержит дискретные значения, например: 1, 2, 3..., или несколько текстовых значений. Значения такой переменной называются кодами и могут быть целочисленными или целочисленными с текстовыми эквивалентами. С помощью кодов данные разбиваются на две группы, и далее проверяется гипотеза о равенстве средних в этих уже независимых группах.

Проверьте, равны ли средние пробега автомобилей до первой серьезной поломки, до капитального ремонта двигателя и после капитального ремонта для автомобилей европейского и японского производства. В меню процедуры **Basic Statistic Tables** выберите команду **t-test, independent, by groups**. Откроется рабочее окно процедуры. Нажмите кнопку **Variables** и определите группирующую переменную *Произв.* и зависимые переменные *Пробег1, Пробег2, Пробег3*. После нажатия на **OK** программа вернется в рабочее окно модуля, укажите в нем коды группирующей переменной *Europe* и *Japan*. Щелкните кнопкой **Summary: t-tests**, откроется таблица с результатами анализа (рис. 4.6). По данным таблицы можно сделать вывод, что средние отличаются существенно. Вывод статистически достоверен, так как верна гипотеза о равенстве дисперсий (*p Variances* значительно больше, чем 0,05).

Variable	T-tests; Grouping: Произв. (Auto) Group 1: Europe Group 2: Japan Mean Europe	Mean Japan	t-value	df	p	Std.Dev. Europe	Std.Dev. Japan	F-ratio Variances	p Variances
Пробег1	71,43	102,50	-7,98	13	0,00	7,48	7,56	1,02	1,00
Пробег2	250,71	304,38	-7,81	13	0,00	14,27	12,37	1,33	0,71
Пробег3	232,86	295,00	-5,71	13	0,00	9,51	27,26	8,21	0,02

Рис. 4.6

Воспользуйтесь этой же процедурой и проверьте, равны ли средние пробега автомобилей до первой серьезной поломки, до капитального ремонта двигателя и после капитального ремонта для автомобилей с дизельным и бензиновым топливом. Для этого определите группирующую переменную *Тип. топл.* и выберите коды *P* и *D*. Щелкните кнопкой **Summary: t-tests**, откроется таблица с результатами анализа (рис. 4.7). По данным таблицы можно сделать вывод, что средние не отличаются существенно при справедливости гипотезы о равенстве дисперсий.

	T-tests; Grouping: Тип.топл. (Auto) Group 1: P Group 2: D							
Variable	Mean P	Mean D	t-value	p	Std.Dev. P	Std.Dev. D	F-ratio Variances	p Variances
Пробег1	77,00	88,00	-0,91	0,39	19,87	18,23	1,19	0,87
Пробег2	256,00	282,00	-1,36	0,21	31,30	29,28	1,14	0,90
Пробег3	244,00	274,00	-1,07	0,31	37,15	50,30	1,83	0,57

Рис. 4.7

Если предположить, что значения величин пробега в столбцах *Пробег2* и *Пробег3* получены по разным выборкам (тестировались различные группы автомобилей), для сравнения средних можно применить процедуру **t-test, independent, by variables.** После выбора этой команды откроется рабочее окно модуля. Укажите имена анализируемых переменных и щелкните по **OK**. По данным таблицы результатов (рис. 4.8) можно сделать вывод, что верна гипотеза о равенстве средних, при этом также верна гипотеза о равенстве дисперсий. Ранее при использовании процедуры **t-test, dependent samples** был получен противоположный результат. Такое несоответствие результатов как раз и объясняется большими дисперсиями величин пробега в анализируемых группах. При применении модуля **t-test, dependent samples** эти дисперсии не учитываются, и получается более верный результат.

	T-test for Independent Samples (Auto) Note: Variables were treated as independent samples							
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-val	p	Std.Dev Group 1	Std.Dev Group 2	F-ratio Varianc	p Varian
Пробег2 vs. Пробег3	279,33	262,67	1,47	0,15	30,52	31,73	1,08	0,89

Рис. 4.8

4.4. Группировка и однофакторная *ANOVA*

Для сравнения средних в более чем двух группах необходимо воспользоваться модулем дисперсионного анализа *ANOVA*.

Модуль **Breakdown & one-way ANOVA** (группировка и однофакторный дисперсионный анализ *ANOVA*) [2] определяет внутригрупповые описательные статистики и корреляции для зависимых переменных в каждой из нескольких групп, определенных одной или большим числом группирующих (независимых) переменных. Сравнивает средние и определяет, в каких именно группах средние отличаются между собой. В качестве нулевой гипотезы предполагается, что средние в группах равны.

Работу данного модуля проследим на примере данных из таблицы, приведенной на рис. 4.3. Выделите в модуле **Basic Statistic Tables** команду **Breakdown & one-way ANOVA**. Откроется рабочее окно команды. Нажмите кнопку **Variables** и выберите **Grouping variables** (группирующие переменные) *Произв.* и *Тип. топл.* и **Dependent variables** (зависимые переменные) *Пробег1*, *Пробег2*. Нажмите **OK** и, вернувшись в исходное окно, щелкните кнопкой **Codes for grouping variables** (коды для группирующих переменных). Выберите коды для группирующих переменных в диалоговом окне **Select codes for indep. vars (factors)** (коды для независимых факторов).

Чтобы выбрать все коды переменной, можно либо ввести номера кодов в соответствующем поле ввода, либо нажать на кнопку **All** (все), либо поставить * на соответствующем поле ввода. Нажатие **OK** без задания каких-либо значений эквивалентно определению всех значений для всех переменных. Если перед выбором кодов необходимо посмотреть значения переменной, нажмите кнопку **Zoom** (информация), которая откроет окно **Value/Stats** (значение/статистики). В этом окне будет выведен отсортированный список значений переменной (при этом будут отображаться все значения независимо от условий выбора наблюдений).

Нажмите **OK** в диалоговом окне **Statistics by Groups (Breakdown)**. Откроется новое диалоговое окно (рис. 4.9) **Statistics by Groups-Results** (внутригрупповые описательные статистики — результаты), которое предоставляет различные процедуры и настройки для анализа данных внутри групп. Цель такого анализа — лучшее понимание различий между группами. Информационная часть окна сообщает, что зависимых — две переменные: *Пробег1*, *Пробег2*; группирующих — две переменные: *Произв.* с двумя кодами (*Europe*, *Japan*) и *Тип. топл.* с тремя кодами (*P*, *G* + *P*, *D*). На рисунке активизирована вкладка **Quick**. На ней находятся следующие кнопки:

- **Summary: Table of statistics** — итоговая таблица средних;
- **Detailed two-way tables** — подробные двухвходовые таблицы;
- **Analysis of Variance** — дисперсионный анализ;
- **Interaction plot** — графики взаимодействий;
- **Categorized box & whisker plot** — категоризованные диаграммы размаха.

Щелкните кнопкой **Summary: Table of statistics**, появится таблица результатов (рис. 4.10). В приведенной таблице имеются описательные

статистики для выбранных переменных, разбитых на группы. Так в столбцах 3 и 5 показаны средние (*means*) переменных *Пробег1* и *Пробег2*, в столбцах 4 и 7 — количество автомобилей, в столбцах 5, 8 — среднеквадратические отклонения *(Std.Dev)*.

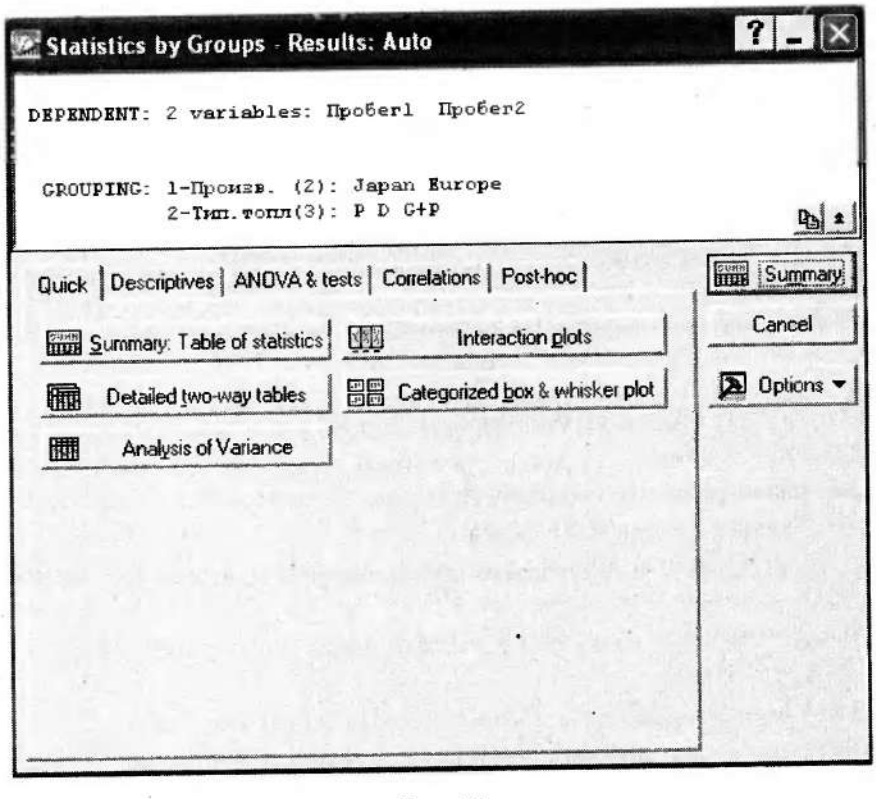


Рис. 4.9

Breakdown Table of Descriptive Statistics (Auto)
N=15 (No missing data in dep. var. list)

Произв.	Тип.топл.	Пробег1 Means	Пробег1 N	Пробег1 Std.Dev.	Пробег2 Means	Пробег2 N	Пробег2 Std.Dev.
Japan	P	110,0000	1	0,00000	310,0000	1	0,00000
Japan	D	107,5000	2	3,53553	312,5000	2	3,53553
Japan	G+P	99,0000	5	7,41620	300,0000	5	14,14214
Europe	P	68,7500	4	8,53913	242,5000	4	9,57427
Europe	D	75,0000	3	5,00000	261,6667	3	12,58306
Europe	G+P		0			0	
All Groups		88,0000	15	17,60682	279,3333	15	30,52322

Рис. 4.10

Для проверки значимости различий в средних указанных групп надо использовать процедуру **Analysis of Variance** (анализ дисперсий). Щелкните кнопкой **Analysis of Variance** на вкладке **ANOVA & tests**. Откроется таблица результатов **Analysis of Variance** (рис. 4.11). Из таблицы видно, что можно отвергнуть гипотезу о равенстве средних переменных *Пробег1*, *Пробег2* в группах. Так как число групп более двух, то из таблицы не видно, какие группы вызвали значительное отличие средних. Процедура **Post-hoc** (апостериорные сравнения средних) позволяет устранить этот недостаток.

	Analysis of Variance (Auto) Marked effects are significant at p < ,05000							
Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Пробег1	3838,75	4	959,688	501,250	10	50,1250	19,14589	0,000111
Пробег2	11639,17	4	2909,792	1404,167	10	140,4167	20,72259	0,000079

Рис. 4.11

Следует выделить вкладку **Post-hoc**, на которой представлены различные апостериорные процедуры:

- **LSD test or planned comparison** (критерий наименьшей значимости (НЗР));
- **Scheffe test** (критерий Шеффе);
- **Newman-Keuls test & critical ranges** (критерий Ньюмана-Кеулса и критические размахи);
- **Duncan's multiple range test & critical ranges** (критерий Дункана и критические размахи);
- **Tukey honest significant difference (HSD)** (критерий Тьюки ДЗР);
- **Tukey HSD for unequal N (Spjotvoll/Stoline)** (критерий Тьюки ДЗР для неравных *N*).

Можно назначить **p-level for highlighting** (*p-уровень* значимости для выделения).

Нажмите, например, кнопку **LSD test or planne domparison**. Появится таблица, состоящая из вероятностей (рис. 4.12).

	LSD Test; Variable: Пробег1 (Auto) Marked differences are significant at p < ,05000					
Произв. Тип.топл.	{1} M=110,00	{2} M=107,50	{3} M=99,000	{4} M=68,750	{5} M=75,000	{6} M=0,00
Japan P {1}		0,778993	0,186500	0,000395	0,001608	
Japan D {2}	0,778993		0,181823	0,000087	0,000515	
Japan G+P {3}	0,186500	0,181823		0,000081	0,000919	
Europe P {4}	0,000395	0,000087	0,000081		0,274618	
Europe D {5}	0,001608	0,000515	0,000919	0,274618		
Europe G+P {6}						

Рис. 4.12

Если вероятность, стоящая в таблице на пересечении строки и столбца с соответствующими номерами групп, больше, чем 0,05, то гипотезу о равенстве средних этих групп принимаем, в противном случае — отвергаем. Из таблицы видно, что верна гипотеза о равенстве средних в группах: 1, 2; 1, 3; 2, 3; 4, 5. Не верна гипотеза о равенстве средних в группах: 4, 1; 4, 2; 4, 3; 5, 1; 5, 2; 5, 3.

Различия средних можно увидеть на графиках, доступных в диалоговом окне **Statistics by Groups-Results**. Например, щелкните по кнопке **Categorized box & whisker plot**, которая находится на вкладке **Descriptives**. Откроется диалоговое окно **Box-Whisker Type.** В этом окне выделите одну из опций, например *Mean/SE/SD*. Программа построит диаграммы размаха (рис. 4.13), визуализирующие степень сходства и различия средних в анализируемых группах.

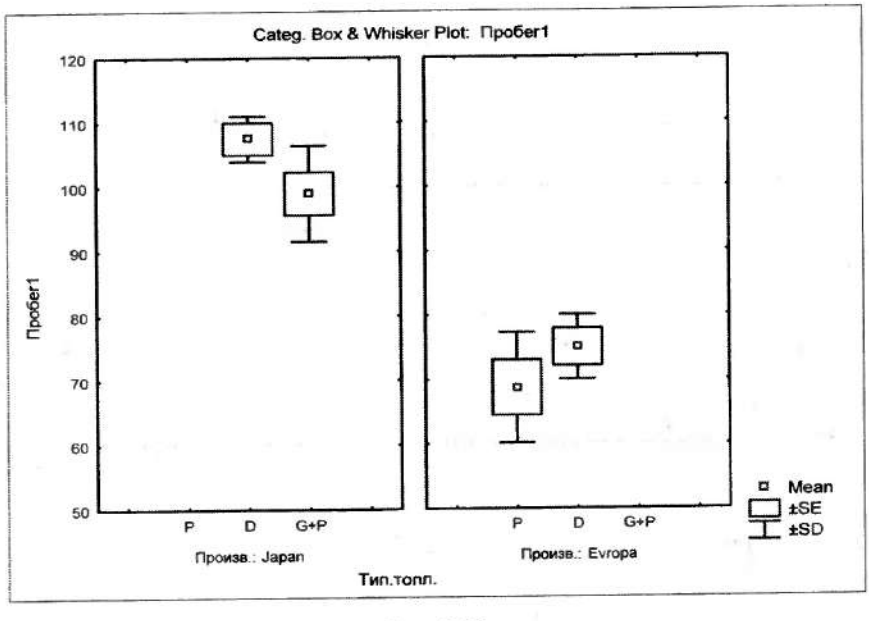


Рис. 4.13

Из приведенных результатов можно сделать вывод, что средний пробег японских автомобилей до обращения на СТО одинаков для различных типов топлива. Аналогичный вывод справедлив для автомобилей европейского производства. Но средний пробег японских автомобилей значительно больше пробега европейских автомобилей для любых типов топлива. Другими словами, пробег автомобилей до обращения на СТО не зависит от типа топлива, но зависит от страны производителя.

Корреляции измеряют степень зависимости между переменными. Если данные разбиты на однородные группы, то есть надежда, что зависимости станут более отчетливыми. Если имеется массив данных, то можно начать с группировки данных. Если данных мало, то поле действий резко сокращается. Проведем группировку данных по двум группирующим переменным, рассмотрим зависимости внутри групп и сравним с результатами для негруппированных наблюдений.

Щелкните по вкладке **Correlations**. На открывшемся диалоговом окне находятся кнопки **Within-group correlations & covariances** (внутригрупповые корреляции и ковариации); **Categ. Scatterplos** (категоризованные диаграммы рассеяния). Нажмите кнопку **Within-group correlations & covariances**. Откроется диалоговое окно **Select groups** (выберите группу), в котором надо выбрать группу или **All Groups** (все группы) для корреляционных матриц.

Дважды щелкните на строке **All Groups** и программа построит совокупность корреляционных матриц. На рис. 4.14 приведена корреляционная матрица переменных *Пробег1* и *Пробег2* для группы, состоящей из японских автомобиле с типом топлива *P* + *G*. Как видно из матрицы, зависимость между переменными слабая, причем величины коэффициентов корреляции статистически незначимы. В то же время зависимость между переменными сильная, если анализируется вся совокупность автомобилей (рис. 4.15).

Variables	Within-Group Correlatior Group: Произв.: Japa Marked correlations are Пробег1	Пробег2
Пробег1	1,000000	0,238366
Пробег2	0,238366	1,000000

Рис. 4.14

Variable	Correlations (Auto) Marked correlations are N=15 (Casewise deletior Пробег1	Пробег2
Пробег1	1,00	0,91
Пробег2	0,91	1,00

Рис. 4.15

Если просмотреть весь список корреляционных матриц, то можно заметить, что корреляции в отдельных группах заметно отличаются друг от друга. Следовательно, в разных группах зависимости между анализируемыми величинами проявляются по-разному.

Внутригрупповые зависимости можно представить графически. Для этого надо воспользоваться кнопкой **Categ. Scatterplos**.

Глава 5
Частотный анализ

5.1. Таблицы частот

Таблицы частот, или одновходовые таблицы, представляют собой простейший метод анализа категориальных (номинальных) переменных [2]. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке. Например, в социологических опросах таблицы частот могут отображать число мужчин и женщин, выразивших симпатию тому или иному политическому деятелю, число респондентов из определенных этнических групп, голосовавших за того или иного кандидата и т.д. В медицинских исследованиях табулируют пациентов с определенными симптомами, в маркетинговых исследованиях — покупательский спрос на товары разного типа у разных категорий населения. Чтобы открыть диалоговое окно **Frequency tables** (таблицы частот), надо из стартовой панели **Basic Statistics/Tables** выбрать команду **Frequency tables**. Это диалоговое окно предлагает множество настроек, позволяющих изменять вид и группировку в таблицах частот, а также проверять нормальность распределения, в том числе и графическими способами.

Рассмотрим функциональное назначение некоторых кнопок на вкладке **Quick**.

Variables. Открывается диалоговое окно выбора одного списка переменных для анализа.

Summary: Frequency tables. Открываются итоговые таблицы частот для выбранных переменных.

Histograms. Программа строит последовательность гистограмм для выбранных переменных. Одна гистограмма — для одной переменной.

Descriptive Statistics. Программа строит таблицу результатов с описательными статистиками для выбранных переменных.

3D histograms, bivariate distributions. Программа строит каскад трехмерных гистограмм для пар выбранных переменных, один график на каждую пару. После нажатия этой кнопки программа попросит пользователя выбрать два набора переменных (из списка выбранных ранее с помощью кнопки **Variables**). Гистограммы будут построены для каждой пары переменных, из разных списков.

Рассмотрим возможности вкладки **Advanced** (рис. 5.1). Установки опций под общим названием **Categorization methods for tables & graphs** (методы категоризации для таблиц и графиков) определяют, как будут сгруппированы или табулированы выбранные переменные в таблицах частот и гистограммах, как обрабатываются наблюдения при вычислении.

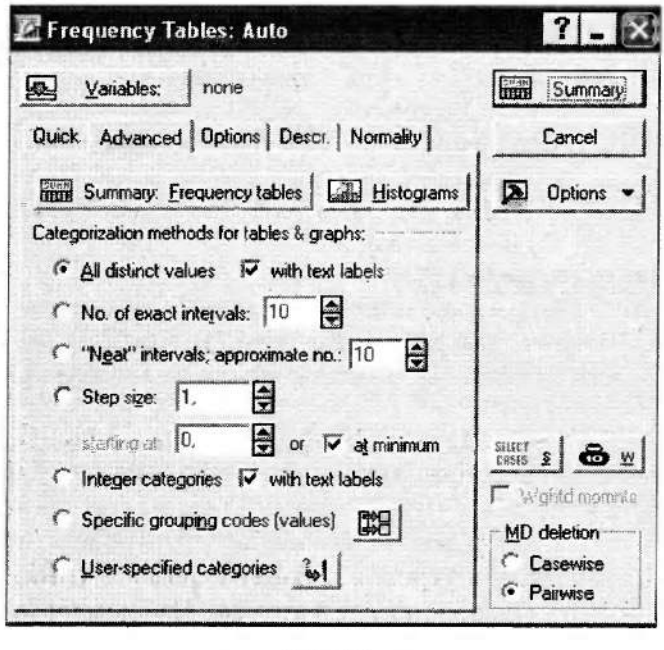


Рис. 5.1

Если установить флажок на *All distinct values* (все различные значения), то таблицы частот будут строиться с учетом всех различных значений анализируемых переменных.

Если установить флажок на *With text labels* (с текстовыми значениями), то таблицы частот будут строиться с учетом всех различных текстовых значений выбранных переменных.

Если установить флажок на *No of exact intervals* (число равных интервалов), то диапазон значений каждой переменной будет разделен на указанное число интервалов.

Если установить флажок на *«Neat» intervals; approximate no* (приблизительное число интервалов; приближенное число), то программа построит приближенные интервалы и выберет приближенный размер шага так, что последняя десятичная цифра будет 0 или 5 (например, 10,5, 11,0, 11,5 и т.д.). Такие интервалы легче интерпретировать, чем интервалы, содержащие много десятичных знаков (например, 10,12423, 10,13533 и т.д.).

Если установить флажок на *Step size* (размер шага), то программа задает ширину интервала категоризации в таблицах частот и гистограммах.

Если установить флажок на *At minimum* (начать с минимального значения), то группировка начинается с минимального значения переменной (первый интервал группировки включает это значение). В противном случае левая граница первого интервала группировки задается пользователем в соответствующем поле.

Если установить флажок на *Integer categories* (целые категории), то границами интервалов категоризации в таблицах частот (гистограммах) будут целые числа, а размер шага равен наименьшему целому значению. Все нецелые значения переменных будут проигнорированы программой.

Если установить флажок на *With text labels* (с текстовыми значениями), то категории при выборе *Frequency tables* и *Histogram* будут помечены текстовыми значениями (например, *Japan, Europe*), а не целыми значениями (например, 1, 2), которые доступны в текущем файле данных.

Если установить флажок на *Specific grouping codes(values)* (заданные группирующие коды (значения)), то таблицы частот (и гистограммы) будут построены целочисленными кодами, определенными пользователем с помощью расположенной рядом с флажком кнопки. Все нецелые значения переменных будут проигнорированы программой.

Если установить флажок на *User-specified categories* (определенные пользователем категории), то логическими условиями можно отнести наблюдения к определенной категории (до 16) в таблице частот. Логические условия могут быть сложными и включать различные переменные файла данных, а также наблюдения. Для каждого наблюдения условия выбора наблюдений (отнесения к определенной категории) проверяются последовательно. Наблюдение относится к той категории, в какую оно попало первым (где соответствующее логическое условие приняло значение истины).

Рассмотрим возможности вкладки **Options** (опции).

Если установить флажок на *Display options for frequency tables* (опции отображения), то можно определить, как будут сгруппированы или табулированы выбранные переменные в таблицах частот (см. ранее). В зависимости от выбора в поле **MD Deletion** (пропущенные данные) программа включает пропущенные данные в обработку или исключает из нее.

Если установить флажок на *Cumulative frequency tables* (кумулятивные частоты), то будут вычислены кумулятивные (накопленные) частоты.

Если установить флажок на *Percentages (relative frequencies)* (проценты (относительные частоты)), то программа вычислит относительные частоты (проценты).

Если установить флажок на *Cumulative percentages* (кумулятивные проценты), то будут вычислены кумулятивные или накопленные проценты.

Если установить флажок на *100% minus cumulative percentage* (100% минус кумулятивные проценты), то будут вычислены 100 минус кумулятивные проценты.

Если установить флажок на *Logit transformed proportions* (логит-преобразование частот), то для частот каждой группы будет осуществлено логит-преобразование

$$логит\ (i) = ln(n_i/(1-n_i)),$$

где n_i — относительная частота группы i.

Если установить флажок на *Probit transformed proportions* (пробит- преобразование частот), то для кумулятивных частот каждой группы будет осуществлено пробит-преобразование. Вычисляются *z-значения*, связанные с вероятностью в соответствующей ячейке.

Если установить флажок на *Count and report missing data (MD)* (подсчет и учет пропущенных данных (ПД)), то программа приводит статистику ПД.

Если установить флажок на *Count and report MD & non-selected cases* (подсчет и учет ПД и невыбранных наблюдений), то программа приводит статистику ПД и невыбранных наблюдений.

Вкладка **Descriptive** предназначена для анализа основных статистик исследуемых переменных.

Вкладка **Normality** предназначена для проверки соответствия закона распределения нормальному.

5.2. Таблицы кросстабуляции и таблицы флагов и заголовков

Кросстабуляция (сопряжение) — процесс объединения двух (или нескольких) таблиц частот так, что каждая ячейка (клетка) в построенной таблице представляется единственной комбинацией значений или уровней табулированных переменных [2]. Таким образом, кросстабуляция позволяет совместить частоты появления наблюдений на разных уровнях рассматриваемых факторов. Исследуя эти частоты, можно определить связи между табулированными переменными. Обычно табулируются категориальные переменные или переменные с относительно небольшим числом значений. Если надо табулировать непрерывную переменную (например, доход), то вначале ее следует перекодировать, разбив диапазон изменения на небольшое число интервалов (например, доход: низкий, средний, высокий). Для построения таблиц кросстабуляции надо из стартовой панели **Basic Statistics/Tables** выбрать процедуру **Tables and banners**. Откроется диалоговое окно **Crosstabulation Tables** (таблицы кросстабуляции). На вкладке **Crosstabulation (Stub-and-banners)** надо нажать кнопку **Specify tables.** Программа запросит ввести имена переменных для анализа. После возвращения в исходное диалоговое окно нужно установить флажок на *Use selected grouping codes only,* нажать кнопку **Codes** и ввести коды. После нажатия на **OK** откроется окно **Crosstabulation Tables Results**.

Рассмотрим функциональные назначения кнопок вкладки **Advanced** (рис. 5.2).

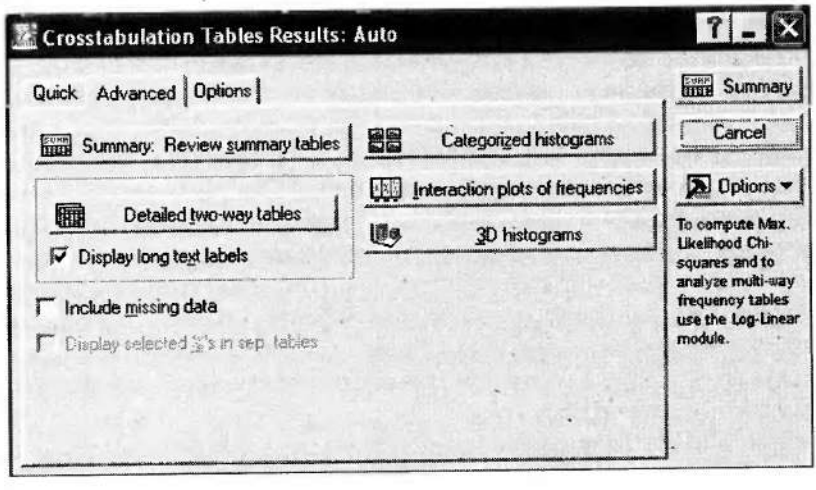


Рис. 5.2

Summary: Review summary tables (просмотреть итоговые таблицы). Программа строит итоговую таблицу для многовходовых таблиц сопряженности. Если определено более одной таблицы, то открывается диалоговое окно, в котором можно выбрать таблицы для просмотра. Если выбрано *All tables*, будет построен каскад таблиц результатов или графиков. Итоговая таблица сопряженности содержит частоты исходной таблицы; последняя переменная (фактор) будет табулирована в столбцах, все оставшиеся факторы — в строках. Если присутствуют больше двух факторов, то таблица может рассматриваться как совокупность нескольких двухвходовых таблиц (для последних двух факторов). Таким образом, таблица до 6-го порядка может быть просмотрена в одной таблице данных. Детали вывода определяются параметрами в поле **Compute tables** на вкладке **Options**. Отметим, что *Expected frequencies* (ожидаемые частоты) или *Residual frequencies* (остаточные частоты) вычисляются с помощью итеративной подгонки.

Detailed two-way tables (подробные двухвходовые таблицы). Программа строит таблицы результатов для двухвходовых таблиц. Если определено более одной таблицы, то открывается диалоговое окно, в котором можно выбрать таблицы для отображения на экране. Для таблиц с более чем двумя переменными (факторами) строится каскад таблиц результатов для последних двух факторов внутри уровней других факторов. Детали результатов определяются установками полей **Statistics for two-way tables** и **Compute tables** на вкладке **Options**. Если выбраны все статистики, то для каждой из таблиц вычисляются все запрашиваемые статистики.

Stub-and-banner table (таблица флагов и заголовков). Процедура доступна, если она выбрана в предыдущем окне. В таблице результатов один список переменных будет табулирован в столбцах, другой — в строках.

Display long text labels (отображать длинные метки значений). Программа отобразит длинные метки значений в первом столбце двухвходовой таблицы результатов. Если соответствующий фактор не имеет длинной метки, опция игнорируется.

Display selected %'s in sep. tables (отображать выбранные % в отдельных таблицах). Процедура доступна, если выбрана одна из опций **Percentages...** (проценты...) в поле **Compute tables.** По умолчанию, если запрошено вычисление процентов, то они будут отображены в одной таблице с частотами. Если выделена данная опция, то таблицы с процентами будут выведены на экран отдельно.

Categorized histograms (категоризованные гистограммы). Если происходит вычисление более одной таблицы, можно выбрать, какие именно таблицы отобразить на графике. После этого *STATISTICA* произведет расчет и построение категоризованных гистограмм для выбранных таблиц. Отметим, что каждый график может содержать информацию максимум о трех переменных (факторах). То есть одна категоризованная гистограмма может отобразить до трех таблиц — последний (измененный) фактор будет представлен в столбце графика; предшествующий ему будет представлен различными гистограммами, табулированными горизонтально; третий фактор будет представлен гистограммами, табулированными вертикально. Для таблиц с более чем тремя факторами будут построены каскады категоризованных гистограмм.

Interaction plots of frequencies (графики взаимодействий частот). С помощью линейного графика (графика взаимодействий) будет представлено распределение частот между тремя переменными (факторами). Если в таблице имеется больше трех факторов, то программа построит последовательность графиков взаимодействия на каждой комбинации уровней оставшихся факторов. После нажатия кнопки можно выбрать таблицу (или несколько таблиц) для графика (графиков). Программа построит график (графики) взаимодействия для выбранных таблиц.

3D histograms. Программа построит *3D* гистограммы для выбранных таблиц. После нажатия кнопки можно выбрать таблицу (или несколько таблиц) для графика (графиков). Каждая гистограмма представляет совместное распределение частот двух выбранных переменных.

Рассмотрим вкладку **Options** (рис. 5.3).

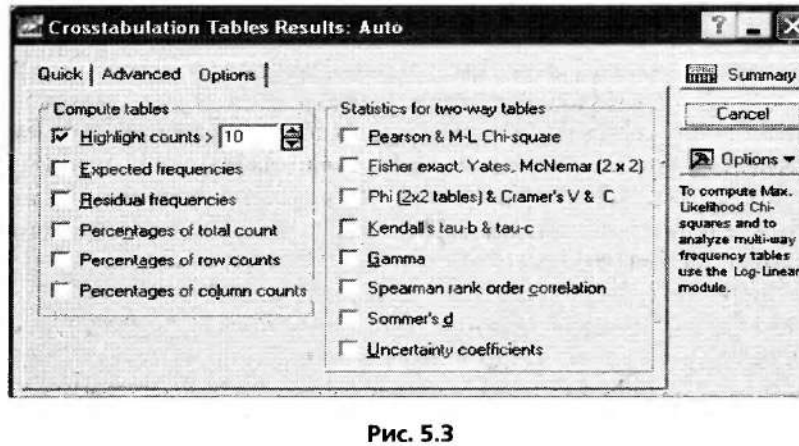


Рис. 5.3

Установки в рамке **Compute tables** (подсчитать таблицы) определяют подробности вывода результатов в **Detailed two-way tables**, **Summary tables** и некоторые — в **Stub-and-banner table**.

Highlight counts > (выделить частоты). Все частоты по строкам, которые превышают введенное значение, будут выделены цветом.

Expected frequencies (ожидаемые частоты). Для всех двухвходовых таблиц будут вычислены ожидаемые частоты в предположении независимости всех факторов (переменных) в таблице. Таким образом, для **Detailed two-way tables** ожидаемые частоты вычисляются на основе маргинальных частот двух факторов; для **Summary tables** с более чем двумя факторами ожидаемые частоты вычисляются на основе маргинальных частот всех факторов (предполагается, что между факторами нет взаимодействий).

Residual frequencies (остаточные частоты). Для всех двухвходовых таблиц и итоговой таблицы будут вычислены остаточные частоты — наблюдаемые частоты минус ожидаемые частоты.

Percentages of total count (проценты от общего числа). Программа вычислит проценты для каждой ячейки в **Summary tables**, **Detailed two-way tables** и в **Stub-and-banner table**. Проценты вычисляются от общего числа наблюдений в таблице. Если не установлен флажок на *Display selected %'s in sep. tables*, то проценты будут выведены в тех же таблицах, что и частоты.

Percentages of row counts (проценты по строке). Программа определит проценты относительно общего количества наблюдений в соответствующей строке текущей таблицы, для каждой ячейки в **Summary tables**, **Detailed two-way tables** и в **Stub-and-banner table.** Если не установлен флажок на *Display selected %'s in sep. tables*, то проценты будут выведены в тех же таблицах, что и частоты.

Percentages of column counts (проценты по столбцу). Аналогично предыдущему, только применительно к столбцам таблиц.

В рамке **Statistics for two-way tables** (статистики для двухвходовых таблиц) приведены основные статистики двухвходовых таблиц.

Pearson Chi-square (критерий χ^2 Пирсона). Это наиболее простой критерий проверки значимости связи между двумя категоризованными переменными. Критерий Пирсона основан на том, что в двухвходовой таблице ожидаемые частоты при гипотезе «между переменными нет зависимости» можно вычислить непосредственно. Величина статистики χ^2 и ее уровень значимости зависят от общего числа наблюдений и количества ячеек в таблице. В соответствии с принципами статистики относительно малые отклонения наблюдаемых частот от ожидаемых будут доказывать значимость связи, если число наблюдений велико. Имеется только одно существенное ограничение использования критерия (кроме очевидного предположения о случайном выборе наблюдений), которое состоит в том, что ожидаемые частоты не должны быть очень малы. Это связано с тем, что критерий χ^2 по своей природе проверяет вероятности в каждой ячейке, и если ожидаемые частоты в ячейках становятся малыми, например меньше 5, то эти вероятности нельзя оценить с достаточной точностью с помощью имеющихся частот.

M-L Chi-square (максимум правдоподобия χ^2) предназначен для проверки той же самой гипотезы относительно связей в таблицах сопряженности, что и критерий χ^2 Пирсона. Однако его вычисление основано на методе максимального правдоподобия. На практике эта статистика очень близка по величине к обычной статистике χ^2 Пирсона.

Yates (поправка Йетса). Аппроксимация статистики χ^2 для таблиц 2 × 2 с малым числом наблюдений в ячейках может быть улучшена уменьшением абсолютного значения разностей между ожидаемыми и наблюдаемыми частотами на величину 0,5 перед возведением в квадрат (так называемая поправка Йетса). Поправка Йетса, делающая оценку более умеренной, обычно применяется в тех случаях, когда таблицы содержат только малые частоты, например, когда некоторые ожидаемые частоты становятся меньше 10.

Fisher exact (точный критерий Фишера). Этот критерий применим только для таблиц 2 × 2. Критерий Фишера вычисляет точную вероятность появления наблюдаемых частот при нулевой гипотезе (отсутствие связи между табулированными переменными). В таблице результатов приводятся как односторонние, так и двусторонние уровни критерия.

McNemar (2×2) (критерий Макнемара). Данный критерий применим, когда частоты в таблице 2 × 2 представляют зависимые выборки. Например, наблюдения одних и тех же индивидуумов до и после эксперимента. В частности, можно подсчитать число студентов, имеющих минимальные успехи по математике в начале и в конце семестра, или предпочтения одних и тех же респондентов до и после рекламы. Процедура вычисляет два значения критерия: *A/D* и *B/C*. Первый проверяет гипотезу о том, что частоты в ячейках *A* и *D* (верхняя левая, нижняя правая) одинаковы, второй — гипотезу о равенстве частот в ячейках *B* и *C* (верхняя правая, нижняя левая).

Phi (2x2 tables) & Cramer's V & C (коэффициент Фи). Коэффициент представляет собой меру связи между двумя переменными в таблице 2 × 2. Его значения изменяются от 0 (нет зависимости между переменными) до 1 (абсолютная зависимость между двумя факторами в таблице).

Spearman rank order correlation (*R* Спирмена). Статистику *R* Спирмена можно интерпретировать так же, как и корреляцию Пирсона, в терминах объясненной доли дисперсии (имея, однако, в виду, что статистика Спирмена вычислена по рангам). Предполагается, что переменные измерены как минимум в порядковой шкале.

Kendall's tau-b & tau-c (тау Кендалла). Статистика тау Кендалла эквивалентна статистике *R* Спирмена при выполнении некоторых основных предположений. Также эквивалентны их мощности. Однако обычно значения *R* Спирмена и тау Кендалла различны, потому что они отличаются как своей внутренней логикой, так и способом вычисления. Обычно вычисляются два варианта статистики. Эти статистики различаются только способом обработки совпадающих рангов. В большинстве случаев их значения довольно похожи. Если возникают различия, то, по-видимому, самый безопасный способ — рассматривать наименьшее из двух значений.

Sommer's d (коэффициент *d* Соммера: *d(X|Y), d(Y|X)*). Статистика *d* Соммера представляет собой несимметричную меру связи между двумя переменными. Эта статистика близка к статистике тау Кендалла.

Gamma (гамма-статистика). Если в данных имеется много совпадающих значений, эта статистика предпочтительнее *R* Спирмена или тау Кендалла. С точки зрения основных предположений она эквивалентна статистике *R* Спирмена или тау Кендалла. Причем ее интерпретация и вычисление более похожи на статистику тау Кендалла, чем на статистику *R* Спирмена.

Uncertainty coefficients S(X|Y) u S(Y|X) (коэффициенты неопределенности). Эти коэффициенты измеряют информационную связь между факторами (строками

и столбцами таблицы). Статистики измеряют количество информации в переменной *Y* относительно переменной *X* или в переменной *X* относительно переменной *Y*.

Рассмотрим построение и анализ таблиц сопряженности для данных на рис. 4.3. После осуществления последовательности действий, описанных в начале данного раздела, выделите имена переменных для анализа: в списке *1 — Производ.*, в списке *2 — Тип. Топл.* На вкладке **Options** установите флажки на *Percentages of column counts, Highlight counts, Percentages of row counts* и на все опции рамки **Statistics for two-way tables**.

Нажмите **OK** в диалоговом окне **Crosstabulation Tables**. Программа построит всевозможные подмножества (сопряжения) из значений (кодов) указанных переменных и произведет подсчет частот элементов выборки (членов группы) при соответствующих значениях переменных.

Статистики двухвходовой таблицы (рис. 5.4) и проценты по строкам (рис. 5.5) и столбцам итоговой таблицы позволят установить наличие взаимосвязи между категориальными переменными *Производ.* и *Тип. Топл.* и исследовать характер взаимосвязи.

	Statistics: Произв.(2) x Тип.топл.(3)		
Statistic	Chi-square	df	p
Pearson Chi-square	6,964286	df=2	p=,03075
M-L Chi-square	8,993558	df=2	p=,01115
Phi	,6813852		
Contingency coefficient	,5630925		
Cramer's V	,6813852		
Kendall's tau b & c	b=-,617213	c=-,711111	
Sommers D(X|Y), D(Y|X)	X|Y=-,5333	Y|X=-,7142	
Gamma	-,869565		
Spearman Rank R	-,654654	t=-3,122	p=,00809
Uncertainty coefficient	X=,4338908	Y=,2728763	X|Y=,33504

Рис. 5.4

Из таблицы **Statistics:** следует, что между страной производителя автомобилей и типом топлива существует статистически значимая взаимосвязь — уровни значимости *p* критериев *Pearson & M-L Chi-square* меньше 0,05. Из значений остальных статистик следует, что связь между переменными умеренная или сильная (гамма-статистика больше 0,75), отрицательного знака.

Чтобы понять смысл зависимости обратимся к процентам по строкам и столбцам. Для японских автомобилей процент машин с типом топлива *P* минимальный (12,5%), с типом топлива *G* + *P* — максимальный (62,5%). Для европейских автомобилей, наоборот — с типом топлива *P* процент машин максимальный (57,14%), с типом топлива *G* + *P* — минимальный (0%). Проценты по столбцам показывают, что на автомобили с топливом *P* приходится 20% японских и 80% европейских машин, на автомобили с топливом *G* + *P* приходится 100% японских и 0% европейских автомобилей. Таким образом, зависимость между страной производителя и типом топлива автомобилей проявляется в том, что на европейские машины предпочитают устанавливать двигатели на бензине, а на японские — со смешанным топливом: бензин-газ.

Произв.	2-Way Summary Table: Observed Frequencies (Marked cells have counts > 10			
	Тип.топл. P	Тип.топл. D	Тип.топл. G+P	Row Totals
Japan	1	2	5	8
Column %	20,00%	40,00%	100,00%	
Row %	12,50%	25,00%	62,50%	
Evropa	4	3	0	7
Column %	80,00%	60,00%	0,00%	
Row %	57,14%	42,86%	0,00%	
Totals	5	5	5	15

Рис. 5.5

Иногда отдельные строки и столбцы таблицы удобно представлять в виде графиков. Полезно также отобразить целую таблицу на отдельном графике. Другой способ визуализации таблиц сопряженности — построение категоризованной гистограммы (рис. 5.6), в которой каждая переменная представлена индивидуальными гистограммами на каждом уровне другой переменной.

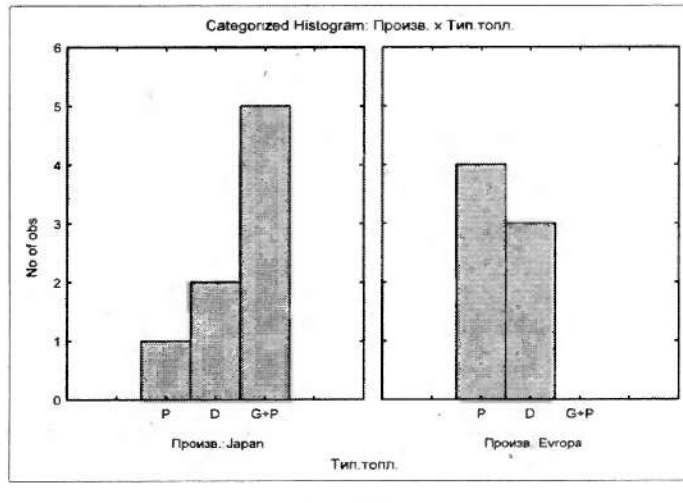


Рис. 5.6

Таблицы с двумя входами можно изобразить на *3D* гистограмме. Преимущество *3D* гистограммы заключается в том, что она позволяет представить на одном графике таблицу целиком. Достоинство категоризованного графика в том, что он дает возможность точно оценить отдельные частоты в каждой ячейке.

5.3. Многомерные отклики

Переменные типа многомерных откликов возникают в ситуациях, когда исследователя интересуют не только «простые» частоты событий, но и некоторые (часто неструктурированные) качественные свойства событий. Типичным примером является опрос общественного мнения, где вопросы, по крайней мере, частично,

имеют так называемые открытые концы (не подразумевает однозначного ответа), и респондент делает выбор из неограниченного списка ответов. Природу многомерных переменных лучше всего рассмотреть на примере.

Пример основан на данных опроса студентов относительно степени проявления интереса к дисциплинам. Каждый из пятнадцати студентов должен был составить список, включающий пять наиболее интересных дисциплин. Дисциплины располагаются в порядке убывания интереса к ним. Результаты опроса приведены в таблице данных на рис. 5.7.

	1 1-место	2 2-е место	3 3-е место	4 4-место	5 5-е место
1	умф	мат.ан.	диф.ур.	теор.вер.	архит.эвм
2	архит.эвм	физ-ра	медицина	кое	умф
3	умф	архит.эвм	скт	диф.ур.	мат.ан.
4	скт	диф.ур.	числ.мет.	архит.эвм	excel
5	теор.вер.	умф	мат.ан.	функ.ан.	диф.ур.
6	функ.ан.	умф	скт	excel	архит.эвм
7	дискр.прог.	функ.ан.	архит.эвм	скт	теор.вер.
8	умф	архит.эвм	дискр.прог.	скт	теор.вер.
9	скт	архит.эвм	числ.мет.	дискр.прог.	теор.вер.
10	умф	архит.эвм	диф.ур.	функ.ан.	дискр.прог.
11	функ.ан.	скт	умф	теор.вер.	диф.ур.
12	умф	медицина	excel	скт	теор.вер.
13	скт	архит.эвм	теор.вер.	дискр.прог.	дискр.прог.
14	теор.вер.	теор.вер.	дискр.прог.	скт	excel
15	скт	архит.эвм	теор.вер.	дискр.прог.	excel

Рис. 5.7

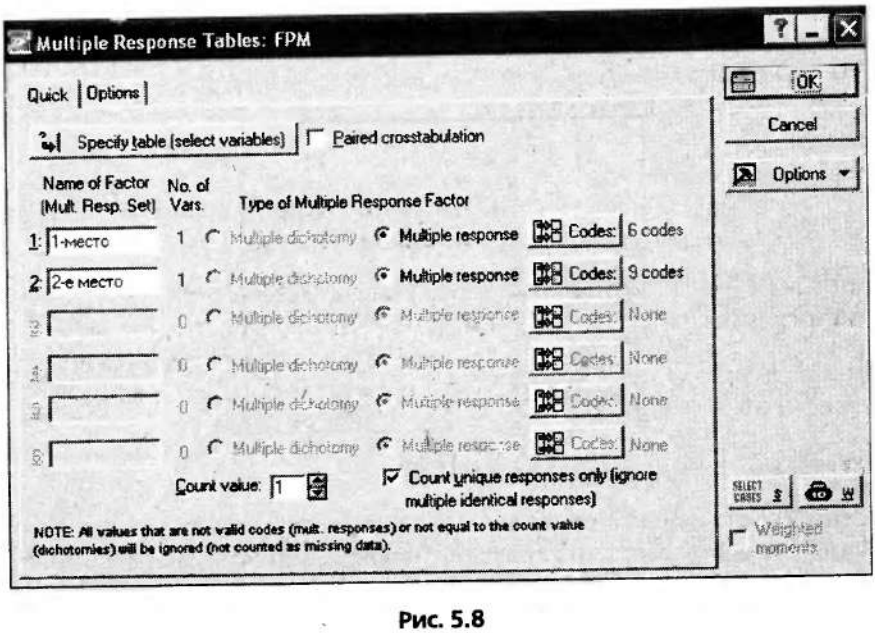


Рис. 5.8

Из стартовой панели модуля **Basic Statistics/Tables** выберите команду **Multiple response tables** (таблицы многомерных откликов). В диалоговом окне **Multiple response tables** (рис. 5.8) задайте переменные, например: *1-е место*; *2-е место*.

После чего необходимо задать коды — имена дисциплин, которые есть в столбце данной переменной. Щелкните по кнопке **OK**. Программа построит таблицу многомерных откликов. На рис. 5.9 приведен фрагмент таблицы.

Summary Table for all Multiple Response Items (FPM)
Totals/percentages based on number of responses
Multiple identical responses were ignored

N=15 1-место	2-е место мат.ан.	2-е место архит.эвм	2-е место скт	2-е место функ.ан.	2-е место физ-ра	2-е место диф.ур.
архит.эвм	0	0	0	0	1	0
скт	0	3	0	0	0	1
дискр.прог.	0	0	0	1	0	0
функ.ан.	0	0	1	0	0	0
умф	1	3	0	0	0	0
теор.вер.	0	0	0	0	0	0
All Grps	1	6	1	1	1	1

Рис. 5.9

В таблице рассмотрены пересечения двух переменных: *1-е место* и *2-е место* и подсчитаны частоты, с которыми пересеченные дисциплины встречаются соответственно на 1-м и 2-м местах в таблице исходных данных. Так, например, *архит. эвм* и *физ-ра*, занимающие1-е и 2-е места соответственно, встречаются 1 раз; *скт.* и *архит. эвм* встречаются 3 раза; *функ. ан.* и *скт* — 1 раз и т.д. По аналогии можно построить пересечения 3, 4, 5 и 6 чисел переменных.

Провести анализ исходных данных многомерных откликов можно также при помощи **Frequency table** (таблицы частот). Например, можно построить таблицу (рис. 5.10), в которой будет приведено распределение частот дисциплин, занявших 1-е место. Из таблицы видно, что пять респондентов (33%) отдали предпочтение *умф*, четыре респондента (26%) — *скт.*, два — *функ. ан.* и т.д.

N=15 Category	Frequencies (Identical resp. v Variable: 1-место (Simple Grouping Variable) Count	Prcnt.of Responses	Prcnt.of Cases
архит.эвм	1	6,67	6,67
скт	4	26,67	26,67
дискр.прог.	1	6,67	6,67
функ.ан.	2	13,33	13,33
умф	5	33,33	33,33
теор.вер.	2	13,33	13,33
Totals	15	100,00	100,00

Рис. 5.10

Если построить аналогичные таблицы частот для дисциплин, занявших соответственно 2-е место, 3-е и т.д., и выделить дисциплины, которым соответствует наибольшая частота, то получим последовательность предпочтений дисциплин студентами: *умф, архит. эвм, скт, дискр. пр., теор. вер.*

Глава 6

Непараметрическая статистика

6.1. Корреляционный анализ

В § 4.3 было замечено, что если условия применения параметрических критериев не выполнены, необходимо воспользоваться непараметрическими критериями. Условия могут быть не выполнены, если закон распределения переменных не удается аппроксимировать нормальным законом либо из-за малого объема выборки, либо из-за свойств переменной.

Альтернативой коэффициента корреляции Пирсона в непараметрической статистике являются коэффициенты Спирмена, тау Кендалла и гамма. Возможность применения перечисленных коэффициентов связана со шкалой, в которой измерены признаки объектов [12].

Коэффициент Спирмена рекомендуется использовать, если переменные — количественные (закон распределения которых не известен или не является нормальным) и (или) качественные (порядковые).

Коэффициент тау Кендалла рекомендуется использовать, если хотя бы одна переменная — качественная (порядковая).

Коэффициент гамма рекомендуется использовать, если переменные содержат много повторяющихся значений.

Рассмотрим процедуру вычисления коэффициентов корреляции на примере таблицы данных из рис. 4.3.

Необходимо определить, есть ли зависимость между производителями автомобилей и типом топлива; между производителями автомобилей, типом топлива и величинами *Пробег1, Пробег2, Пробег3*; между величинами *Пробег1, Пробег2, Пробег3*.

В верхнем меню **Statistica** выберите команду **Nonparametrics,** откроется окно с меню команды. После выбора команды **Correlations (Spearman, Kendall tau, gamma)** откроется окно диалога (рис. 6.1). Щелкните по кнопке **Variables** и задайте имена анализируемых переменных двумя списками — в каждом списке *Пробег1 — Пробег3*. Нажмите кнопку **Speaman rank R**, результаты анализа появятся в виде таблицы (рис. 6.2).

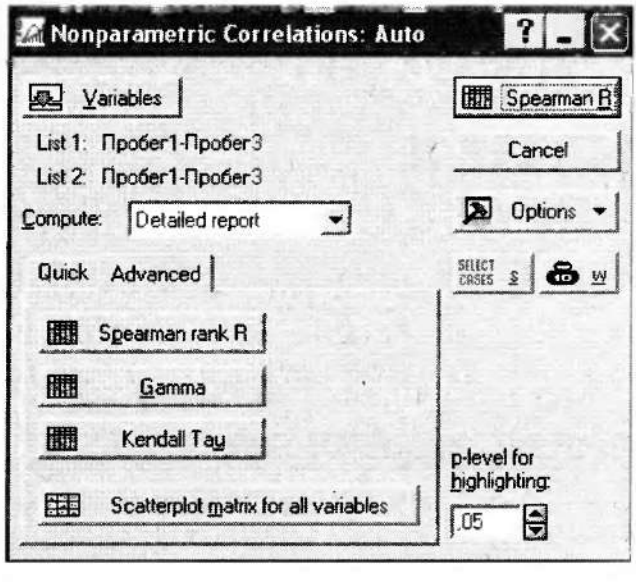


Рис. 6.1

Из рис. 6.2 видно, что между переменными *Пробег1, Пробег2, Пробег3* корреляция сильная и статистически значимая.

Pair of Variables	Valid N	Spearman R	t(N-2)	p-level
Spearman Rank Order Correlations (Aut MD pairwise deleted Marked correlations are significant at p <				
Пробег1 & Пробег2	15	0,853262	5,899496	0,000052
Пробег1 & Пробег3	15	0,857035	5,997169	0,000045
Пробег2 & Пробег3	15	0,685962	6,688044	0,000011

Рис. 6.2

На рис. 6.3–6.5 приведены корреляции между переменными *Тип. топл.* и *Произв.*; *Тип. топл.*, *Произв.* и переменными *Пробег1*, *Пробег2* и *Пробег3*. Из результатов анализа следует, что существует статистически значимая сильная зависимость между типом топлива автомобиля и местом его производства (эта зависимость проявляется в том, что все автомобили с топливом $G + P$ и большинство с топливом D произведены в Японии); существует статистически значимая сильная зависимость между местом производства автомобилей и пробегом до первой серьезной поломки, до капитального ремонта и после капитального ремонта; существует статистически значимая умеренная зависимость между типом топлива автомобилей и пробегом до первой серьезной поломки, до капитального ремонта и после капитального ремонта.

Pair of Variables	Valid N	Kendall Tau	Z	p-level
Kendall Tau Correlations (Auto) MD pairwise deleted Marked correlations are significant at p <				
Тип.топл. & Произв.	15	-0,617213	-3,20713	0,001341

Рис. 6.3

Pair of Variables	Valid N	Kendall Tau	Z	p-level
Kendall Tau Correlations (Auto) MD pairwise deleted Marked correlations are significant at p <				
Произв. & Пробег1	15	-0,752101	-3,90803	0,000093
Произв. & Пробег2	15	-0,755929	-3,92792	0,000086
Произв. & Пробег3	15	-0,767772	-3,98946	0,000066
Тип.топл. & Пробег1	15	0,394576	2,05028	0,040337
Тип.топл. & Пробег2	15	0,489898	2,54558	0,010909
Тип.топл. & Пробег3	15	0,414644	2,15455	0,031197

Рис. 6.4

Pair of Variables	Gamma Correlations (Auto) MD pairwise deleted Marked correlations are significant at p Valid N	Gamma	Z	p-level
Произв. & Пробег1	15	-1,00000	-3,90803	0,000093
Произв. & Пробег2	15	-1,00000	-3,92792	0,000086
Произв. & Пробег3	15	-1,00000	-3,98946	0,000066
Тип.топл. & Пробег1	15	0,48571	2,05028	0,040337
Тип.топл. & Пробег2	15	0,58333	2,54558	0,010909
Тип.топл. & Пробег3	15	0,49296	2,15455	0,031197

Рис. 6.5

Так как переменные *Тип топл.* и *Произв.* содержат большое число повторяющихся значений, желательно корреляции посчитать по критерию гамма. Из результатов, приведенных на рис. 6.4 и 6.5, видно, что существуют незначительные различия между значениями параметров критериев.

6.2. Непараметрические критерии сравнения средних

Для зависимых (сравниваемые группы основываются на одной и той же совокупности наблюдений) и независимых групп (выборок) применяют различные непараметрические критерии сравнения средних. При сравнении средних в двух и более чем в двух группах также желательно использовать различные непараметрические критерии. Непараметрические критерии сравнения средних достаточно подробно описаны в [10; 13]. Практически все непараметрические критерии сравнения средних предполагают, что анализируемая переменная измерена как минимум в порядковой шкале.

Для сравнения средних в более чем двух независимых группах применяют критерии *Kruskal-Wallis test* (Краскела-Уоллиса) и *Median test* (медианный тест), которые являются непараметрическими альтернативами однофакторного дисперсионного анализа. Файл должен содержать группирующую переменную. Критерий *Kruskal-Wallis test* основан на рангах и проверяет гипотезу, имеют ли сравниваемые выборки одно и тоже распределение или же распределения с одинаковыми медианами. В критерии *Median test* подсчитывается число наблюдений каждой группы, которые попадают правее или левее общей медианы выборок. При справедливости нулевой гипотезы — все группы извлечены из популяций с равными медианами, ожидается, что примерно половина всех наблюдений попадает левее (правее) общей медианы.

Для сравнения средних в двух независимых группах данных используют критерии *Wald-Wolfowoitz test* (Вальда-Вольфовица), *Kolmogorov-Smirnov test* (Колмогорова-Смирнова), *Mann-Whitney test* (Манна-Уитни), являющиеся непараметрическими альтернативами *t-критерия* для двух независимых выборок

и проверяет нулевую гипотезу, что две независимые выборки извлечены из двух популяций, которые могут отличаться не только средними но и формой распределения. Файл должен содержать группирующую переменную. Критерий *Wald-Wolfowoitz test* упорядочивает наблюдения по возрастанию или убыванию признака и исследует распределение серий (серией называется цепочка значений признака, соответствующих одной группе и примыкающих друг к другу в вариационном ряду) признака, относящихся к одной и той же группе. Если верна нулевая гипотеза, то число и длина серий, относящихся к одной и той же группе, будут более или менее случайными.

Критерий *Mann-Whitney test* основан на подсчете общего числа наблюдений, для которых значения признака в одной выборке превосходят значения признака в другой выборке. Этот критерий — наиболее мощная непараметрическая альтернатива *t-критерию*, а в некоторых случаях он имеет даже большую мощность, чем *t-критерий*.

Критерий *Kolmogorov-Smirnov test* основан на сравнении эмпирических функций распределения двух выборок, поэтому он чувствителен к различию форм распределений двух выборок (например, асимметрия, эксцесс).

Для сравнения средних в более чем двух зависимых группах используют критерий *Friedman ANOVA test* (Фридмана), который является непараметрической альтернативой однофакторному дисперсионному анализу с повторными измерениями.

Для сравнения средних в двух зависимых выборках используют критерии *Sign test* (критерий знаков), *Wilcoxon test* (Вилкоксона), которые являются непараметрической альтернативой *t-критерия* сравнения средних в двух зависимых выборках. Критерий *Sign test* основан на подсчете количества положительных разностей между значениями переменных до и после повторных измерений. Для применения этого критерия требуются очень слабые предположения, например об однозначной определенности медианы разностей.

Критерий *Wilcoxon test* основан на ранжировании значений рассматриваемого признака. Подсчитывается сумма рангов значений второй выборки в общем вариационном ряду двух выборок. Требования к применимости этого критерия более строгие, чем для критерия знаков. Но если эти требования выполнены, то критерий *Wilcoxon test* имеет большую мощность, чем критерий *Sign test*.

Проверим равенство средних пробега европейских и японских автомобилей до первого ремонта, до капитального ремонта и после капитального ремонта до очередной поломки двигателя.

Так как число групп — более двух и группы зависимые (рассматриваются пробеги одних и тех же автомобилей), в окне **Nonparametrics Statistics** выберите команду **Comparing multiple dep. Samples (variables)**, откроется рабочее окно **Friedman ANOVA by Ranks** (рис. 6.6).

После ввода имен анализируемых переменных щелкните кнопкой **OK**, появится таблица с результатами анализа (рис. 6.7). Так как уровень значимости критерия p значительно меньше 0,05 ($p < 0{,}000$), верна альтернативная гипотеза о неравенстве средних в трех группах.

Для того чтобы определить, в каких группах средние неравны, воспользуйтесь критериями знаков и Вилкоксона. Для этого в окне **Nonparametrics Statistics** выберите процедуру **Comparing two dep. variables**. В открывшемся окне (рис. 6.8) после выбора имен переменных последовательно нажмите кнопки с названиями критериев. На рис. 6.9–6.11 приведены таблицы с результатами попарного сравнения средних по критерию знаков.

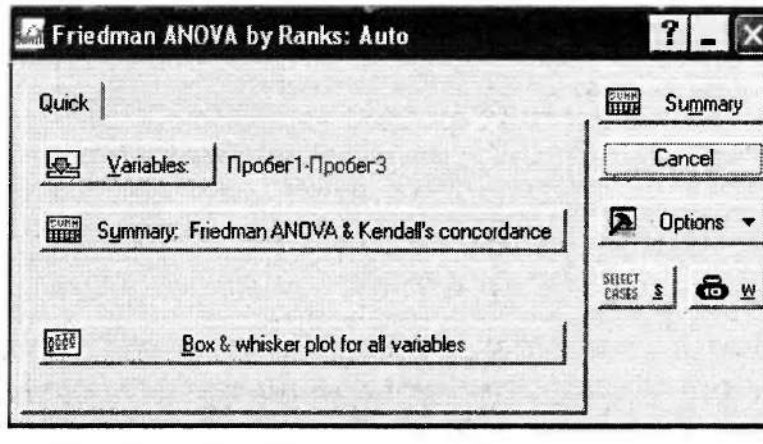


Рис. 6.6

Так как во всех таблицах p меньше 0,05, можно утверждать, что при попарном сравнении средних также верна альтернативная гипотеза о неравенстве средних, причем *среднее Пробег2 > среднее Пробег3 > среднее Пробег1.*

Friedman ANOVA and Kendall Coeff. of Concordance (Auto)
ANOVA Chi Sqr. (N = 15, df = 2) = 27,19298 p < ,00000
Coeff. of Concordance = ,90643 Aver. rank r = ,89975

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
Пробег1	1,000000	15,00000	88,0000	17,60682
Пробег2	2,833333	42,50000	279,3333	30,52322
Пробег3	2,166667	32,50000	266,0000	37,94733

Рис. 6.7

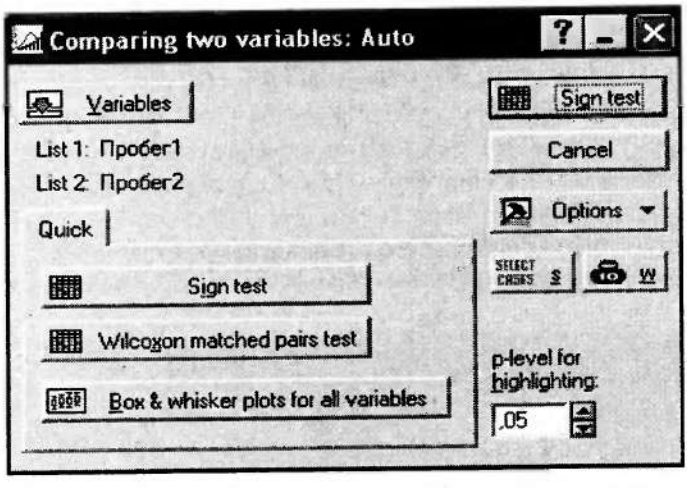


Рис. 6.8

Pair of Variables	Sign Test (Auto) Marked tests are significant at p <,05000			
	No. of Non-ties	Percent v < V	Z	p-level
Пробег1 & Пробег2	15	100,0000	3,614784	0,000301

Рис. 6.9

Pair of Variables	Sign Test (Auto) Marked tests are significant at p <,05000			
	No. of Non-ties	Percent v < V	Z	p-level
Пробег1 & Пробег3	15	100,0000	3,614784	0,000301

Рис. 6.10

Pair of Variables	Sign Test (Auto) Marked tests are significant at p <,05000			
	No. of Non-ties	Percent v < V	Z	p-level
Пробег2 & Пробег3	12	8,333333	2,598076	0,009375

Рис. 6.11

Проверьте, есть ли различие в пробегах автомобилей в зависимости от типа топлива. Так как сравниваются группы различных автомобилей, необходимо использовать критерии сравнения средних для независимых групп. В окне **Nonparametrics Statistics** выберите процедуру **Comparing multiple indep. Samples (groups),** откроется окно критериев Краскела-Уоллиса и медианного теста (рис. 6.12). Укажите имена переменных: группирующая — *Тип. топл.*; зависимая — *Пробег2*. Таблицы с результатами расчетов приведены на рис 6.13–6.14. Из таблиц следует, что верна гипотеза о равенстве средних в трех группах (уровни значимости критериев *p* больше, чем 0,05).

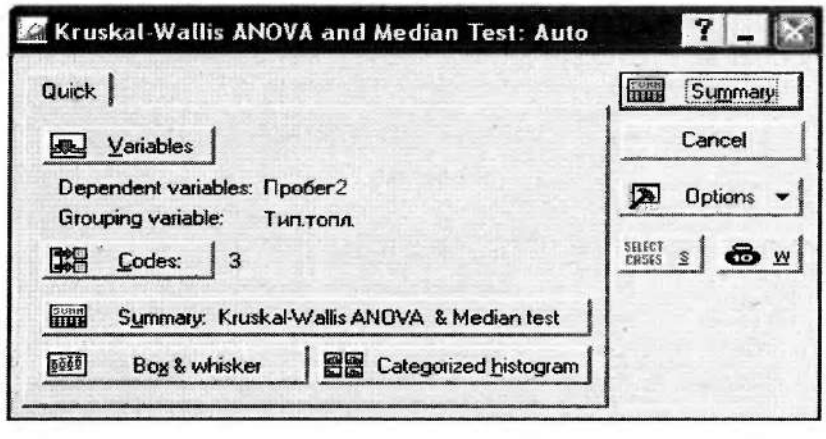


Рис. 6.12

Kruskal-Wallis ANOVA by Ranks; Пробег2
Independent (grouping) variable: Тип.топл.
Kruskal-Wallis test: H (2, N= 15) =4,609074 p =,0998

Depend.: Пробег2	Code	Valid N	Sum of Ranks
P	1	5	23,50000
D	2	5	43,50000
G+P	3	5	53,00000

Рис. 6.13

Dependent: Пробег2	Median Test, Overall Median = 280,000; Independent (grouping) variable: Тип.топл Chi-Square = 3,750000, df = 2, p = ,1534			
	P	D	G+P	Total
<= Median: observed	4,00000	3,000000	1,00000	8,00000
expected	2,66667	2,666667	2,66667	
obs.-exp.	1,33333	0,333333	-1,66667	
> Median: observed	1,00000	2,000000	4,00000	7,00000
expected	2,33333	2,333333	2,33333	
obs.-exp.	-1,33333	-0,333333	1,66667	
Total: observec	5,00000	5,000000	5,00000	15,00000

Рис. 6.14

Проверьте, есть ли различие в пробегах автомобилей в зависимости от места производства. Так как речь идет о сравнении средних в двух группах, надо воспользоваться процедурой **Comparing two indep. Samples (groups).** Выберите в качестве группирующей переменной переменную *Произв.*, а в качестве зависимых — переменные *Пробег1 — Пробег3*. Далее можете воспользоваться одним из критериев: Вальда-Вольфовица, Колмогорова-Смирнова или Манна-Уитни (рис. 6.15). Из результатов расчетов по критерию Вальда-Вольфовица, приведенных в таблице (рис. 6.16), следует, что альтернативная гипотеза о неравенстве средних подтверждается. При этом пробег автомобилей японского производства

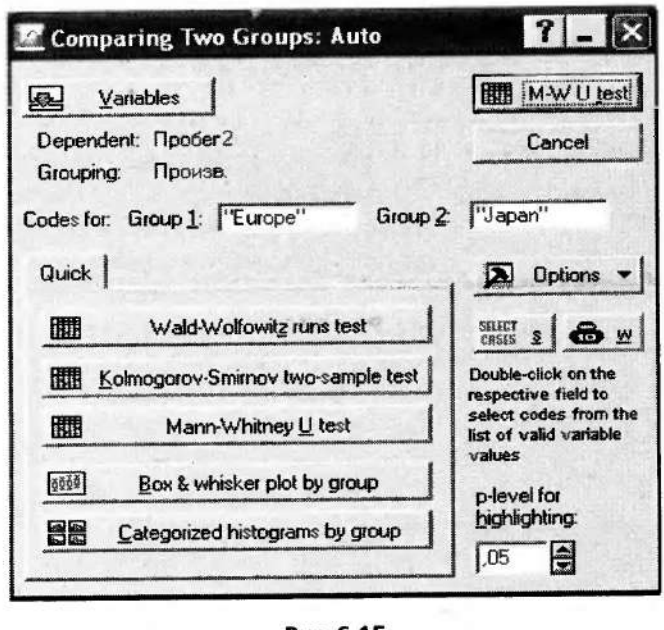


Рис. 6.15

до первой серьезной поломки, до капитального ремонта и после капитального ремонта превосходит соответствующие пробеги автомобилей европейского производства.

Обратите внимание, что результаты сравнения средних, осуществленных методами непараметрической статистики, в основном совпали с результатами, полученными по процедурам критерия Стьюдента в § 4.3.

	Wald-Wolfowitz Runs Test (A uto2) By variable Произв. Marked tests are significant at p <,05000							
Variable	Valid N Evropa	Valid N Japan	Mean Evropa	Mean Japan	Z	p-level	Z adjstd	p-le
Пробег1	7	8	71,4286	102,5000	-3,48210	0,000498	3,212863	0,001
Пробег2	7	8	250,7143	304,3750	-3,48210	0,000498	3,212863	0,001

Рис. 6.16

Глава 7

Основные законы распределения

7.1. Вероятностный калькулятор

Вероятностный калькулятор – процедура, предназначенная для работы с наиболее известными законами распределения [2, 14]. Используя ее, можно строить графики интегральной и дифференциальной функций распределения, для непрерывных случайных величин – вычислить процентные точки, определить вероятность попадания значений в заданный интервал, для дискретных случайных величин – вычислить вероятности и строить ряды распределения. Прежде чем перейти к изложению основных принципов работы с данной процедурой, рассмотрим некоторые из наиболее часто используемых законов распределения [15].

Нормальное распределение — наиболее важный закон распределения непрерывных случайных величин. С помощью нормального распределения можно описать большинство явлений окружающего мира, например, распределение некоторых физических параметров представителей животного, растительного мира. Нормальное распределение используется для моделирования экономических процессов — распределение заработной платы, налоговых поступлений, продолжительности жизни и т.д. Нормальный закон также находит широкое применение для приближения распределения дискретных случайных величин — объемов производства или продаж того или иного вида продукции, числа посетителей тех

или иных учреждений и т.д. Иногда это распределение называют распределением ошибок, так как ошибки всевозможных измерений также приближаются нормальным законом. Главной особенностью нормального распределения, выделяющего его среди других, является то, что оно — предельный закон, к которому приближаются другие законы распределения при выполнении определенных условий. Если из значений нормально распределенной случайной величины вычесть математическое ожидание и разделить на стандартное отклонение, то полученные случайные величины имеют стандартное нормальное распределение.

Распределение χ^2 (Пирсона). Случайная величина, имеющая распределение χ^2 с k степенями свободы, определяется как сумма квадратов k независимых случайных величин со стандартным нормальным распределением. В частном случае, когда $k = l$, случайная величина χ^2 равна квадрату стандартной нормальной величины. Это распределение асимметрично, обладает положительной правосторонней асимметрией (сосредоточено только на положительной полуоси). При увеличении числа степеней свободы пик плотности распределения уменьшается и смещается вправо. Это распределение играет важную роль при проверке зависимостей в таблицах сопряженности и в критериях согласия.

Распределение Стьюдента (*t-распределение*). Случайная величина, имеющая *t-распределение с k* степенями свободы, определяется как отношение случайной величины со стандартным нормальным распределением на корень квадратный из среднего арифметического квадратов k случайных величин, имеющих также нормальное стандартное отклонение. Кривая *t-распределения*, как и стандартная нормальная кривая, симметрична относительно оси ординат, но по сравнению с нормальной более пологая. При увеличении k это распределение приближается к нормальному. Данное распределение применяется при оценке среднего, в регрессионном анализе, при использовании временных рядов.

F-распределение Фишера. Случайная величина, имеющая *F-распределение* с парой степеней свободы m и n, определяется как отношение двух независимых случайных величин, имеющих распределение χ^2 со степенями свободы m и n, умноженным на нормировочный сомножитель n/m. Распределение асимметрично, обладает положительной правосторонней асимметрией. При увеличении m и n распределение приближается к нормальному. Распределение Фишера используется при оценке дисперсии случайной величины, в регрессионном, дисперсионном и дискриминантом анализе, а также в других видах многомерного анализа данных.

Логарифмически-нормальное распределение. Неотрицательная случайная величина X имеет логарифмически-нормальное (логнормальное) распределение, если случайная величина $ln(X)$ имеет нормальное распределение. Кривая плотности распределения асимметрична и располагается на положительной полуоси. Логнормальное распределение используется для описания распределения доходов, банковских вкладов, месячной заработной платы, посевных площадей под разные культуры, долговечности изделий в режиме износа и старения и др.

Биномиальное распределение. Биномиальное распределение представляет собой закон распределения числа наступлений m некоторого события A в n

независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p ($P_n(m) = C_n^m p^n (1-p)^{n-m}$). Этот закон распределения широко используется в теории и практике статистического контроля качества продукции, при моделировании систем массового обслуживания, в теории стрельбы и других областях.

Распределение Пуассона. Дискретная случайная величина X имеет распределение Пуассона, если она принимает значения 0, 1, 2... с вероятностями $P(X = m) = \lambda^m e^{-\lambda}/m!$, где $m = 0, 1, 2...$ Закон распределения Пуассона является хорошим приближением биномиального распределения при достаточно больших n и малых значениях вероятности p (при условии, что произведение np — постоянная величина). По закону Пуассона распределены, например, число рождений четверней, число сбоев на автоматической линии, число отказов сложной системы в «нормальном режиме», число требований на обслуживание, поступивших в единицу времени в системах массового обслуживания и т.д.

Распределение Бернулли. Случайная величина имеет распределение Бернулли, если она принимает значения 0 или 1 с вероятностями $P(X = m) = p^m(1-p)^{1-m}$, где $m \in \{0,1\}$, p — вероятность наступления события. Это распределение наилучшим образом описывает ситуации, где результатами являются успех или неуспех.

Геометрическое распределение. Дискретная случайная величина имеет геометрическое распределение, если она принимает значения $m = 1, 2...$ с вероятностями $P(X = m) = (1-p)^{m-1}p$, где m — число неуспехов; p — вероятность успеха в одном испытании. Это распределение используют тогда, когда моделируют ситуации, в которых испытания проводятся до первого наступления успеха.

Рассмотрим основные принципы работы процедуры **Probability calculator** (вероятностный калькулятор). Для запуска процедуры надо в модуле **Basic Statistics/Tables** выбрать команду **Probability calculator Distributions**. Откроется рабочее окно команды **Probability Distribution Calculator** (калькулятор вероятностных распределений). В левой части расположен список распределений **Distribution** (распределение). Многие стандартные распределения в этом окне можно выбрать, высвечивая их названия в списке слева: Бета, Коши, χ^2, нормальное, логнормальное, распределение Стьюдента и т.д. Выберите, например, в списке нижнюю строчку *Z Normal* (нормальное распределение). Автоматически справа появятся поля, где можно задать параметры нормального распределения: *mean* (среднее) и *std.dev* (стандартное отклонение). По умолчанию система запишет в них стандартные значения: среднее = 0, стандартное отклонение = 1. Данные значения можно изменить: надо поместить курсор мыши в эти поля, щелкнуть левой кнопкой и ввести с клавиатуры нужные величины. Одновременно с выбором распределения в левом списке справа в калькуляторе появятся графики нормальной плотности и функции распределения: *Density Function* (функция плотности), *Distribution Function* (функция распределения). В поле p надо задать уровень вероятности, при этом флажок автоматически установится на *Inverse* (инверсия). После нажатия на кнопку *Compute* (подсчет) (в правом верхнем углу калькулятора) в поле Z появится значение квантиля, соответствующее выбранному уровню вероятности. То же

можно сделать и в обратную сторону — по заданному значению Z вычислить уровень вероятности p. Для этого надо задать значение квантиля, щелкнуть по кнопке *Compute*; в поле p появится значение вероятности, соответствующее данному значения Z. Если установить флажок на *Create Graph* (создать график) и нажать на кнопку *Compute*, то на экране появятся графики плотности и функции распределения (рис. 7.1) с выделенными на них значениями вероятности и квантили.

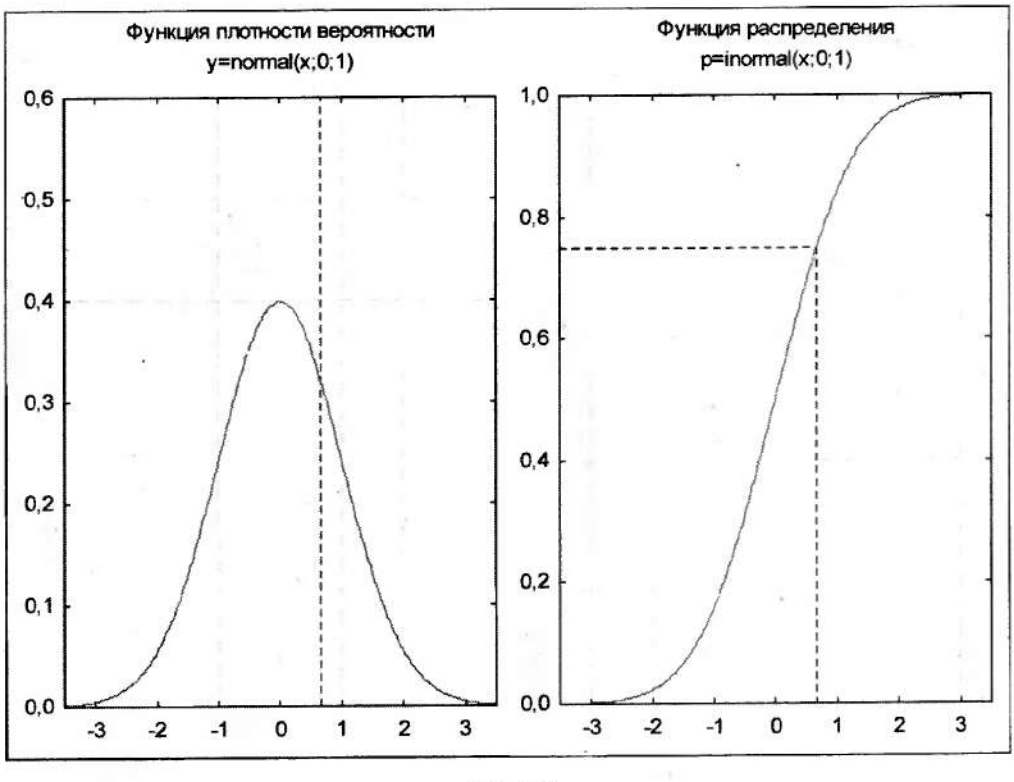


Рис. 7.1

Если установить флажок на *two-tailed* (двойной критерий), то расчет будет проведен для отрезка $[m - x; m + x]$ (где m — среднее значение), в противном случае — для отрезка $[-\infty; x]$. Если установить флажок на *1-Cumulative p* (*1-совокупный p*), то расчет будет проведен для отрезка, противоположного указанному. Так, например, если рассматривался отрезок $[-\infty; x]$, то расчет будет проведен для отрезка $[x; +\infty]$.

Флажок на *Fixed Scaling* (фиксированная шкала) под списком распределений *Distributions* указывает, что выбрана фиксированная шкала.

Помимо вычисления уровня вероятности, квантили и построения кривых распределений, **Probability calculator** может быть использован для изучения поведения кривых распеделений при изменении параметров распределений, а также

для решения некоторых задач. Так, например, увеличивая *mean* (среднее) нормального распределения, можно увидеть, как кривая плотности нормального распределения сдвигается по оси ординат вправо. При увеличении стандартного отклонения плотность нормального распределения расплывается или рассеивается относительно среднего значения. При уменьшении она, наоборот, сжимается, концентрируясь возле одной точки — точки максимального значения.

Рассмотрим примеры решения задач. Известно, что рост студентов имеет нормальное распределение со средним 175,6 см и стандартным отклонением 7,63 см. Произвольным образом выбирается студент, например, первый вошедший в аудиторию. Какова вероятность, что рост этого студента не больше 185 см и не меньше 175 см?

Выберите в списке распределений *Z Normal*. Задайте в поле *mean* — 175,6, в поле *std.dev.* — 7,63. В поле *X* — 185. Нажмите кнопку *Подсчет*. В поле *p* появится значение 0,891022. Запомните его как p_1.

В поле *X* задайте 175. Нажмите кнопку *Compute*. В поле *p* появится значение 0,468661. Запомните это значение как p_2. Вычтите p_2 из p_1. Получите 0,422361. Итак, с вероятностью 0,422361 случайный студент имеет рост не ниже 175 и не выше 185 см.

Рассмотрим решение задачи с использованием дискретного распределения. В среднем 30% студентов сдают экзамен по дискретному программированию на отлично. Найдите вероятность того, что в группе, состоящей из 15 человек, не более 5 человек получат отлично.

Создайте пустую электронную таблицу. В первом столбце переменной *Var1* проставьте возможное число студентов, сдавших на отлично (количество испытаний). Дважды щелкните по имени переменной *Var2*. Откроется диалоговое окно спецификации переменной *Var2*. В нижней части окна в поле **Long name** запишите функцию с указанием параметров. Запись должна начинаться со знака =. Функцию можно также выбрать из списка, нажав на кнопку **Functions**. В списке предложенных функций выберите нужную функцию (в данном случае *Binom*) и два раза щелкните по ней. Как видно из подсказки, функция *Binom(x,p,n)* использует три параметра, которые перечислены в круглых скобках через точку с запятой (рис. 7.2). Первый параметр *x* — ссылка на переменную, в строках которой указано количество проводимых испытаний (в нашем случае «*V1*»). Второй параметр *p* — вероятность удачного исхода в одном испытании (в нашем случае *p = 0,3* — вероятность сдать экзамен на отлично). Третий параметр *n = 15* — количество испытаний. Нажмите **OK**. Согласно формуле Бернулли программа вычислит вероятности успеха и занесет их в столбец таблицы, соответствующий второй переменной *Var2*. Для определения искомой вероятности надо выделить курсором мыши первые 6 элементов столбца *Var2*, далее щелкнуть правой кнопкой мыши и в открывшемся контекстном меню щелкнуть на *Statistics of Block Data* → *Block Columns* → *Sums* (статистика блока данных → блок столбцов → суммы). Получите значение вероятности *p = 0,72* того, что не более 5 студентов сдадут экзамен на отлично.

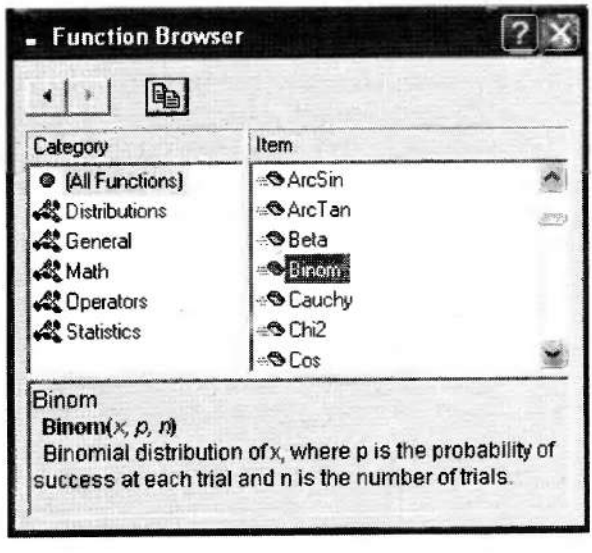


Рис. 7.2

Рассмотрим решение задачи с использованием распределения Пуассона. На факультете прикладной математики обучаются 685 студентов. Какова вероятность того, что 7 октября является днем рождения одновременно 7 студентов факультета, днем рождения более чем 2 студентов. Вероятность того, что день рождения студента 7 октября, равна $p = 1/365$. Так как вероятность $p = 1/365$ — мала, а $n = 685$ велико, применим формулу Пуассона при $\lambda = np = 6851/365 \approx 1{,}88$. Создайте пустую электронную таблицу. В первом столбце переменной *Var1* проставьте возможное число студентов — 0, 1, 2, ..., 7. Дважды щелкните по имени переменной *Var2*. Откроется диалоговое окно спецификации переменной *Var2*. В нижней части окна в поле **Long name** запишите функцию *Poisson* с указанием параметров. Этих параметров всего два. Первый, как и в предыдущем примере, — ссылка на количество успешных испытаний (*Var1*), а второй — $\lambda \approx 1{,}88$. Нажмите **OK**. Согласно формуле Пуассона программа вычислит вероятности и занесет их в столбец, соответствующий второй переменной *Var2*. Получите, что вероятность того, что день рождения семи студентов из 685 придется на 7 октября мала и составит $p = 0{,}0025$. *Для нахождения вероятности того, что 7 октября* является днем рождения более чем у 2 студентов, надо из единицы вычесть сумму первых трех элементов столбца *Var2*. Получите вероятность $p = 0{,}29$.

Рассмотрим вариант решения задачи с использованием геометрического распределения. Для изучения вкусов и предпочтений студентов на факультете прикладной математики проведены маркетинговые исследования. Исследования показали, что 65% студентов предпочитают по утрам пить растворимый кофе, 15% — натуральный кофе, остальные 20% пьют чай. Компания *Nescafe* решила провести повторные исследования среди любителей растворимого кофе для определения того, каким сортам кофе студенты отдают наибольшее предпочтение. Потенциальных участников опроса выбирали случайным образом. Какова вероятность того,

что только *k-й* из опрошенных является любителем растворимого кофе (*k* может принимать любое из значений 1, 2, 3...). Только *k-й* означает, что все опрошенные до него, начиная с 1-го и заканчивая *k-1-м*, не являются любителями кофе.

Создайте пустую электронную таблицу. Для обозначения числа неуспехов используйте столбец переменной *Var1*. Впишите в него значения 0, 1, ...9. Число неуспехов может быть сколь угодно большим, но ограничимся наибольшим значением 9. В столбец *Var2* программа запишет посчитанные вероятности. Дважды щелкните по имени переменной *Var2*. Откроется диалоговое окно спецификации переменной *Var2*. В нижней части окна в поле **Long name** запишите функцию *Geom(x;p)* с указанием параметров. Этих параметров всего два. Первый параметр x — ссылка на переменную, в строках которой указано количество неуспешных испытаний (в нашем случае «*V1*»), второй — вероятность $p = 0{,}65$. Нажмите **OK**. По формуле геометрического распределения программа вычислит вероятности и занесет их в столбец *Var2*.

Столбец *V0* будет содержать число проведенных испытаний, завершившихся упехом. Для нашего примера — это число опрошенных, последний из которых окажется любителем растворимого кофе. Так, например, вероятность 0,0097 соответствует случаю, когда всего опросили 5 человек, причем первые 4 — не любители кофе, а последний оказался любителем.

7.2. Подбор закона распределения

При обработке экспериментальных данных иногда возникает необходимость аппроксимировать эмпирическое распределение тем или иным известным законом распределения. Для этой цели в *STATISTICA* предназначен модуль **Distribution Fitting** (подгонка распределения). Для изучения этого модуля воспользуемся файлом **Turtles.sta** из библиотеки **Examples**. Чтобы запустить модуль **Distribution Fitting**, необходимо в главном меню **Statistics** выбрать одноименную команду. В открывшемся окне **Distribution Fiting** надо указать природу случайной величины, т.е. *Continuous Distribution* (непрерывная) или *Discret Distribution* (дискретная), а также предполагаемый закон распределения, которому случайная величина подчиняется. Для непрерывных случайных величин предложено шесть законов распределения, а для дискретных — четыре. Выберите, например, *Lognormal* и нажмите **OK**. В открывшемся окне укажите переменную *WIDTH* (ширина), для которой будет производиться исследование. Слева вверху становится активным выпадающее меню, и можно выбрать другой закон распределения. По умолчанию активной является вкладка **Quick**. На этой вкладке есть две кнопки: **Summary: Observed and expected distributions** (результат: наблюдаемые и ожидаемые распределения) и **Plot of Observed and expected distributions** (график наблюдаемых и ожидаемых распределений). Если нажать на первую кнопку, то программа отобразит численные характеристики в виде таблицы. Каждая строка этой таблицы характеризует интервал, в который попадают значения исследуемой переменной. В первом столбце *Observed Frequency* (наблюдаемая частота) для каждого рассмотренного интервала указано количество значений,

попавших в этот интервал. Во втором столбце *Cumulative Observed* (совокупный наблюдаемый) для каждого интервала приведено количество значений, попавших в этот и все предшествующие интервалы (накопленные частоты). В третьем и четвертом столбцах *Percent Observed* (процент наблюдаемый) и *Cumul. %* (суммарный процент) указаны те же величины, что и в предыдущих двух, но исчисленные в процентах. В пятом столбце *Frequency Expected* (ожидаемая частота) даны теоретические частоты, соответствующие логнормальному распределению.

При нажатии на вторую кнопку будет построена кривая теоретического закона распределения и гистограмма эмпирического, построенного по имеющимся данным (рис. 7.3). Над гистограммой выведен заголовок, в котором указана анализируемая переменная, предполагаемый закон распределения, а также три числовых параметра, которые рассмотрим подробнее. Первый параметр — это значение критерия χ^2. Чем меньше это значение, тем больше вероятность того, что проверяемая случайная величина имеет предполагаемый закон распределения. Второй параметр df — число степеней свободы. Определяется как $df = n\text{-}l\text{-}1$, где n — число интервалов, на которые разбит диапазон изменения случайной величины; l — число оцениваемых параметров распределения (для логнормального распределения $l = 2$). Третий параметр *p-уровень* значимости критерия, который определяет вероятность ошибки при отклонении гипотезы о нормальности. Так как вероятность ошибки достаточно велика, примерно 0,5 (что значительно больше 0,05), гипотезу о соответствии закона распределения логнормальному принимаем.

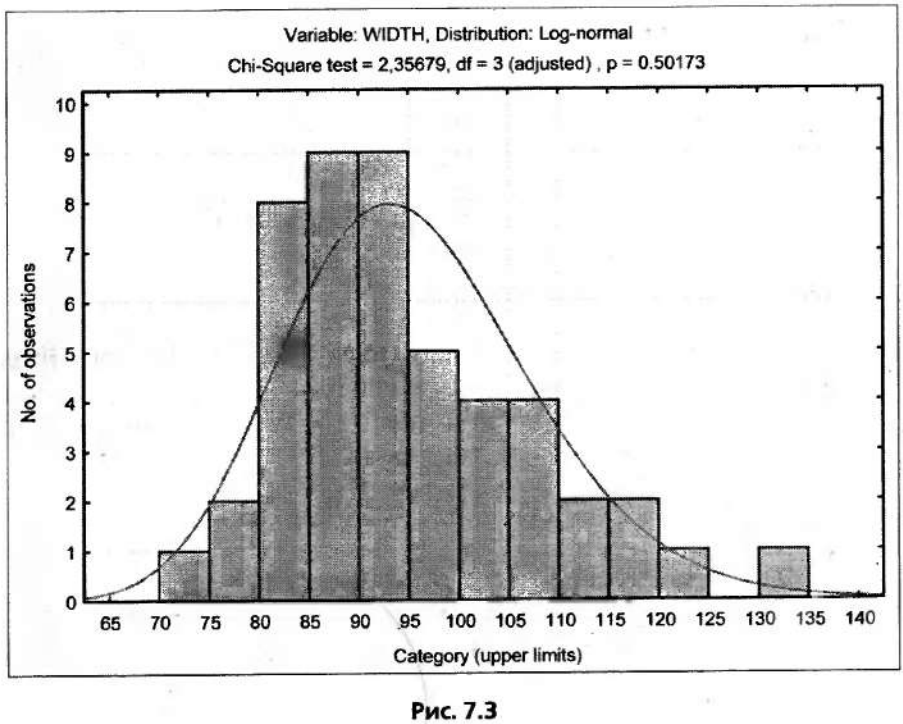


Рис. 7.3

Кратко опишем другие вкладки рабочего окна рассматриваемого модуля.

На вкладке **Parameters** приведены значения параметров предполагаемого закона распределения. Кнопка **Set To Default** — установить значения по умолчанию. Среди приведенных здесь параметров три являются общими для всех распределений, а остальные зависят от выбора распределений.

Number of categories (количество категорий) — количество интервалов, на которое будет разбита выборка.

Lower Limit, Uper Limit (нижний и верхний пределы). По умолчанию берутся, минимальное и максимальное значения выборки соответственно, однако изменив эти параметры, можно исключить из рассмотрения все значения, не попадающие в интересующий нас интервал.

Mean and Variance (среднее и дисперсия) — только для нормального распределения. Эти параметры определяются программой автоматически, но их можно переопределить и вручную, если, например, требуется не определить закон распределения, а проверить, насколько распределение случайной величины отличается от закона распределения с заданными параметрами.

На вкладке **Options** отображены четыре настройки.

Kolmogorov-Smirnov test — критерий Колмогорова-Смирнова проверки гипотезы о соответствии выборочных данных тому или иному закону распределения. Статистика Колмогорова равна максимальной абсолютной разности между гипотетической функцией распределения и эмпирической функцией распределения. Можно выбрать три опции: тест не вычисляется, вычисляется по группированным (интервальным) данным, вычисляется по негруппированным данным. При вычислении по негруппированным данным программа отсортирует наблюдаемые данные и вычислит кумулятивную ожидаемую частоту в каждой точке (очевидно, при этом возрастет время реализации критерия).

Chi-Square test. Критерий χ^2 (Пирсона) проверки гипотезы о соответствии выборочных данных тому или иному закону распределения. Если в интервал попало менее 5 значений, при установке флажка на *Combine Categories* он объединяется с соседним и т.д., пока количество значений в интервалах будет не менее 5. Для нового разбиения вычисляется значение критерия χ^2. В противном случае интервалы не объединяются.

Graph Plot Distributions. Для вкладки **Quick** (кнопка **Plot of Observed and expected distributions**) устанавливается тип графика. Если установить флажок на *Frequency distribution*, то программа построит график плотности распределения. Если флажок установить на *Cumulative distributions*, будет построен график функции распределения.

Plot row frequencies or %. Если установить флажок на *Raw frequencies*, на вертикальной оси графика будут отложены значения относительных частот, в противном случае — их процентные отношения.

Рассмотрим пример приближения эмпирического распределения пуассоновским. Фирма **L&L** является дистрибьютором компании **Desa**. Ежедневно десятки автомобилей фирмы в течение дня развозят продукцию компании в различные торговые организации. Для уменьшения времени простоя автомобилей в очереди

при загрузке было решено изучить закон распределения количества автомобилей, подъезжающих к складским помещениям в течение часа, что позволит определить оптимальное количество кладовщиков, грузчиков, погрузочных площадок для организации эффективной работы складов. С этой целью 200 раз было подсчитано количество автомобилей, подъехавших в течение часа к складским помещениям. Данные файла **L&L**, которые представляют собой один столбец с 200 элементами, приведены на рис. 7.4. в виде прямоугольной таблицы.

	Количество автомобилей, подъехавших в течении часа									
	1	2	3	4	5	6	7	8	9	10
1	9	6	7	5	8	3	4	6	3	3
2	8	5	5	3	5	5	5	7	6	6
3	6	3	7	5	6	3	4	5	6	6
4	4	8	5	3	7	1	5	6	3	3
5	3	9	5	5	5	5	4	6	5	5
6	5	4	5	8	5	7	7	9	4	4
7	5	2	6	8	5	6	5	5	6	6
8	5	7	5	5	13	3	7	3	3	3
9	8	3	3	5	8	2	4	7	3	3
10	6	7	2	7	5	3	6	4	2	2
11	4	8	6	10	4	4	3	3	6	6
12	6	4	7	1	4	3	2	4	6	6
13	4	3	3	3	4	3	3	4	7	7
14	3	5	3	4	3	4	5	3	9	9
15	6	4	5	5	6	5	6	2	9	9
16	7	4	6	5	2	1	6	5	4	4
17	9	8	7	8	6	8	7	5	6	6
18	4	2	3	3	4	8	3	7	6	6
19	10	7	2	5	5	6	7	4	3	3
20	4	7	3	4	10	2	5	6	9	9

Рис. 7.4

В главном меню **Statistics** выберите команду **Distribution Fitting**. В открывшемся окне надо указать вид случайной величины, а именно — *Discret Distribution* (дискретная). В списке дискретных распределений укажите предполагаемый закон распределения, например *Poisson*. Щелкните по **OK**. В открывшемся окне на вкладке **Quick** нажмите кнопку **Plot of Observed and expected distributions.** Программа построит гистограмму — график эмпирической плотности распределения — и обозначит красной линией кривую предполагаемого теоретического распределения (рис. 7.5). Уровень значимости критерия p = 0,22 принимает достаточно большое значение, чтобы можно было отвергнуть гипотезу о соответствии закона распределения пуассоновскому. В информационном поле указано значение параметра = 5,11, равное среднему числу автомобилей, находящихся в течение часа в очереди на загрузку. Так как модели теории массового обслуживания предполагают соответствие закона распределения входного потока заявок на обслуживание распределению Пуассона, проведенное исследование позволит использовать методы

теории массового обслуживания для минимизации очереди автомобилей при загрузке товара. В рассмотренном примере программа анализировала закон распределения для данных, представленных в виде несгруппированного ряда — перечислены значения случайной величины в порядке их появления в выборке. В то же время программа может анализировать данные, прошедшие обработку и записанные в виде сгруппированного ряда — перечислены ранжированные значения случайной величины и соответствующие им частоты в выборке.

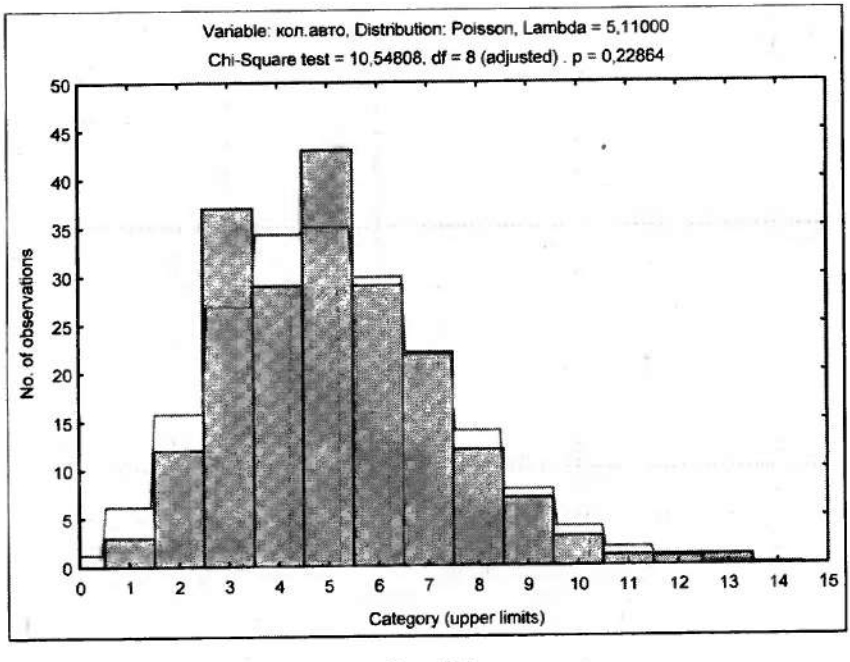


Рис. 7.5

Предположим, надо проанализировать закон распределения количества ошибок при написании студентами коллоквиума по математическому анализу. Сгруппированный ряд количества ошибок представлен в виде таблицы на рис. 7.6. Как и в предыдущем случае в главном меню **Statistics** выберите команду **Distribution Fitting.** В открывшемся окне укажите вид случайной величины — **Discret Distribution**. В списке дискретных распределений укажите предполагаемый закон распределения, например, распределение Пуассона (*Poisson*). Щелкните по **OK**. В открывшемся окне **Select the variable for analysis** надо указать имя переменной *Кол.ошиб.* После нажатия на **OK** программа возвратится в исходное рабочее окно **Fitting Discrete Distributions**, в котором надо щелкнуть по кнопке **W**(веса). В открывшемся окне в поле **Weight variable** наберите имя переменной *Кол.студ.* и произведите установки опций в соответствии с рис. 7.7.

	1 Кол.ошиб.	2 Кол.студ.
1	0	2
2	1	10
3	2	27
4	3	32
5	4	23
6	5	6

Рис. 7. 6

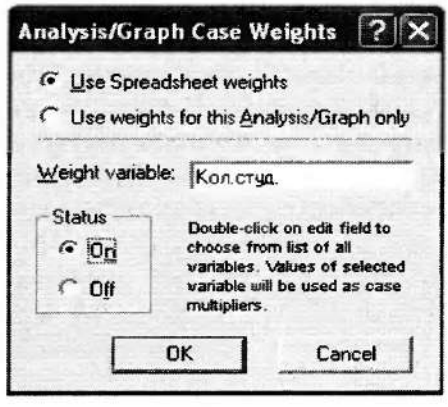


Рис. 7.7

Щелкните по **OK** и, вернувшись в окно **Fitting Discrete Distributions,** нажмите кнопку **Plot of Observed and expected distributions.** Программа построит гистограмму (рис. 7.8) эмпирического распределения с нанесенной на нее ломаной линией пуассоновского распределения.

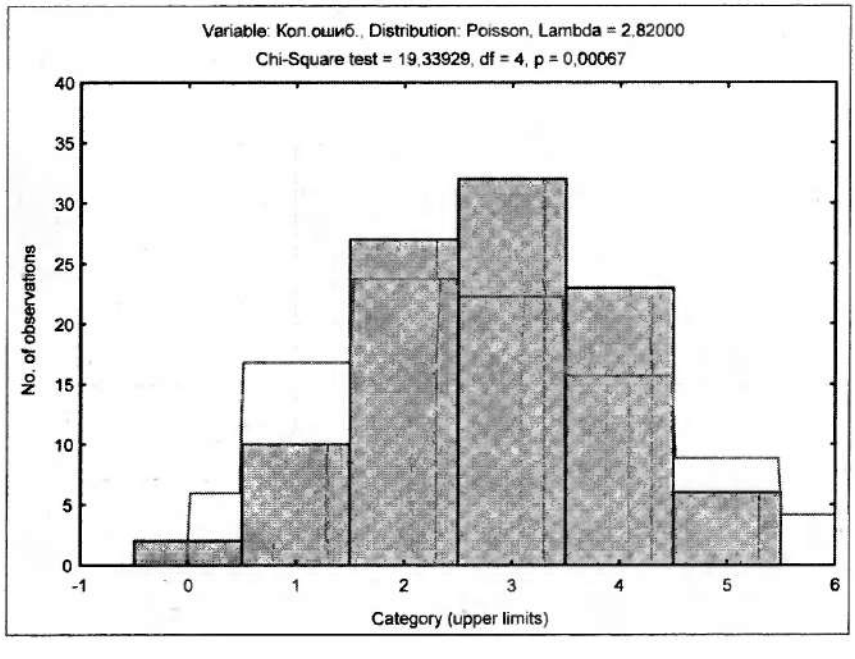


Рис.7.8

Из графиков, значений $\chi^2 = 19{,}33$ и уровня значимости вероятности $p = 0{,}00007$ следует вывод о несоответствии закона распределения количества ошибок распределению Пуассона. Повторим всю изложенную процедуру, указав в списке дискретных распределений — *Binomial*. Программа построит новый график (рис. 7.9), из которого видно соответствие гистограмм; значение $\chi^2 = 0{,}205$; $p = 0{,}978$. Малое значение χ^2 и большое значение p говорят о большой вероятности ошибки, если отвергнуть гипотезу о соответствии закона распределения количества ошибок биномиальному закону. Поэтому гипотезу принимаем.

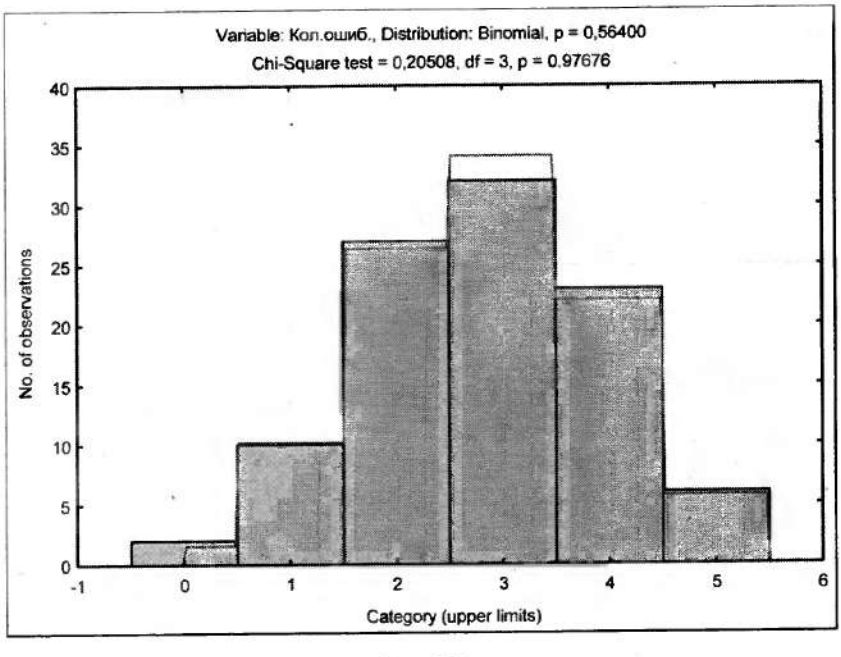


Рис. 7.9

7.3. Генерация случайных чисел

В программе *STATISTICA 6.0* имеется возможность генерировать случайные числа, подчиняющиеся равномерному, нормальному и пуассоновскому законам распределения. Известно, что с увеличением объема выборки возрастает соответствие эмпирического закона распределения теоретическому. Так, например, если количество генерируемых чисел более 1000, то отклонение эмпирического закона от теоретического практически незаметно. Если генерируется примерно 500 чисел, то видны значительные отклонения от закона распределения. При количестве случайных величин менее 100 охарактеризовать выборку каким-либо законом распределения весьма затруднительно.

Для генерации случайных чисел надо дважды щелкнуть в таблице данных (в которой предполагается записать сгенерированные числа) на имени переменной.

В окне спецификации переменной нажмите кнопку **Functions**. В открывшемся окне надо выделить *All Functions* и выбрать нужную функцию.

RND(X) (генерация равномерно распределенных чисел). Эта функция имеет только один параметр — *X*, который задает правую границу интервала, содержащего случайные числа. При этом 0 является левой границей. Аналогичные действия выполняет и функция *Uniform(X)*. Чтобы вписать общий вид функции *RND(X)* в окно спецификации переменной, достаточно дважды щелкнуть на имени функции в окне **Function Browser**. После указания числового значения параметра *X* надо нажать **OK**. Программа выдаст сообщение о правильности написания функции и запросит подтверждение о пересчете значения переменной. После подтверждения соответствующий столбец заполняется случайными числами. Для визуализации распределения сгенерированных чисел и оценки соответствия закону распределения можно воспользоваться модулем **Distribution Fitting**. Для изменения значения математического ожидания на величину **g** надо заменить общий вид функции на *RND(X)* + *g*.

RNDNormal (генерация нормально распределенных чисел). Эта функция имеет один параметр — *X*, соответствующий стандартному отклонению случайной величины с математическим ожиданием 0. Запись *RND-Normal(X)* + *g* означает генерацию чисел с математическим ожиданием *g*.

RndPoisson (генерация чисел, соответствующих распределению Пуассона). Функция имеет один параметр — *X*, соответствующий среднему значению. При необходимости генерирования случайных чисел других законов распределения надо воспользоваться известными в теории вероятностей и математической статистике соотношениями.

Глава 8
Дисперсионный анализ

Сравнение средних является одним из способов выявления зависимостей между переменными. Так, например, если при разбиении объектов исследования на подгруппы при помощи категориальной независимой переменной (предиктора) верна гипотеза о неравенстве средних некоторой зависимой переменной в подгруппах, то это означает, что существует стохастическая взаимосвязь между этой зависимой переменной и категориальным предиктором. Наиболее общим методом сравнения средних является дисперсионный анализ — **ANOVA (Analysis of Variance)**. В терминологии дисперсионного анализа категориальный предиктор называется фактором.

Дисперсионный анализ можно определить как статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования экспериментов.

Таким образом, в дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов).

Рассмотрим сначала основные идеи однофакторного дисперсионного анализа [11]. Представим исходные данные в виде таблицы, строки и столбцы которой отображают различные уровни фактора, а в ячейках таблицы расположены значения анализируемого признака (зависимой переменной). Такая таблица называется планом эксперимента.

Фактор	Значения переменной			
Группа 1	x_{11}	x_{12}	...	x_{1n}
Группа 2	x_{21}	x_{22}	...	x_{2n}
... ...	...	...	...	...
Группа m	x_{m1}	x_{m2}		x_{mn}

Однофакторная, дисперсионная модель имеет следующий вид:

$$x_{ij} = \mu + F_i + \varepsilon_{ij},$$

где x_{ij} — значение исследуемой переменной, соответствующей *i-й* группе (*i-му* уровню фактора) с *j-м* порядковым номером ($i = 1, ..., m; j = 1, ..., n$), μ — общая средняя, F_i — эффект, обусловленный влиянием *i-го* уровня фактора, ε_{ij} — случайная компонента, или возмущение, вызванное влиянием неконтролируемых факторов, т.е. вариацией переменных внутри отдельного уровня факторов.

Предположим, что элементы строк таблицы — реализации случайных величин $X_1, X_2, ..., X_m$, имеющих нормальный закон распределения с математическими ожиданиями $a_1, a_2, ..., a_m$ и одинаковыми дисперсиями σ^2 Тогда задача сравнения средних в группах сведется к проверке нулевой гипотезы —

$$H_0: a_1 = a_2 = ... = a_m.$$

Обозначим выборочные средние в группах $\bar{x}_{i\bullet}$, а общую выборочную среднюю — $\bar{x}$. Тогда

$$\bar{x}_{i\bullet} = \sum_{j=1}^{n} x_{ij}/n, \qquad \bar{x} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}/mn = \sum_{i=1}^{m} \bar{x}_{i\bullet}/m.$$

Можно показать, что сумму квадратов Θотклонений наблюдений x_{ij} от общей средней $\bar{x}$ можно представить следующим образом:

$$\Theta = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\bar{x}_{i\bullet} - \bar{x}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 = n\sum_{i=1}^{m}\left(\bar{x}_{i\bullet} - \bar{x}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2.$$

Обозначим слагаемые в правой части равенства, соответственно Θ_1 и Θ_2. Получим соотношение $\Theta = \Theta_1 + \Theta_2$. Здесь Θ — общая, или полная, сумма квадратов отклонений, Θ_1 — межгрупповая (факторная) сумма квадратов отклонений, Θ_2 — внутригрупповая (остаточная) сумма квадратов отклонений. Полученное равенство показывает, что общая изменчивость признака, измеренная величиной Θ, состоит из двух компонент, одна из которых характеризует изменчивость признака между группами (Θ_1), вторая — изменчивость внутри групп (Θ_2). В дисперсионном анализе используются не сами суммы квадратов отклонений Θ_1, Θ_2, а усредненные квадраты отклонений S_1, S_2, получающиеся делением последних на число степеней свободы. Число степеней свободы определяется как общее

число наблюдений минус число связывающих их уравнений. Для Θ_1 число степеней свободы равно $l_1 = m - 1$, для $\Theta_2 - l_2 = mn - m$. Таким образом, $S_1 = \Theta_1 / m - 1$, $S_2 = \Theta_2 / mn - m$. В терминах модуля **ANOVA**, Θ_1 называют эффектом, а Θ_2 называют ошибкой. S_1, S_2 называют, соответственно, *MS* эффекта и *MS* ошибки. Можно показать, что проверка нулевой гипотезы сводится к проверке существенности различия *MS* эффекта и *MS* ошибки, которые являются оценками дисперсии σ^2. *MS* эффекта и *MS* ошибки можно сравнить с помощью *F-критерия*. Гипотеза H_0 отвергается, если $F = S_1/S_2$ больше табличного F_α, l_1, l_2.

Если сравниваются средние в двух выборках, дисперсионный анализ даст тот же результат, что и обычный *t-критерий* для независимых переменных (если сравниваются две независимые группы наблюдений) или *t-критерий* для зависимых выборок (если сравниваются две переменные на одном и том же множестве наблюдений).

Основная причина, по которой использование дисперсионного анализа предпочтительнее повторного сравнения двух выборок при разных уровнях факторов с помощью серий *t-критерия*, заключается в том, что дисперсионный анализ существенно более эффективен и для малых выборок более информативен. Еще одно преимущество дисперсионного анализа по сравнению с *t-критерием* заключается в том, что он позволяет обнаруживать взаимодействия между факторами и, следовательно, изучать более сложные модели [16].

Идея однофакторного дисперсионного анализа перенесена в многофакторный анализ. Более сложными становятся *S* факторный план эксперимента и процедуры вычисления *MS* эффекта и *MS* ошибки. Так, например, для двухфакторного дисперсионного анализа факторный план можно представить в виде табл. 8.1.

Таблица 8.1

Факторы	Группа 1*	Группа 2*		Группа *k**
Группа 1	$x_{111}, \ldots, x_{11j}$	$x_{121}, \ldots, x_{12j}$	...	$x_{1k1}, \ldots, x_{1kj}$
Группа 2	$x_{211}, \ldots, x_{21j}$	$x_{221}, \ldots, x_{22j}$	...	$x_{2k1}, \ldots, x_{2kj}$
.....	...	...	...	...
Группа m	$x_{m11}, \ldots, x_{m1j}$	$x_{m21}, \ldots, x_{m2j}$		$x_{mk1}, \ldots, x_{mkj}$

Применение дисперсионного анализа целесообразно, если анализируемые признаки измерены минимум в интервальной шкале и имеют нормальное распределение внутри сравниваемых групп, дисперсии в группах однородны. Но следует заметить, что *F-критерий*, применяемый в дисперсионном анализе, устойчив к отклонению от нормальности и однородности дисперсий. Если условия применимости дисперсионного анализа не выполнены, можно воспользоваться непараметрическими критериями сравнения средних или модулем «Общие линейные модели» (*GLM*).

Ключевыми понятиями дисперсионного анализа являются главные эффекты и сложные эффекты, которые позволяют обнаружить и исследовать взаимодействия между факторами. Для иллюстрации этих понятий рассмотрим простой

пример. Студенты группы А специализируются на кафедре математического моделирования, студенты группы В — на кафедре прикладной математики, а студенты группы С — на кафедре численного анализа. Количество студентов в группах одинаковое. Дисциплины «Теория вероятностей и математическая статистика», «Уравнения математической физики», «Функциональный анализ» ведутся, соответственно, преподавателями кафедр: прикладной математики, математического моделирования, численного анализа. После завершения изучения этих дисциплин проведено тестирование знаний студентов. В качестве показателя качества усвоенных знаний использовали суммарное количество правильных ответов. Результаты тестирования представлены в табл. 8.2.

Таблица 8.2

	Группа *A*	Группа *B*	Группа *C*
УМФ	70	20	30
ТВиМС	10	80	20
Функциональный анализ	15	25	70

Данная таблица представляет двухфакторный план, факторами являются группа студентов и изучаемый предмет. Существует ли взаимодействие этих факторов? И если да, то какой характер оно приобретает? Анализируя таблицу, можно сделать вывод, что студенты отдают большее предпочтение дисциплинам, которые ведутся преподавателями кафедр, на которых они специализируются. Этот вывод является главным эффектом взаимодействия факторов.

Можно усложнить задачу, добавив еще один фактор, например, разделив студентов на юношей и девушек. Если число факторов больше двух, то объяснить взаимодействия высших порядков значительно сложнее. В этом случае надо воспользоваться графиками средних, которые позволяют легко интерпретировать сложные эффекты.

В общем случае взаимодействие между факторами описывается в виде изменения одного эффекта под воздействием другого. В рассмотренном ранее примере двухфакторное взаимодействие можно описать как изменение главного эффекта фактора, характеризующего изучаемый предмет, под воздействием фактора, описывающего принадлежность студентов к определенной группе.

8.1. Описание процедуры *Factorial ANOVA*

Рассмотрим работу процедуры **ANOVA,** используя файл **Crabs** (крабы) из библиотеки **Examples** (рис. 8.1). В файле приведены данные по количеству спутников (*SATELLTS*) — особей мужского пола у особей женского пола в зависимости от их цвета (*COLOR*), состояния клешней (*SPINE*), размеров (*CATWIDTH*,

WIDTH — ширина) и веса (*WEGHT*). Если число спутников больше 0, то переменная *Y* в первом столбце принимает значение 1, в противном случае — 0. Общее число наблюдений (крабов) равно 173.

Для запуска программы в верхнем меню **Statistics** надо выбрать команду **ANOVA**, что переводится как анализ вариаций или дисперсионный анализ. Появится стартовая панель **General ANOVA/MANOVA** (рис. 8.2).

	Number of crab satellites by female's color, spine condition, width,						
	1 Y	2 COLOR	3 SPINE	4 WIDTH	5 SATELLTS	6 WEIGHT	7 CATWIDT
1	1	medium	bothworn	28,3	8	3,05	28,
2	0	darkmed	bothworn	22,5	0	1,55	22,
3	1	lightmed	bothgood	26,0	9	2,30	25,
4	0	darkmed	bothworn	24,8	0	2,10	24,
5	1	darkmed	bothworn	26,0	4	2,60	25,
6	0	medium	bothworn	23,8	0	2,10	23,
7	0	lightmed	bothgood	26,5	0	2,35	26,
8	0	darkmed	oneworn	24,7	0	1,90	24,

Рис. 8.1

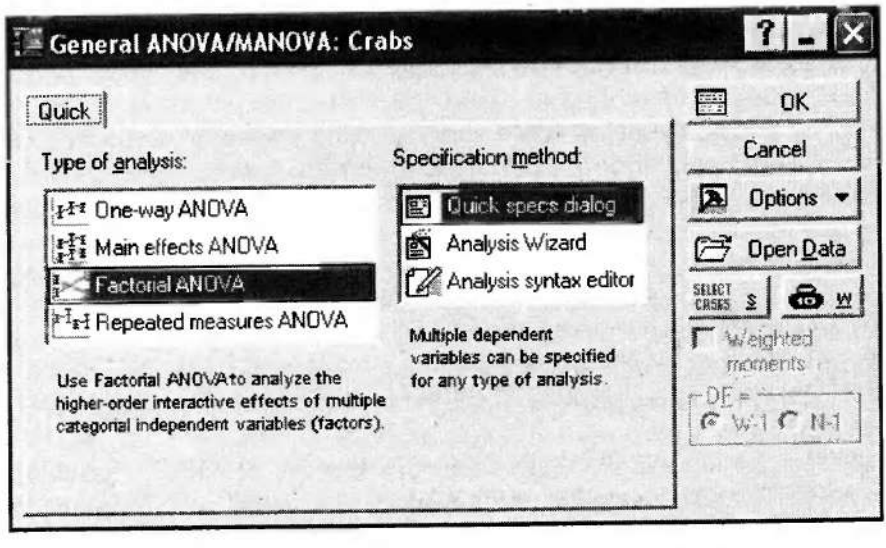


Рис. 8.2

Данный диалог содержит два списка **Type of analysis** (вид анализа) и **Specification method** (задание метода).

Список **Type of analysis** состоит из четырех элементов, представляющих собой различные модели дисперсионного анализа:

- **One-way ANOVA** (однофакторный дисперсионный анализ);
- **Main effects ANOVA** (дисперсионный анализ главных эффектов);

- **Factorial ANOVA** (многофакторный дисперсионный анализ);
- **Repeat measures ANOVA** (дисперсионный анализ повторных измерений).

Список **Specification method** позволяет задать три типа интерфейса дисперсионного анализа в *STATISTICA*:

- **Quick Specs Dialog** (диалог быстрых спецификаций);
- **Analysis Wizard** (мастер анализа);
- **Analysis syntax editor** (редактор кода).

В диалоге **Quick Specs Dialog** можно задать зависимые переменные и категориальные переменные (предикторы). Вариация числа и типа переменных зависит от выбранного вида анализа в списке **Type of analysis**.

Диалог **Analysis Wizard** предназначен для задания по шагам интересующего анализа в рамках выбранной модели. В конце анализа можно вычислить результаты или использовать **Analysis syntax editor** для дальнейшей настройки при помощи встроенных команд, открыть существующий файл с командами или сохранить для дальнейшего использования.

Диалог **Analysis syntax editor** позволяет полностью настроить как параметры плана, так и параметры вычислительных процедур. В случае необходимости можно сохранить файл с готовым кодом анализа для дальнейшего использования или открыть уже существующий.

После выбора диалога **Specification method** можно задать **Type of analysis**.

One-way ANOVA позволяет оценить эффект одной группирующей переменной (одного межгруппового фактора) на одну или более зависимых переменных.

Для анализа **Main effects ANOVA** в диалоге **Quick Specs Dialog** можно задать до четырех категориальных предикторов. Затем программа произведет оценку модели главных эффектов. Данный тип планов часто используется в анализе и планировании промышленных экспериментов для оценки большого набора факторов в сильно раздробленных планах. Также данный тип планов используется при анализе сбалансированных неполных планов.

В отличие от рассмотренных типов анализа, в **Factorial ANOVA** учитывается еще один возможный источник изменчивости — взаимодействие факторов. Планы содержат переменные, которые представляют комбинации различных уровней двух или более категориальных предикторов. В частности, полные факторные планы представляют все возможные комбинации уровней категориальных предикторов. Полный факторный план с двумя категориальными предикторами *A* и *B*, каждый из которых имеет по два уровня, будет являться 2 x 2 полным факторным планом. В диалоге **Quick Specs Dialog** также можно задать до четырех категориальных предикторов. Данные планы часто используются в анализе и планировании промышленных экспериментов.

В **Repeat measures ANOVA** зависимые переменные содержат значения одного фактора повторных измерений. В диалоге **Quick Specs Dialog** также можно задать до четырех категориальных предикторов и две или более зависимые переменные, которые будут проинтерпретированы программой как повторные измерения одного фактора.

Для всех типов дисперсионного анализа, если необходимо использовать пять или более категориальных предикторов, надо воспользоваться модулем **GLM** (общие линейные модели).

Для того чтобы задать план факторного дисперсионного анализа, выберите **Factorial ANOVA** в качестве вида анализа и **Quick Spec Dialog** в списке **Specification method** на вкладке **Quick** стартовой панели дисперсионного анализа. Откроется диалоговое окно **ANOVA/MANOVA Factorial ANOVA**. На вкладке **Quick** нажмите кнопку **Variables**. В появившемся окне выберите группирующие переменные *COLOR* и *SPINE*, зависимые *WIDTH, WEIGHT* (рис. 8.3). Различные цвета и состояния клешней крабов являются межгрупповыми факторами. Если число зависимых переменных — более 1, то программа осуществит многомерный дисперсионный анализ. Чтобы вручную задавать коды для межгрупповых факторов, нажмите кнопку **Factor Codes** (коды факторов). Необязательно коды задавать вручную, так как программа задаст по умолчанию все коды выбранных переменных. Кодами предиктора *COLOR* являются цвета крабов: *medium* (серый); *lightmed* (светло-серый); *dark* (темный); *darkmed* (темно-серый). Кодами предиктора *SPINE* являются состояния клешней крабов: *bothgood* (обе клешни целые); *oneworr* (одна клешня повреждена); *bothworr* (обе клешни повреждены).

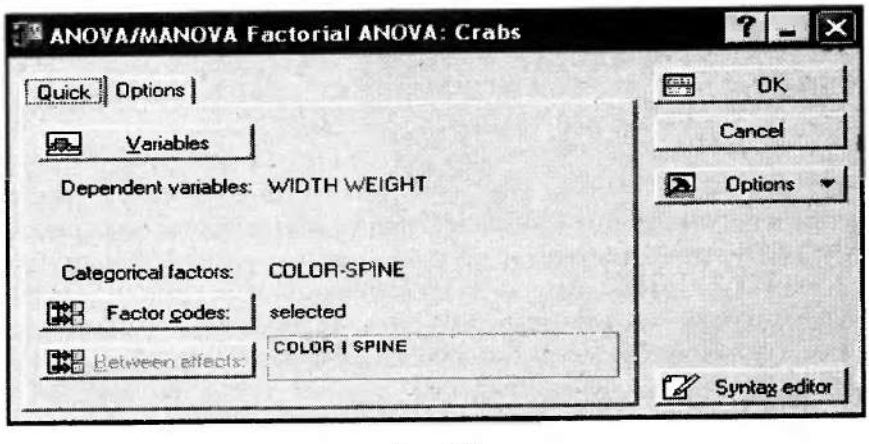


Рис. 8.3

Кнопка **Options** используется для задания параметров вычисления, кнопка **Syntax editor** (редактор кода) позволяет произвести дальнейшие настройки модели при помощи синтаксиса анализа.

Щелкните по кнопке **ОК**, откроется диалоговое окно (рис. 8.4) **ANOVA Results 1** (результаты анализа) с набором вкладок, которые позволяют всесторонне отобразить результаты анализа в виде таблиц и графиков.

На вкладке **Means** можно указать различные способы вычисления средних.

Наблюдаемые невзвешенные (*Unweighted*) средние вычисляются при помощи усреднения средних по уровням и комбинациям уровней факторов, которые не использовались в таблице (или графике) маргинальных средних, после этого производится деление полученного значения на количество средних.

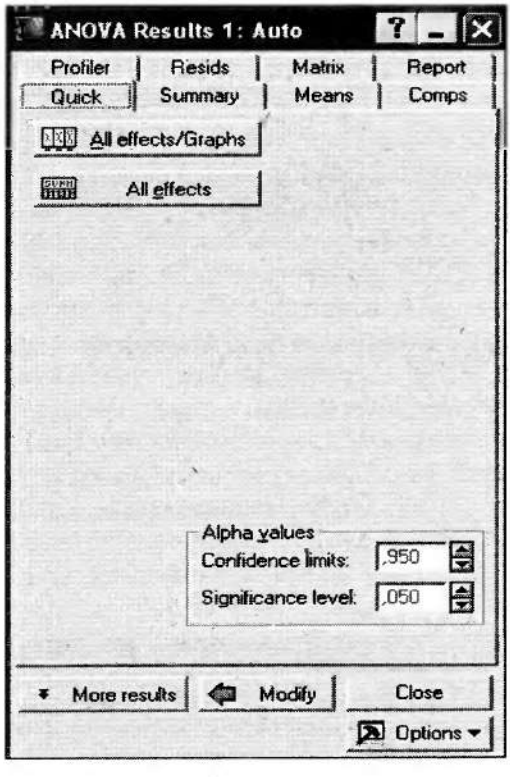


Рис. 8.4

Наблюдаемые взвешенные (*Weighted*) средние вычисляются как стандартные средние значения для соответствующих комбинаций уровней факторов. Поэтому итоговые средние являются взвешенными маргинальными средними, поскольку они «взвешиваются» соответствующими количествами наблюдений в каждой ячейке плана. В полных факторных планах взвешенные маргинальные средние можно также вычислить, усреднив средние значения ячеек для каждого маргинального среднего, а затем взвесив их с помощью количеств наблюдений в соответствующих ячейках.

Оценки МНК (средние наименьшие квадраты — *Least squares*) являются ожидаемыми маргинальными средними генеральной совокупности для текущей модели. Отметим, что для полных факторных планов без пропущенных ячеек средние частных наименьших квадратов идентичны наблюдаемым невзвешенным средним. Средние наименьших квадратов также иногда называются предсказанными средними, поскольку они являются предсказанными значениями, когда все факторы в модели равны соответствующим средним или уровням факторов для соответствующих средних.

На вкладке **Quick** нажмите кнопку **All effects/Graphs** (все эффекты/графики). Данный диалог **Table of All Effects** (таблицы всех эффектов) (рис. 8.5) содержит результаты и используется для просмотра выбранных из данной таблицы эффектов в виде графиков средних или таблиц.

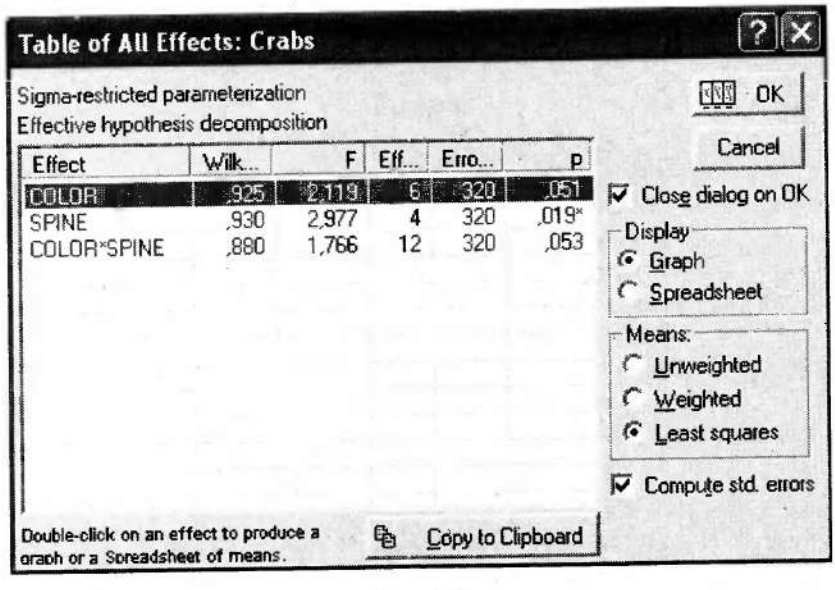


Рис. 8.5

Значимые эффекты ($p<0,05$) в таблице *Table of All Effects* помечены *. Видно, что гипотеза о неравенстве средних верна только для предиктора *SPINE*. Для предиктора *COLOR* и взаимодействия предикторов *COLOR*SPINE* уровень значимости незначительно превосходит 0,05. Можно изменить значимость критерия, введя необходимое значение параметра *Alpha* (альфа) в поле **Significance level** (уровень значимости) вкладки **Quick** окна **ANOVA Results 1**. Выделите, например, опцию *Spreadsheet* (таблица) в рамке **Display** (отображать) и два раза щелкните на эффекте *SPINE* или, выделив эффект *SPINE*, нажмите **OK**. Появится таблица (рис. 8.6) со значениями средних всех зависимых переменных и другими статистиками в группах, соответствующих межгрупповым факторам (трем уровням категориального предиктора) *SPINE – bothgood; oneworr; bothworr*.

Вернитесь в окно **Table of All Effects** и выделите опцию *Graph* (график) в рамке **Display**, нажмите **OK**. В появившемся окне выберите, например, зависимую переменную *WIDTH*. Программа построит график средних переменной *WIDTH* (рис. 8.7).

Cell	SPINE	WIDTH Mean	WIDTH Std.Err.	WIDTH -95,00%	WIDTH +95,00%	SATELLTS Mean	SATELLTS Std.Err.	SATELLTS -95,00%	SATELLTS +95,00%	WE Me
	SPINE; LS Means (Crabs) Wilks lambda=,91021, F(8, 316)=1,9024, p=,05907 Effective hypothesis decomposition									
1	bothgood	26,80	0,60	25,62	27,99	3,27	0,95	1,40	5,139039	
2	oneworn	24,11	0,67	22,78	25,43	2,00	1,06	-0,10	4,102355	
3	bothworn	25,94	0,51	24,92	26,95	1,94	0,81	0,34	3,547349	

Рис. 8.6

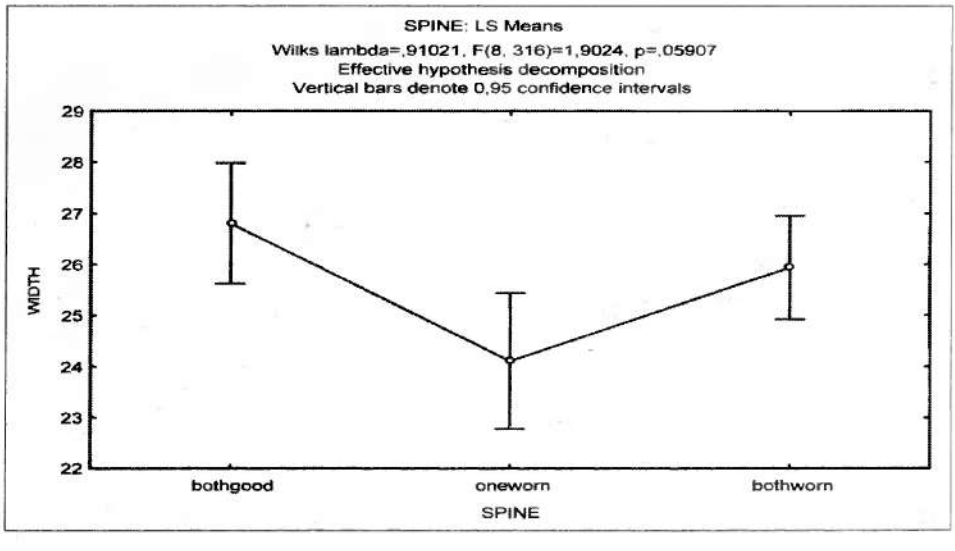


Рис. 8.7

Из графика видно, что средняя ширина крабов с двумя целыми клешнями превосходит ширину крабов с двумя поврежденными клешнями и значительно превосходит ширину крабов с одной поврежденной клешней. По-видимому, более широкие крабы обладают большей силой, и это позволяет им сохранить клешни в целости.

Выделите опцию *Graph* в рамке **Display** и нажмите **OK**, в открывшемся окне (рис. 8.8) **Dependent vars for the...** укажите имя переменной *WIDTH*. Щелкните **OK,** появится окно (рис. 8.9) **Arrangement of Factors** (расположение факторов), в котором можно указать порядок выбора взаимодействующих факторов. Выберите *COLOR* в поле *x-axis, upper* (ось x, верх) и *SPINE* в поле *line pattern* (шаблон линии). Нажмите кнопку **OK**, появятся графики средних (рис. 8.10).

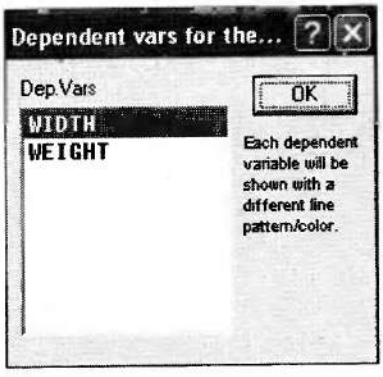


Рис. 8.8

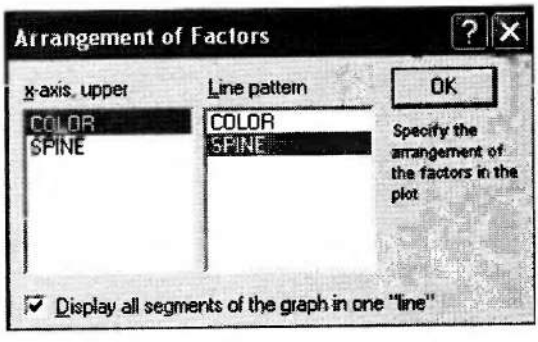


Рис. 8.9

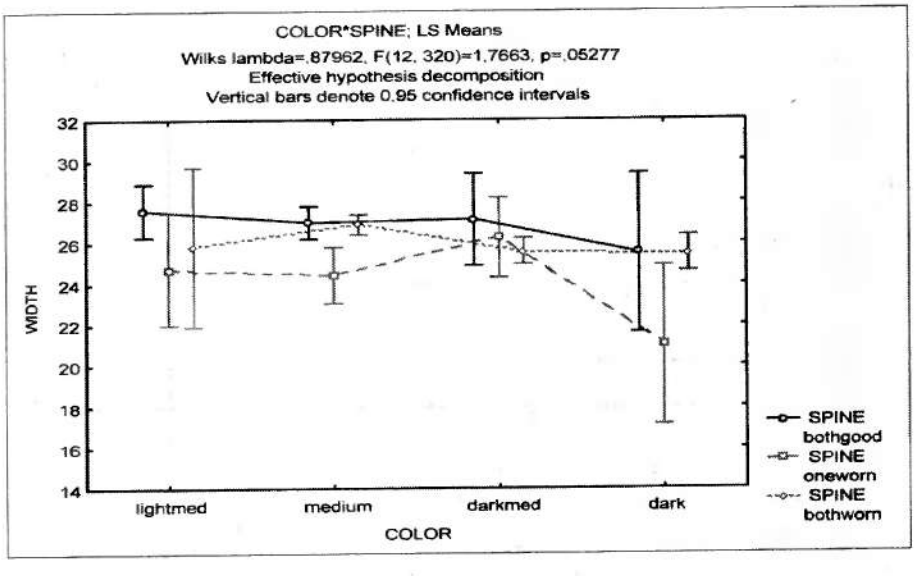


Рис. 8.10

Из приведенных графиков видно, что у крабов светло-серого цвета с двумя целыми клешнями и у крабов темного цвета с одной поврежденной клешней средняя ширина соответственно больше и меньше, чем во всех остальных группах. Независимо от цвета средняя ширина крабов с одной поврежденной клешней меньше, чем средняя ширина крабов с двумя целыми клешнями. Приведенные результаты показывают, что существуют различия между средними в группах, соответствующих различным межгрупповым факторам. Но значимы ли эти различия? Для ответа на этот вопрос нужно использовать апостериорные сравнения для проверки разности средних.

В диалоге **ANOVA Results 1** нажмите кнопку **More results** и в открывшемся окне выберите вкладку **Post-hoc**, на которой представлены различные апостериорные критерии (рис. 8.11). Все эти критерии позволяют сравнивать средние

при отсутствии априорной гипотезы относительно этих средних. Большое количество критериев минимизирует вероятность случайных результатов.

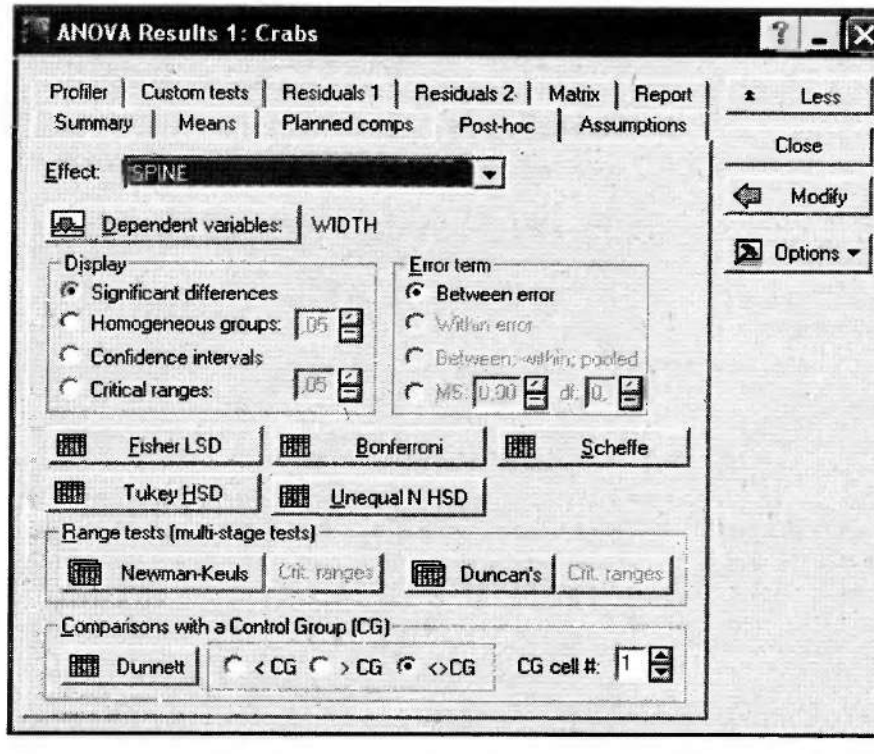


Рис. 8.11

Выберите зависимую переменную *WIDTH*, эффект *SPINE* и нажмите кнопку **Fisher LSD**. В открывшейся таблице (рис. 8.12) в первой строке приведены значения средних, в столбце 1 — названия групп, в остальных ячейках — уровни значимости. Из таблицы следует, что неверна гипотеза о равенстве средних, т.е. средняя ширина крабов статистически значимо отличается во всех группах, соответствующих различным уровням предиктора *SPINE*.

Более интересный результат получится, если в диалоге **ANOVA Results 1** для исследования взаимодействия предикторов выбрать эффект *COLOR*SPINE*. Так, из таблицы, изображенной на рис. 8.13, следует, что средняя ширина крабов светло-серого цвета с обеими целыми клешнями (27,58) значимо больше, чем средняя ширина крабов серого цвета с одной поврежденной клешней (24,42). Средняя ширина крабов темного цвета с одной поврежденной клешней (27,58) значимо больше, чем средняя ширина крабов серого цвета с обеими целыми клешнями (26,99).

Cell No.	LSD test; variable WIDTH (Crabs) Probabilities for Post Hoc Tests Error: Between MS = 3,8563, df = 161,00 SPINE	{1} 27,111	{2} 24,727	{3} 26,245
1	bothgood		0,000110	0,020209
2	oneworn	0,000110		0,005320
3	bothworn	0,020209	0,005320	

Рис. 8.12

Cell	LSD test; variable WIDTH (Crabs) Probabilities for Post Hoc Tests Error: Between MS = 3,8563, df = 161,00 COLOR	SPINE	{1} 27,58	{2} 24,75	{3} 25,80	{4} 26,99	{5} 24,42	{6} 26,88	{7} 27,13	{8} 26,25	{9} 25,58	{10} 25,50	{11} 21,00	{12} 25,485
1	lightmec	bothgood		0,07	0,39	0,44	0,00	0,31	0,73	0,26	0,01	0,31	0,00	0,01
2	lightmec	oneworn	0,07		0,66	0,12	0,83	0,13	0,19	0,38	0,56	0,76	0,12	0,61
3	lightmec	bothworn	0,39	0,66		0,55	0,51	0,59	0,56	0,84	0,91	0,91	0,09	0,88
4	medium	bothgood	0,44	0,12	0,55		0,00	0,81	0,91	0,48	0,01	0,46	0,00	0,01
5	medium	oneworn	0,00	0,83	0,51	0,00		0,00	0,04	0,13	0,13	0,61	0,10	0,20
6	medium	bothworn	0,31	0,13	0,59	0,81	0,00		0,83	0,53	0,00	0,49	0,00	0,01
7	darkmec	bothgood	0,73	0,19	0,56	0,91	0,04	0,83		0,56	0,19	0,47	0,01	0,18
8	darkmec	oneworn	0,26	0,38	0,84	0,48	0,13	0,53	0,56		0,52	0,73	0,02	0,48
9	darkmec	bothworn	0,01	0,56	0,91	0,01	0,13	0,00	0,19	0,52		0,97	0,02	0,86
10	dark	bothgood	0,31	0,76	0,91	0,46	0,61	0,49	0,47	0,73	0,97		0,11	0,99
11	dark	oneworn	0,00	0,12	0,09	0,00	0,10	0,00	0,01	0,02	0,02	0,11		0,03
12	dark	bothworn	0,01	0,61	0,88	0,01	0,20	0,01	0,18	0,48	0,86	0,99	0,03	

Рис. 8.13

Для проверки предположений, лежащих в основе метода дисперсионного анализа, необходимо воспользоваться вкладкой **Assumptions** (предположения) в окне **ANOVA Results 1** (рис. 8.14). На вкладке представлены различные критерии проверки гипотезы однородности дисперсий (критерий Кохрана, Хартли, Бартлетта, критерий Левена, М критерий Бокса), графические средства проверки соответствия закона распределения переменной нормальному закону (гистограммы, диаграммы рассеяния, нормальные вероятностные графики).

Выберите эффект *SPINE* и нажмите кнопку **Histograms**. В появившемся окне выберите переменную *WIDTH* и укажите группу, если нужно проанализировать распределение внутри каждой группы. Если выбрать *All* (все), то программа построит (рис. 8.15) гистограмму частот для всех групп. Видно, что общее распределение соответствует нормальному закону. Нажмите кнопку **Levens test (ANOVA)**, появится таблица (рис. 8.16) с результатами проверки гипотезы об однородности дисперсий для зависимых переменных *WIDTH* и *WEGHT*. Из таблицы следует, что во всех группах, соответствующих уровням категориального предиктора *SPINE* дисперсии однородны, т.е. верна гипотеза о равенстве дисперсий.

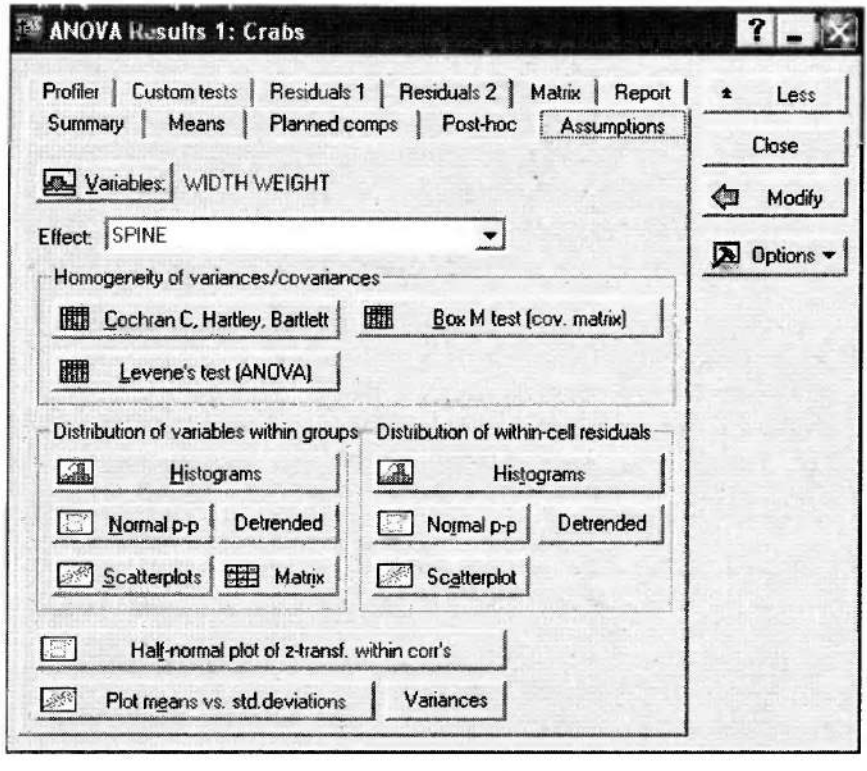


Рис. 8.14

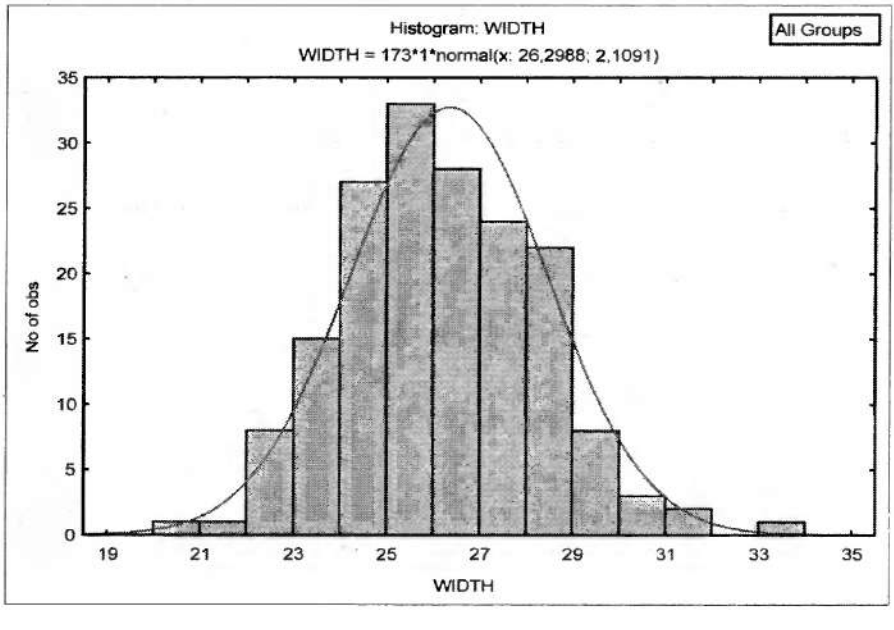


Рис. 8.15

| | Levene's Test for Homogeneity of Variances (
Effect: SPINE
Degrees of freedom for all F's: 2, 170 | | | |
|---|---|---|---|---|
| | MS Effect | MS Error | F | p |
| WIDTH | 2,95 | 1,55 | 1,91 | 0,15 |
| WEIGHT | 0,30 | 0,11 | 2,65 | 0,07 |

Рис. 8.16

Еще одним дополнительным условием применимости дисперсионного анализа является отсутствие корреляции между средними и стандартными отклонениями 16. На вкладке **Assumptions** нажмите кнопку **Plot means vs.std.deviations**. Из диаграммы рассеяния, изображенной на рис. 8.17, видно, что средние и стандартные отклонения коррелируют незначительно.

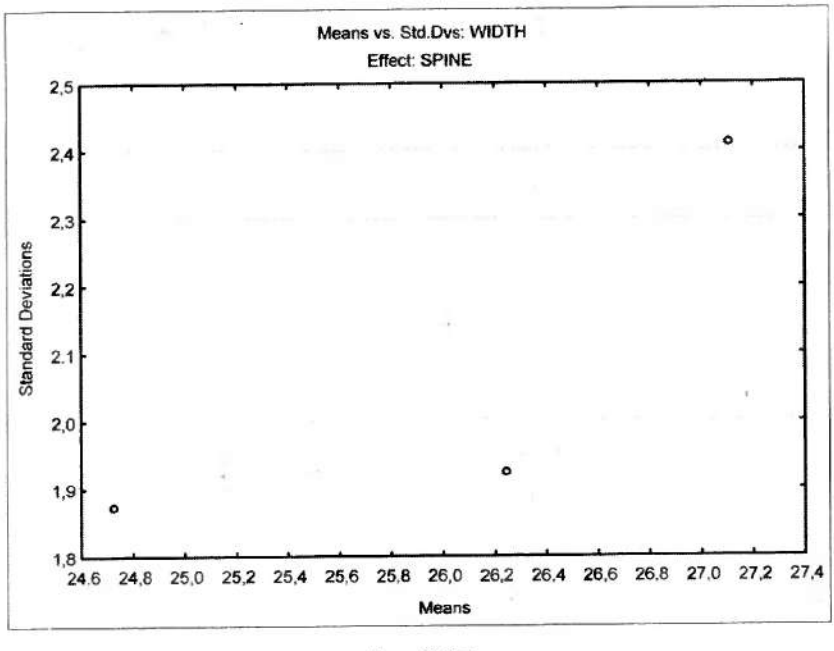


Рис. 8.17

Таким образом, основные условия применимости дисперсионного анализа выполнены, что подтверждает достоверность полученных результатов.

8.2. Описание процедуры *Repeat measures ANOVA*

Предположим, что зависимая переменная *CATWIDTH* является результатом повторного измерения переменной *WIDTH*, и рассмотрим дисперсионный анализ с повторными измерениями. На стартовой панели **General ANOVA/MANOVA** (рис. 8.2) в списке **Type of analysis** выделите **Repeat measures ANOVA**; в списке **Specification method** выберите **Quick Specs Dialog**. Щелкните по **OK**, откроется окно диалога (рис. 8.18) **ANOVA/MANOVA Repeat measures ANOVA**.

На вкладке **Quick** нажмите кнопку **Variables**. В появившемся окне выберите группирующие переменные *COLOR* и *SPINE*, зависимые *WIDTH*, *CATWIDTH*. Если нажать на кнопку **OK**, то появятся результаты многомерного дисперсионного анализа без учета повторных измерений, т.е. переменные *WIDTH*, *CATWIDTH* будут проинтерпретированы как зависимые переменные. Но, согласно нашему предположению, эти переменные рассматриваются как двухуровневый фактор повторных измерений. Чтобы ввести в программу фактор повторных измерений, нажмите кнопку **Within effects** (внутригрупповые эффекты). Откроется окно (рис. 8.19) **Specify within-subjects factor** (задайте фактор повторных измерений). Данная процедура позволяет ввести только один фактор (переменную, многократно измеренную).

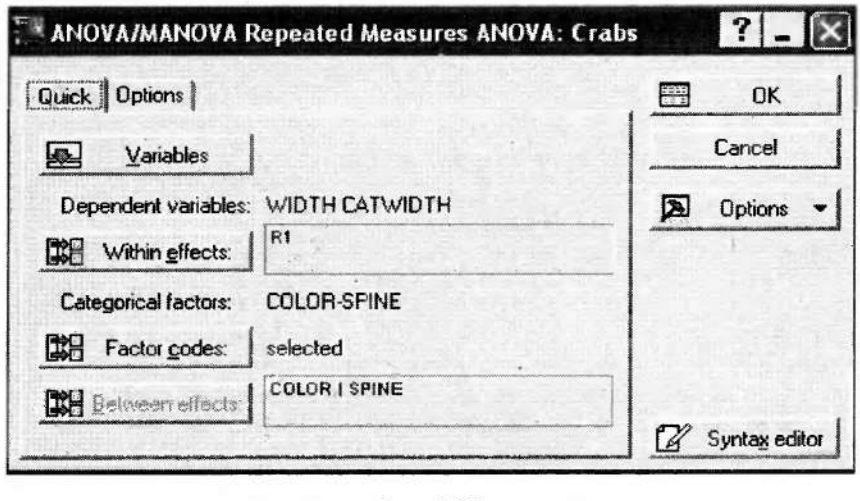


Рис. 8.18

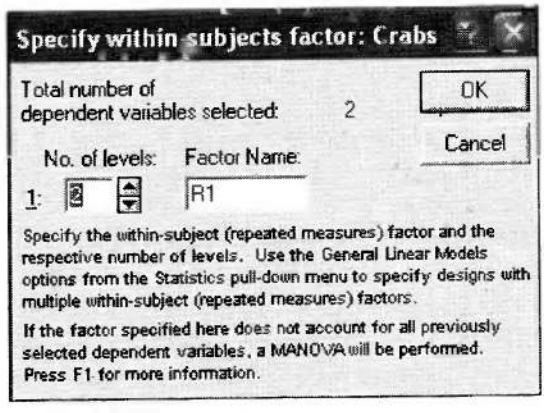


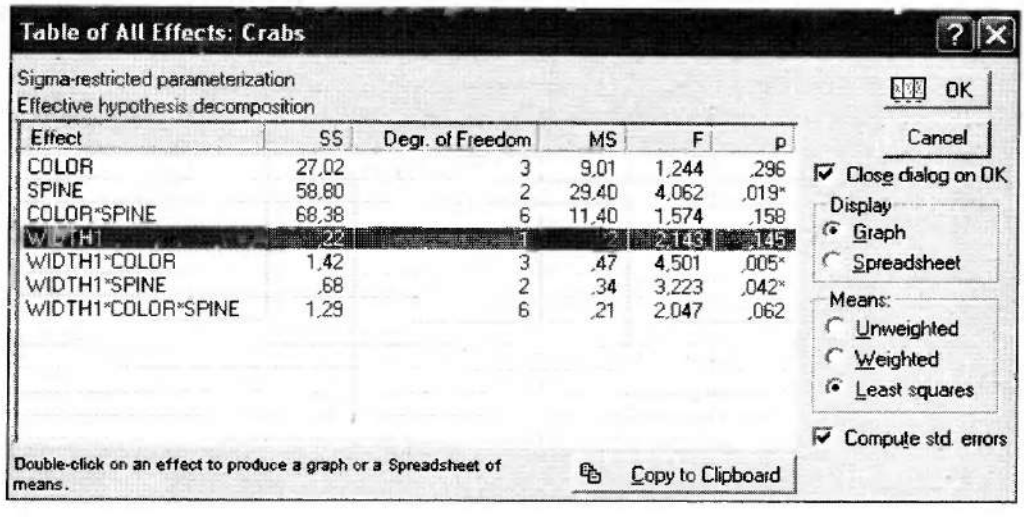
Рис. 8.19

При необходимости проведения анализа с большим числом факторов необходимо воспользоваться модулем **GLM**. Число уровней (*No. of levels*) соответствует количеству повторных измерений. Можно изменить число уровней и задать имя фактора, например, в поле **Factor Name** введите имя *WIDTH1*. Нажмите кнопку **OK**. При помощи кнопки **Factor codes** в диалоге **ANOVA/MANOVA Repeat measures ANOVA** задайте коды уровней категориальных предикторов. Нажмите **OK**. Появится уже знакомое нам окно **ANOVA Results 1**. На вкладке **Quick** нажмите кнопку **All effects**. Из появившейся таблицы **Table of All Effects** (рис. 8.20) видно, что гипотеза о неравенстве средних верна для эффектов *SPINE*, *WIDTH1*COLOR*, *WIDTH1* SPINE*.

	Repeated Measures Analysis of Variance (Crabs) Sigma-restricted parameterization Effective hypothesis decomposition				
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	42625,38	1	42625,38	5888,582	0,000000
COLOR	27,02	3	9,01	1,244	0,295542
SPINE	58,80	2	29,40	4,062	0,019016
COLOR*SPINE	68,38	6	11,40	1,574	0,157795
Error	1165,42	161	7,24		
WIDTH1	0,22	1	0,22	2,143	0,145183
WIDTH1*COLOR	1,42	3	0,47	4,501	0,004624
WIDTH1*SPINE	0,68	2	0,34	3,223	0,042405
WIDTH1*COLOR*SPINE	1,29	6	0,21	2,047	0,062412
Error	16,90	161	0,10		

Рис. 8.20

Чтобы посмотреть средние и их графики, в диалоге **ANOVA Results 1** на вкладке **Quick** нажмите кнопку **All effects/Graphs**. В появившейся таблице **Table of All Effects** (рис. 8.21) высветите какой-либо значимый эффект, содержащий переменную повторных измерений, например, *WIDTH1*COLOR* и нажмите **OK**.



Effect	SS	Degr. of Freedom	MS	F	p
COLOR	27,02	3	9,01	1,244	,296
SPINE	58,80	2	29,40	4,062	,019*
COLOR*SPINE	68,38	6	11,40	1,574	,158
WIDTH1	,22	1	,2	2,143	,145
WIDTH1*COLOR	1,42	3	,47	4,501	,005*
WIDTH1*SPINE	,68	2	,34	3,223	,042*
WIDTH1*COLOR*SPINE	1,29	6	,21	2,047	,062

Рис. 8.21

Видно (рис. 8.22), что средние начальных и повторных измерений ширины крабов незначительно отличаются для крабов всех цветов, кроме темного. Причем средняя ширина крабов темного цвета существенно отличается от средней ширины крабов других цветов как для повторных, так и для начальных измерений. Средняя ширина крабов всех цветов, кроме темного, практически одинакова.

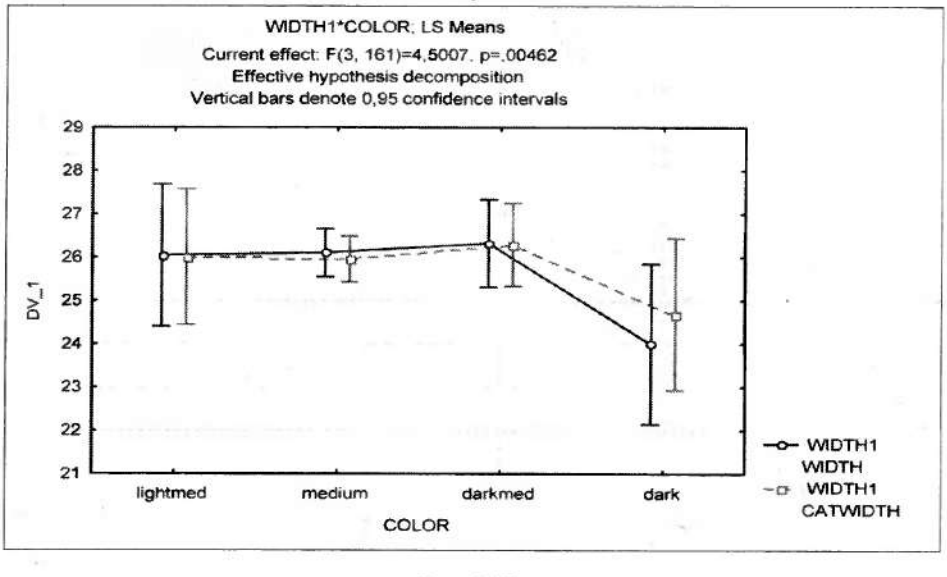


Рис. 8.22

Для того чтобы определить, какие средние статистически значимо различны, проверить выполнение условий применимости дисперсионного анализа,

воспользуйтесь вкладками **Post-hoc** и **Assumptions** в диалоге **ANOVA Results1**, аналогично тому, как это было описано в § 8.1.

Для просмотра графиков средних взаимодействия более высокого порядка, дважды щелкните по строке *WIDTH1*COLOR*SPINE*. Появится промежуточный диалог (рис. 8.23) **Specify the arrangement of factors in the plot**, который используется для настройки расположения факторов на графике.

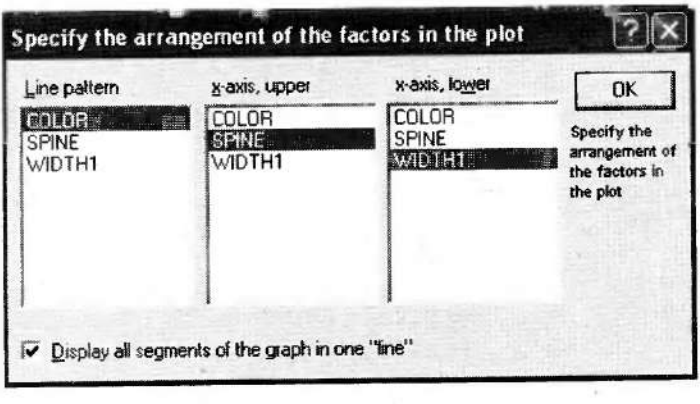


Рис. 8.23

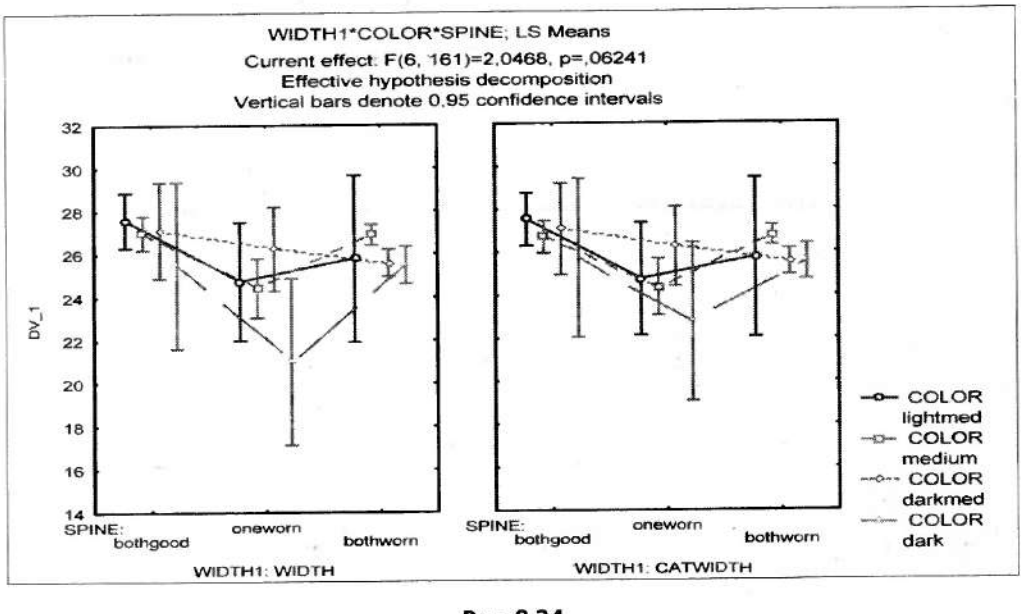


Рис. 8.24

Нажмите **OK**. Программа построит графики средних, разделив на уровни межгрупповых факторов (рис. 8.24).

Видно, что средние начальных и повторных измерений ширины крабов незначительно отличаются для всех уровней межгрупповых факторов. Причем для начальных и повторных измерений характер изменения средних для всех уровней межгрупповых факторов одинаков. Наименьшая средняя ширина при начальном и повторном измерениях соответствует крабам темного цвета с одной поврежденной клешней. Наибольшая средняя ширина при начальном и повторном измерениях соответствует крабам светлого цвета с обеими целыми клешнями.

Глава 9

Линейное многомерное моделирование взаимосвязей

9.1. Линейная регрессионная модель

Анализ взаимосвязей, присущих изучаемым процессам и явлениям, — важнейшая задача многих исследований. В тех случаях, когда речь идет о явлениях и процессах, обладающих сложной структурой и многообразием свойственных им связей, такой анализ представляется сложным. Прежде всего, необходимо установить наличие взаимосвязей и их характер. Вслед за этим возникает вопрос о тесноте взаимосвязей и степени воздействия различных факторов (причин) на интересующий исследователя результат. Если черты и свойства изучаемых объектов могут быть измерены и выражены количественно, то анализ взаимосвязей может вестись с применением математических методов, что позволяет проверить гипотезу о наличии или отсутствии взаимосвязей между теми или иными признаками, выдвигаемую на основе содержательного анализа. Далее, лишь посредством математических методов можно установить тесноту и характер взаимосвязей или выявить силу (степень) воздействия различных факторов на результат. В таких исследованиях широко используются процедуры множественной регрессии.

Регрессионный анализ тесно связан с другими статистическими методами — методами корреляционного и дисперсионного анализа. В отличие от корреляционного анализа, который изучает направление и силу статистической связи признаков, регрессионный анализ изучает вид зависимости признаков, т.е. параметры функции зависимости одного признака от одного или нескольких других признаков. В отличие от дисперсионного анализа, с помощью которого исследуется зависимость количественного признака от одного или нескольких качественных признаков, в регрессионном анализе обычно исследуется зависимость (количественного или качественного признака) от одного или нескольких количественных признаков [12].

Таким образом, в регрессионном анализе рассматривается односторонняя зависимость случайной зависимой переменной от одной или нескольких независимых переменных. Независимые переменные называются факторами, или предикторами, а зависимая переменная — результативным признаком, или откликом.

Если число предикторов равно 1, регрессию называют простой, если число предикторов больше 1 — множественной. Множественная регрессия позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что является лучшим предиктором для...». Например [6], исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими предикторами успешной учебы в средней школе. А психолога мог быть заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида.

Если в ходе количественного анализа выявлена и обоснована зависимость одного явления от других, то задача регрессионного анализа — измерение зависимости, в которой причинно-следственный механизм выступает в наглядной форме. Прогноз в этом случае лучше поддается содержательной интерпретации, становится более ясным воздействие отдельных факторов, и исследователь лучше понимает природу изучаемого явления. Кроме того, регрессии создают базу для расчетного экспериментирования с целью получения ответов на вопросы типа «Что будет, если...?». Регрессионный анализ предполагает решение двух задач.

Первая заключается в выборе независимых переменных, существенно влияющих на зависимую величину, и определении формы уравнения регрессии. Данная задача решается путем анализа изучаемой взаимосвязи.

Вторая задача — оценивание параметров — решается с помощью того или иного статистического метода обработки данных наблюдения.

Функция $F(X)$, описывающая зависимость условного среднего значения результативного признака Y от заданных значений фактора, называется функцией (уравнением) регрессии [9]. Для точного описания уравнения регрессии необходимо знать условный закон распределения результативного признака Y. В статистической практике такую информацию получить обычно не удается, поэтому ограничиваются поиском подходящих аппроксимаций для функции $F(X)$, основанных на исходных статистических данных. Значения переменной X в i-м опыте будем обозначать через x_i, соответствующие им значения величины Y — через y_i, $i = 1, ..., n$.

Рассмотрим самую простую регрессионную модель — линейную. Для линейной модели предполагается, что наблюдаемые величины связаны между собой зависимостью вида

$$y_i = b_0 + b_1 x_i + c_i,$$

где b_0, b_1 — неизвестные параметры (коэффициенты уравнения), c_i — независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Иногда c_i называют ошибками наблюдения. Общая задача регрессионного анализа состоит в том, чтобы по наблюдениям x_i, y_i оценить параметры модели b_1, b_0 «наилучшим образом»; построить доверительные интервалы для b_1, b_0; проверить гипотезу о значимости уравнения и коэффициентов регрессии; оценить степень адекватности полученной зависимости и т.д.

Если под «наилучшим образом» понимать минимальную сумму квадратов расстояний до прямой от наблюдаемых точек, вычисленных вдоль оси ординат, то такой метод построения уравнения регрессии называется методом наименьших квадратов. В качестве меры «наилучшим образом» можно использовать минимум суммы квадратов расстояний от точек до прямой, вычисленных вдоль оси абсцисс; минимум суммы квадратов расстояний длин перпендикуляров, опущенных из точек на прямую и т.д.

Линейная модель с несколькими предикторами называется линейной множественной регрессионной моделью, а именно:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi} + b_0 + c_i,$$

где b_0, b_1, b_2, ..., b_p — неизвестные параметры модели, которые вычисляются при помощи систем нормальных уравнений. Например, система нормальных уравнений для регрессии с двумя предикторами имеет следующий вид:

$$\begin{cases} nb_0 + b_1 \Sigma_i x_{1i} + b_2 \Sigma_i x_{2i} = \Sigma_i y_i \\ b_0 \Sigma_i x_{1i} + b_1 \Sigma_i x_{1i}^2 + b_2 \Sigma_i x_{1i} x_{2i} = \Sigma_i x_{1i} y_i \\ b_0 \Sigma_i x_{2i} + b_1 \Sigma_i x_{1i} x_{2i} + b_2 \Sigma_i x_{2i}^2 = \Sigma_i x_{2i} y_i \end{cases}$$

9.2. Описание модуля Multiple Regression

Кратко рассмотрим основные обозначения и понятия, используемые в модуле **Multiple Regression** (множественная регрессия) [2, 6].

Predictable values (предсказанные значения) — значения Y, вычисленные по уравнению регрессии. Обозначим их PrY_i.

Residuals (остатки) — разность между наблюдаемыми значениями и предсказанными: $Res = Y_i - PrY_i$.

SS (сумма квадратов Y_i, скорректированная на среднее):

$$SS = \Sigma_i (Y_i - Y)^2, \text{ где } Y = \Sigma_i Y_i / n.$$

SSPr (сумма квадратов PrY_i, скорректированная на среднее):

$$SSPr = \Sigma_i (PrY_i - Y)^2.$$

SSRes (сумма квадратов остатков): $SS\ Res = \Sigma_i (PrY_i - Y_i)^2$.

$R^2=1 - SSRes/SS$ (коэффициент детерминации). Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Например, если связь между предиктором X и откликом Y отсутствует, то отношение остаточной изменчивости переменной Y к исходной дисперсии равно 1. Если X и Y связаны функциональной зависимостью, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0. В общем случае отношение будет лежать между этими экстремальными значениями, т.е. между 0 и 1. Коэффициент детерминации R^2 интерпретируется следующим образом. Если, например, $R^2 = 0{,}4$, то изменчивость значений переменной Y около линии регрессии составляет 1 — 0,4 от исходной дисперсии; другими словами, 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R^2 является индикатором степени подгонки модели к данным (значение R^2, близкое к 1, показывает, что модель объясняет почти всю изменчивость соответствующих переменных) [6].

$R = \sqrt{R^2}$ — коэффициент множественной корреляции. Характеризует тесноту связи между предикторами и откликом, а также является оценкой качества предсказания. Изменяется в пределах от 0 до 1.

Adjusted $R^2 = 1 - (1 - R^2)(n/(n - k))$ — скорректированное R^2, где k — число параметров в регрессионном уравнении.

Ограничимся лишь описанием некоторых элементов модуля «Множественная регрессия» при осуществлении регрессионного анализа.

Предположим, что исследователей интересует возможность составления математической модели зависимости месячных объемов продаж (млн дол.) продукта компании «Петлокс» [20] от цены за единицу (дол.), расходов на рекламу в предыдущем месяце (10 тыс. дол.) и количества работников, занятых сбытом продукции. Выборку из восьми месяцев за последние два года можно представить в виде таблицы данных рис. 9.1. Здесь *Объем продаж* — зависимая переменная (функция отклика); *Розн. цена, Расх. на рекламу, Кол. работ.* — независимые переменные (предикторы).

Месяц	1	2	3	4
	Объем продаж	Розн. цена	Расх. на рекламу	Кол. работ.
1	4,00	1,00	8,00	24,00
2	5,20	0,90	9,00	26,00
3	3,80	1,10	6,00	20,00
4	2,90	1,20	5,00	18,00
5	4,60	0,95	7,00	20,00
6	4,50	0,90	6,00	30,00
7	3,70	1,00	6,00	27,00
8	5,00	0,95	10,00	28,00

Рис. 9.1

Проверим, можно ли зависимость между функцией отклика и предикторами описать линейной моделью:

$$Объем\ продаж = b_0 + b_1\ Розн.цена + b_2\ Расх.\ на\ рекламу + b_3\ Кол.\ работ.,$$

где b_0 — свободный член уравнения; b_1, b_2, b_3 — коэффициенты уравнения регрессии.

Для запуска программы в меню **Statistics** выберите команду **Multiple Regression**. Откроется стартовая панель модуля. Выбор переменных осуществляется с помощью кнопки **Variables**. После того как кнопка будет нажата, откроется диалоговое окно **Select dependent and independent variable list** (выбрать из списка зависимых и независимых переменных).

Высветив имя переменной в левой части окна, выберите зависимую переменную *Объем продаж*. Высветив имя переменной в правой части окна, выберите независимые переменные *Розн.цена, Расх. на рекламу, Кол. работ*. То же можно сделать, набрав номера переменных в строках **Dependent variable list** (список зависимых переменных) и **Independent variable list** (список независимых переменных). Нажмите кнопку **OK**, и программа вновь вернется в стартовое окно модуля.

Для задания дополнительных установок надо выбрать вкладку **Advanced**. Здесь можно использовать следующие установки:

- *Advanced options (stepwise or ridge regression)* (расширенные опции — ступенчатая или гребневая регрессия);
- *Review descriptive statistics, descriptive statistics* (обзор описательных статистик, корреляционная матрица);

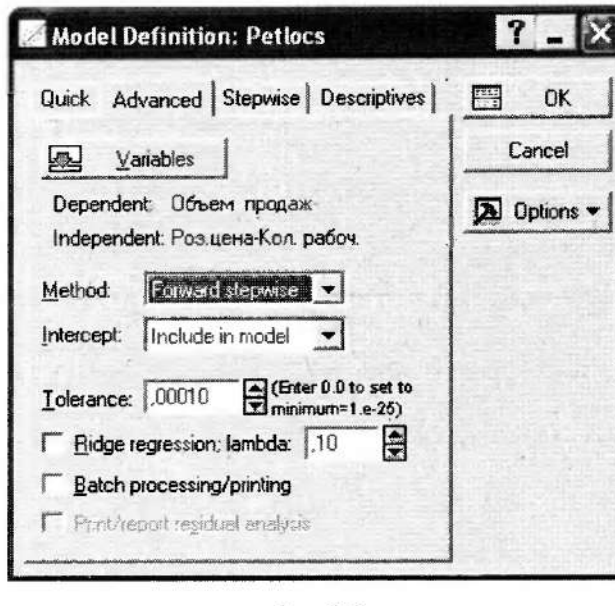


Рис. 9.2

- *Extended precision computation* (повышенная точность вычислений);
- *Batch processing / reporting* (пакетная обработка данных / печать).

Установите флажок на *Advanced options (stepwise or ridge regression)* и нажмите **OK**. Появится диалоговое окно (рис. 9.2) **Model definition** (построение модели).

На вкладке **Quick** этого окна можно указать метод *(Method)*:

- *Standard* (стандартный);
- *Forward stepwise* (пошаговый с включением);
- *Backward stepwise* (пошаговый с исключением).

На вкладке **Advanced** можно выбрать метод, а также произвести оценку свободного члена регрессии *(Intercept)* и сделать другие установки. Выберите процедуру **Forward stepwise** и нажмите кнопку **OK**. Откроется окно результатов (рис. 9.3) **Multiple Regression Results.**

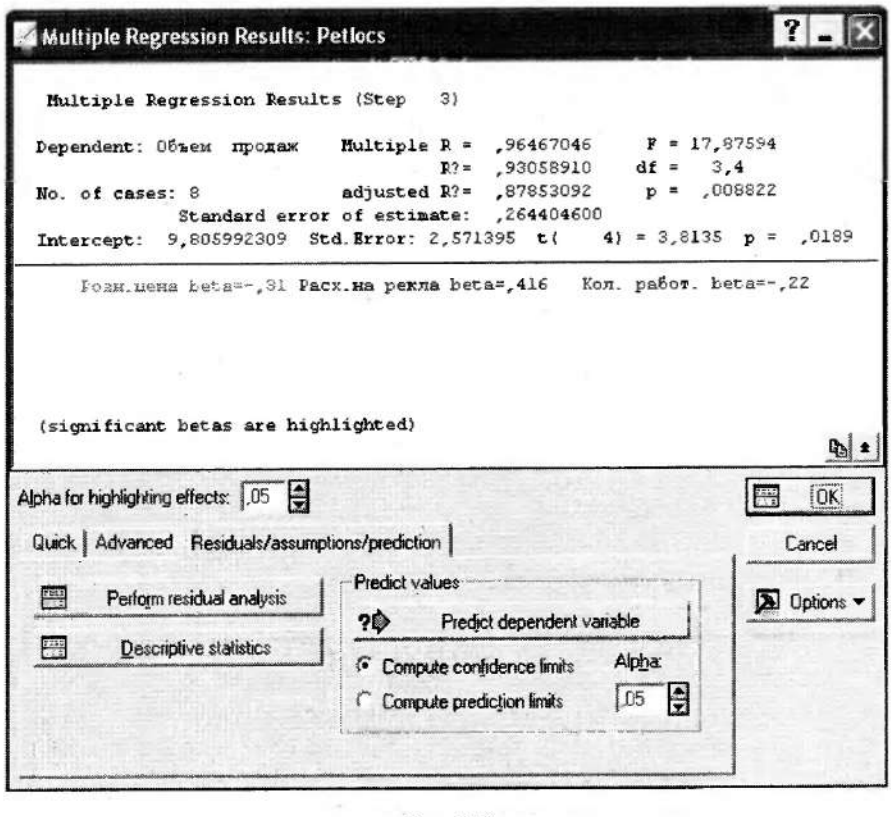


Рис. 9.3

Верх окна результатов — информационный. В первой части содержится основная информация о результатах оценивания, во второй — высвечиваются значимые стандартизованные регрессионные коэффициенты. Внизу окна находятся функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа. В информационной части содержатся краткие сведения о результатах анализа, а именно:

- *Dependent* (имя зависимой переменной), в данном случае — это *Объем продаж*;
- *No. of cases = 8* (число наблюдений, по которым построена регрессионная модель);
- *Multiple R* = 0,9646 (коэффициент множественной корреляции);
- *R-square* = R^2 = 0,9305 (коэффициент детерминации);
- *Adjusted* R^2 = 0,8785 (скорректированный коэффициент детерминации);
- *Standard error of estimate = 0,2644* (стандартная ошибка оценки). Эта статистика — мера рассеяния наблюдаемых значений относительно регрессионной прямой;
- *Intercept* = 9,8059 (оценка свободного члена b_0 регрессии), если выбрана регрессия, включающая свободный член;
- *Std.Error* = 2,5713 (стандартная ошибка оценки свободного члена b_0);
- t = 3,8135, p = 0, 0189 (значение *t-критерия* и уровень значимости p) для проверки гипотезы о равенстве нулю свободного члена b_0;
- F = 17,8759, df = 3, 4, p = 0,0088 (значение *F-критерия*, число степеней свободы и уровень значимости p) используются в качестве общего *F-критерия* для проверки гипотезы о зависимости предикторов и отклика.

Из приведенных результатов анализа следует, что зависимость между откликом и предикторами сильная ($R^2 > 0,75$); построенная линейная регрессия адекватно описывает взаимосвязь между откликом и предикторами, свободный член статистически значим.

Если нажать на кнопку **Summary:regression results**, появится таблица результатов с подробными статистиками (рис. 9.4).

	Regression Summary for Dependent Variable: Объем продаж R= ,96467046 R?= ,93058910 Adjusted R?= ,87853092 F(3,4)=17,876 p<,00882 Std.Error of estimate: ,26440					
N=8	Beta	Std.Err. of Beta	B	Std.Err. of B	t(4)	p-level
Intercept			9,80599	2,571395	3,81349	0,018883
Розн.цена	-0,812418	0,228380	-5,95435	1,673837	-3,55730	0,023643
Расх.на рекламу	0,415881	0,164709	0,18270	0,072358	2,52494	0,065014
Кол. работ.	-0,223980	0,202949	-0,03900	0,035339	-1,10362	0,331683

Рис. 9.4

Таблица содержит стандартизованные (*Beta*) и нестандартизованные (*B*) регрессионные коэффициенты (веса), их стандартные ошибки и уровни значимости. Коэффициенты *Beta* оцениваются по стандартизованным данным, имеющим выборочное среднее, равное 0 и стандартное отклонение, равное 1. Поэтому величины *Beta* позволяет сравнить вклады каждого предиктора в предсказание отклика. Так, в зависимую переменную *Обьем продаж* больший вклад вносит переменная *Розн. цена*, а меньший — *Кол. работ.* Отрицательный знак коэффициентов при этих переменных означает, что с увеличением розничной цены и количества работников, занятых сбытом продукции, объемы продаж падают. Положительный знак коэффициента при переменной *Расх. на рекламу* означает, что с увеличением затрат на рекламу в предыдущем месяце объемы продаж растут. Коэффициенты уравнения регрессии b_1, b_2 и свободный член статистически значимы при уровне значимости $p = 0{,}1$; коэффициент уравнения регрессии b_3 статистически незначим (так как $p > 0{,}1$).

В диалоговом окне **Multiple Regression Results** нажмите кнопку **Partial correlation** (частная корреляция), появится таблица (рис. 9.5). Таблица содержит коэффициенты *Beta*, частные коэффициенты корреляции, получастные коэффициенты корреляции *(Semipart Cor)*, толерантности *(Tolerance)*, коэффициенты детерминации *(R-square)*, значения *t-критерия* и уровни значимости p — вероятности отклонения гипотезы о значимости частного коэффициента корреляции.

	Variables currently in the Equation; DV: Объем продаж (Petlc						
Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(4)	p-level
Розн.цена	-0,812	-0,872	-0,469	0,333	0,667	-3,557	0,024
Расх.на рекламу	0,416	0,784	0,333	0,640	0,360	2,525	0,065
Кол. работ.	-0,224	-0,483	-0,145	0,421	0,579	-1,104	0,332

Рис. 9.5

Частные коэффициенты корреляции (*Partial Cor*) показывают степень влияния одного предиктора на отклик в предположении, что остальные предикторы закреплены на постоянном уровне, т.е. контролируется их влияние на отклик. Например, отклик — длина волос может коррелировать с предиктором — ростом человека (чем выше человек, тем короче волосы), однако эта зависимость становится слабой, если устранить влияние другого предиктора — пола людей, поскольку женщины обычно ниже ростом и чаще имеют более длинные волосы, чем мужчины [6]. Частные коэффициенты корреляции так же, как и стандартизованные коэффициенты *Beta* позволяют провести ранжирование предикторов по степени их влияния на отклик. Кроме того, частные коэффициенты корреляции широко используются при решении проблемы отбора предикторов — целесообразность включения того или иного предиктора в модель определяется величиной частного коэффициента корреляции [17]. Из таблицы следует, что предикторы можно проранжировать по степени влияния на отклик в следующем порядке: *Розн.*

цена, Расх. на рекламу, Кол. работ., причем, первые два предиктора оказывают на отклик сильное влияние, а третий — умеренное.

Semipart Cor (получастная корреляция) — корреляция предиктора и отклика в предположении, что контролируется влияние других предикторов на данный предиктор, но не контролируется влияние предикторов на отклик. Если получастная корреляция мала, в то время как частная корреляция относительно велика, то соответствующий предиктор может иметь самостоятельную «часть» в объяснении изменчивости зависимой переменной, т.е. «часть», которая не объясняется другими предикторами. Из таблицы видно, что предикторы *Розн. цена, Расх. на рекламу* имеют самостоятельную часть в объяснении изменчивости отклика.

R-square (коэффициент детерминации) — квадрат коэффициента множественной корреляции между данной переменной и всеми остальными переменными, входящими в уравнение регрессии. Из таблицы следует, что все коэффициенты детерминации — умеренные, но взаимосвязь между предиктором *Расх. на рекламу* и двумя другими предикторами значительно меньше чем у *Розн. цена* и *Кол. работ.* с двумя другими.

Toleranse (толерантность) — это *1 – R-square*.

t(4) — значение критерия Стьюдента для проверки гипотезы о значимости частного коэффициента корреляции с указанным (в скобках) числом степеней свободы.

p-level (*p-уровень*) — вероятность отклонения гипотезы о значимости частных коэффициентов корреляции. Частные коэффициенты корреляции значимы для переменных *Розн. цена, Расх. на рекламу* при уровне значимости $p = 0{,}1$.

Важной характеристикой регрессионного анализа являются **Residuals** (остатки). Нажмите кнопку **Residual**, откроется рабочее окно **Residuals Analysis** (анализ остатков).

В диалоговом окне **Residuals Analysis** нажмите кнопку **Durbin-Watson statistic** (статистика Дарбина-Уотсона). Эта статистика характеризует наличие или отсутствие сериальной корреляции (зависимости) между остатками для соседних наблюдений [6]. Существование сериальной корреляции может служить доказательством зависимости наблюдений в файле данных. Дело в том, что критерии значимости в множественной регрессии предполагают, что данные являются случайной выборкой из независимых наблюдений. В противном случае оценки коэффициентов уравнения регрессии могут быть более неустойчивыми, чем это гарантируют их уровни значимости. Из таблицы, изображенной на рис. 9.6, видно, что статистика Дарбина-Уотсона имеет небольшое значение (1,9204) при умеренной сериальной корреляции (–0,2505). Это свидетельствует о некоторой зависимости наблюдений, следовательно, можно говорить о недостаточной устойчивости некоторых значений коэффициентов регрессии, а значит о невысокой адекватности модели изучаемому процессу.

Для графического сравнения предсказанных программой значений отклика и наблюдаемых значений надо в диалоговом окне **Residual Analysis** выбрать вкладку **Predicted** (предсказанные) и нажать на кнопку **Predicted vs.independent var** (предсказанные независимые переменные).

Уравнение регрессии можно использовать для прогноза значений отклика — объема розничных продаж по значениям предикторов: розничной цены, расходов на рекламу, количества работников, занятых сбытом. Для этого надо вернуться в окно **Multiple Regression Results**, выбрать вкладку **Residuals/assumptions/prediction** (остатки/оценки/предсказания) и нажать на кнопку **Predict dependent variable** (предсказать зависимую переменную). Далее в открывшемся окне **Specify values for indep.vars** (рис. 9.7) в полях *Розн. цена*, *Расх.на рекламу*, *Кол.работ.* указать значения розничной цены (1,1), расходов на рекламу (6) и количество работкников (30). Нажмите **OK**. Появится таблица результатов предсказания (рис. 9.8). В таблице указан предсказанный (*predicted*) объем продаж –3,1824 (млн долл.) с 95%-м доверительным интервалом (2,2176; 4,1471).

	Durbin-Watson d (Petl and serial correlation c	
	Durbin-Watson d	Serial Corr.
Estimate	1,920437	-0,250565

Рис. 9.6

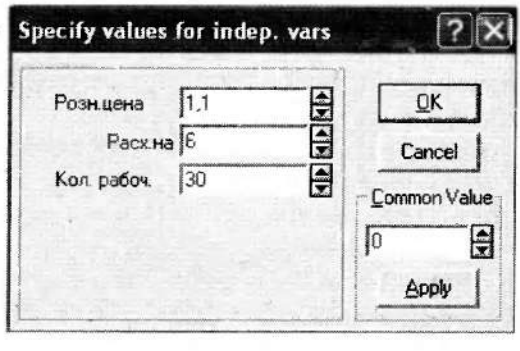


Рис. 9.7

Variable	Predicting Values for (Petlocs) variable: Объем продаж		
	B-Weight	Value	B-Weight * Value
Розн.цена	-5,95435	1,10000	-6,54978
Расх.на рекламу	0,18270	6,00000	1,09621
Кол. работ.	-0,03900	30,00000	-1,17002
Intercept			9,80599
Predicted			3,18240
-95,0%CL			2,21760
+95,0%CL			4,14719

Рис. 9.8

Одним из условий корректного применения регрессионного анализа является соответствие закона распределения остатков нормальному закону [6]. В диалоговом окне **Multiple Regression Results** на вкладке **Residuals/assumptions/prediction** нажмите кнопку **Perform Residual analysis**, в открывшемся окне на вкладке **Residuals** нажмите кнопку **Histogram of residual**. Из построенного графика (рис. 9.9) видно, что из-за очень малого числа наблюдений (8), распределение остатков не соответствует нормальному закону.

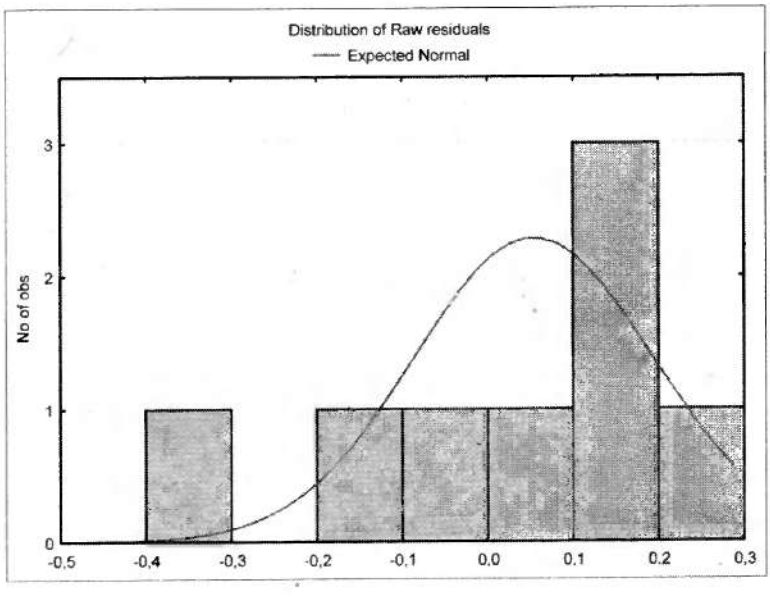


Рис. 9.9

Для визуального анализа распределения остатков можно также использовать нормальные вероятностные графики (вкладка **Probability plots** в диалоговом окне **Residual analysis**).

Из приведенных результатов регрессионного анализа можно сделать вывод о невысокой адекватности построенной линейной модели зависимости объемов продаж компании «Петлокс» от розничной цены продукта, расходов на рекламу в предыдущем месяце, количества работников, занятых сбытом продукции. Линейная модель имеет вид

$$Объем\ продаж = 9{,}8059 - 5{,}9543\ Розн.\ цена + 0{,}1827\ Расх.\ на\ рекламу - 0{,}039\ Кол.\ работ.$$

Последнее слогаемое — *Кол. работ.* из модели можно исключить, так как коэффициент (0,039) статистически незначим, т.е. верна гипотеза о равенстве нулю.

Заметим, что построенная модель будет премлемо точна, при условии, что независимые переменные (предикторы) — *Розн. цена*, *Расх.на рекламу*, *Кол. работ.* лежат в пределах заданных таблицей данных; вне этих пределов модель может оказаться абсолютно ненадежной [20]. При помощи контекстного меню или модуля «Discriptive statistics» по файлу **Petlocs** легко определить пределы изменения предикторов — наименьшие и наибольшие значения: *Розн. цена* изменяется в пределах (см. рис. 9.1) от 0,9 до 1,2; *Расх. на рекламу* — от 5 до 10; *Кол. работ.* — от 18 до 30.

Глава 10

Нелинейное многомерное моделирование взаимосвязей

10.1. Линеаризующие преобразования

Существует несколько причин широкого распространения предположения линейности в регрессионном анализе [18], а именно:

- во многих случаях такое предположение является простейшим, поэтому естественным бывает желание начать исследование с простейшего;
- в некоторых случаях это предположение имеет вполне удовлетворительную, а иногда и высокую адекватность;
- многие математические методы исследования приспособлены к линейной задаче. Это обстоятельство вынуждает исследователей использовать линейные схемы даже в тех случаях, когда есть серьезные основания ожидать, что реальная зависимость значительно отличается от линейной;
- известную роль играет человеческая психология. Линейная зависимость обычно воспринимается человеком с минимальным внутренним сопротивлением. Порой они как бы сами собой подразумеваются.

Как правило, все зависимости, встречающиеся в окружающей нас природе, являются нелинейными. Конечно, имеются зависимости, линейность которых в рассматриваемой области приложений практически достоверна с любой разумной степенью точности. При построении математических моделей гораздо чаще предположение о линейности имеет отчетливый характер допущения, хотя и далеко не всегда формулируется как таковое. Поэтому при моделировании взаимосвязей, присущих изучаемым процессам и явлениям, наряду с линейными регрессионными моделями, целесообразно рассматривать нелинейные регрессионные модели. Как правило, необходимость в нелинейной регрессии появляется, если исследователь получает данные о неадекватности линейной модели, и для уточнения модели в ее уравнение добавляются некоторые нелинейные члены.

В общем случае регрессионная модель может быть записана в следующем виде:

$$Y = F(X_1, X_2, ..., X_n).$$

При любом исследовании возникает вопрос, связаны ли зависимая переменная и набор независимых переменных, если да, то каким образом выглядит эта связь. Для лучшего понимания аспектов линейности и нелинейности обратимся к примерам, рассмотренным в [6].

Физиологами установлено, что зависимость между производительностью некоторого объекта и степенью его физиологического возбуждения выражается следующим уравнением регрессии:

$$Y = b_0 + b_1 X + b_2 X^2,$$

где b_0 — свободный член; b_1 и b_2 — коэффициенты регрессии; Y — величина, характеризующая производительность; X — величина, характеризующая возбуждение.

Нелинейность данной модели выражается членом X^2. Такая модель называется нелинейной по переменным. Она допускает линеаризацию, которую можно осуществить, произведя замены: $X = X_1$, $X^2 = X_2$. Уравнение примет вид

$$Y = b_0 + b_1 X_1 + b_2 X_2,$$

для анализа которого можно использовать множественную регрессию.

Известно, что зависимость между возрастом человека и скоростью его роста не является равномерной (в младшем возрасте скорость роста человека выше, чем в старшем). Ее можно представить в виде экспоненциальной функции $Y = exp(-b_1 X)$, где Y — рост; X — возраст. Данная модель не является линейной и называется нелинейной по параметрам. Она так же, как и предыдущая, может быть линеаризована. Произведем логарифмирование $lnY = -b_1 X$. Произведя замену $lnY = Y_1$, получим линейную модель, для анализа которой также можно использовать множественную регрессию. В табл. 10.1 приведены примеры нелинейных функций $Y = F(X, b_0, b_1)$ и соответствующие им линеаризующие преобразования, приводящие их к линейному виду [19].

Таблица 10.1

№ п/п	Функция	Линеаризующие преобразования			
		Y_1	X_1	a_0	a_1
1	$Y = b_0 + b_1/X$	Y	$1/X$	b_0	b_1
2	$Y = 1/(b_0 + b_1X)$	$1/Y$	X	b_0	b_1
3	$Y = X/(b_0 + b_1X)$	X/Y	X	b_0	b_1
4	$Y = b_0 b_1^X$	lnY	X	lnb_0	lnb_1
5	$Y = 1/(b_0 + b_1e^{-X})$	$1/Y$	e^{-X}	b_0	b_1
6	$Y = b_0X^{b1}$	lnY	lnX	Lnb_0	b_1
7	$Y = b_0 + b_1\,ln(X+1)$	Y	$ln(X+1)$	b_0	b_1
8	$Y = b_0X/(b_1 + b_2X)$	$1/Y$	$1/X$	b_1/b_0	b_2/b_0
9	$Y = b_0e^{b1/X}$	lnY	$1/X$	lnb_0	b_1
10	$Y = b_0 + b_1X^n$	Y	X^n	b_0	b_1

Однако существуют модели, которые не могут быть сведены к линейным. Добавив в рассмотренную ранее модель некоторый параметр-ошибку (обозначим его ε, он будет означать влияние различных случайных факторов на скорость роста), в итоге получим более точную модель, но не подлежащую линеаризации

$$Y = exp(-b_1X) + \varepsilon.$$

Следует отметить, что в выбранном примере параметр (предполагаемая ошибка) вряд ли будет сохранять постоянство при любом значении переменной X. По этой причине более реалистичной моделью, включающей ошибку, явится

$$Y = exp(-b_1X)\varepsilon.$$

В таком виде модель уже подлежит линеаризации. Цель данного примера — показать, что существуют модели, которые нельзя преобразовать в линейные. Для анализа таких моделей можно использовать только нелинейное оценивание.

В программе *STATISTICA* реализован комплекс процедур нелинейного регрессионного анализа:

- модуль **Fixed Nonlinear Regression** (фиксированная нелинейная регрессия), если модель допускает линеаризацию;
- модуль **Nonlinear Estimation** (нелинейное оценивание), если модель не допускает линеаризацию.

10.2. Описание модуля *Fixed Nonlinear Regression*

Так как в модуле **Fixed Nonlinear Regression** реализован множественный линейный регрессионный анализ с линеаризованной моделью, то его диалоговые окна во многом похожи на окна **Multiple Regression**. Поэтому подробно не будем останавливаться на их описании.

Рассмотрим работу модуля на примере данных таблицы (рис. 10.1), в котором приведены размеры (*LENGTH* — длина, *WIDTH* — ширина, *HEIGHT* — высота, *WEIGHT*— вес) 20 черепах. Построим квадратичную зависимость веса черепах от длины, ширины и высоты.

$$WEIGHT = b_0 + b_1 LENGTH + b_2 WIDTH + b_3 HEIGHT + b_4 LENGTH^2 + \\ + b_5 WIDTH^2 + b_6 HEIGHT^2$$

В меню **Statistics** щелкните по **Advanced Linear/Nonlinear Models** и выберите команду **Fixed Nonlinear Regression**. Откроется одноименное окно (рис. 10.2), в котором, нажав на кнопку **Variables**, выберите имена всех переменных для анализа. Установите галочку в опции *Review descriptive statistics, correlation matrix* (показать описательные статистики, корреляционную матрицу), опция *Extended precision computations* означает режим вычисления с повышенной точностью. Нажмите **OK**, появится окно с предусмотренными в программе линеаризующими преобразованиями переменных (рис. 10.3). В столбце *Non-linear Transformation Functions* приведены соответствующие преобразования, в столбце *Valid Range* указаны допустимые диапазоны переменных. Установите галочку на преобразовании *X*2* и щелкните по **OK**.

	1 LENGTH	2 WIDTH	3 HEIGHT	4 WEIGHT
1	98	81	38	155
2	103	84	38	160
3	103	86	42	164
4	105	86	42	165
5	109	88	44	170
6	123	92	50	180
7	123	95	46	180
8	133	99	51	190
9	133	102	51	193
10	133	102	51	190
11	134	100	48	190
12	136	102	49	192
13	138	98	51	192
14	138	99	51	193
15	141	105	53	200
16	147	108	57	210
17	149	107	55	206
18	153	107	56	210
19	155	115	63	220
20	155	117	60	220

Рис. 10.1

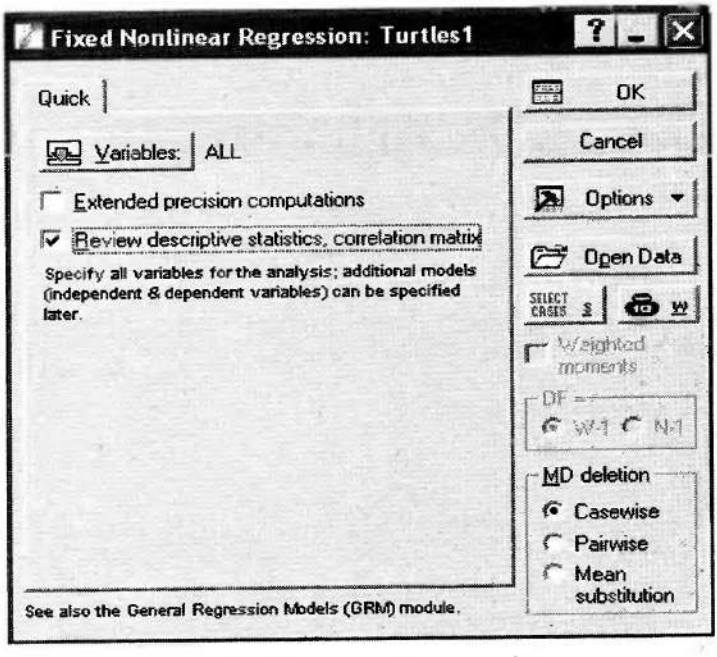


Рис. 10.2

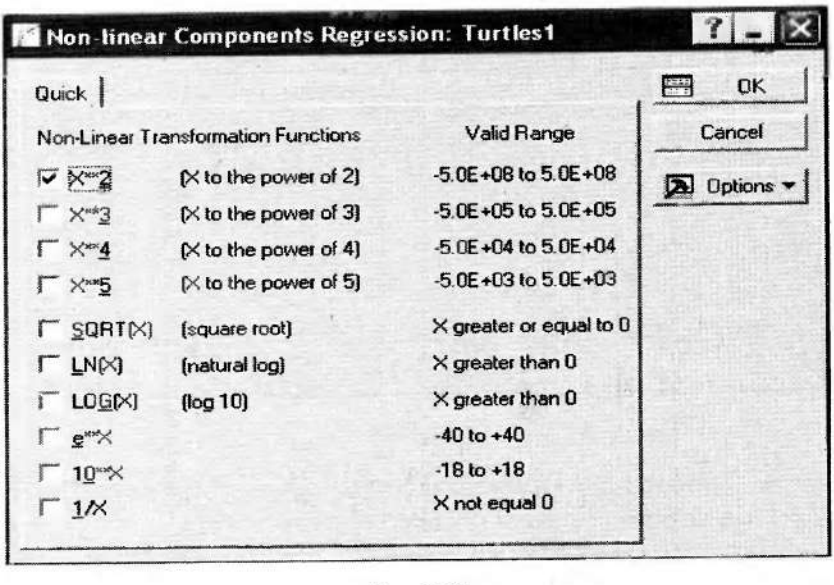


Рис. 10.3

Откроется окно **Review Descriptive Statistics** (рис. 10.4), в котором нажмите на кнопку **Correlations**. Появится корреляционная матрица (рис. 10.5), в которой можно предварительно, до включения в модель множественной регрессии,

просмотреть корреляции предикторов (*LENGTH, WIDTH, HEIGHT, LENGTH*^{2}, *WIDTH*^{2}, *HEIGHT*^{2}) и функции отклика (*WEIGHT*).

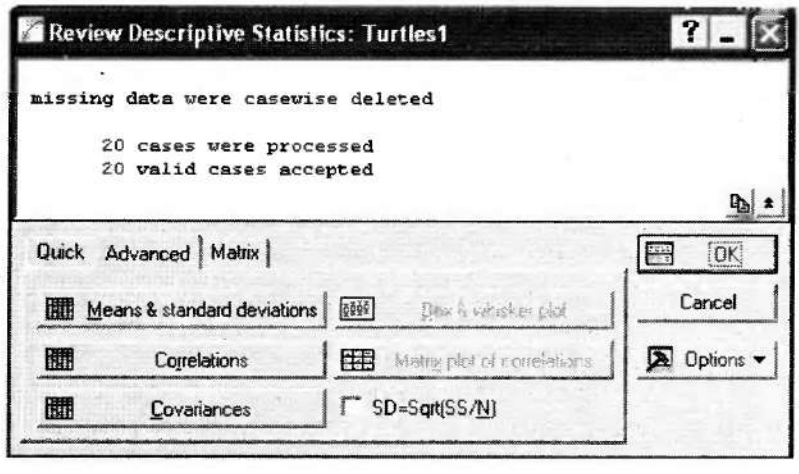


Рис. 10.4

При помощи кнопок **Means & standard deviations**, **Correlations**, **Covariances** можно также просмотреть средние и среднеквадратические отклонения переменных, корреляционную и ковариационную матрицу переменных.

	Correlations (Turtles1)							
Variable	LENGTH	WIDTH	HEIGHT	VEIGHT	V1**2	V2**2	V3**2	V4**2
LENGTH	1,000	0,970	0,957	0,979	0,998	0,961	0,944	0,970
WIDTH	0,970	1,000	0,962	0,980	0,973	0,998	0,958	0,977
HEIGHT	0,957	0,962	1,000	0,975	0,960	0,961	0,997	0,972
WEIGHT	0,979	0,980	0,975	1,000	0,983	0,977	0,970	0,999
V1**2	0,998	0,973	0,960	0,983	1,000	0,967	0,952	0,977
V2**2	0,961	0,998	0,961	0,977	0,967	1,000	0,961	0,977
V3**2	0,944	0,958	0,997	0,970	0,952	0,961	1,000	0,972
V4**2	0,970	0,977	0,972	0,999	0,977	0,977	0,972	1,000

Рис. 10.5

Вернитесь в окно **Review Descriptive Statistics** и щелкните по **OK**. Откроется окно **Model definition**, идентичное одноименному окну в модуле **Multiple Regression** (рис. 9.2). Выберите метод *Forward stepwise* (пошаговый с включением). Нажмите на кнопку **Variables** и в открывшемся окне **Select dependent and independent variable lists** высветите имя функции отклика и независимых переменных (предикторов), как это сделано на рис. 10.6, и щелкните по **OK**.

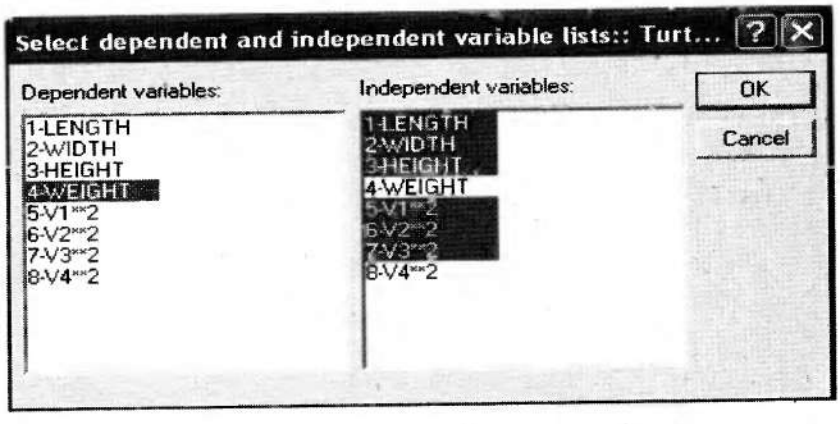


Рис. 10.6

Откроется окно **Multiple Regression Results** (см. рис. 9.3), нажмите на кнопку **Summary:regression results**, появится таблица результатов, в которой приведены стандартизованные (*Beta*) и нестандартизованные (*B*) регрессионные коэффициенты, их стандартные ошибки и уровни значимости. Как видно из таблицы на рис. 10.7, коэффициент множественной корреляции и коэффициент детерминации приняли значения, близкие к 1 (0,9910; 0,9821). В уравнение нелинейной регрессии метод *Forward stepwise* не включил переменные *LENGTH*, *HEIGHT*, но включил их квадраты — $LENGTH^2$, $HEIGHT^2$ и переменную *WIDTH*.

N=20	Regression Summary for Dependent Variable: WEIGHT (Turtles1) R= ,99105603 R2= ,98219205 Adjusted R2= ,97885306 F(3,16)=294,16 p<,00000 Std.Error of estimate: 2,8917					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(16)	p-level
Intercept			80,39644	19,96380	4,027112	0,000975
V1**2	0,447722	0,152004	0,00191	0,00065	2,945463	0,009500
V3**2	0,268806	0,122070	0,00796	0,00361	2,202068	0,042675
WIDTH	0,286999	0,162999	0,56705	0,32205	1,760739	0,097380

Рис. 10.7

Для оценки адекватности построенной модели перейдите на вкладку **Residuals/assumptions/prediction** и нажмите на кнопку **Perform Residual analysis**, в открывшемся окне на вкладке **Residuals** нажмите кнопку **Histogram of residual**. Из построенного графика (рис. 10.8) следует, что из-за малого числа наблюдений (20) распределение остатков не вполне соответствует нормальному закону. Таким образом, построена адекватная нелинейная (квадратичная) модель зависимости веса черепах от линейных размеров — длины, ширины, высоты, имеющая следующий вид:

$$WEIGHT = 80{,}3964 + 0{,}567WIDTH + 0{,}0019LENGTH^2 + 0{,}0079HEIGHT^2.$$

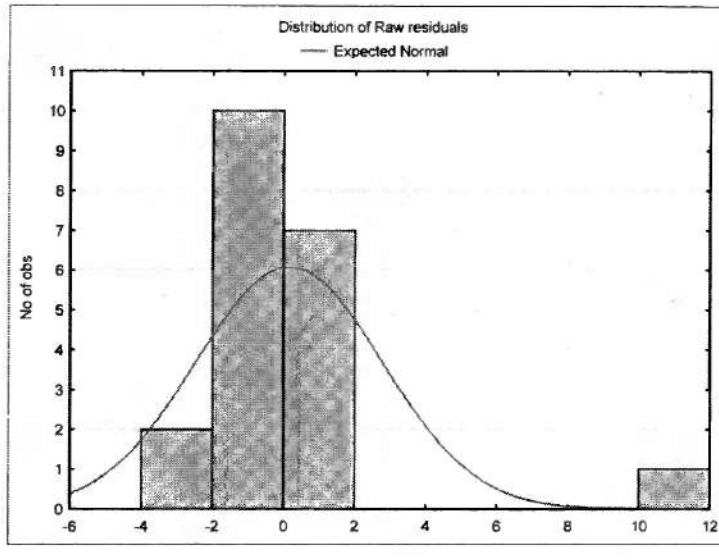


Рис. 10.8

10.3. Модели бинарных откликов. Описание модуля *Nonlinear Estimation*

Модуль нелинейное оценивание включает бинарные (логит и пробит), экспоненциальную, заданную пользователем модели.

Бинарные модели применяют, если зависимая переменная (отклик) бинарна по своей природе, т.е. может принимать только два значения [6]. Например, пациент может выздороветь, а может не выздороветь. Кандидат на работу может пройти тест, а может провалить и т.д. Во всех этих случаях представляет интерес поиск зависимостей между одной или несколькими «непрерывными» переменными и одной зависимой от них бинарной переменной. Так как технически достаточно сложно смоделировать бинарную функцию от непрерывных аргументов, задачу регрессии формулируют иначе. Вместо предсказания бинарной переменной предсказывают непрерывную переменную со значениями на отрезке [0, 1].

Наибольшее распространение получили логит и пробит модели, которые реализованы в программе *STATISTICA*.

В логит модели отклик принимает значения из отрезка [0, 1]. Это достигается применением регрессионного уравнения

$$Y = exp(b_0 + b_1X_1 + ... + b_nX_n)/\{1 + exp(b_0 + b_1X_1 + ... + b_nX_n)\}.$$

Легко заметить, что вне зависимости от коэффициентов регрессии и значений X значения отклика Y, предсказанные этой моделью, всегда будут принадлежать отрезку [0, 1]. Покажем это для случая $n = 1$.

$$Y = exp(b_0 + b_1X)/\{1 + exp(b_0 + b_1X)\} = \{exp(b_0 + b_1X) + 1 - 1\}/\{1 + exp(b_0 + b_1X)\} = 1 - 1/\{1 + exp(b_0 + b_1X)\}.$$

Очевидно, что при $exp(b_0 + b_1X) \to \infty$, $Y \to 1$; при $exp(b_0 + b_1X) \to 0$, $Y \to 0$.

Так как Y принимает значения из [0, 1], можно предположить, что Y — некоторая вероятность, т.е. $Y = p \in [0, 1]$. Преобразуем p следующим образом:

$$p' = ln(p/(1-p)).$$

Такое преобразование называют логит преобразованием. Логит преобразование является линеаризующим. Покажем это. Пусть

$$Y = exp(b_0 + b_1X)/\{1 + exp(b_0 + b_1X)\}.$$

Проведем логит преобразование:

$$p' = ln\{exp(b_0 + b_1X)/\{1 + exp(b_0 + b_1X)\}/\{1 - exp(b_0 + b_1X)/\{1 + \\ + exp(b_0 + b_1X)\}\}\} = b_0 + b_1X.$$

Получим: $p' = b_0 + b_1X$. Для общего случая: $p' = b_0 + b_1X_1 + ... + b_nX_n$.

Логистическая регрессия является обобщением логит регрессии. В логит регрессии отклик принимает лишь два значения: 0 либо 1. В логистической регрессии зависимая переменная может принимать несколько значений [2].

В пробит регрессии бинарная зависимая переменная рассматривается как отклик некоторой нормированной нормально распределенной переменной Y, принимающей любое действительное значение.

Рассмотрим регрессионную модель

$$Y = b_0 + b_1X_1 + ... + b_nX_n,$$

где $Y \in R$ и имеет нормированное нормальное распределение. Тогда в качестве бинарного отклика рассмотрим функцию распределения вероятностей переменной Y, принимающей значения из [0, 1],

$$P(Y<y) = F(y) = \Phi\left(\frac{y - \mu Y}{\sigma Y}\right) = \frac{1}{\sqrt{2\pi}} \int_0^{\frac{y-\mu Y}{\sigma Y}} e^{\frac{-1}{2}t^2} dt,$$

где y — значение переменной Y; μY и σY — соответственно математическое ожидание и среднеквадратическое отклонение Y. Данное уравнение называют «пробит» регрессионной моделью.

Таким образом, в пробит и логит регрессии бинарный отклик моделируют как непрерывную переменную, принимающую значения из интервала [0, 1]. Из такой переменной легко получить бинарную переменную, например, при помощи следующего правила:

$$если\ Y \in [0; 0,5],\ то\ Y = 0;\ если\ Y \in (0,5; 1],\ то\ Y = 1.$$

Для запуска модуля **Nonlinear Estimation** надо в меню **Statistics** выбрать команду **Advanced Linear/Nonlinear Models** (линейные/нелинейные модели). В открывшемся меню выбрать команду **Nonlinear Estimation** (нелинейная оценка).

В окне модуля представлены шесть видов нелинейного оценивания (рис. 10.9):

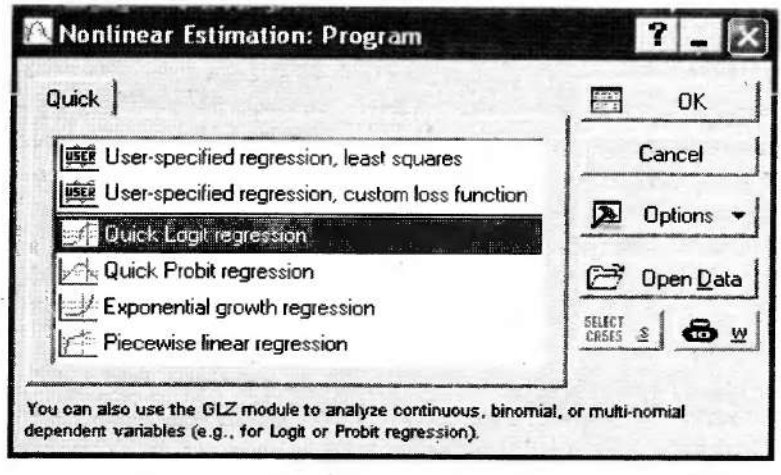


Рис. 10.9

- **User-specified regression, least squares** (определенная пользователем регрессия с квадратичной функцией ошибки);
- **User-specified regression, custom loss function** (регрессия и функция ошибки, определенные пользователем);
- **Quick Logit regression** (логит регрессия);
- **Quick Probit regression** (пробит регрессия);
- **Exponential growth regression** (экспоненциальная регрессия);
- **Piecewise linear regression** (кусочно-линейная регрессия).

Рассмотрим работу с командой **Quick Logit regression**. В качестве данных рассмотрим файл **Program.sta** из библиотеки **Example**, содержащий информацию о тестировании программистов. Наблюдения — это имена программистов. Переменная *Experience* отображает стаж программиста. Переменная *Success* принимает значения *failure* (провал) или *success* (удача) в зависимости от результатов теста. Необходимо построить регрессионную модель зависимости бинарного отклика *Success* от непрерывной переменной *Experience*.

Выберите в диалоговом окне команду **Quick Logit regression**. Откроется окно диалога **Logistic regression**. Для того чтобы начать анализ, следует выбрать зависимую и независимые переменные из списка переменных, щелкнув кнопкой **Variables**. Зависимой переменной (откликом) выберите *Success*, независимой — *Experience*. Нажмите **OK**. Программа возвратится в начальное диалоговое окно. С помощью строки *Input File contains* (введите содержимое файла) можно выбрать один из двух вариантов:

- *Codes and no count* (только коды);
- *Codes and count* (цифровые значения и коды).

Выберите пункт *Codes and no count* и вновь нажмите на **OK**. Откроется окно **Model Estimation**. В верхней информационной части содержится информация о модели: название модели; название зависимой и независимой переменных; коды бинарного отклика; число наблюдений. В нижней информационной части окна можно выбрать процедуру оценивания — *Estimation method*. Список включает следующие процедуры: *quasi-Newton, Simplex, Simplex and quasi-Newton, Hooke-Jeeves pattern moves, Hooke-Jeeves and quasi-Newton, Rosenbrock pattern search* и *Rosenbrock and quasi-Newton*. В данном окне можно назначить параметры процедуры *Maximum number of iterations* (максимальное количество итераций метода), **Convergence criterion** (критерий сходимости метода), *Initial step sizes* (размер шага), *Start values* (стартовые значения параметров). Все эти возможности относятся к вкладке **Advanced**. Вкладка **Review** (обзор) содержит стандартный набор кнопок для предварительного просмотра различных сравнительных характеристик:

- *Means & standard deviations* (среднее значение и стандартное отклонение);
- *Matrix plot for all variables* (матричные графики всех переменных);
- *Box & whisker plot for all var* («ящики с усами» для всех переменных).

После того как все параметры выбраны, изучены различные статистические характеристики, можно перейти непосредственно к оцениванию. В окне **Model Estimation** нажмите **OK**. Если процесс оценивания сошелся за указанное количество итераций, то появится диалоговое окно **Results**.

Окно **Results** состоит из информационной и функциональной частей. Из первой части видно, что значение параметра *Chi-square* достаточно велико, а значение p — мало. Это говорит о достаточной адекватности выбранной модели. Наиболее полно графическая информация о результатах моделирования приведена на вкладке **Residuals** (остатки).

Остатки представляют собой разницу между исходными величинами и предсказанными с помощью модели. Все кнопки этой вкладки (кроме трех) предназначены для графической визуализации результатов, кнопка **Histogram of residuals** — для визуализации гистограммы остатков. Гистограмма остатков дана в сравнении с плотностью нормального распределения. Из рис. 10.10 видно, что гистограмма достаточно «хорошо» приближается кривой плотности нормального распределения. Это также свидетельствует об адекватности модели. Соответствие распределения остатков нормальному закону можно оценить также при помощи кнопок **Normal probability plot of residuals** (нормальный вероятностный график) и **Half-normal probability plot** (полунормальный вероятностный график).

Остальные кнопки позволяют просмотреть и при желании сохранить числовые значения результатов. Кнопка **Observed, predicted, residual vals** предназначена для просмотра наблюдаемых (*observed*) значений отклика, предсказанных значений (*predicted*), остатков (*residual vals*).

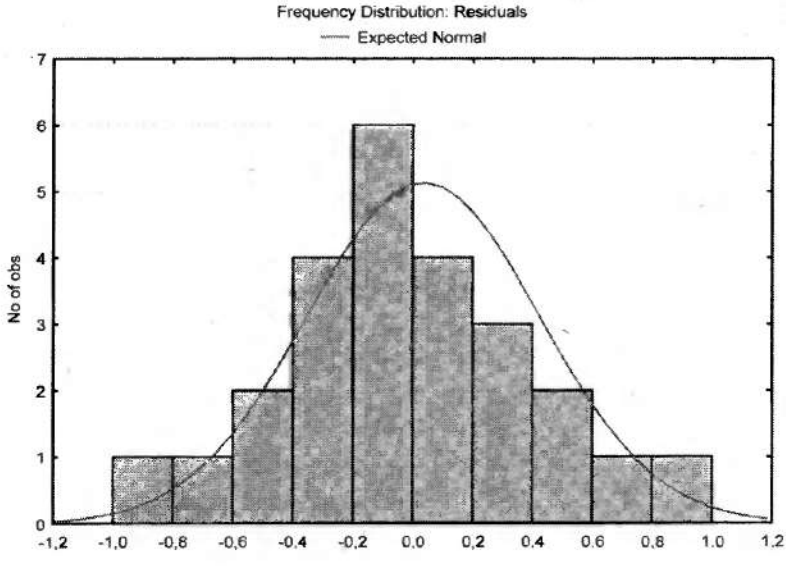


Рис. 10.10

Заметим, что предсказанные значения не переведены в бинарные коды, так как в противном случае невозможно было бы провести анализ остатков.

Кнопка **Save predicted and residual vals** позволяет сохранить прогнозируемые и остаточные значения.

Кнопка **Summary. Parameter estimates** на вкладке **Advanced** предназначена для визуализации предсказанных значений параметров b_0, b_1 уравнения логит регрессии.

Из результатов проведенного исследования следует, что существует достаточно тесная взаимосвязь между переменными модели, а именно стажем (*Experience*) программиста и результатами тестирования (*Success*). Эту взаимосвязь можно аппроксимировать уравнением логит регрессии:

$$Success = exp(-3{,}06 + 0{,}16Experience)/(1 + exp(-3{,}06 + 0{,}16Experience)).$$

Чем выше стаж, тем больше значение предсказанной величины результата тестирования. Таким образом, показано, что чем опытнее программист, тем больше вероятность того, что тест, а значит и вся порученная ему работа будут выполнены успешно.

В **Quick Probit Regression** все рассмотренные окна и методы представлены в таком же виде, как и в **Quick Logit regression.**

10.4. Экспоненциальная регрессия. Описание процедуры *Exponential growth regression*

Exponential growth regression (регрессии экспоненциального роста) соответствует модель вида

$$Y = c + exp(b_0 + b_1X + b_2X^2 + ...) + \varepsilon,$$

где c, b_0, b_1,... — параметры, которые необходимо оценить. Таблица данных изображена на рис. 10.11. Переменная *N город* — это размерность решаемой задачи коммивояжера (число городов); *N операций* – число опреаций, необходимое выполнить некоторым комбинаторным алгоритмом (например, «ветвей и границ») для нахождения оптимального в некотором смысле маршрута. Надо определить, существует ли взаимосвязь между размерностью задачи и количеством операций (трудоемкостью алгоритма), можно ли эту взаимосвязь представить моделью экспоненционального роста, найти параметры модели. В качестве зависимой переменой выберите *N операций*, а в качестве независимой — соответственно *N город*. После выбора переменных в окне **Exponential Growth** нажмите **OK**. Откроется окно **Model Estimation** (рис.10.12).

	1 N город	2 N операций
1	5	140
2	6	403
3	7	1097
4	8	2981
5	9	8000
6	10	22026
7	11	59874
8	12	162755
9	13	442000
10	14	1202604

Рис. 10.11

Выберите метод оценивания, например **Rosenbrock and quasi-newton**, и если нужно установите параметры вычислительной процедуры, щелкните по **OK**. Появится диалоговое окно результатов **Results** (рис. 10.13). Из информационной части окна следует, что итерационный процесс завершился успешно, коэффициент множественной корреляции R составил 0,9999.

Нажмите кнопку **Observed, predicted, residual vals**. Откроется таблица для просмотра наблюдаемых (*observed)* и предсказанных значений (*predicted*) трудоемкости алгоритма, остатков (*residual vales*) (рис. 10.13).

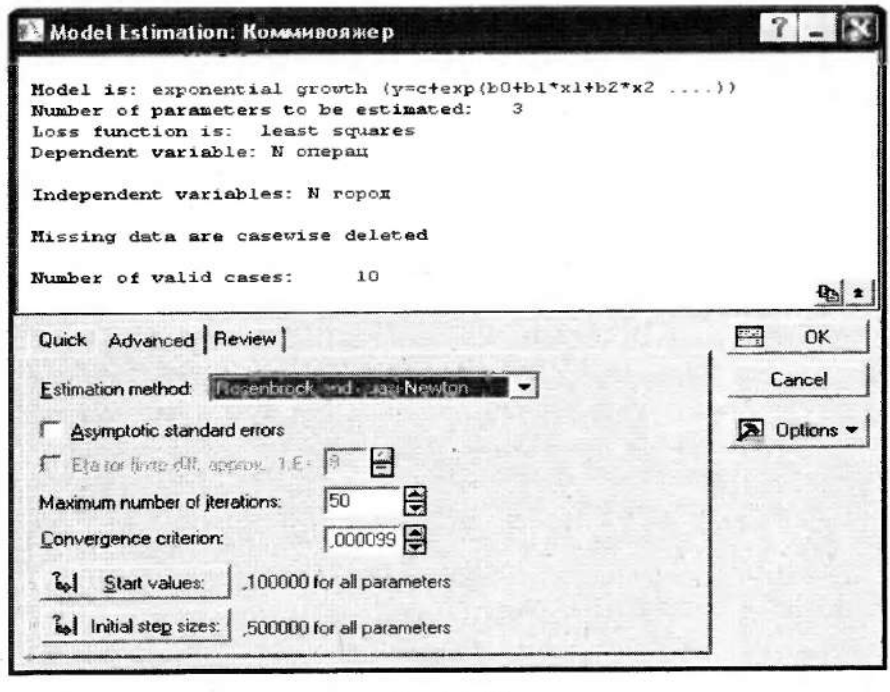


Рис. 10.12

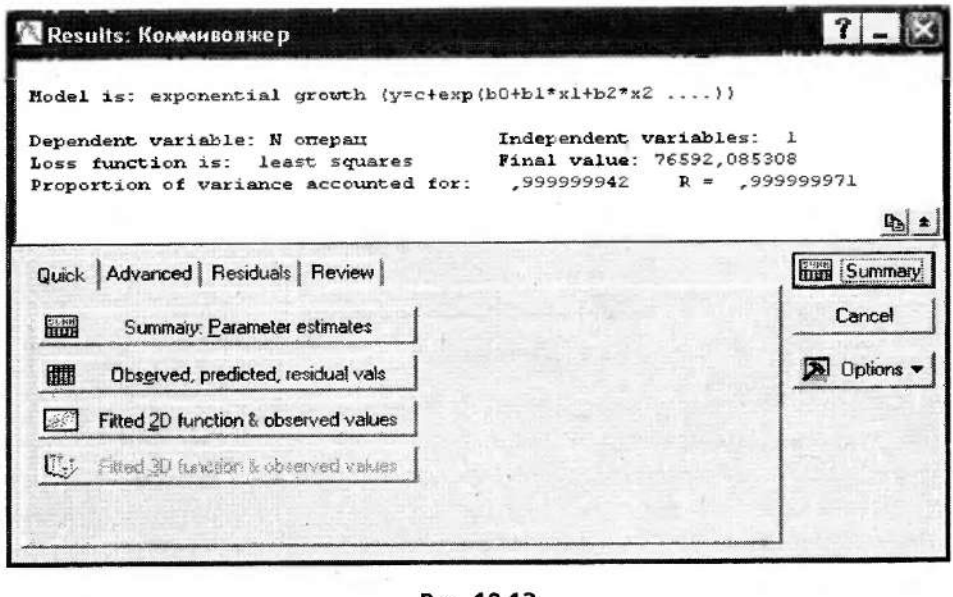


Рис. 10.13

Из таблицы (рис. 10.14) видно, что наблюдаемые и предсказанные значения незначительно отличаются друг от друга. Несмотря на большие значения сравниваемых величин, наибольший остаток не превышает по модулю 500. Дополнительно

убедиться в адекватности модели можно при помощи вкладки **Residuals**. Нажмите кнопку **Summary. Parameter estimates**. Появится таблица (рис. 10.15) со значениями параметров регрессионной модели.

	Model is: (Коммивояжер) Dep. Var. : N операций		
	Observed	Predicted	Residuals
1	140	166	-25,847
2	403	420	-16,353
3	1097	1110	-13,804
4	2981	2989	-7,942
5	8000	8098	-97,989
6	22026	21994	32,653
7	59874	59788	86,124
8	162755	162582	173,170
9	442000	442162	-162,186
10	1202604	1202572	32,028

Рис. 10.14

	Model is: exponential growth (y=c+exp Dependent variable: N операций Lc Final loss: 76592,085308 R=1,0000 Va		
N=10	Const.C	Const.B0	N город
Estimate	18,19634	-0,007985	1,000567

Рис. 10.15

По данным таблицы можно составить уравнение регрессии экспоненциального роста:

$$N\,операц = 18{,}19634 + exp(-0{,}007985 + 1{,}000567N\,город).$$

Используя приведенное соотношение, легко по заданному значению числа городов оценить трудоемкость алгоритма поиска решения задачи коммивояжера.

Нажмите кнопку **Fitted function & observed vales**. Появится изображение графика с нанесенными наблюдаемыми значениями (рис. 10.16).

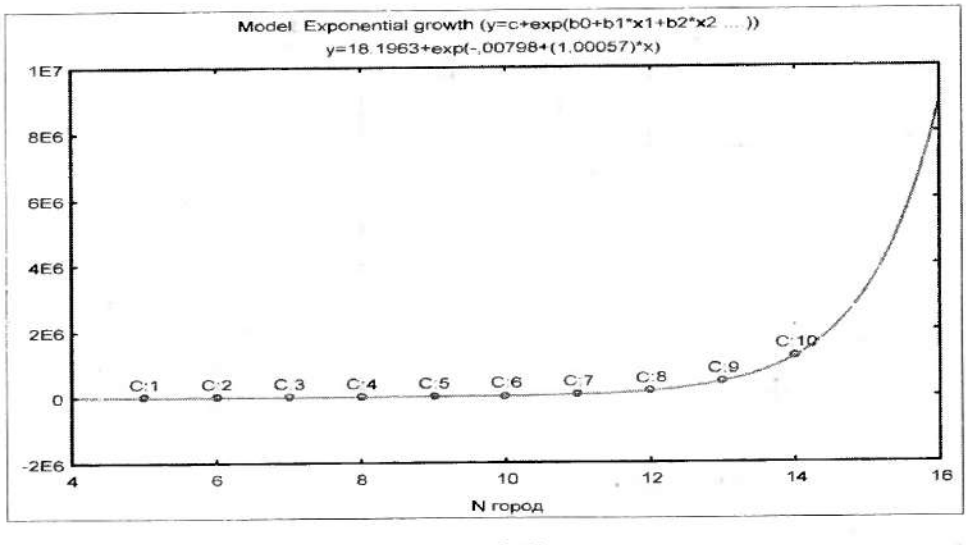


Рис. 10.16

Из графика видно значительное соответствие регрессионной модели эмпирической зависимости между числом городов и трудоемкостью алгоритма.

Об адекватности модели также свидетельствует гистограмма остатков (рис. 10.17), из которой следует соответствие (несмотря на малое число наблюдений) распределения остатков нормальному закону.

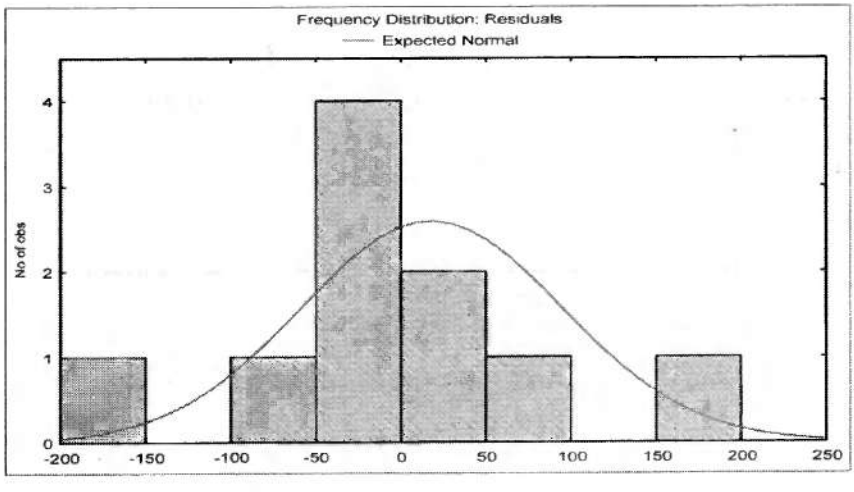


Рис. 10.17

10.5. Кусочно-линейная регрессия. Описание процедуры *Piecewise linear regression*

Кусочно-линейной регрессии соответствует модель:

$$Y = (b_{01} + b_{11}x_1 + ... + b_{m1}x_m)(Y <= Y^*) + (b_{02} + b_{12}x_1 + ... + b_{m2}x_m)(Y > Y^*),$$

где Y^* — точка разрыва, которая может быть либо выбрана пользователем, либо оценена программой. Такой выбор предлагается в стартовом диалоговом окне разрывной регрессии в поле **Breakpoint (Estimated by program** — оценивание программой, **User-defined** — определение пользователем).

Такая модель оценивания очень удобна в том случае, когда зависимая переменная при достижении некоторого критического значения резко меняется. Тогда до достижения критического момента оценивание производится одной моделью, а после достижения — другой.

После выбора рабочего файла и зависимой и независимой переменных дальнейший диалог с программой осуществляется аналогично рассмотренным ранее моделям с той лишь разницей, что в таблицах результатов кроме предсказанных значений и оцененных параметров будет содержаться оцененное значение точки разрыва (в случае, если было указано оценивание этой точки программой). Будут выведены две различные оценки одного и того же параметра — до критического момента и после.

Рассмотрим работу с командой **Piecewise linear regression**. В качестве данных возьмем файл *Акции*, содержащий информацию о помесячной стоимости акций компании *Газпром* в апреле, мае 1997 г. (рис. 10.18). Из таблицы видно, что с 1-е по 10-е наблюдение стоимость акций растет, затем происходит резкое снижение курса и далее вновь с 11-е по 20-е наблюдение продолжается рост курса (рис. 10.19). Необходимо построить регрессионную модель зависимости стоимости акций от месяца. Целесообразно применить кусочно-линейную регрессию.

В окне **Breakpoint Regression** выберите зависимую переменную — *Курс*, а независимую — *N даты* и запустите процесс оценивания, указав в открывшемся окне **Model Estimation** метод **Quasi-Newton**. После успешного завершения итерационного процесса (коэффициент множественной корреляции R равен 0,9344) откроется окно **Results** (Рис. 10.20). Нажмите кнопку **Summary. Parameter estimates**. В результате оценивания получим по два значения параметров — до точки разрыва и после (рис. 10.21).

	1 N даты	2 Дата	3 **Курс**
1	1	14/05/97	1850,000
2	2	15/05/97	1870,000
3	3	16/05/97	1890,000
4	4	17/05/97	1900,000
5	5	18/05/97	1950,990
6	6	19/05/97	2000,000
7	7	20/05/97	2044,800
8	8	21/05/97	2100,680
9	9	22/05/97	2150,760
10	10	23/05/97	2203,760
11	1	24/05/97	1716,160
12	2	25/05/97	1739,556
13	3	26/05/97	1756,160
14	4	27/05/97	1809,800
15	5	28/05/97	1853,685
16	6	29/05/97	1888,950
17	7	30/05/97	1943,640
18	8	31/05/97	2018,578
19	9	01/06/97	2050,400
20	10	02/06/97	2096,659

Рис. 10.18

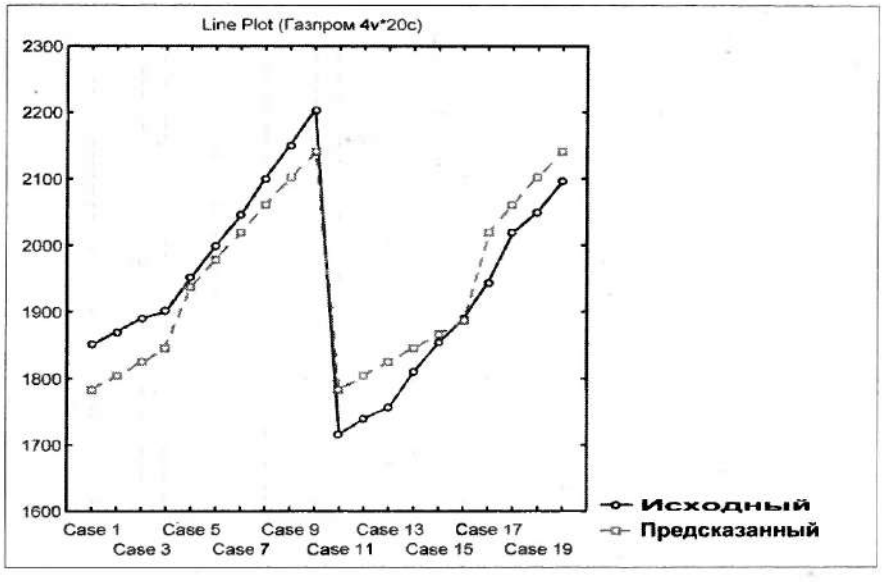


Рис. 10.19

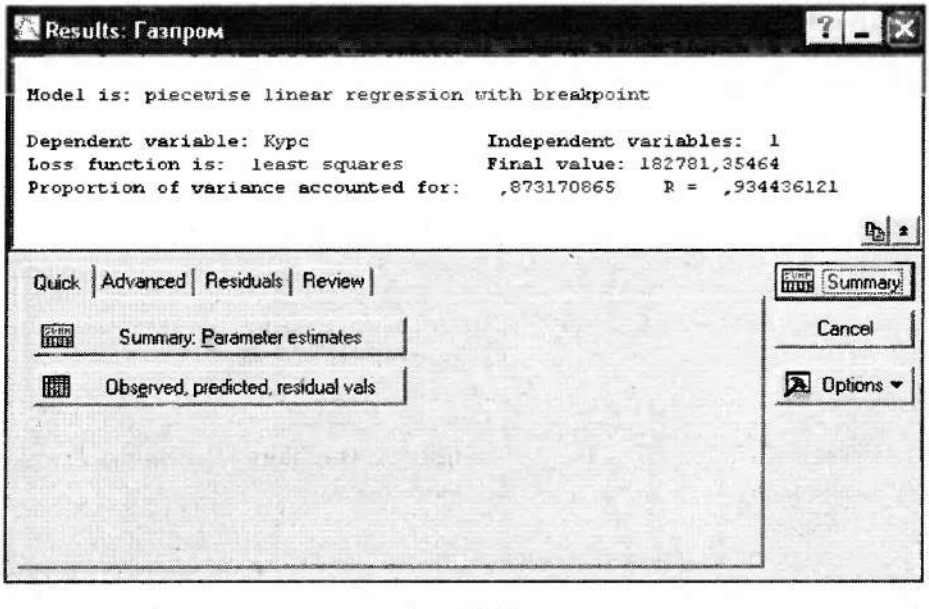


Рис. 10.20

	Model is: Piecewise linear regression with breakpc Dependent variable**Курс** Loss: Least squares Final loss: 52131,169922 R=,92595 Variance expl				
N=20	Const.B0	N даты	Const.B0	N даты	Breakpt.
Estimate	1763,350	20,6712	1733,486	40,82785	1941,729

Рис. 10.21

Соответственно будет построено две линейные модели — до точки разрыва **Breakpt** (1941,729) и после:

1) *Курс* = 1763,350 + 20,6712*N даты;*

2) *Курс* = 1733,486 + 40,8278*N даты.*

Обратите внимание, что точка разрыва 1941,729 принимает некоторое промежуточное значение между «критическими» значениями курса 2203,760 и 1716,160.

Нажмите кнопку **Observed. predicted. residual vales**, появится таблица с исходными, предсказанными значениями курса и остатками. Линейный график предсказанных значений приведен на рис. 10.18. Как видно из графиков исходных и предсказанных значений, основная тенденция изменения курса акций сохранена при незначительном отличии предсказанных и исходных значений.

Таким образом, построена адекватная кусочно-линейная регрессионная модель изменения курса акций.

10.6. Определенная пользователем регрессия

Теперь рассмотрим регрессию, определенную пользователем, — **User-specified regression**. В данной версии пользовательская регрессия разделена на два пункта: **User-specified regression, least squares** (пользовательская регрессия с квадратичной функцией потерь) и **User-specified regression, custom loss function** (регрессия с определенной пользователем функцией потерь). В первом случае независимо от заданной модели квадратичная функция потерь задана программой. Во втором же случае функцию потерь задает пользователь. В том и другом случае регрессионную модель определяет пользователь. Перебором различных моделей можно найти наилучшую регрессию, которой соответствует минимальное значение функции потерь.

В § 10.4 зависимость между размерностью задачи коммивояжера и количеством операций (трудоемкостью алгоритма) аппроксимировали моделью регрессионного роста с достаточно высокой степенью точности. Аппроксимируем эту зависимость моделью другого типа, например, полиномиальной.

В стартовом окне модуля «Нелинейное оценивание» выберите команду **User-specified regression, least squares**. В появившемся окне нажмите кнопку **Function to be Estimated**. Откроется окно **Estimated function**, в одноименном поле окна надо указать тип пользовательской регрессии, например $V2 = b1V1^{b2}$. В нижней части окна даны пояснения и приведены примеры составления регрессионной

модели. Нажмите **OK**. В открывшемся окне **Nonlinear Least Squares Model Estimation** задайте метод вычисления коэффициентов модели — **Levenberg-Marquardt** и при необходимости на вкладке **Advanced** — параметры итерационной процедуры *(Maximum number of iterations* = 200). После запуска и завершения вычислительной процедуры откроется диалоговое окно результатов **Results**. Из информационной части окна следует, что процедура завершилась успешно — R = 0,9998, но велико окончательное значение функции потерь *(Loss function Final value* = 316873072,94) и значительно отличие предсказанных значений от исходных (рис. 10.22), в особенности при малых значениях числа городов (*N город*).

Для визуализации коэффициентов уравнения регрессии нажмите кнопку **Summary. Parameters estimate.** Появится таблица со значениями коэффициентов регрессии *b1* и *b2* (рис. 10.23).

	Model is: v2=b1*v1**b2 Коммивояжер Dep. Var. : N операций		
	Observed	Predicted	Residuals
1	140	1	138,50
2	403	17	386,78
3	1097	127	969,24
4	2981	743	2238,38
5	8000	3516	4483,91
6	22026	14130	7896,21
7	59874	49730	10144,35
8	162755	156856	5898,70
9	442000	451264	-9264,25
10	1202604	1200420	2184,04

Рис. 10.22

	Model is: v2=b1*v1**b2 (Коммивояжер) Dep. Var. : N операций Caution: Degenerated result, the values may not correct !!!					
	Estimate	Standard error	t-value df = 8	p-level	Lo. Conf Limit	Up. Conf Limit
b1	0,00000	0,000000	0,00	0,00	0,00	0,00
b2	13,20201	0,155248	0,00	0,00	0,00	0,00

Рис. 10.23

Таким образом, для приведенного файла данных точность приближения полиномиальной регрессии значительно уступает экспоненциальной. Данный результат не противоречит известному в теории сложности алгоритмов факту, что задача коммивояжера является, по-видимому, труднорешаемой, так как не существует алгоритма, решающего ее быстрее, чем за время, экспоненциально зависящее от размерности задачи.

Глава 11

Анализ взаимосвязей между списками переменных

11.1. Канонический анализ

Во многих модулях программы *STATISTICA* можно вычислить парные коэффициенты корреляции для выражения зависимости между двумя переменными, а также матрицы парных коэффициентов корреляции. Например, коэффициент корреляции Пирсона (r) показывает степень линейной зависимости между двумя переменными. Модуль **Nonparametric Statistics** (непараметрические статистики) предлагает различные статистики, основанные на рангах исследуемых переменных. Модуль **Multiple Regression** (множественная регрессия) позволяет оценить зависимость между зависимой переменной (откликом) и множеством предикторных переменных. Модуль **Correspondence Analysis** (анализ соответствий) дает возможность исследовать зависимости внутри множества категориальных переменных.

Модуль **Canonical Analysis** (канонический анализ) предназначен для анализа зависимостей между списками переменных [6]. Если говорить точнее, он позволяет исследовать зависимость между двумя множествами переменных, тем самым развивая

возможности других модулей. Например, исследователь в сфере образования может оценить зависимость между навыками по трем учебным дисциплинам и оценками по пяти школьным предметам. Социолог может определить зависимость между прогнозами социальных изменений, печатаемыми в двух газетах, и реальными изменениями, оцененными с помощью четырех различных статистических признаков. Медик может изучить зависимость между различными неблагоприятными факторами и появлением определенной группы симптомов заболевания. Во всех этих случаях нас интересует зависимость между двумя множествами переменных, для анализа которой и предназначен модуль **Canonical Analysis**.

Канонический анализ является обобщением множественной корреляции как меры связи между одной случайной величиной и множеством других случайных величин [6]. Но, как известно, множественная корреляция есть максимальная корреляция между одной случайной величиной и линейной функцией других случайных величин. Эта концепция была обобщена на случай связи между множествами случайных величин. Канонический анализ очень полезен, если имеется два больших множества величин и необходимо определить взаимосвязь между ними. При этом достаточно ограничиться рассмотрением небольшого числа наиболее коррелированных линейных комбинаций из каждого множества.

В качестве примера использования анализа канонических корреляций рассмотрим файл данных **Factor.sta** из библиотеки **Examples**. В файле собраны результаты опросов 100 респондентов относительно степени их удовлетворенности жизнью. В файле даны значения следующих 10 переменных:

- *WORK 1* (удовлетворенность работой), первая компонента;
- *WORK 2* (удовлетворенность работой), вторая компонента;
- *WORK 3* (удовлетворенность работой), третья компонента;
- *HOBBY 1* (удовлетворенность свободным временем), первая компонента;
- *HOBBY 2* (удовлетворенность свободным временем), вторая компонента;
- *HOME 1* (удовлетворенность домашней жизнью), первая компонента;
- *HOME 2* (удовлетворенность домашней жизнью), вторая компонента;
- *HOME 3* (удовлетворенность домашней жизнью), третья компонента;
- *MISCEL 1* (общая удовлетворенность), первая компонента;
- *MISCEL 2* (общая удовлетворенность), вторая компонента.

Многомерность каждой переменной объясняется тем, что рассматриваются различные аспекты удовлетворенности, например, можно быть удовлетворенным зарплатой, получаемой на работе, но не удовлетворенным коллективом или тем, сколько времени затрачивается, чтобы до нее добраться, и т.д. Значения переменных преобразованы таким образом, что средней степени удовлетворенности соответствуют значения близкие к 100, низкой и высокой степени удовлетворенности — значения соответственно меньше и больше 100.

Предположим, что нас интересует характер связи между удовлетворенностью работой (группа переменных *WORK*) и удовлетворенностью свободным временем (группа переменных *HOBBY*). Проще всего просуммировать значения откликов по двум множествам вопросов и посчитать корреляцию полученных сумм.

Если полученная корреляция статистически значима, можно заключить, что существует зависимость между удовлетворением от работы и удовлетворением от свободного времени. При этом не получили информации о связи различных форм удовлетворенности свободным временем и удовлетворенности работой. Например, если значения двух откликов второго множества соответствуют удовлетворению от пребывания в публичных местах (музеи, театры, рестораны) и удовлетворению от общения с диваном в свободное время, то складывать их все равно, что складывать яблоки с апельсинами. По сути дела, упрощая задачу и суммируя отклики, мы теряем важную информацию. И, возможно, просто «разрушаем» существующие зависимости между переменными.

Для исправления положения разумно немного изменить изучаемые объекты. Вместо рассмотрения обычных сумм по множествам полезно рассматривать взвешенные суммы, чтобы веса, приписанные отдельным слагаемым, соответствовали реальной «структуре» переменных, т.е. их взаимной значимости. Например, если на удовлетворение, получаемое от работы, мало влияет удовлетворение от общения с диваном, но сильно влияет удовлетворение от увлечения театром, первому следует придать меньший вес, чем второму. Эту общую идею можно выразить следующим соотношением:

$$a_1y_1 + a_2y_2 + ... + a_py_p = b_1x_1 + b_2x_2 + ... + b_qx_q,$$

где под знаком « = » подразумевается наличие стохастической взаимосвязи между линейными комбинациями переменных обоих множеств.

Таким образом, если имеем два множества, содержащие p и q переменных соответственно, будем исследовать зависимость между взвешенными суммами переменных из каждого множества (т.е. между линейными комбинациями p и q переменных соответственно).

После того как сформулировали в общем виде «уравнение модели» для канонической корреляции, надо определить веса для двух наборов переменных. Взвешенные суммы, слабо коррелированные друг с другом, не представляют никакого интереса для исследователя, поэтому при подборе весовых коэффициентов нужно исходить из условия максимальной коррелированности двух множеств.

По терминологии анализа канонической корреляции взвешенные суммы определяют канонический корень, или каноническую переменную. Эти канонические переменные (взвешенные суммы) можно рассматривать как обозначения некоторых «скрытых» переменных, лежащих в основе наблюдаемых явлений. При этом канонический анализ практически всегда приводит к вычислению более чем одной пары взвешенных сумм. Если быть точным, число канонических корней, вычисляемых программой, равно числу переменных в меньшем множестве. В нашем примере, когда анализируемые группы содержат три и две переменные соответственно, число канонических корней будет равно двум.

Как было отмечено, при вычислении корней программа рассматривает все максимально коррелированные взвешенные суммы (максимизирует значение корреляции между каноническими переменными). При вычислении более чем одного корня каждая последующая пара канонических переменных объясняет свою уникальную долю изменчивости в этих двух наборах переменных.

При этом последовательно получаемые пары канонических переменных не коррелированы друг с другом и объясняют все меньшую и меньшую долю изменчивости.

При вычислении канонических корней *STATISTICA* использует общую корреляционную матрицу [10], которая состоит из подматриц R_{pp}, R_{pq}, R_{qq}, R_{qp}:

$$R = \begin{bmatrix} Rpp & Rpq \\ Rqp & Rqq \end{bmatrix}$$

где R_{pp} — матрица корреляций между переменными 1-й группы; $R_{pq} = R^{T}_{qp}$ — матрица корреляций между переменными 1-й и 2-й групп; R_{qq} — матрица корреляций между переменными 2-й группы.

Как функция от общей корреляционной матрицы строится матрица B размерности pp:

$$B = R^{-1}_{pp} R_{pq} R^{-1}_{qq} R_{qp}.$$

При проведении анализа программа вычислит столько собственных значений матрицы B, сколько имеется канонических корней, т.е. столько, сколько переменных содержит наименьшее множество. Если извлечь квадратный корень из полученных собственных значений, получим набор чисел, который можно проинтерпретировать как коэффициенты корреляции. Поскольку они относятся к каноническим переменным, их также называют каноническими корреляциями. Поэтому собственные значения матрицы B, ранжированные по убыванию, равняются квадратам канонических корреляций.

На первом шаге, после того как найдено первое собственное значение, программа вычислит веса, максимизирующие корреляцию между взвешенными суммами по двум множествам, и определит соответствующее им значение первого корня. На втором и последующих шагах (по числу канонических корней) программа найдет следующую пару канонических переменных, имеющих максимальную корреляцию и не коррелированных с предыдущими парами, и вычислит соответствующее ей значение канонического корня. Как и собственные значения, корреляции между последовательно выделяемыми на каждом шаге каноническими переменными убывают. Поэтому в выводимом на экран отчете о коррелированности между множествами переменных часто приводят лишь первое, т.е. максимальное, значение. Однако другие канонические переменные также могут быть значимо коррелированны, и эти корреляции часто допускают достаточно осмысленную интерпретацию.

Критерий значимости канонических корреляций сравнительно несложен. Канонические корреляции оцениваются одна за другой в порядке убывания. Только те корни, которые оказались статистически значимыми, остаются для последующего анализа. Хотя на самом деле вычисления происходят немного иначе. Программа сначала оценивает значимость всего набора корней, затем значимость набора, остающегося после удаления первого корня, второго корня, и т.д.

Исследования показали, что используемый критерий обнаруживает большие канонические корреляции даже при небольшом размере выборки (например, $n = 50$). Слабые канонические корреляции (например, $R = 0{,}3$) требуют больших размеров выборки ($n > 200$) для обнаружения в 50% случаев.

После определения числа значимых канонических корней возникает вопрос об интерпретации каждого (значимого) корня. Напомним, что каждый корень в действительности представляет две взвешенные суммы, по одной на каждое множество переменных. Одним из способов толкования «смысла» каждого канонического корня является рассмотрение весов, сопоставленных каждому множеству переменных. Эти веса называются каноническими весами. При анализе обычно учитывают, что чем больше приписанный вес (т.е. абсолютное значение веса), тем больше вклад соответствующей переменной в значение канонической переменной. Таким образом, рассмотрение канонических весов позволяет понять «значение» каждого канонического корня, т.е. увидеть, как конкретные переменные в каждом множестве влияют на взвешенную сумму (т.е. каноническую переменную).

Канонические веса также могут использоваться для вычисления значений канонических переменных. Для этого достаточно сложить исходные переменные с соответствующими весовыми коэффициентами. Напомним, что канонические веса обычно определяются для стандартизированных (z-преобразованных) переменных.

Еще одним способом интерпретации канонических корней является рассмотрение обычных корреляций между каноническими переменными (или факторами) и переменными из каждого множества. Эти корреляции называются каноническими нагрузками факторов. Считается, что переменные, сильно коррелированные с канонической переменной, имеют с ней много общего. Поэтому при описании смысла канонической переменной следует исходить в основном из реального смысла этих сильно коррелированных переменных. Такой способ интерпретации канонических переменных похож на метод, используемый в факторном анализе.

Иногда канонические веса для переменной оказываются близкими к нулю, а соответствующие им нагрузки очень велики. Также возможна обратная ситуация, когда канонические веса велики, а нагрузки небольшие. В таких случаях вывод может оказаться достаточно противоречивым. Однако следует помнить, что канонические значения соответствуют уникальному вкладу, вносимому соответствующей переменной во взвешенную сумму или каноническую переменную; нагрузки канонических факторов отражают полную корреляцию между соответствующей переменной и взвешенной суммой.

Коэффициенты канонической корреляции соответствуют корреляции между взвешенными суммами по двум множествам переменных. Они не говорят ничего о том, какую часть изменчивости (дисперсии) каждый канонический корень объясняет в переменных. Однако можно сделать заключение о доле объясняемой дисперсии, рассматривая нагрузки канонических факторов. Если возвести эти корреляции в квадрат, то полученные числа будут отражать долю дисперсии, объясняемую каждой переменной. Для каждого корня можно вычислить среднее значение этих долей. В результате получится средняя доля изменчивости, объясненной в этом множестве на основании соответствующей канонической переменной. Другими словами, можно вычислить среднюю долю дисперсии, извлеченной каждым корнем.

Каноническая корреляция при возведении в квадрат дает долю дисперсии, общей для сумм по каждому множеству (канонической переменной). Если умножить эту долю на долю извлеченной дисперсии, то получится мера избыточности множества переменных, т.е. величина, показывающая, насколько избыточно одно

множество переменных, если задано другое множество. Избыточность может быть записана следующим образом:

$$Избыточность_{лев} = [(нагрузки_{лев}^2)/p]R_i^2;$$
$$Избыточность_{прав} = [(нагрузки_{прав}^2)/q]R_i^2.$$

В этих уравнениях, p обозначает число переменных в первом (левом) множестве переменных, а q — число переменных во втором (правом) множестве. Величина R_i^2 соответствует квадрату соответствующей канонической корреляции. Поскольку последовательно извлекаемые канонические корни не коррелированы между собой, то можно просто просуммировать избыточности по всем (или только по значимым) корням, получив при этом общий коэффициент избыточности.

Для измерения избыточности также бывает полезным определение практической значимости канонических корней. При больших размерах выборки канонические корреляции со значением $R = 0{,}3$ могут оказаться статистически значимыми. Если возвести этот коэффициент в квадрат ($R^2 = 0{,}09$) и использовать формулу для избыточности, становится ясным, что такие канонические корни объясняют лишь незначительную долю изменчивости переменных. Хотя, разумеется, окончательное решение о практической значимости принимается на основании субъективной позиции исследователя. Однако для получения правдоподобных оценок того, насколько реальная изменчивость переменных объясняется конкретным каноническим корнем, не следует забывать о мере избыточности, т.е. о том, насколько реальная изменчивость в одном множестве переменных объясняется другим множеством.

Рассмотрим наиболее важные предположения анализа канонической корреляции, выполнение которых обеспечивает получение достоверных и обоснованных результатов.

Применение критерия значимости при анализе канонической корреляции основано на предположении, что переменные в выборке имеют многомерное нормальное распределение. Как и большинство других модулей пакета *STATISTICA*, модуль **Canonical Analysis** позволяет провести графический анализ данных, т.е. построить гистограмму частот с наложенной на нее нормальной кривой или вывести на экран диаграмму рассеяния наблюдаемой переменной. Теоретически последствия нарушения этого предположения мало изучены. Однако при очень больших размерах выборки результаты анализа канонической корреляции достаточно устойчивы.

При наличии больших корреляций между данными (например, $R > 0{,}7$) даже малые размеры выборки (например, $n = 50$) позволяют в большинстве случаев обнаружить эти корреляции. Однако для получения достоверных оценок нагрузок канонических факторов (если нужно интерпретировать только наиболее значимый корень) рекомендуется использовать как минимум в 20 раз больше наблюдений, чем число переменных, используемых в анализе. Для получения достоверных оценок для двух канонических корней рекомендуется использовать в 40–60 раз больше наблюдений, чем число исследуемых переменных.

Наличие выбросов может существенно влиять на значение коэффициентов корреляции, а значит, и на вычисление канонических корреляций. Конечно, чем

больше размер выборки, тем меньшее влияние оказывают один или два выброса. Однако при проведении анализа надо выявить и исключить выбросы.

Еще одним предположением является требование, чтобы переменные в обоих множествах не были полностью избыточными. Например, если включить одну и ту же переменную дважды в одно из множеств, то окажется непонятным, какие ей следует придать веса. С вычислительной точки зрения такая избыточность нарушает ход анализа. При наличии полной коррелированности между наблюдаемыми переменными ($R = 1$) корреляционная матрица не может быть обращена, и вычисления, необходимые для анализа канонической корреляции, таким образом, не могут быть завершены. Подобные корреляционные матрицы называются плохо обусловленными.

11.2. Описание модуля *Canonical Analysis*

Рассмотрим последовательность шагов при работе с модулем **Canonical Analysis**. В верхнем меню **Statistics** щелкните по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выберите команду **Canonical Analysis**. Откроется стартовая панель модуля. Рассмотрим функциональное назначение основных кнопок.

После нажатия кнопки **Variables** открывается стандартное окно выбора переменных, в котором можно выбрать переменные для анализа. Только переменные, выбранные в этом окне, впоследствии доступны для канонического анализа. Матрица корреляции будет вычислена для всех переменных, выбранных в этом окне. Выберите переменные *WORK 1 – WORK 3, HOBBY 1, HOBBY 2*.

Поле списка **Input File** имеет два возможных значения: необработанные исходные данные *Raw Data* и матрица корреляции *Correlation Matrix*. Если выбрана первая опция, программа ожидает на входе файл с необработанными исходными данными. Если выбрана вторая опция, в качестве файла данных необходимо указать файл, содержащий соответствующую матрицу корреляции в стандартном формате матричного файла *STATISTICA*. Файлы корреляционных матриц могут быть созданы в различных модулях системы (например, в модулях **Basic Statistics/Tables**, **Factor Analysis**, **Multiply Regression**). Выберите *Raw Data*.

Установите флажок на *Review descriptive statistics and correlation matrix* (отображать описательные статистики и корреляционные матрицы), чтобы после выхода из стартовой панели открыть окно **Review Descriptive statistics** (просмотр описательных статистик). Это окно позволяет просмотреть описательные статистики для выбранных переменных. Нажмите **OK**. В открывшемся окне **Model Definition** (определение модели) нажмите кнопку **Variables** и выберите в первом списке переменных *WORK 1–WORK 3*, во втором – *HOBBY 1–HOBBY 2*.

Если установить флажок на *Bath processing/reporting* (пакетная обработка/печать) и выбрать вывод на принтер, все выходные результаты будут распечатаны без вмешательства пользователя.

Если нажать кнопку **Descriptives**, откроется окно просмотра описательных статистик. Не будем описывать это окно, так как о нем достаточно подробно шла речь в предыдущих разделах. Нажмите **OK**, появится окно **Canonical Analyses Results** (результаты анализа канонической корреляции) (рис. 11.1). Наиболее

существенные результаты анализа приведены в верхней информационной части окна. Каноническая корреляция $R = 0{,}76$, приведенная в верхней строке окна, соответствует корреляции между первыми каноническими переменными (взвешенными суммами). Она равна максимальному извлеченному каноническому корню. Ее значение свидетельствует о наличии сильной зависимости между группами переменных. Значения *Chi-Square* (χ^2) = 0,86 и уровень значимости $p = 0{,}00$ показывают значимость *R. Number of valid cases* (число наблюдений) — это количество людей (100), которые отвечали на вопросы анкеты.

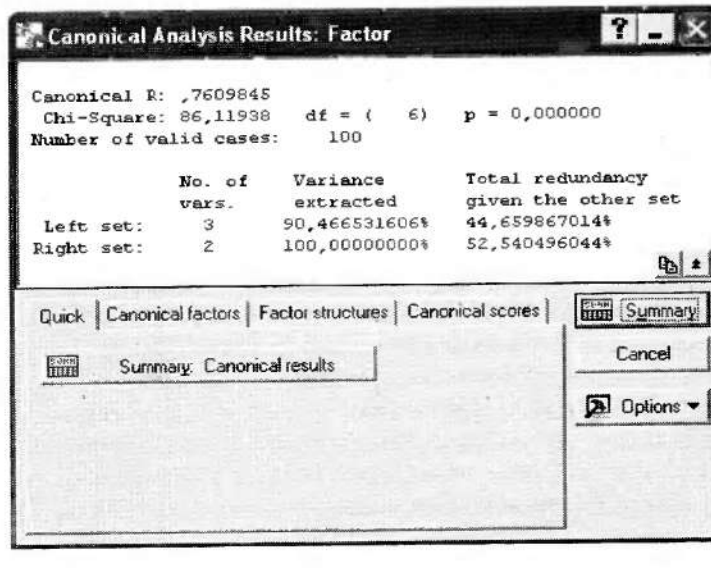


Рис. 11.1

В нижней информационной части окна приведены *No. Of vars* (число переменных в левом и правом множествах); *Variance extracted* (процентное число извлеченных дисперсий из левого и правого множеств переменных); *Total redundancy given the other set* (общая избыточность при заданном втором множестве).

Нажмите кнопку **Summary. Canonical results** (итоговые результаты). На экран будет выведена таблица результатов с итоговыми значениями статистик для текущего анализа (рис. 11.2).

	Canonical Analysis Si Canonical R: ,76098 Chi2(6)=86,119 p=0,0	
N=100	Left Set	Right Set
No. of variables	3	2
Variance extracted	90,4665%	100,000%
Total redundancy	44,6599%	52,5405%
Variables: 1	WORK_1	HOBBY_1
2	WORK_2	HOBBY_2
3	WORK_3	

Рис. 11.2

Variance extracted означает долю дисперсии (изменчивости), объясняемую каждым множеством переменных.

Total redundancy — величина, показывающая, насколько реальная изменчивость в одном множестве переменных объясняется другим множеством.

На вкладке **Canonical factors** (рис. 11.3) нажмите кнопку **Eigenvalues.** Появится таблица результатов (рис. 11.4), содержащая собственные значения, соответствующие каноническим корням. Канонические корни не имеют определенных значений, они определяются собственными значениями. Отметим, что квадратный корень из собственного значения равен соответствующему каноническому коэффициенту корреляции.

Если нажать на кнопку **Plot of Eigenvalues**, программа предоставит изображение кусочно-линейного графика убывающих собственных значений (рис. 11.5).

Щелкните по кнопке **Chi-Square tests** (критерий χ^2). Появится таблица результатов (рис. 11.6), содержащая для каждого канонического корня значения R, R^2, χ^2, число степеней свободы, *p-уровень* и значение лямбда.

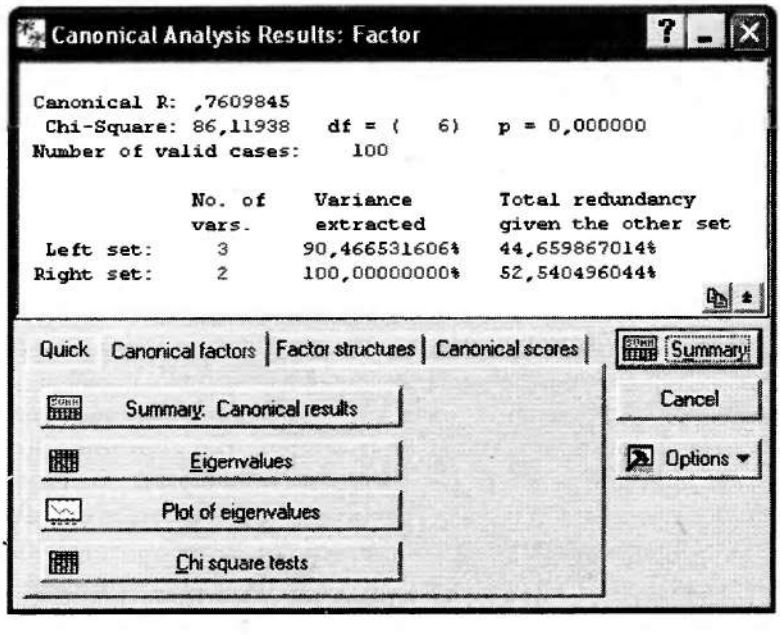


Рис. 11.3

	Eigenvalues (Factor	
Root	Left Root 1	Right Root 2
Value	0,579097	0,031225

Рис. 11.4

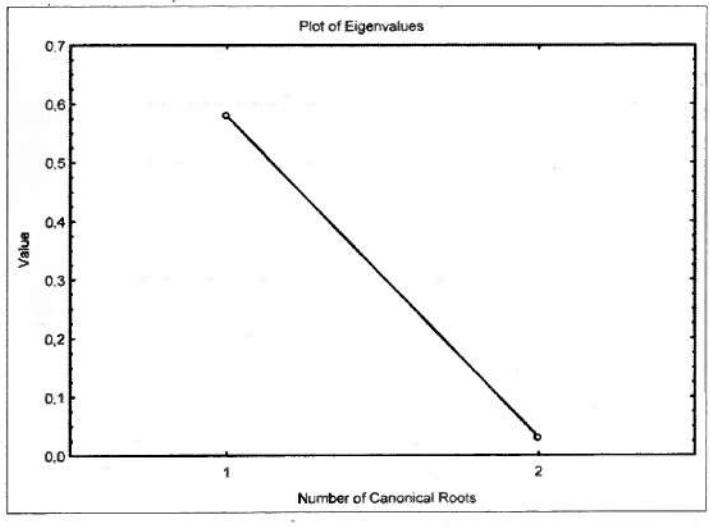


Рис. 11.5

Root Removed	Chi-Square Tests with Successive Roots Remc				
	Canonicl R	Canonicl R-sqr.	Chi-sqr.	df	p
0	0,760984	0,579097	86,11938	6	0,000000
1	0,176706	0,031225	3,04541	2	0,218137

Рис. 11.6

Результаты этой таблицы показывают, какие канонические корни следует считать статистически значимыми, чтобы использовать их для дальнейшего рассмотрения (т.е. для интерпретации). Так называемый последовательный критерий значимости работает следующим образом. Сначала рассматриваются все канонические корни вместе, т.е. производится анализ без удаления корней. Полученное значение критерия — χ^2 и *p-уровень* выводятся в первой строке таблицы результатов. Если это значение значимо, то можно заключить, что хотя бы один канонический корень является статистически значимым. Далее удаляется первый (т.е. наиболее значимый) корень и определяется статистическая значимость оставшихся корней. Полученное значение критерия выводится во второй строке таблицы результатов. Если оно значимо, то как минимум два корня статистически значимы. Если нет, то на этом шаге следует остановиться, оставив для интерпретации только первый корень. Если и второй корень окажется значимым, то он также удаляется, а процедура определяет статистическую значимость оставшихся корней и т.д.

Из данных, приведенных в таблице, можно сделать вывод о статистической значимости только первого канонического корня. Значит, целесообразно рассматривать лишь первую пару канонических переменных.

Выделите вкладку **Factor structures** (рис. 11.7) и щелкните по кнопке **Correlations with & between sets** (корреляции внутри и между множествами). Появятся таблицы с корреляциями между переменными из одного и из разных

множеств (рис. 11.8–11.10). В первом множестве наибольшая зависимость между переменными *WORK 2* и *WORK 3*. Во втором множестве сильная зависимость между переменными *HOBBY 1* и *HOBBY 2*. Между переменными *WORK 2* и *HOBBY 2* наибольшая корреляции для переменных из разных множеств.

Эти корреляционные матрицы используются для составления общей корреляционной матрицы двух множеств переменных.

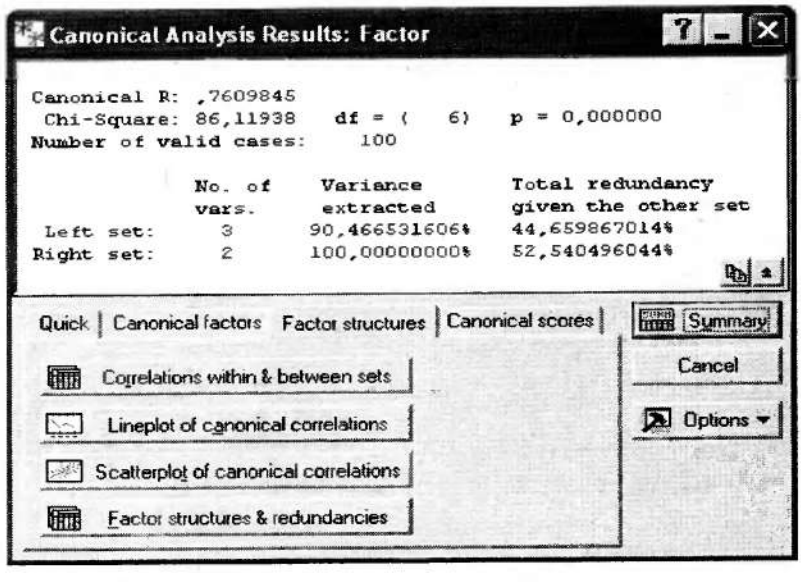


Рис. 11.7

Root Removed	Correlations, left set (Factor)		
	WORK_1	WORK_2	WORK_3
WORK_1	1,000000	0,647401	0,652615
WORK_2	0,647401	1,000000	0,731893
WORK_3	0,652615	0,731893	1,000000

Рис. 11.8

Root Removed	Correlations, right set	
	HOBBY_1	HOBBY_2
HOBBY_1	1,000000	0,804694
HOBBY_2	0,804694	1,000000

Рис. 11.9

Root Removed	Correlations, left set w	
	HOBBY_1	HOBBY_2
WORK_1	0,598122	0,521100
WORK_2	0,688546	0,697782
WORK_3	0,636948	0,630030

Рис. 11.10

Нажмите кнопку **Factor structures & redundancies** (факторная структура и избыточности). Программа выведет таблицы результатов с нагрузками канонических факторов и коэффициентами избыточности (извлеченной дисперсией) для обоих множеств переменных. Нагрузки канонических факторов (рис. 11.11, 11.12) можно интерпретировать так же, как и в факторном анализе. Они представляют собой корреляции между переменными из множества и соответствующими каноническими переменными.

Root Variable	Factor Structure, right s	
	Root 1	Root 2
HOBBY_1	-0,956218	-0,292654
HOBBY_2	-0,943209	0,332201

Рис. 11.11

Root Variable	Factor Structure, left se	
	Root 1	Root 2
WORK_1	-0,77735(	-0,627867
WORK_2	-0,958288	0,169565
WORK_3	-0,87646(	0,015931

Рис. 11.12

Variance extractd (извлеченная дисперсия) вычисляется как сумма квадратов нагрузок канонических факторов по всем переменным множества, деленная на число переменных в этом множестве (например, $0{,}763591 = (0{,}77735^2 + 0{,}95828^2 + 0{,}87646^2)/3$). Полученное значение можно интерпретировать как среднюю долю дисперсии, объясняемой соответствующим корнем. Общая доля извлеченной дисперсии, которая приводится в верхней части окна **Canonical Analyses Results**, может интерпретироваться как доля дисперсии, объясняемая всеми каноническими корнями (например, $90{,}4665\% = (0{,}763591 + 0{,}141074)100\%$). Коэффициенты избыточности *(Reddncy)* (рис. 11.13, 11.14) вычисляются умножением извлеченной дисперсии на квадрат канонической корреляции (например, $0{,}442194 = 0{,}763591 \times 0{,}7609845^2$). Коэффициенты избыточности (для конкретного корня) можно интерпретировать как среднюю долю дисперсии, объясняемую в переменных соответствующего множества, исходя из значения корня, при заданных значениях переменных другого множества.

Root Factor	Variance Extracted (Propo	
	Variance extractd	Reddncy.
Root 1	0,763591	0,442194
Root 2	0,141074	0,004405

Рис. 11.13

Root Variable	Variance Extracted (Propc	
	Variance extractd	Reddncy.
Root 1	0,901998	0,522345
Root 2	0,098002	0,003060

Рис. 11.14

Общая избыточность, показываемая в верхней части окна **Canonical Analyses Results**, равна сумме коэффициентов избыточности по всем корням, умноженная на 100% (например, 44,6598% = (0,442194 + 0,004405)100%).

Нажмите кнопку **Lineplot of canonical correlations** (график канонических корреляций). Программа построит кусочно-линейный график канонических корреляций (рис. 11.15), которые равны квадратным корням из собственных значений.

Нажмите кнопку **Scatterplot of canonical correlations** (диаграмма рассеяния канонических корреляций).

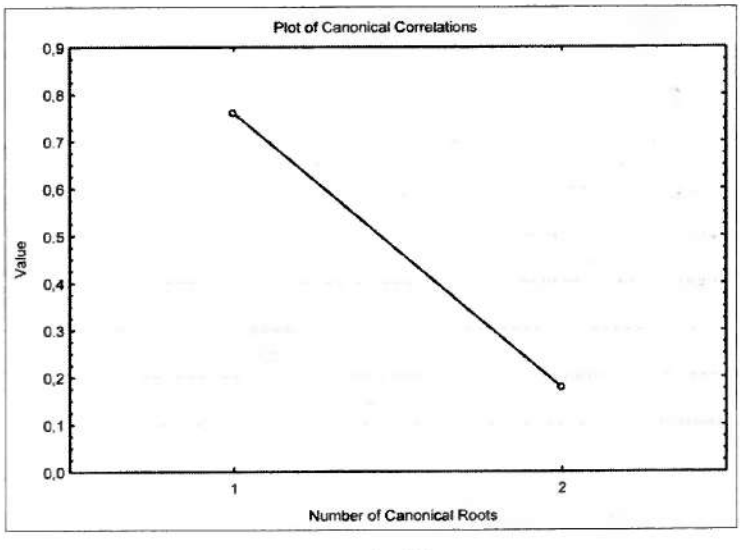


Рис. 11.15

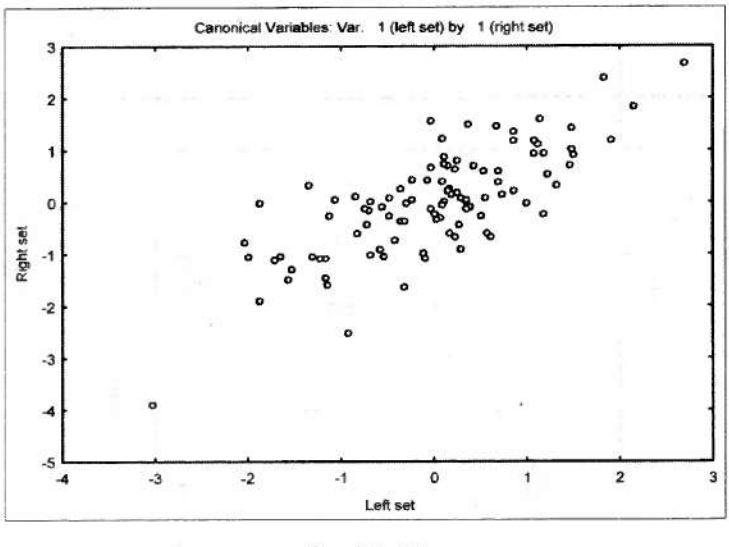


Рис. 11.16

Эта опция доступна, только если проводится анализ необработанных исходных данных (т.е. в стартовой панели модуля **Canonical Analysis** выбран файл исходных данных). В открывшемся окне укажите канонические корни, в соответствии с которыми будут выбраны канонические переменные обоих множеств переменных (левое и правое) (например, *root1, root1)*.

Щелкните по **OK**. Появится диаграмма рассеяния значений канонических переменных левого (*Left set)* и правого множеств (*Right set*), соответствующих первому каноническому корню (рис. 11.16). Из диаграммы видно, что зависимость между каноническими переменными левого и правого множеств близка к линейной. Для второго корня линейная зависимость не просматривается.

Выделите вкладку **Canonical scores** (канонические значения) и нажмите кнопку **Left & right set canonical weights** (канонические веса для левого и правого множества). На экран будут выведены таблицы результатов с каноническими весами для каждого множества переменных (рис.11.17, 11.18). Веса соответствуют нормированным переменным, их можно использовать для вычисления канонических значений для каждого канонического корня, для каждого множества переменных, а также для интерпретации канонических корней. Чем больше абсолютное значение веса, тем больше вклад соответствующей переменной в значение канонической переменной.

Из таблиц видно, что для левого множества наибольший вклад в значение первой канонической переменной вносит переменная *WORK 2*, а для правого множества — *HOBBY 1*.

Variable	Canonical Weights, le	
	Root 1	Root 2
WORK_1	-0,174860	-1,35928
WORK_2	-0,618370	0,83701
WORK_3	-0,309764	0,29041

Рис. 11.17

Variable	Canonical Weights, ri	
	Root 1	Root 2
HOBBY_1	-0,559552	-1,58872
HOBBY_2	-0,492941	1,61064

Рис. 11.18

Кнопка **Save canonical scores** (сохранить канонические значения) доступна, если в стартовой панели модуля **Canonical Analysis** выбран ввод необработанных исходных данных. После нажатия на эту кнопку пользователю предлагается ввести имя файла, в котором будут сохранены значения. В появившемся затем диалоговом окне пользователь может выбрать переменные, которые необходимо сохранить вместе с каноническими значениями.

Таким образом, между списками переменных *WORK 1, WORK 2, WORK 3* и *HOBBY 1, HOBBY 2* существует сильная зависимость, и наибольший вклад в зависимость вносят переменные *WORK 2* и *HOBBY 1*.

Глава 12

Классификационный анализ с обучением

12.1. Дискриминантный анализ

Кластерный и дискриминантный анализ наиболее ярко отражают черты многомерного анализа в классификации, факторный анализ — в исследовании связи.

Дискриминантный анализ как раздел многомерного статистического анализа включает статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками (классификация с обучением) [9].

Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров) объекта классифицировать его, т.е. отнести к одной из нескольких групп (классов) некоторым оптимальным способом. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. Этот вид статистического анализа является многомерным, так как использует несколько параметров объекта.

Широкий круг задач, возникающих на практике и связанных с классификацией, можно решить методами дискриминантного анализа, типичные области применения которого медицина, управление производством, экономика, геология, контроль качества.

В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является построение *k-мерного* случайного вектора $X = (x_1, x_2, ..., x_k)$. Требуется установить правило, согласно которому по значениям координат вектора X объект относят к одной из возможных совокупностей π_i, $i = 1, 2, ..., n$. Для построения правила дискриминации все выборочное пространство R значений вектора X разбивается на области R_i, $i = 1, 2, ..., n$, так что при попадании X в R_i объект относят к совокупности π_i.

Правило дискриминации выбирается в соответствии с определенным принципом оптимальности на основе априорной информации о вероятностях p_i извлечения объекта из π_i. При этом следует учитывать размер убытка от неправильной дискриминации. Априорная информация может быть представлена как в виде некоторых сведений о функциях *k-мерного* распределения признаков в каждой совокупности, так и в виде выборок из этих совокупностей. Априорные вероятности p_i могут быть либо заданы, либо нет. Очевидно, что рекомендации будут тем точнее, чем полнее исходная информация.

Обычно в задаче различения переходят от вектора признаков, характеризующих объект, к линейной функции от них, дискриминантной функции-гиперплоскости, наилучшим образом разделяющей совокупность выборочных точек.

Методы дискриминации можно условно разделить на параметрические и непараметрические.

В параметрических известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Здесь естественно в дискриминантной функции заменить неизвестные параметры распределения их наилучшими оценками, произведенными на основе выборочных точек. Правило дискриминации можно основывать на отношении правдоподобия.

Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений.

Таким образом, параметрический дискриминантный анализ применяется при выполнении ряда предположений:

- предположения о том, что наблюдаемые величины — измеряемые характеристики объекта имеют нормальное распределение. Это предположение следует проверять. В модуле имеются специальные опции, позволяющие быстро построить гистограммы и нормальные вероятностные графики. Умеренные отклонения от этого предположения допустимы;
- предположения об однородности дисперсий и ковариаций наблюдаемых переменных в разных классах. Умеренные отклонения от этого предположения также допустимы.

Наиболее важным критерием правильности построенного классификатора является практика.

В модуле **Discriminant Analysis** пакета *STATISTICA* имеется широкий набор средств, обеспечивающих проведение дискриминантного анализа данных, визуализации и интерпретации результатов. Модуль позволяет проводить классификационный анализ с пошаговым включением или исключением переменных или вводить в модель заданные пользователем блоки переменных. В дополнение к многочисленным графикам и статистикам, описывающим дискриминирующую функцию, программа содержит также большой набор средств и статистик для классификации старых и новых наблюдений (для оценки качества модели). Программа выполняет полный канонический анализ и выдает все собственные значения, их уровни значимости, коэффициенты дискриминантной функции, структурной матрицы и т.д. Встроенные средства графической поддержки включают гистограммы, диаграммы рассеяния, большой набор категоризованных графиков, позволяющий исследовать распределение и взаимосвязи между зависимыми переменными для разных групп и мн. др.

В целом модуль **Discriminant Analysis** — это обучающая система и очень полезный инструмент для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп; классификации наблюдений в различные группы.

Модели, реализованные в модуле, являются линейными, а функции классификации и дискриминантные функции — линейными комбинациями наблюдаемых величин.

12.2. Описание модуля *Discriminant Analysis*

Возможности модуля и основные принципы работы с ним продемонстрируем на классическом примере анализа цветов ириса [14]. Задача состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветка отнести ирис к одному из трех типов: *SETOSA, VERSICOL, VIRGINIC.*

Для запуска модуля в верхнем меню **File** надо выбрать команду **Open** и открыть файл данных **Irisdat** из библиотеки **Examples**. В файле содержатся результаты измерений 150 цветков ириса, по 50 цветков каждого типа.

В меню **Statistics** необходимо щелкнуть по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выбрать команду **Discriminant Analysis**. Откроется стартовая панель модуля (рис. 12.1).

Чтобы выбрать переменные для анализа, нужно нажать кнопку **Variables**. Появится окно, в котором необходимо выбрать группирующие и независимые переменные. В качестве *Grouping variable* выберите переменную *IRISTYPE* (сорт ириса). Группирующая переменная не должна входить в список независимых переменных. В качестве *Independent variable list* (список независимых переменных) выберите переменные *SEPALLEN* (длина чашелистика), *SEPALWID* (ширина чашелистика), *PETALLEN* (длина пестика), *PETALWID* (ширина пестика) и щелкните по **OK**. Далее надо задать коды для значений группирующей переменной. Нажмите кнопку **Codes for grouping variables**. Откроется окно, где можно задать коды для названия групп, к которым принадлежит объект. В качестве кодов группирующих переменных

выберите типы цветов с помощью нажатия кнопки **All**. Щелкните **OK**. Если в диалоге **Discriminant Function Analysis** была установлена галочка на опции *Advanced options*, откроется окно диалога **Model Definition** (рис. 12.2). Перейдите на вкладку **Advanced**. В поле **Method** можно указать метод дискриминантного анализа:

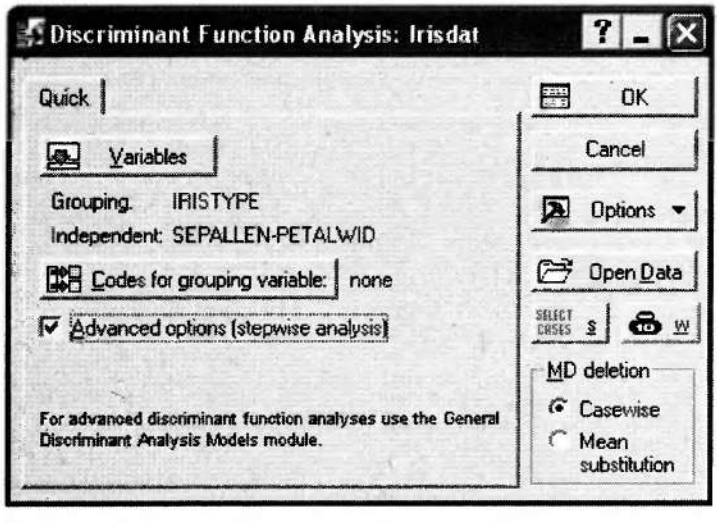


Рис. 12.1

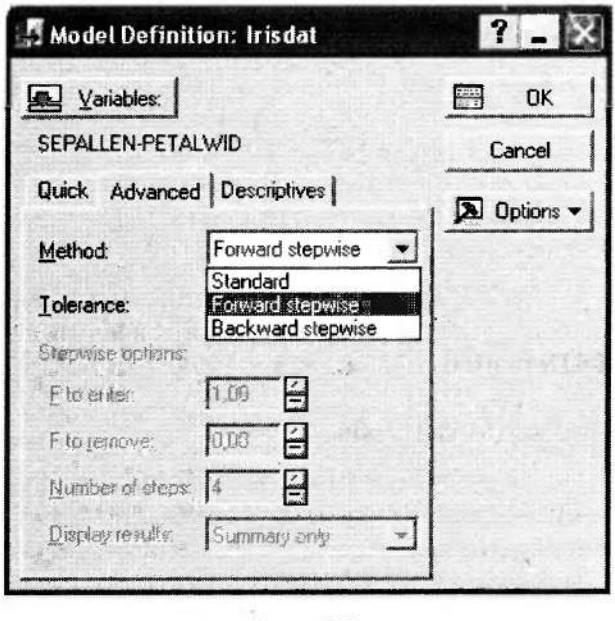


Рис. 12.2

- *Standard* (стандартный). При этом методе все выбранные переменные будут одновременно включены в модель (уравнение);
- *Forward stepwise* (пошаговый вперед). Программа на последовательных шагах включает переменные в модель;
- *Backward stepwise* (пошаговый назад). Программа включает в модель все выбранные переменные и затем удаляет на каждом шаге переменные из модели.

Опция *Tolerance* (толерантность) задает нижнюю границу толерантности. Толерантность, как уже указывалось, является мерой избыточности переменных. Чем меньше значение толерантности, тем избыточнее переменная в модели, так как переменная несет малую дополнительную информацию. Переменные с толерантностью меньше заданного значения в модель не включаются.

Можно выделить следующие опции диалога **Stepwise options** для методов пошагового анализа (*Forward stepwise, Backward stepwise*).

F to enter (*F-включить*), *F to remove* (*F-исключить*). В пошаговом анализе дискриминантной функции переменные включают в модель, если соответствующее им значение *F* больше, чем значение *F-включить*, переменные удаляют из модели, если соответствующее им значение *F* меньше, чем значение *F-исключить*. Заметим, что значение *F-включить* всегда должно быть больше, чем значение *F-исключить*. Если при проведении пошагового анализа с включением необходимо включить все переменные, надо установить в поле **F to enter** значение, равное малому числу (например, 0,0001), а в поле **F to remove** — значение 0. Если при проведении пошагового анализа с исключением необходимо исключить все переменные из модели, надо установить в поле **F to enter** значение, равное очень большому числу (например, 9999), а в поле **F to remove** — меньшее значение того же порядка (например, 9998).

Number of steps (число шагов) определяет максимальное количество шагов, которое будет осуществлено. Эта опция имеет приоритет перед значениями **F to enter**, **F to remove**. Пошаговый метод будет остановлен при достижении максимального числа шагов, несмотря на то, следует ли еще включать или исключать переменные на основе значений *F*.

Dysplay results (вывод результатов). Если в предлагаемом программой списке выбрать *Summary only* (только итог), то программа выполнит все шаги пошагового анализа и только потом появится окно результатов. При выборе *At each step* (на каждом шаге) программа будет выводить результаты анализа на каждом шаге.

В диалоге **Model Definition** выберите метод *Standard* и щелкните по **OK**, откроется окно результатов (рис. 12.3).

Информационная часть окна сообщает, что:

- *Number of variables in model* (число переменных в модели) равно 4;
- *Wilks' Lambda* (значение лямбда Уилкса) равно 0,0234386;
- *approx.F*(8,288) (приближенное значение *F-статистики* с числом степеней свободы 8 и 288) равно 199,1454;
- *p* (уровень значимости *F-критерия*) меньше 0,0000.

Статистика лямбда Уилкса (λ) вычисляется как отношение детерминанта матрицы внутригрупповых дисперсий/ковариаций к детерминанту общей

ковариационной матрицы. Значения λ принадлежат интервалу [0,1]. Значения λ, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения λ, лежащие около 1, свидетельствуют о плохой дискриминации.

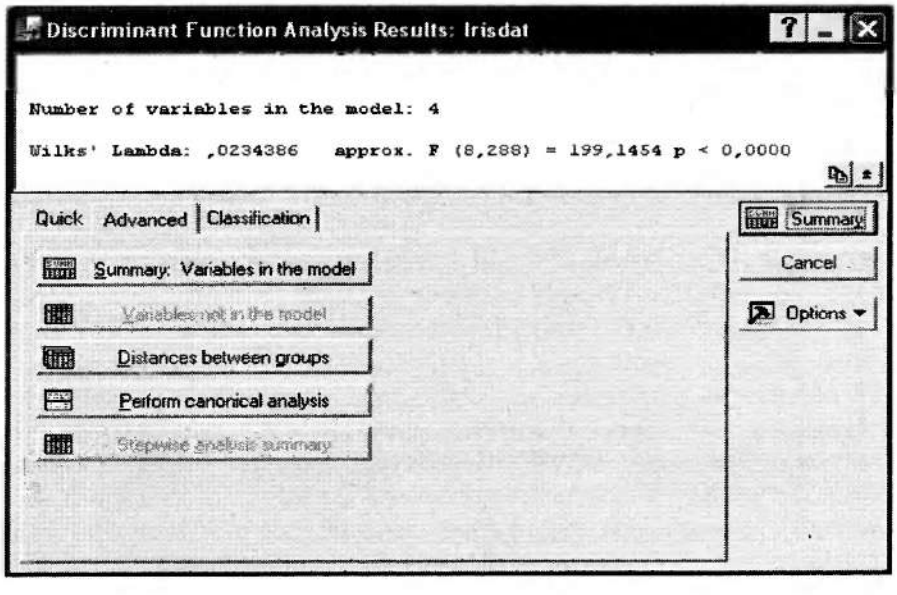


Рис. 12.3

Нажмите кнопку **Summary:Variables in the model** (итоги: переменные, включенные в модель). Появится итоговая таблица анализа данных (рис. 12.4):

	Discriminant Function Analysis Summary (Irisdat) No. of vars in model: 4; Grouping: IRISTYPE (3 grps) Wilks' Lambda: ,02344 approx. F (8,288)=199,15 p<0,0000					
N=150	Wilks' Lambda	Partial Lambda	F-remove (2,144)	p-level	Toler.	1-Toler. (R-Sqr.)
SEPALLEN	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007
SEPALWID	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALLEN	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
PETALWID	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686

Рис. 12.4

В первом столбце таблицы приведены значения *Wilks Lambda*, являющиеся результатом исключения соответствующей переменной из модели [6]. Чем больше значение λ, тем более желательно присутствие этой переменной в процедуре дискриминации.

Значение *Partial Lambda* (частная лямбда) есть отношение лямбда Уилкса после добавления соответствующей переменной к лямбде Уилкса до добавления этой переменной. Частная лямбда характеризует единичный вклад соответствующей

переменной в разделительную силу модели. Чем меньше статистика лямбда Уилкса, тем больше вклад в общую дискриминацию. из таблицы видно, что переменная *PETALLEN* дает вклад больше всех, переменная *PETALWID* — вторая по значению вклада, переменная *SEPALWID* — третья по значению вклада, а переменная *SEPALLEN* вносит в общую дискриминацию вклад меньше всех. Поэтому на этой стадии исследования можно заключить, что размеры пестика являются главными переменными, которые позволяют производить дискриминацию между различными типами ирисов.

F-remove (*F-исключить*) — это значения *F-критерия*, связанные с соответствующей частной лямбда Уилкса. Значения *p-level* — это уровни значимости критериев *F-remove*.

Толерантность (*Toler*) определяется как $1\text{-}R^2$, где R^2 — это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Как уже отмечалось, толерантность является мерой избыточности переменной в модели.

Кнопка **Distances between groups** (расстояние между группами) предназначена для вывода таблицы с расстояниями между группами. По данным этой таблицы можно судить о качестве дикриминации наблюдений и о степени различия (неоднородности) групп.

Для получения дальнейших результатов о природе дискриминации следует провести канонический анализ. Чтобы увидеть, как четыре переменные разделяют различные совокупности (типы ирисов), надо вычислить дискриминантную функцию. Каждая последующая дискриминантная функция будет вносить все меньший и меньший вклад в общую дискриминацию. Максимальное число оцениваемых функций равно числу переменных (4) или числу совокупностей (3) минус один, в зависимости от того, какое число меньше. В нашем случае оцениваются две дискриминантные функции.

Щелкните по кнопке **Perform canonical analysis** (выполнение канонического анализа), программа вычислит независимые (ортогональные) дискриминантные функции. В открывшемся окне **Canonical Analysis** нажмите кнопку **Summary: Chi square tests of successive roots** (итоги: χ^2-критерий последовательности корней). Появится таблица результатов с пошаговым критерием для канонических корней — дискриминантных функций (см. § 11.2). Первая строка дает критерий значимости для всех корней. Вторая строка содержит значимость корней, оставшихся после удаления первого корня и т.д. Таким образом, таблица позволит оценить, сколько значимых корней нужно интерпретировать (рис. 12.5). Как видно из таблицы, обе дискриминантные функции статистически значимы.

	Chi-Square Tests with Successive Roots Removed (Iris					
Roots Removed	Eigen-value	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	32,19193	0,984821	0,023439	546,1153	8	0,000000
1	0,28539	0,471197	0,777973	36,5297	3	0,000000

Рис. 12.5

Если нажать кнопку **Coefficients for canonical variables** (коэффициенты канонических переменных, см. § 11.1), появятся две таблицы с коэффициентами дискриминантных (канонических) функций. В первой таблице даны исходные (нестандартизованные) коэффициенты дискриминантных функций. Эти коэффициенты могут быть использованы для вычисления значений канонических переменных для каждого наблюдения каждой дискриминантной функции. Во второй таблице (рис.12.6) приведены стандартизованные коэффициенты дискриминантных функций. Эти коэффициенты, основанные на стандартизованных переменных, принадлежат к одной и то же шкале измерений (абсолютной), поэтому их можно сравнивать, чтобы определить величины и направления вкладов переменных в каждую каноническую функцию. из таблицы видно, что наибольший вклад в дискриминантную функцию 1 вносят переменные *PETALLEN, PETALWID*, в дискриминантную функцию 2 — *SEPALWID, PETALWID*. В таблицах приведены собственные значения (*Eigenval*) для каждой дискриминантной функции и кумулятивная доля объясненной дисперсии (*Cum. Prop.*), накопленной каждой функцией. Как видно, функция 1 ответственна за 99,1% объясненной дисперсии, т.е. 99,1% всей дискриминирующей мощности определяется этой функцией. Поэтому эта функция наиболее «важна».

В диалоговом окне **Canonical Analysis** выберите вкладку **Advanced**. В открывшемся окне щелкните кнопкой **Factor structure** (факторная структура). В появившейся таблице (рис. 12.7) приведены объединенные внутригрупповые корреляции переменных с соответствующими дискриминантными функциями. Эти корреляции называют еще структурными коэффициентами. Обычно структурные коэффициенты используют для содержательной интерпретации функций, в отличие от коэффициентов дискриминантной функции, которые обозначают вклад каждой переменной в функции. У переменных *PETALLEN, PETALWID* наибольшие корреляции с дискриминантной функцией 1, у переменных *SEPALWID, PETALWID* — наибольшие корреляции с дискриминантной функцией 2.

Variable	Standardized Coeffici for Canonical Variable	
	Root 1	Root 2
SEPALLEN	0,42695	0,012408
SEPALWID	0,52124	0,735261
PETALLEN	-0,94726	-0,401038
PETALWID	-0,57516	0,581040
Eigenval	32,19193	0,285391
Cum.Prop	0,99121	1,000000

Рис. 12.6

Variable	Factor Structure Matr Correlations Variables (Pooled-within-groups	
	Root 1	Root 2
SEPALLEN	-0,222596	0,310812
SEPALWID	0,119012	0,863681
PETALLEN	-0,706065	0,167701
PETALWID	-0,633178	0,737242

Рис. 12.7

Нажмите кнопку **Means of canonical variables** (средние канонических переменных). Программа выведет таблицу (рис. 12.8) со средними значениями для дискриминантных функций, которые позволяют определить группы, лучше всего

идентифицируемые конкретной дискриминантной функцией. Из таблицы видно, что дискриминантная функция 1 идентифицирует в основном сорта *SETOSA* (значение среднего значительно отличается от других средних), а дискриминантная функция 2 – сорт *VERSICOL*. Но дискриминантная функция 2 определяет всего лишь 0,879% дискриминирующей мощности (100%–99,121%).

	Means of Canonical Variables (Irisdat)	
Group	Root 1	Root 2
SETOSA	7,60760	0,215133
VERSICOL	-1,82505	-0,727900
VIRGINIC	-5,78255	0,512767

Рис. 12.8

Перейдите на вкладку **Canonical scores** (канонические значения), щелкните кнопкой **Canonical scores for each case** (канонические значения для каждого наблюдения). Появится таблица (рис. 12.9) со значениями дискриминантных функций для каждого наблюдения.

	Unstandardized Canonical Scores (I		
Case	Group	Root 1	Root 2
1	SETOSA	7,67197	-0,13489
2	VIRGINIC	-6,80015	0,58090
3	VERSICOL	-2,54868	-0,47220
4	VIRGINIC	-6,65309	1,80532
5	VIRGINIC	-3,81516	-0,94299
6	SETOSA	7,21262	0,35584
7	VIRGINIC	-5,10556	1,99218
8	VERSICOL	-3,49805	-1,68496

Рис. 12.9

Наблюдения (ирисы), определяемые программой как объекты, принадлежащие одной группе, должны иметь близкие значения дискриминантных функций. Чтобы сохранить эти значения, надо нажать на кнопку **Save canonical scores** (сохранить канонические значения).

Кнопка **By group** предназначена для вывода гистограммы канонических значений по группам. Кнопка **All groups combined** выведет комбинированную гистограмму для всех групп.

По таблице, изображенной на рис. 12.9, трудно судить о результатах разделения программой наблюдений по группам. Нажмите кнопку **Scatterplot of canonical scores** (диаграмма рассеяния для канонических значений). Появится диаграмма рассеяния (рис. 12.10) канонических значений для пар значений дискриминантных функций. на диаграмме видно, что наблюдения (ирисы), принадлежащие одинаковым группам (сортам), локализованы в определенных областях плоскости,

при этом расстояние между центроидами групп *VERSICOL* и *VIRGINIC* намного меньше, чем расстояния между центроидами групп *SETOSA* и *VERSICOL*, *SETOSA* и *VIRGINIC*. Это может говорить о том, что сорта *VERSICOL* и *VIRGINIC* наиболее схожи между собой, а сорт *Setosa* значительно отличается от них обоих.

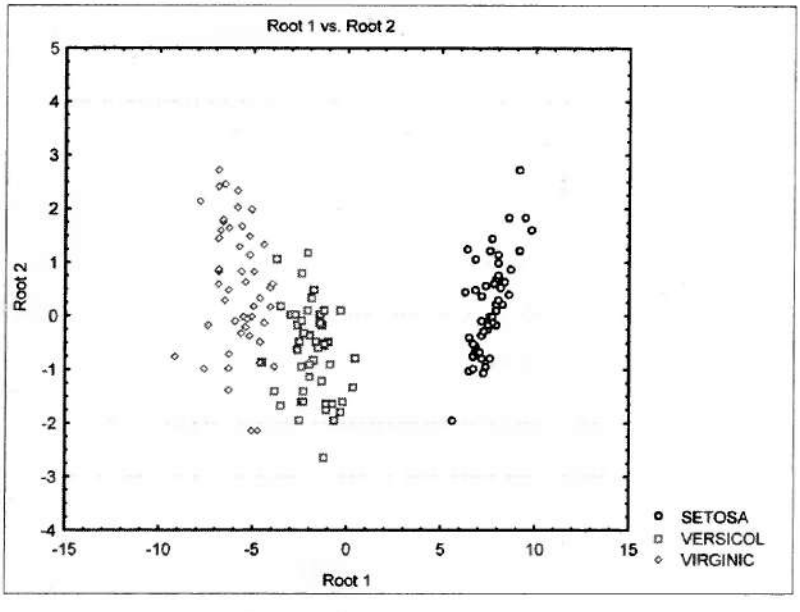


Рис. 12.10

Из диаграммы видно, что цветы сорта *Setosa* представлены на диаграмме точками далеко справа, т.е. этим цветам соответствуют бульшие значения корня 1 (*Root 1*). Поэтому дискриминантная функция 1 главным образом дискриминирует цветы между этим сортом и двумя другими. Дискриминантная функция 2, по-видимому, дает основную дискриминацию между цветками сорта *VERSICOL* (которые преимущественно имеют бульшие отрицательные значения корня 2) и двумя другими сортами. Однако дискриминация здесь не настолько отчетлива, как это имеет место для дискриминантной функции 1. Из диаграммы видно, что дискриминация по дискриминантной функции 1 более сильная, чем по дискриминантной функции 2.

Дискриминантная функция 1 имеет отрицательные коэффициенты (см. рис. 12.6) для ширины (*PETALWID*) и длины (*PETALLEN*) пестиков и положительные коэффициенты для ширины (*SEPALWID*) и длины (*SEPALLEN*) чашелистиков. Таким образом, чем шире и длиннее пестики, короче и уже чашелистики, тем менее вероятно, что это цветки сорта *SETOSA*.

Вернитесь в окно результатов **Discriminant Function Analysis Results** и активизируйте вкладку **Classification** (классификация). Откроется окно результатов классификации (рис. 12.11).

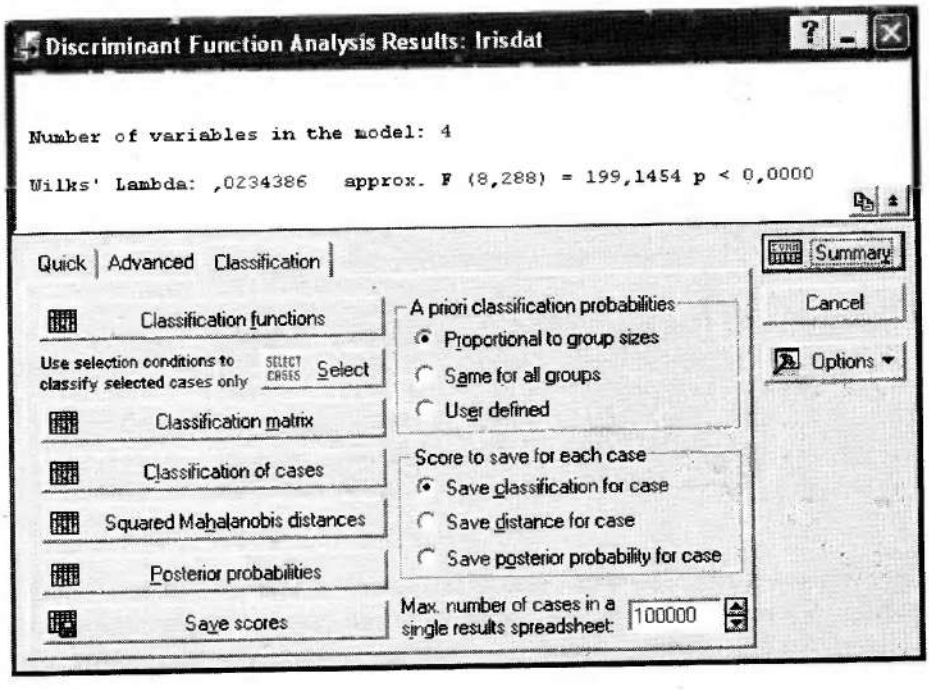


Рис. 12.11

В рамке **A priori classification probabilites** (априорные вероятности классификации) приведены различные опции задания априорных вероятностей того, что наблюдение при классификации попадет в одну из групп: *Proportional to group sizes* (пропорциональные размерам групп); *Same for all groups* (одинаковые для всех групп); *User defined* (заданные пользователем). Априорные вероятности могут существенно влиять на точность классификации. Если есть предварительные сведения (оценки) о возможном количественном соотношении наблюдений в группах, то желательно выбрать опцию *User defined*, если таких сведений нет и число наблюдений в группах примерно одинаково, то надо выбрать *Same for all groups*, в противном случае — *Proportional to group sizes*.

В рамке **Score to save for each case** (сохранить для каждого наблюдения) приведены опции, при выборе которых можно сохранить тот или иной результат классификации: *Save classification for case* (сохранить результаты классификации); *Save distance for case* (сохранить расстояния); *Save posterior probability for case* (сохранить апостериорные вероятности).

В строке *Max. number of cases in a single results spreadsheet* (максимальное число наблюдений в таблице результатов) можно указать максимальное число наблюдений в таблице результатов. Если наблюдений больше указанного числа, то результаты будут выведены несколькими таблицами. Нажмите кнопку **Classification functions** (функции классификации). Функции классификации — это линейные функции, которые вычисляются для каждой группы и могут быть использованы для классификации наблюдений. Наблюдение приписывают той

группе, для которой классификационная функция имеет наибольшее значение. В таблице, изображенной на рис. 12.12, приведены коэффициенты и свободные члены при переменных линейных функций. Например, классификационное уравнение для группы *SETOSA* имеет вид

$$SETOSA = 23{,}54SEPALEN + 23{,}58SEPALWID - 16{,}43PETALLEN - \\ - 17{,}39PETALWID - 86{,}30.$$

	Classification Functions; grouping:		
Variable	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
SEPALLEN	23,5442	15,6982	12,446
SEPALWID	23,5879	7,0725	3,685
PETALLEN	-16,4306	5,2115	12,767
PETALWID	-17,3984	6,4342	21,079
Constant	-86,3085	-72,8526	-104,368

Рис. 12.12

Нажмите кнопку **Classification matrix** (матрица классификации). Матрица (рис.12.13) содержит информацию о количестве и проценте корректно классифицированных наблюдений в каждой группе. Строки матрицы — исходные классы, столбцы — предсказанные классы.

	Classification Matrix (Irisdat) Rows: Observed classifications Columns: Predicted classifications			
Group	Percent Correct	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
SETOSA	100,0000	50	0	0
VERSICOL	96,0000	0	48	2
VIRGINIC	98,0000	0	1	49
Total	98,0000	50	49	51

Рис. 12.13

Нажмите кнопку **Classification of cases** (классификация наблюдений). Программа выведет таблицу классификации для каждого наблюдения (рис. 12.14). Классификации упорядочены по первому, второму и третьему выбору. Столбец 1 содержит первый классификационный выбор, т.е. группу, для которой соответствующее наблюдение имеет наивысшую апостериорную вероятность и наибольшее значение классификационной функции. Наблюдения, которые не удалось правильно классифицировать, помечены *.

Щелкните кнопкой **Squared Mahalanobis distances** (квадраты расстояний Махаланобиса). Будет выведена таблица квадратов расстояний Махалонобиса каждого наблюдения от центроида группы (рис. 12.14).

Case	Observed Classif.	Squared Mahalanobis Distances from Group Centroic Incorrect classifications are marked with * SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
1	SETOSA	0,2419	90,6602	181,5587
2	VIRGINIC	208,5713	27,3188	1,8944
3	VERSICOL	105,2663	2,2329	13,0720
4	VIRGINIC	207,9180	31,7492	4,4506
* 5	VIRGINIC	133,0668	5,2529	7,2359
6	SETOSA	1,3337	84,0118	170,0569
7	VIRGINIC	173,1838	26,5620	11,0484

Рис. 12.14

Эти расстояния аналогичны квадратам евклидовых расстояний, но учитывают корреляции между переменными в модели. Наблюдение приписывают группе, к которой оно ближе всего. Наблюдения, которые не удалось правильно классифицировать, также помечены *.

Нажмите кнопку **Posterior probabilities** (апостериорные вероятности). В открывшейся таблице (рис. 12.15) каждому наблюдению будет поставлена в соответствие вероятность принадлежности к группам. Эта вероятность определяется посредством расстояний Махалонобиса и априорных вероятностей. Чем дальше наблюдение расположено от центра группы, тем менее вероятно, что оно принадлежит этой группе. Наблюдение приписывают той группе, для которой имеется наибольшая апостериорная вероятность классификации. Априорные вероятности могут быть заданы пользователем, могут быть равны для всех групп, могут быть пропорциональны размерам групп.

На данном этапе удобно рассмотреть возможность классификации новых наблюдений. Для этого добавьте в таблицу исходных данных новое наблюдение, например, так, как это показано на рис. 12.16.

Case	Observed Classif.	Posterior Probabilities (Irisdat) Incorrect classifications are marked with * SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
1	SETOSA	1,000000	0,000000	0,000000
2	VIRGINIC	0,000000	0,000003	0,999997
3	VERSICOL	0,000000	0,995590	0,004410
4	VIRGINIC	0,000000	0,000001	0,999999
* 5	VIRGINIC	0,000000	0,729388	0,270612
6	SETOSA	1,000000	0,000000	0,000000
7	VIRGINIC	0,000000	0,000428	0,999572

Рис. 12.15

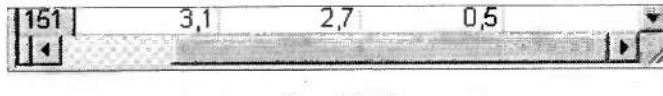


Рис. 12.16

Для того чтобы понять, к какому классу относится этот объект, нажмите кнопку **Posterior probabilities**. Появится таблица с апостериорными вероятностями, к которой будет добавлена строка (рис. 12.17).

Рис. 12.17

Максимальное значение вероятности соответствует группе *SETOSA*. Значит, новое наблюдение (цветок) с вероятностью 0,999874 можно отнести к типу *SETOSA*. Нажмите кнопку **Squared Mahalanobis distances.** Появится строка таблицы с расстояниями от нового случая до центроидов групп (рис.12.18).

Рис. 12.18

Расстояние от нового цветка до центроидов групп минимально для группы *SETOSA*. Это дополнительное подтверждение того, что новый цветок ириса относится к сорту *SETOSA*.

Если выделить вкладку **Descriptives** (описания) и нажать на кнопку **Review Descriptives Statistices**, то программа предоставит пользователю широкие возможности анализа описательных статистик исходных данных, которые можно использовать для проверки выполнения предположений применения параметрической дискриминации. Так, на вкладке **Quick** можно посмотреть **Pooled within-groups covariances& correlations** (объединенные внутригрупповые ковариации и корреляции) и **Means& number of cases** (средние и число наблюдений)

На вкладке **Within** (внутри) можно посмотреть:

- **Within-groups standard deviations** (внутригрупповые стандартные отклонения);
- **Categorized histogram by group** (категоризованные гистограммы по группам);
- **Box plot of means by group** (диаграммы размаха);
- **Categorized scatterplot by group** (категоризованные диаграммы рассеяния);
- **Categorized normal probability plot by group** (категоризованный нормальный график по группам).

Вкладка **All cases** (все наблюдения) предоставит следующие данные:

- **Total covariances& correlations** (полные ковариации и корреляции);
- **Plot of total correlations** (график полной корреляции);
- **Box plot of means** (диаграмма размаха средних).

12.3. Общие модели дискриминантного анализа

Если не выполняются условия применимости модуля **Discriminant Analysis** (*DA*) – независимые переменные (предикторы) должны быть измерены как минимум в интервальной шкале, их распределение должно соответствовать нормальному закону, необходимо воспользоваться модулем **General Discriminant Analysis Models** (общие модели дискриминантного анализа, *GDA*). Модуль имеет такое название, потому что в нем для анализа дискриминантных функций используется общая линейная модель (*GLM*). В этом модуле анализ дискриминантных функций рассматривается как общая многомерная линейная модель, в которой категориальная зависимая переменная (отклик) представляется векторами с кодами, обозначающими различные группы для каждого наблюдения. Например, в файле **Irisdat**, который был рассмотрен нами при изучении модуля *DA*, категориальная переменная *IRISTIPE* – сорт цветков ириса принимает значения *SETOSA, VERSICOL, VIRGINIC* и в модуле *DA* эти значения кодируются целыми числами 1, 2, 3. А в модуле *GDA* эти значения будут закодированы векторами (1, 0, 0), (0, 1, 0), (0, 0, 1) (рис. 12.19).

Таким образом, категориальная зависимая переменная преобразуется в три различные независимые переменные, каждая из которых содержит значение 1 (если соответствующее наблюдение принадлежит отдельной группе) и 0 (в противном случае).

Модуль *GDA* обладает рядом существенных преимуществ перед модулем *DA* [6]:

Сорт	Коды		
SETOSA	1	0	0
VERSICOL	0	1	0
VIRGINIC	0	0	1

Рис. 12.19

1. Не устанавливается никаких ограничений на тип используемого предиктора (категориальный или непрерывный) или на тип определяемой модели. В модуле *GDA* категориальные предикторы называются также факторами.
2. Предусмотрены опции для пошагового выбора предикторов и выбора наилучшего подмножества предикторов, на основе статистик *F-включить* и *p-включить* (эти статистики связаны с многомерной статистикой лямбда Уилкса).
3. В случае наличия в файле данных кросс-проверочной выборки выбор наилучшего подмножества предикторов можно провести на основе долей ошибочной классификации для кросс-проверочной выборки. Другими словами, после оценки дискриминантных функций для данного множества предикторов

вычисляются оценки ошибочной классификации для кросс-проверочной выборки и выбирается модель (подмножество предикторов), которая соответствует наименьшей доле ошибочной классификации для кросс-проверочной выборки. Этот способ выбора модели позволяет получать в итоге высокую точность прогноза, избегая при этом переобучения.

4. Другой уникальной особенностью *GDA* является наличие опции для построения и анализа профилей предсказанных значений отклика (зависимой переменной) и показателя желательности. Программа вычисляет предсказанные значения отклика для каждого значения зависимой переменной, а полученные значения объединяются в один показатель желательности. Чтобы наглядно показать «поведение» предсказанных откликов и показателя желательности, для различных диапазонов значений предикторов строятся различные графики — профили. В модуле *GDA* строятся профили апостериорных вероятностей предсказания. Профили позволяют анализировать, насколько различные уровни предикторов влияют на классификацию наблюдений, что в конечном итоге дает возможность определить комбинации значений предикторов, которые максимизируют правдоподобие того, что соответствующее наблюдение принадлежит тому или иному классу.

5. В модуле *GDA* предоставлены функциональные возможности, которые делают этот модуль продвинутым средством для классификации и добычи данных. Однако в большинстве книг применение анализа дискриминантных функций ограничено в области простого или пошагового анализа с непрерывными предикторами. В *GDA* предусмотрена возможность включения категориальных «*ANOVA-подобных*» эффектов (для категориальных предикторов, см. Гл. 1) в сложные *ANOVA-подобные* модели для предикторов [6].

6. *GDA* предоставляет опции для проведения поиска наилучшего подмножества предикторов даже для сложных *ANOVA-подобных* эффектов. Доступны несколько критериев для выбора эффектов предиктора, включаемых в модель. Один из критериев заключается в том, что эффекты предиктора включаются в модель, если они имеют наименьшую долю ошибочной классификации (на основе апостериорных вероятностей классификации). Можно вычислить эти доли ошибочной классификации для анализируемой выборки (т.е. для наблюдений, которые включены в вычисления оценок параметров) и для кросс-проверочной выборки (т.е. для наблюдений, которые исключаются из вычисления оценок параметров). Этот метод используется в задачах добычи данных, когда требуется построить модели с хорошей точностью прогноза для классификации новых наблюдений. Он также используется для предотвращения переобучения моделей. Подобные ситуации возникают, когда рассматривается задача классификации наблюдений из анализируемой выборки, а размер выборки очень большой. В этом случае часто включаются эффекты предиктора, которые незначительно улучшают подгонку модели к анализируемой выборке, но не имеют соответствующей достоверности предсказания в кросс-проверочной выборке.

Для того чтобы рассмотреть работу модуля, создадим соответствующий ему по структуре файл данных, преобразовав файл **Irisdat** из **Example**. Изменение файла будет заключаться во введении новых категориальных предикторов и переменной идентификатора выборки для определения подмножества наблюдений для кросс-проверки.

Категориальные предикторы *SEPALLEN1* и *PETALLEN1* создадим, перекодировав значения переменных *SEPALLEN* и *PETALLEN* при помощи команды **Recode Variables** из меню **Data** на панели инструментов.

Для переменной *PETALLEN* правила перекодирования следующие:

- *если* $1 <= v5$ and $v5 < 3$, то присвоить $v5$ значение 0;
- *если* $3 <= v5$ and $v5 < 5$, то присвоить $v5$ значение 1;
- *если* $5 <= v5$ and $v5 < 7$, то присвоить $v5$ значение 2.

Для переменной *SEPALLEN* правила перекодирования следующие:

- *если* $4{,}3 <= v2$ and $v2 < 5{,}5$, то присвоить $v2$ значение 0;
- *если* $5{,}5 <= v2$ and $v2 < 6{,}7$, то присвоить $v2$ значение 1;
- *если* $6{,}7 <= v2$ and $v2 < 8$, то присвоить $v2$ значение 2.

Коды 0, 1, 2 соответствуют следующим текстовым значениям: *малая*, *средняя*, *большая*.

Переменную идентификатор выборки *IDENTI* создадим более простым способом — наблюдениям с номерами от 0 до 100 присвоим значение 0, а переменным с номерами от 101 до 150 — значение 1. После преобразования файл данных сохраните в **Example** под именем **Irisdat1.** Фрагмент файла приведен на рис. 12.20.

Для запуска модуля в верхнем меню **Statistics** надо щелкнуть по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выбрать команду **General Discriminant Analysis Models**. Откроется стартовая панель модуля (рис. 12.21).

Данный диалог содержит два списка: **Type of analysis** (вид анализа) и **Specification method** (задание анализа).

	Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris							
	1 SEPAL	2 SEPAL1	3 SEPALW	4 PETAL	5 PETAL1	6 PETALW	7 IRISTYPE	8 IDENT
1	5,0	малая	3,3	1,4	малая	0,2	SETOSA	0
2	6,4	средняя	2,8	5,6	большая	2,2	VIRGINIC	0
3	6,5	средняя	2,8	4,6	средняя	1,5	VERSICOI	0
4	6,7	большая	3,1	5,6	большая	2,4	VIRGINIC	0
5	6,3	средняя	2,8	5,1	большая	1,5	VIRGINIC	0

Рис. 12.20

Список **Type of analysis** состоит из двух элементов, представляющих собой различные модели дисперсионного анализа:

- **Traditional discriminant analysis** (классический дискриминантный анализ);
- **General discriminant analysis** (общий дискриминантный анализ).

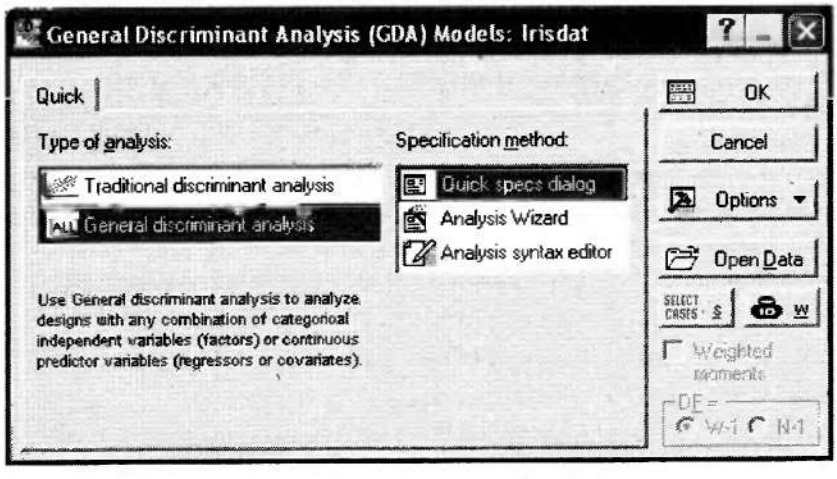


Рис. 12.21

Список **Specification method** состоит из трех элементов:

- **Quick specs dialog** (диалог быстрых спецификаций);
- **Analysis Wizard** (мастер анализа);
- **Analysis syntax editor** (редактор кода).

Диалог **Traditional discriminant analysis** определяет анализ с простыми непрерывными предикторами, аналогичный описанному в §12.2. Однако кроме этого *GDA* предоставляет опции для выбора наилучшего подмножества предикторов. В диалоге **Quick specs dialog** можно определить один или несколько непрерывных предикторов и одну категориальную зависимую переменную.

Диалог **General discriminant analysis** определяет *ANOVA*-подобные планы для непрерывных или категориальных предикторов и эффектов. Можно определить пошаговый выбор эффектов предиктора (с включением или с исключением), а также выбор наилучшего подмножества эффектов. При выполнении пошагового выбора эффектов предиктора или при выборе наилучшего подмножества эффектов, эффекты со многими степенями свободы не преобразуются в предикторы с одной степенью свободы.

Диалог **Quick specs dialog** позволяет выбрать категориальную зависимую переменную и категориальные и/или непрерывные предикторы (в зависимости от выбранной опции в списке **Type of analysis**), а после этого построить стандартную модель.

Диалог **Analysis Wizard** по шагам с помощью последовательных диалогов определяет анализ со сложным планом. После выбора этого диалога в списке **Type of analysis** автоматически выбирается **General discriminant analysis**. Нельзя использовать **Analysis Wizard** для другого типа анализа.

Диалог **Analysis syntax editor** используется для создания и редактирования программы, задающей анализ. После выбора **Analysis syntax editor** в списке **Type of analysis** автоматически выбирается **General discriminant analysis**. Нельзя использовать **Analysis syntax editor** для другого типа анализа.

Выберите в списке **Type of analysis** диалог **General discriminant analysis**, а в списке **Specification method** – диалог **Quick specs dialog** и нажмите **OK**. Откроется окно **General Discriminant Analysis**. На вкладке **Quick** щелкните по кнопке **Variables** и в открывшемся окне **Select dependent variable, categorical, and continuous predictors** выделите зависимую переменную, категориальные и непрерывные предикторы в соответствии с рис. 12.22. Щелкните по **OK**, программа вернется в диалоговое окно **General discriminant analysis** (рис. 12.23).

Нажмите кнопки **Dep.var.codes**, **Factor codes** и выберите коды зависимой переменной *IRISTYPE*, факторов (категориальных предикторов) *SEPALLEN1* и *PETALLEN1*. Если предполагается выбрать все коды, то можно, не пользуясь этими кнопками, сразу нажать на **OK**, программа по умолчанию, автоматически осуществит выбор всех кодов.

Если нажать кнопку **Effects in design** (эффекты плана), будет вызван диалог **GDA Effects in Design** для определения эффектов межгруппового плана, т.е. для непрерывных и категориальных предикторов (рис. 12.24). О планах и эффектах достаточно подробно говорилось в гл. 8, применительно к данному модулю под эффектами надо понимать воздействие (влияние) предикторов на зависимую категориальную переменную (значения переменной). Поэтому эффекты предикторов, включенных в план, имеют те же названия, что и предикторы.

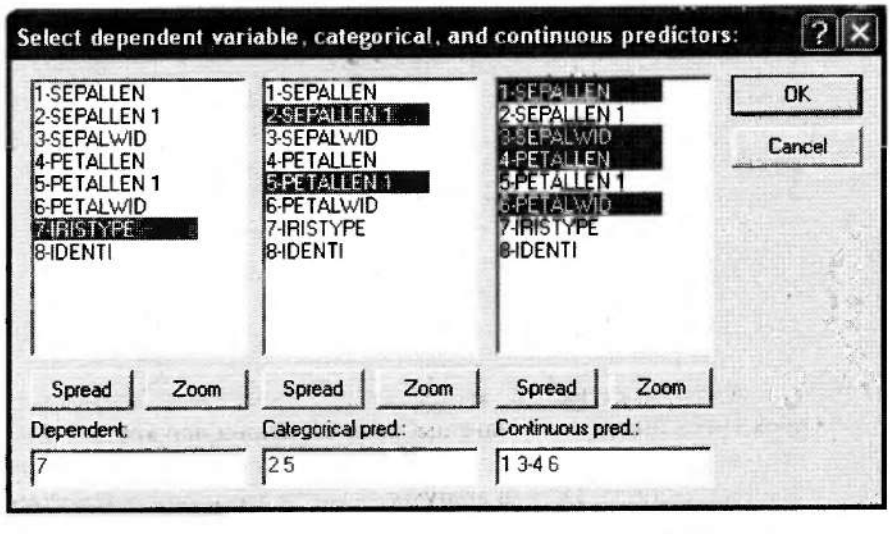


Рис. 12.22

Опция *Use default effects for the design* (использовать стандартные эффекты плана) предполагает, что в качестве эффектов плана будут выбраны все категориальные и непрерывные предикторы. В этом случае будет создан соответствующий стандартный план для текущих категориальных и непрерывных предикторов (например, полный факторный план для категориальных предикторов и главные эффекты только для непрерывных предикторов).

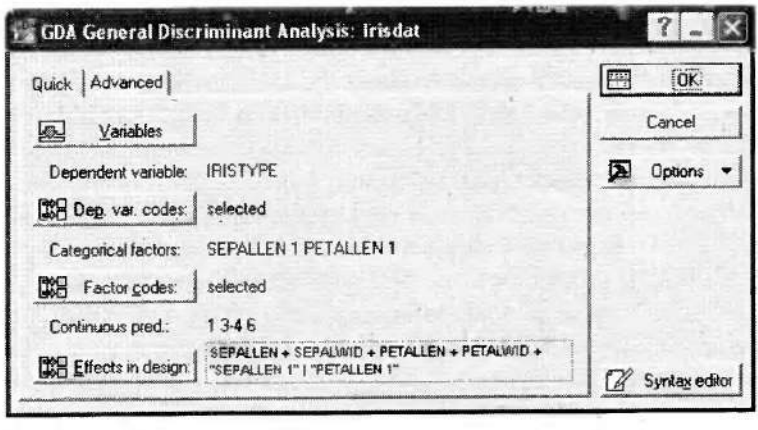


Рис. 12.23

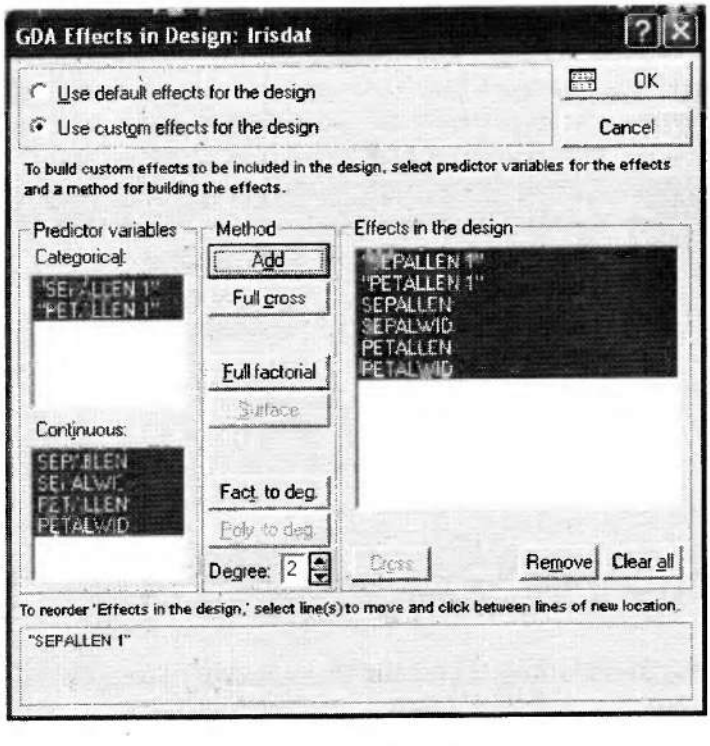


Рис. 12.24

По умолчанию эти эффекты прописаны в поле справа от кнопки **Effects in design** в диалоге **General Discriminant Analysis**. При выборе этой опции будет запрещен доступ к опциям в рамках **Predictor variables** (предикторные переменные), **Method** (метод), **Effects in design** (эффекты в плане).

При выборе опции *Use custom effects for the design* (использовать пользовательские эффекты плана) пользователь может сам определить специальные эффекты

для выбранных переменных, которые необходимо включить в модель, так как становятся активными рамки **Predictor variables**, **Method**, **Effects in design**. Категориальные или непрерывные предикторы, которые не будут выделены в рамке **Predictor variables**, но были выбраны с помощью кнопки **Variables** (рис. 12.23), будут автоматически удалены из анализа. Для создания пользовательских эффектов надо в рамке **Predictor variables** выбрать имена категориальных и непрерывных предикторов и нажать соответствующую кнопку в рамке **Method**: **Add** (добавить), **Full cross** (полное взаимодействие), **Full factorial** (полный факторный), **Fact. to deg.** (факторный) для выбора плана. Чтобы изменить порядок эффектов в рамке **Effects in design**, выберите (высветите) переменную, которую надо переместить, затем установите курсор мыши на соответствующее положение в списке, при этом курсор изменит свой вид на две горизонтальные линии, перечеркнутые вертикальной стрелкой, и нажмите левую кнопку мыши.

При помощи кнопок **Remove**, **Clear all** можно удалить названия эффектов в рамке **Effects in design**.

Выберите опцию *Use default effects for the design* и нажмите **OK**. Программа вернется в диалог **General Discriminant Analysis**. Перейдите на вкладку **Advanced** (рис. 12.25).

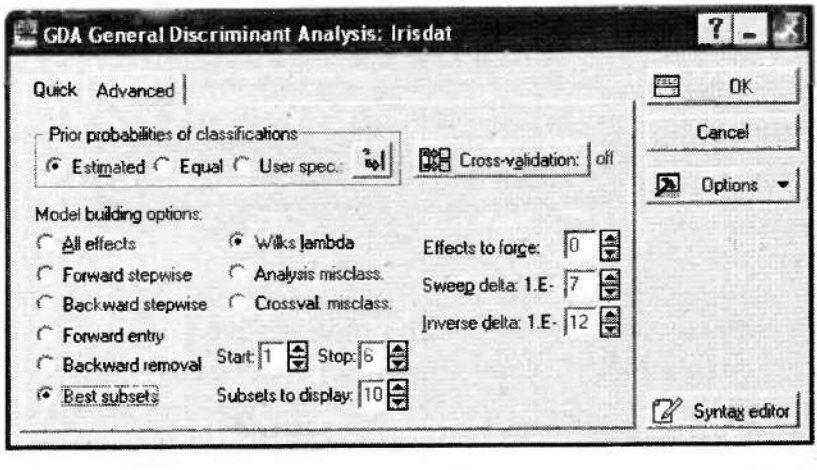


Рис. 12.25

В рамке **Prior probabilities of classifications** задаются способы вычисления априорных вероятностей, которые используются для классификации наблюдений на основе текущего множества предикторов. Опция *Estimated* (вычисленные) означает, что априорные вероятности пропорциональны размерам классов, которые определяются зависимой (группирующей) переменной. Например, если имеется три класса наблюдений и в каждом классе соответственно 20, 30 и 50 наблюдений, то априорные вероятности будут равны соответственно 0,2, 0,3, 0,5. Опция *Equal* (равные) присваивает априорным вероятностям одинаковые значения. Например, если три группы наблюдений, то априорные вероятности будут равны 1/3 для каждой группы. Опция *User spec.* предполагает задание априорных вероятностей

пользователями. Эта опция не будет доступна, если на вкладке **Quick** определили коды для зависимой переменной.

Кнопка **Cross-validation** вызывает одноименный диалог, в котором можно задать категориальную переменную-идентификатор и кодовое значение, определяющее наблюдения, которые необходимо использовать в вычислениях при подгонке модели. Все другие наблюдения с соответствующими значениями всех предикторов и зависимая переменная будут автоматически сопоставлены проверочной выборке.

Нажмите кнопку **Cross-validation**, в появившемся диалоге нажмите кнопку **Sample Identifier Variable**, выберите переменную-идентификатор *IDENTI*, в поле **Code for analysis sample** задайте код, определяющий наблюдения для анализа. В соответствии с рис. 12.26 все наблюдения (ирисы) с кодом идентификатора 0 предназначены для дискриминантного анализа, а наблюдения с кодом 1 будут кросспроверочными. В рамке **Status** (состояние) выберите опцию **ON** (включить) и нажмите **OK**, программа вернется к диалогу **General Discriminant Analysis**.

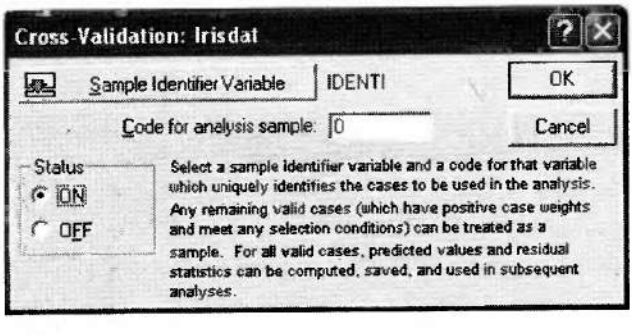


Рис. 12.26

В поле **Model building options** (опции построения модели) можно задать способы построения моделей для дискриминантного анализа:

- *All effects* (все эффекты). Все эффекты вводятся в текущий план;
- *Forward stepwise* (пошаговый с включением). В процессе реализации метода переменные и эффекты добавляются в модель на основе текущих значений параметров *p* или *F*;
- *Backward stepwise* (пошаговый с исключением). Начальная модель будет состоять из всех предикторов и эффектов, которые затем в процессе реализации метода будут последовательно удаляться на основе текущих значений параметров *p* или *F*;
- *Forward entry* (только с включением). Переменные или эффекты будут только включаться в модель;
- *Backward removal* (только с исключением). Переменные или эффекты будут только исключаться из модели;
- *Best subsets* (лучшие подмножества). Среди всех допустимых подмножеств предикторов, заданных в текущем плане анализа, выбирается

лучшее подмножество. Если в модели очень много эффектов, общее число всевозможных подмножеств может быть очень большим. В этом случае с помощью опций *Start* (старт) и *Stop* (стоп) приходится анализировать много выборок с большим объемом. Поэтому поиск наилучшего подмножества необходимо проводить очень осторожно.

В поле **Effects to force** (дословно — эффекты принуждать) указывается количество эффектов, принудительно включенных в каждую построенную программой модель. Под моделями в данном случае понимаются различные подмножества эффектов, которые строит программа, с дальнейшим выбором наилучшего. Если значение *Effects to force* = *k* больше 0, то первые *k* эффектов в плане (например, из указанных в поле **Effects in design** на рис. 12.23) будут принудительно добавлены во все рассматриваемые модели.

Sweep delta (дельта выметания), *Inverse delta* (дельта обращения) — параметры математических моделей, которые используются соответственно для построения матрицы выметания и проверки сингулярности при обращении матрицы.

Опции *Wilks lambda, Analysis misclass, Crossval. misclass* определяют критерии выбора наилучшего подмножества — по значению параметра лямбда Уилкса, по долям ошибочной классификации всех наблюдений и по кросспроверочной выборке наблюдений.

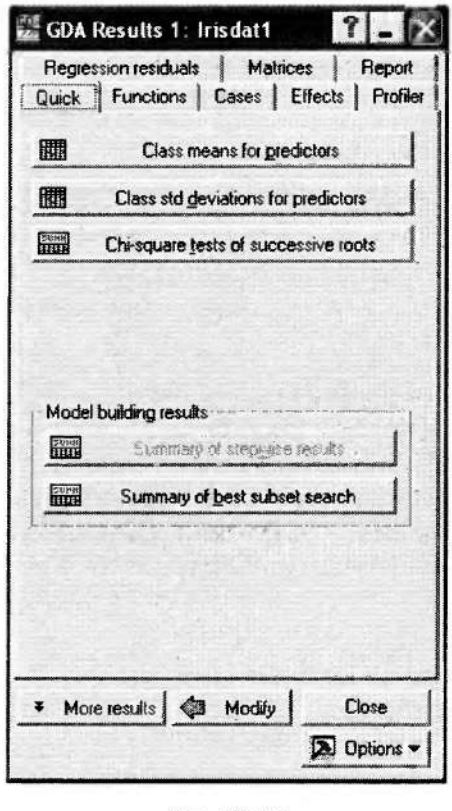


Рис. 12.27

Произведите установки опций в соответствии с диалогом **General Discriminant Analysis** (рис. 12.25) и нажмите **OK**. Откроется окно **GDA Results** (рис. 12.27), нажмите кнопку **Summary of best subset search**, появится таблица с результатами (рис. 12.28). Чем ближе значение показателя к 1, тем более значим предиктор в модели. Из 1-й строки видно, что наиболее значимыми предикторами в 1-й модели являются предикторы *PETALWID* — ширина пестика, *SEPALWID* — ширина чашелистика. Вторая модель уже не содержит предиктора *SEPALLEN* — длина чашелистика. Если в диалоге **General Discriminant Analysis** в поле **Effects to force** установить значение 3, то во все модели принудительно будут включены эффекты *SEPALLEN, SEPALWID, PETALLEN* (рис. 12.29).

	Summary of best subsets; variable: IRISTYPE (Irisdat1) Wilks lambda and tolerances for the effects in each submodel					
Subset No.	Wilks Lambda	No. of Effects	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	0,006772	5	0,203506	0,546611	0,294431	0,62596
2	0,006871	4		0,579327	0,516273	0,63094
3	0,008698	4	0,205126	0,627055	0,317737	
4	0,008880	3		0,659347	0,561083	
5	0,009142	4	0,215686		0,294749	0,71808
6	0,009896	3			0,552830	0,71809
7	0,010649	4	0,356840	0,547202		0,67550

Рис. 12.28

	Summary of best subsets; variable: IRISTYPE (Irisdat1) Wilks lambda and tolerances for the effects in each submodel							
Subset No.	Wilks Lamb	No. of Effects	SEPAL	SEPALW	PETAL	PETALW	SEPAL1	SEPAL1
1	0,007	5	0,204	0,547	0,294	0,626		
2	0,009	4	0,205	0,627	0,318			
3	0,018	5	0,194	0,591	0,363	0,633	0,315	0,699
4	0,023	4	0,197	0,672	0,414		0,316	0,712
5	0,026	4	0,372	0,623	0,388	0,647		
6	0,228	6	0,189	0,546	0,289	0,624	0,100	0,054

Рис. 12.29

В диалоге **GDA Results** (рис. 12.27) перейдите на вкладку **Functions**. Рассмотрим функциональное назначение кнопок этого окна (рис. 12.30):

- **Class means for predictors** (средние в классах для предикторов). Для всех классов зависимой переменной *SETOSA, VERSICOL, VIRGINIC* программа отобразит средние значения предикторов;
- **Class std. deviations for predictors** (стандартные отклонения в классах для предикторов). Для всех классов зависимой переменной программа отобразит стандартные отклонения предикторов;
- **Chi-Square tests of successive roots** (χ^2 критерий для последовательности корней).

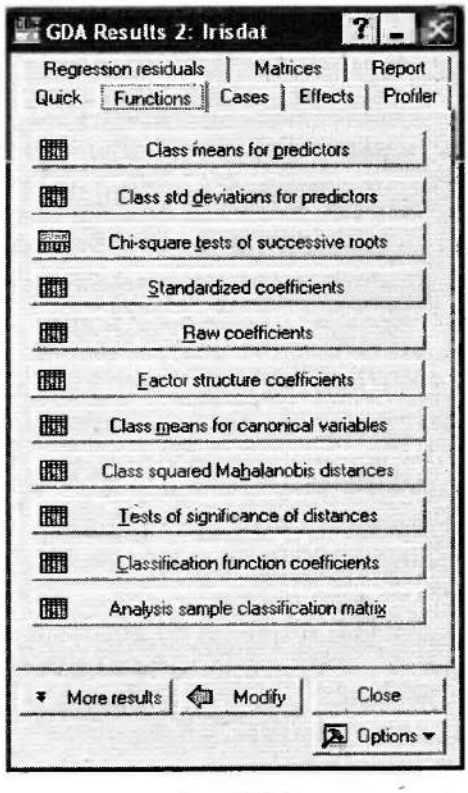


Рис. 12.30

Программа выведет таблицу (рис. 12.31), в которой в строке 1 содержится критерий значимости для всех комбинаций корней, в строке 2 — критерий значимости корней, оставшихся после удаления первого корня, и т.д. По таблице можно определить необходимое количество корней.

	Chi-Square Tests with Successive Roots Removed (Irisdat1) Sigma-restricted parameterization					
Removed	Eigen-value	Canonicl R	Wilk's Lambda	Chi-Sqr.	df	p-level
0	55,31365	0,991081	0,006772	467,0305	16,00000	0,000000
1	1,62227	0,786544	0,381348	90,1379	7,00000	0,000000

Рис. 12.31

В столбцах таблицы указано число удаленных корней, собственные значения, канонические корреляции, значения лямбды Уилкса, значение критерия χ^2 и соответствующий ему уровень значимости p. Из данных таблицы следует, что оба канонических корня статистически значимы;

- **Standardized coefficients** (стандартизованные коэффициенты). Программа отобразит таблицу со стандартизованными коэффициентами дискриминантных функций;
- **Raw coefficients** (исходные коэффициенты). Появится таблица (рис. 12.32), содержащая коэффициенты исходных дискриминантных функций и соответствующие им собственные значения. В дискриминантных функциях коэффициенты при *SEPALLEN* имеют наименьшее значение, а коэффициенты при *PETALWID* — наибольшее, что не противоречит значимости этих предикторов из 1-й модели на рис. 12.28;
- **Factor structure coefficients** (коэффициенты факторной структуры). Программа отобразит объединенные межклассовые коэффициенты корреляции для предикторов с соответствующими дискриминантными функциями. В терминах факторного анализа эти корреляции называются факторными нагрузками для предикторов в дискриминантных функциях;
- **Class means for canonical variables** (средние в классах для канонических функций). Будет выведена таблица со средними для дискриминантных функций для каждого класса. По этим средним можно определить классы, которые наилучшим образом разделяются дискриминантными функциями;
- **Class squared Mahalanobis distances** (квадраты расстояний Махоланобиса). Программа отобразит таблицу с квадратами расстояний Махоланобиса между центрами классов. Расстояние Махаланобиса похоже на стандартное евклидово расстояние, за исключением того, что учитываются корреляции между переменными;
- **Tests of significance of distances** (критерии значимости расстояний).

	Исходные коэффициенты канонической дискриминантной функции (Irisdat1)			
Эффект	Уровень Эффекта	Столбец	Функция 1	Функция 2
Св.член		1	-0,05562	4,88977
SEPALLEN		2	-0,69302	-0,26184
SEPALWID		3	-1,75378	-1,20530
PETALLEN		4	2,17983	0,95982
PETALWID		5	2,36108	-1,44386
SEPALLEN1	малая	6	-0,45542	-0,93199
SEPALLEN1	средняя	7	-0,40002	-0,58572
PETALLEN1	малая	8	-1,03742	2,84163
PETALLEN1	средняя	9	0,00000	0,00000
SEPALLEN1*PETALLEN1	1	10	-3,38495	-3,12933
SEPALLEN1*PETALLEN1	2	11	1,14353	0,63518
SEPALLEN1*PETALLEN1	3	12	0,00000	0,00000
SEPALLEN1*PETALLEN1	4	13	0,00000	0,00000
Собс.знач.			56,87439	1,51655

Рис. 12.32

Будет отображена таблица с критериями значимости расстояний;

- **Classification function coefficients** (коэффициенты функций классификации);
- **Analysis sample classification matrix** (матрица классификации). Программа выведет таблицу с указанием количества ошибок классификации.

Перейдите на вкладку **Cases** (наблюдения), откроется окно (рис. 12.33), опции которого позволят вычислить различные статистики классификации для каждого наблюдения.

В рамке **Sample** при помощи опций *Analysis, Cross-val., Both, Pred.* (предсказанные) можно выбрать различные выборки наблюдений:

- анализируемая, состоящая из наблюдений, на основе которых вычисляются оценки параметров текущей модели;
- кросс-проверки, состоящая из наблюдений, не включенных в вычисление оценок параметров;
- полная, состоящая из всех наблюдений;
- предсказанная, состоящая из наблюдений с корректными данными предикторов, но с учетом пропущенных данных для зависимой переменной.

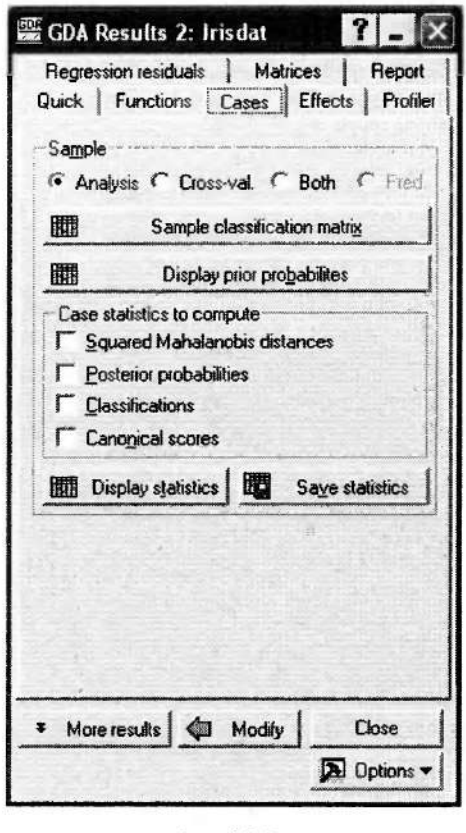


Рис. 12.33

Class	Classification Matrix (Irisdat1) Rows: Observed classifications Columns: Predicted classifications Percent Correct	SETOSA p=,3400	VERSICOL p=,3000	VIRGINIC p=,3600
SETOSA	100,0000	50,00000	0,00000	0,00000
VERSICOL	96,0000	0,00000	48,00000	2,00000
VIRGINIC	94,0000	0,00000	3,00000	47,00000
Total	96,6667	50,00000	51,00000	49,00000

Рис. 12.34

Выберите, например, опцию *Both* и нажмите кнопку **Sample classification matrix**, появится таблица с указанием количества ошибок классификации (рис. 12.34) для всей выборки наблюдений.

Если нажать кнопку **Display prior probabilities** (отобразить априорные вероятности), появится таблица с априорными вероятностями для каждого класса.

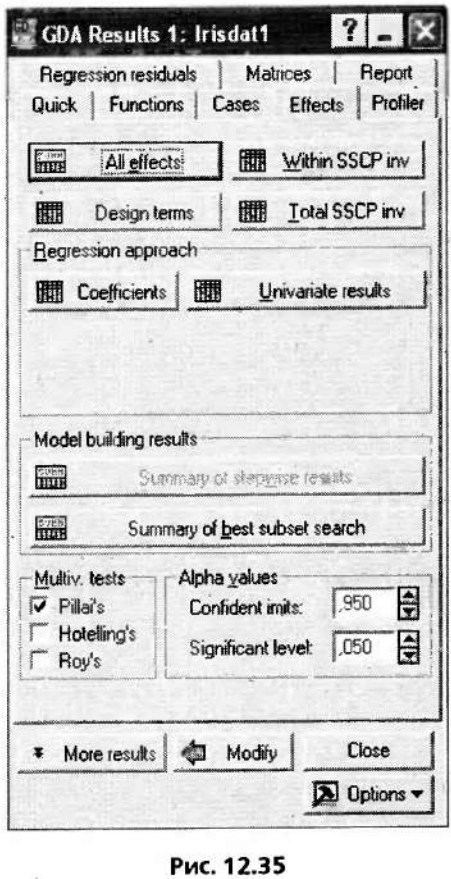


Рис. 12.35

В рамке **Case statistics to compute** (вычислить статистики наблюдений) можно указать те статистики наблюдений, которые надо вычислить. Просмотреть и сохранить эти статистики можно при помощи кнопок **Display statistics** и **Save statistics**. По вычисленным статистикам — квадратам расстояний Махаланобиса, апостериорным вероятностям и исходной классификации наблюдений — можно судить о результатах программной классификации каждого наблюдения.

В диалоге **GDA Results** (рис. 12.27) перейдите на вкладку **Effects** (эффекты, рис. 12.35). На этой вкладке можно просмотреть значимость эффектов предикторов. Нажмите кнопку **All effects**. Программа построит таблицу результатов многомерного дискриминантного анализа со статистиками, которые можно выбрать в поле **Multiv. tests**. Выберите, например, критерий *Pillar's* и нажмите кнопку **All effects**. Как видно из таблицы (рис. 12.36), по критериям Уилкса и Пиллая значимыми будут эффекты предикторов *PETALWID, PETALLEN, SEPALWID* (уровни значимости p критерия Фишера (F) меньше, чем 0,05). Это означает, что данные предикторы будут иметь наибольшее воздействие на значения зависимой переменной *IRISTYPE*, а значит, и на определение принадлежности наблюдений к классам — цветков ириса к тому или иному сорту.

	Multivariate Tests of Significance (Irisdat1) Sigma-restricted parameterization Effective hypothesis decomposition					
Effect	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0,682861	20,69919	2	90	0,000000
	Pillai's	0,317139	20,69919	2	90	0,000000
SEPALLEN	Wilks	0,985564	0,65915	2	90	0,519769
	Pillai's	0,014436	0,65915	2	90	0,519769
SEPALWID	Wilks	0,740769	15,74768	2	90	0,000001
	Pillai's	0,259231	15,74768	2	90	0,000001
PETALLEN	Wilks	0,635926	25,76295	2	90	0,000000
	Pillai's	0,364074	25,76295	2	90	0,000000
PETALWID	Wilks	0,778572	12,79810	2	90	0,000013
	Pillai's	0,221428	12,79810	2	90	0,000013
SEPALLEN1	Wilks	1,000000		0		
	Pillai's	0,000000				
PETALLEN	Wilks	1,000000		0		
	Pillai's	0,000000				
SEPALLEN1*PETALLEI	Wilks	0,259063	21,70584	8	180	0,000000
	Pillai's	0,834660	16,29440	8	182	0,000000

Рис. 12.36

Если нажать на кнопку **Design terms** (проект плана), программа построит таблицу (рис. 12.37), содержащую метки (*Label*) для каждого столбца (*Column*) в матрице плана. В столбце *Variable* (переменные) указаны названия переменных. В столбцах *Level* (уровень) *Variable of* и *versus Level* (другой уровень) приведены

в соответствие с сигма-ограниченной параметризацией два уровня факторов, указанных в каждом столбце матрицы плана.

Нажмите кнопку **Coefficients**, программа построит таблицу, содержащую оценки параметров (*Par*), стандартизованные оценки параметров, их стандартные ошибки, уровни значимости (p) и соответствующие статистики. На рис. 12.38 приведен фрагмент преобразованной (сокращенной) таблицы. Уровни значимости показывают, по значениям каких переменных существенно различаются наблюдения из разных классов, оценки параметров — силу и характер вклада переменных в определение принадлежности наблюдений к тому или иному классу.

Label	Column Labels (Irisdat1) Labels for the columns of the design matrix X						
	Column	Variable	Level of Variable	versus Level	Variable	Level of Variable	versus Level
Intercept	1						
SEPALLEN	2	SEPALLEI					
SEPALWID	3	SEPALWII					
PETALLEN	4	PETALLEI					
PETALWID	5	PETALWII					
SEPALLEN1	6	SEPALLEN	малая	большая			
SEPALLEN1	7	SEPALLEN	средняя	большая			
PETALLEN	8	PETALLEN	малая	большая			
PETALLEN	9	PETALLEN	средняя	большая			
SEPALLEN1*PETALLEI	10	SEPALLEN	малая	большая	PETALLEN	малая	большая
SEPALLEN1*PETALLEI	11	SEPALLEN	малая	большая	PETALLEN	средняя	большая
SEPALLEN1*PETALLEI	12	SEPALLEN	средняя	большая	PETALLEN	малая	большая
SEPALLEN1*PETALLEI	13	SEPALLEN	средняя	большая	PETALLEN	средняя	большая

Рис. 12.37

Например, переменные *SEPALWID, PETALLEN* являются определяющими признаками для *SETOSA* ($p < 0{,}05$). При этом переменные *SEPALWID, PETALLEN* для *SETOSA* имеют преимущественно большие (параметр имеет положительный знак) и меньшие (параметр имеет отрицательный знак) значения по сравнению с другими сортами.

Переменная *PETALWID* является определяющим признаком для *VERSICOL*, причем переменная *PETALWID* для *VERSICOL* имеет преимущественно меньшее значение по сравнению с другими сортами. Переменная *PETALWID* — определяющий признак для *VIRGINIC*, но переменная *PETALWID* для *VIRGINIC* имеет преимущественно большее значение по сравнению с другими сортами. А *SEPALLEN* не является определяющим признаком ни для какого сорта, т.е. этот признак неинформативен в плане дискриминации цветков ириса по сортам. Для того чтобы убедиться в правильности сделанных выводов, можете обратиться к файлу исходных данных.

Effect	Parameter Estimates (Irisdat1) Sigma-restricted parameterization								
	Level of Effect	Column	Commen (B/Z/P)	SET Par	p	VER Par	p	VIR Par	p
Intercept		1		0,33	0,02	1,45	0,00	-0,78	0,05
SEPALLEN		2		0,04	0,26	-0,10	0,39	0,06	0,53
SEPALWID		3		0,14	0,00	-0,17	0,09	0,02	0,79
PETALLEN		4		-0,17	0,00	0,15	0,13	0,03	0,73
PETALWID		5		-0,08	0,05	-0,33	0,02	0,42	0,00
SEPALLEN1	малая	6	Poolec						
SEPALLEN1	средняя	7	Poolec						
PETALLEN	малая	8	Poolec						
PETALLEN	средняя	9	Poolec						
SEPALLEN1*PETALLE	1	10		0,27	0,00	-0,53	0,00	0,26	0,00
SEPALLEN1*PETALLE	2	11		-0,02	0,67	-0,16	0,23	0,17	0,12
SEPALLEN1*PETALLE	3	12		0,01	0,84	0,04	0,77	-0,05	0,68
SEPALLEN1*PETALLE	4	13		-0,09	0,00	0,30	0,00	-0,21	0,00

Рис. 12.38

Для сложных или неполных планов в таблице может также отображаться столбец *Comment* (комментарий). Ячейки в этом столбце либо могут быть пустыми, либо содержать соответствующие метки *B* (смещенный), *Z* (нулевой) или *P* (объединенный).

Если нажать на кнопки **Within SSCP inv** (обратная внутригрупповая *SSCP*), **Total SSCP inv** (общая внутригрупповая *SSCP*), появятся таблицы, содержащие информацию о связях между анализируемыми переменными.

Нажмите кнопку **Univariate results** (одномерные результаты), программа построит таблицу (рис. 12.39) стандартного дисперсионного анализа, по которой также можно судить о роли переменных в определении принадлежности наблюдений к тому или иному классу.

Перейдите на вкладку **Profiler** (профили), откроется диалог (рис. 12.40), в котором можно просмотреть значения и графики профилей и функций желательности для всех классов.

При помощи кнопки **Classes** (классы) можно выбрать группы зависимых переменных при построении профиля.

Нажмите кнопку **Options** (опции) и в открывшемся окне (рис. 12.41) в поле **Options for profile plots** (параметры профилей отклика) дополнительно установите флажок *Show Spreadsheets with plotted values* (показывать таблицы с графиками), остальные установки: *Label conf/pred limits for pred. values* (отмечать доверительные/предсказанные пределы для предсказанных значений); *Show text labels for factor settings* (показывать текстовые метки факторов) оставьте без изменения (по умолчанию).

GENERAL Effect	Univariate Results for Each DV (Irisdat1) Best subset solution Effective hypothesis decomposition						
	Degr. of Freedom	SET SS	SET p	VER SS	VER p	VIR SS	VIR p
Intercept	1	0,04	0,02	0,74	0,00	0,21	0,05
SEPALLEN	1	0,01	0,26	0,06	0,39	0,02	0,53
SEPALWID	1	0,18	0,00	0,24	0,09	0,00	0,79
PETALLEN	1	0,27	0,00	0,19	0,13	0,01	0,73
PETALWID	1	0,03	0,05	0,41	0,02	0,65	0,00
SEPALLEN1	0						
PETALLEN1	0						
SEPALLEN1*PETALLEN1	4	1,39	0,00	9,05	0,00	3,61	0,00
Error	91	0,60		7,19		5,10	
Total	99	22,44		21,00		23,04	

Рис. 12.39

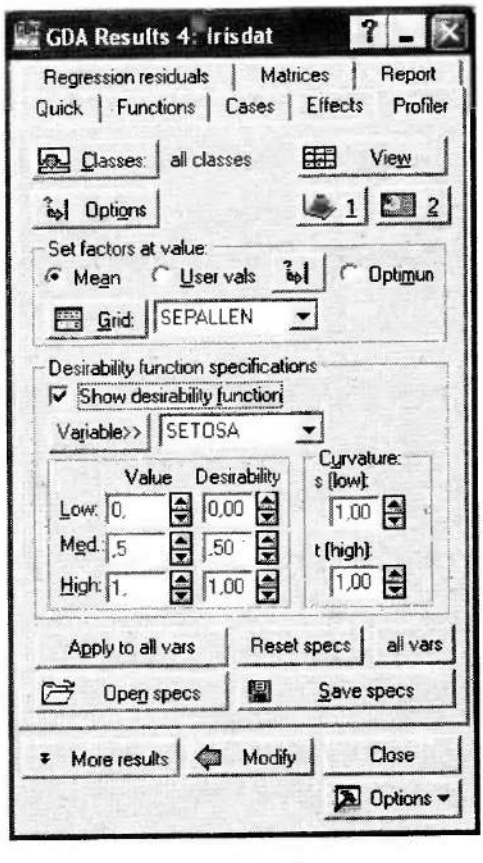


Рис. 12.40

В нижней рамке окна **Fit options for desirability surface/contours** можно выбрать метод подгонки поверхности к значениям желательности: *Quadratic surface* (квадратичная поверхность), *Least squares* (наименьших квадратов), *Negative exponential* (обратная экспоненциальная), *Spline* (сплайны). Опции *Show the grid points* (показать сетку), *Area contours* (показать контуры) предназначены для визуализации сетки и контура. Выберите опцию *Quadratic surface*, установите флажок на *Area contours* и нажмите **OK**. Программа вернется в окно диалога на рис. 12.40.

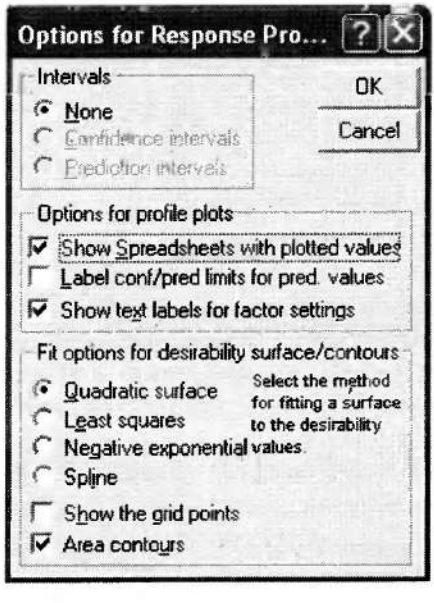


Рис. 12.41

Профили — это графики зависимостей апостериорных вероятностей принадлежности наблюдений к классам зависимой (группирующей) переменной от дискретных значений предиктора (уровней фактора) при фиксированных на определенном уровне текущих значений других предикторов.

Для построения профилей независимые переменные — предикторы — разбиваются на диапазоны и для вычисления апостериорных вероятностей рассматриваются дискретные значения предикторов (границы диапазонов), которые называются уровнями факторов. Число уровней фактора равно количеству диапазонов плюс 1.

Для того чтобы задать количество диапазонов, нажмите кнопку **Grid** (сетка), откроется окно **Specification for Fac...** (рис. 12.42). При помощи кнопок **Prev.**, **Next** изменяется имя фактора в поле **Factor**. В полях **Minimum**, **Maximum** программа прописывает наименьшее и наибольшее значения выбранного фактора. В рамке **No. of steps:** (число шагов) задается количество диапазонов. Если нажать на кнопку **Apply to all**, выбранное значение **No. of steps** будет одинаковым для всех факторов. Для категориальных предикторов число шагов будет соответствовать числу значений (уровней) предиктора, причем настройки в этом диалоге

можно задать только так, чтобы уровни совпадали с действительными (наблюдаемыми) значениями категориальных предикторов.

Рис. 12.42

В этом диалоге уровни категориальных предикторов всегда обозначаются соответствующими целыми числами (например, 1, 2, 3), независимо от используемых в программе кодов. Точки сетки выполняют две функции. Во-первых, они определяют точки факторов на графике предсказанного профиля. Во-вторых, они также определяют уровни факторов для зависимых переменных при выборе опции *Optimum*. При использовании опции *Grid* не рекомендуется выбирать большое количество точек сетки, чтобы сократить длительные вычисления.

Выберите значение *No. of steps*, равное 4. Нажмите кнопку **Apply to all**, далее нажмите **OK**, программа вернется в окно на рис. 12.40.

В рамке **Set factors at value** можно задать уровень текущих (фиксированных) значений предикторов. Опция *Mean* означает, что в качестве уровней приняты средние значения предикторов, опция *User vals* предполагает, что уровни могут быть заданы пользователем, опция *Optimum* означает, что текущий уровень каждого предиктора равен значению, оптимизирующему желательность отклика. При выборе опции *Optimum* активной станет рамка **Desirability function specifications** (спецификации функции желательности), появится галочка перед строкой *Show desirability function*. Выберите опцию *Mean* и установите флажок на опцию *Show desirability function*.

В рамке **Desirability function specifications** задаются параметры функции желательности для класса, название которого указано в поле **Variable**. Эти параметры определяют значения функции желательности (в интервале от 0,0 до 1,0) в соответствии с предсказанными значениями отклика (апостериорных вероятностей в классе). Эти спецификации задаются в полях **Value**, **Desirability**. Функция желательности характеризует соответствие предсказанных значений функции отклика (апостериорных вероятностей), вычисленных по уровню выбранного предиктора и текущим уровням других предикторов, наблюдаемым значениям функции отклика — определенному классу зависимой переменной, т.е. характеризует, насколько типичной является комбинация уровней предикто-

ров для предсказанного класса, которому соответствует максимальное значение апостериорных вероятностей. Например, если уровню предиктора *PETALLEN*, равному 1,963465, соответствуют при средних фиксированных значениях других предикторов апостериорные вероятности 1, 0, 0 для сортов *SETOSA, VERSICOL, VIRGINIC* и желательность равна 1, то это значит, что такая комбинация значений предикторов является типичной для цветков сорта *SETOSA*.

В поле **Value** указываются уровни значений апостериорных вероятностей, а в поле **Desirability** – соответствующие им значения желательности, т.е. таким образом выражается степень предпочтения одних значений вероятностей (отклика) по сравнению с другими. Значения в поле **Value** должны быть заданы в порядке возрастания (или равны). Например, если в поле **Value** выбраны значения 0,0; 0,5; 1,0, а в поле **Desirability** – значения 0,0; 0,1; 0,0, то это означает, что среднему значению вероятности 0,5 отдано большее предпочтение, чем малым и большим значениям. Такая функция желательности имеет 3 точки перегиба – нижнее значения класса, ниже которого отклик является нежелательным; верхнее значение класса, выше которого отклик также является нежелательным; среднее значение класса, в котором отклик становится более желательным по мере приближения к целевому значению. По умолчанию значения нижней *(Low)*, средней *(Med)* и верхней *(High)* точек используют простой тип функции желательности «чем больше, тем лучше» только с двумя точками перегиба. Нижняя точка устанавливается равной минимальному наблюдаемому значению в классе, верхняя точка – равной максимальному наблюдаемому значению в классе, а средняя точка – равной среднему значению интервала. Пользователь может определить любой тип функции желательности, задав соответствующие значения в полях *Low, Med, High*.

В рамке **Curvature** (кривизна) в полях *s(low)* и *t(high)* вводятся значения кривизны желательности откликов между точками перегиба. По умолчанию значения параметров *s* и *t* равны значению 1,0, представляющему линейный «характер» изменения желательности между точками перегиба *Med* и *Low*, а также между точками перегиба *Med* и *High*.

Кнопка **1** предназначена для построения графика поверхности функции желательности вместе с точками сетки для классов наблюдений. График позволяет проследить динамику изменения желательности отклика от комбинаций точек различных пар независимых переменных при условии, что остальные переменные зафиксированы на текущих уровнях. Как ранее было отмечено, при помощи кнопки **Options** можно выбрать различные методы подгонки поверхности к значениям желательности. Кнопка **2** позволит построить карту линий уровня функции желательности.

Кнопка **Apply to all vars** (применить для всех переменных) распространит настройки параметров желательности на все независимые (предикторы) и зависимые (классы) переменные.

Кнопка **Reset specs** предназначена для возврата измененных параметров в стандартное (по умолчанию) состояние, кнопка **all vars** – для возврата измененных параметров в стандартное (по умолчанию) состояние для всех переменных.

Произведите установки, соответствующие диалогу на рис. 12.40, и нажмите кнопку **View** (вид). Программа построит ряд таблиц и профили для апостериорных вероятностей и желательности. В таблице на рис. 12.43 будут отображены выбранные значения в полях **Value**, **Desirability** для всех классов (сортов ириса).

	Desirability function parameters (Irisdat1) Desirability function settings for each dependent variable				
Variable	Low Value	Desirbty Value	Medium Value	Desirbty Value	High Value
SETOSA	0,00	0,000000	0,500000	0,500000	1,000000
VERSICOL	0,00	1,000000	0,500000	0,500000	1,000000
VIRGINIC	0,00	1,000000	0,500000	0,500000	1,000000

Рис. 12.43

В таблице на рис. 12.44 в столбце 2 указаны уровни фактора, название которого приведено в столбце 1 при текущих значениях (средних) других факторов, т.е. заданы параметры некоторого виртуального цветка ириса. В столбцах 3–5 указаны вычисленные для этих параметров цветка значения апостериорных вероятностей принадлежности этого цветка ириса к классу, указанному в названии столбца. В последнем столбце приведены значения желательности, т.е. степень соответствия предсказанного сорта виртуального цветка (ему соответствует максимальное значение апостериорных вероятностей) действительному сорту с этим названием при такой комбинации значений параметров цветка. Красным цветом (на мониторе) выделены значения уровней факторов, при которых достигается наибольшее значение функции желательности. Например, для цветка ириса (строка 10) со значением *SEPALWID*, равным 3,943, и текущими средними значениями *SEPALLEN, PETALLEN, PETALWID*, равными соответственно 5,821, 3,765, 1,199, вычисленное значение апостериорной вероятности для сорта *SETOSA* равно 0,999.

Это означает, что для данного цветка ириса программа предсказывает сорт *SETOSA*. Так как желательность принимает значение, близкое к 1 (0,999), то это означает, что цветок ириса с такими параметрами может реально принадлежать сорту *SETOSA*. А вот для цветка ириса (строка 5) со значением *SEPALLEN*, равным 7,538, и текущими средними значениями *SEPALWID, PETALLEN, PETALWID*, равными соответственно 3,076, 3,765, 1,199, вычисленное значение апостериорной вероятности для сорта *VERSICOL* равно 0,917, но желательность при этом равна 0,189. Это означает, что цветок с такими параметрами не может реально принадлежать прогнозируемому сорту *VERSICOL*. Напомним, что вычисленные оптимальные значения желательности зависят от задаваемых пользователем значений в полях **Value**, **Desirability**.

Factor	Factor levels and predicted responses (Irisdat1) Predicted responses at each level of each factor holding all other factors constant at their current setting				
	Factor Level	Prob. SETOSA	Prob. VERSICOL	Prob. VIRGINIC	Desirbty Value
SEPALLEN	4,103521	0,000000	1,000000	0,000000	0,000003
SEPALLEN	4,962261	0,000000	1,000000	0,000000	0,000018
SEPALLEN	5,821000	0,000003	0,999997	0,000000	0,000214
SEPALLEN	6,679739	0,000511	0,999488	0,000001	0,006395
SEPALLEN	7,538479	0,082645	0,917354	0,000001	0,189734
SEPALWID	2,208578	0,000000	0,999988	0,000012	0,000000
SEPALWID	2,642289	0,000000	0,999998	0,000002	0,000003
SEPALWID	3,076000	0,000003	0,999997	0,000000	0,000214
SEPALWID	3,509711	0,359826	0,640174	0,000000	0,505897
SEPALWID	3,943422	0,999991	0,000009	0,000000	0,999994
PETALLEN	0,161928	1,000000	0,000000	0,000000	1,000000
PETALLEN	1,963464	1,000000	0,000000	0,000000	1,000000
PETALLEN	3,765000	0,000003	0,999997	0,000000	0,000214
PETALLEN	5,566536	0,000000	0,985114	0,014886	0,000000
PETALLEN	7,368071	0,000000	0,002106	0,997894	0,000000
PETALWID	-0,34038	1,000000	0,000000	0,000000	1,000000
PETALWID	0,429309	0,999914	0,000086	0,000000	0,999942
PETALWID	1,199000	0,000003	0,999997	0,000000	0,000214
PETALWID	1,968690	0,000000	0,994464	0,005536	0,000002
PETALWID	2,738381	0,000000	0,015310	0,984690	0,000000
SEPALLEN1	малая	0,000000	1,000000	0,000000	0,000001
SEPALLEN1	средняя	0,000003	0,999997	0,000000	0,000214
SEPALLEN1	большая	0,000003	0,999997	0,000000	0,000214
PETALLEN1	малая	0,000000	0,999997	0,000003	0,000000
PETALLEN1	средняя	0,000003	0,999997	0,000000	0,000214
PETALLEN1	большая	0,000000	0,999719	0,000281	0,000000

Рис. 12.44

В таблице на рис. 12. 45 представлены средние уровни предикторов, значения желательности, а также апостериорные вероятности для всех значений зависимой переменной *IRISTYPE*. Из данной таблицы видно, что по значениям апостериорных вероятностей цветок ириса с такими значениями параметров следует отнести к сорту *VERSICOL*, но так как желательность принимает малое значение, то такие значения параметров ириса нетипичны для данного сорта.

	Current factor settings and predicted responses (Irisdat1) Predicted responses at the current level of each factor in the model									
	Level o SEPL	Level of SEPWL	Level o PETL	Level o PETW	Level of SEPL1	Level of PETAL1	Prob. SET	Prob. VERS	Prob. VIRG	Desir Value
Value	5,821	3,076	3,765	1,199	средняя	средняя	0,000	1,000	0,000	0,000

Рис. 12.45

На графиках (профилях), представленных на рис. 12.46, изображены значения апостериорных вероятностей для каждого значения зависимой переменной (сорта ириса), соответствующие различным уровням факторов (предикторов) как непрерывных, так и категориальных. Так, например, шесть графиков в верхней строке — столбцу 2 (*Prob. SETOSA*), графики во 2-й строке — столбцу 3 (*Prob. VERSICOL*), графики 3-й строки — столбцу 4 (*Prob. VIRDGINIC*) таблицы, изображенной на рис. 12.44. Последний седьмой график каждой строки соответствует значениям желательности для каждого сорта ириса. Красная вертикальная линия соответствует фиксированному уровню предикторов (в данном случае — среднему).

Перейдите на вкладку **Regressions residuals** (рис. 12.47). На вкладке можно просмотреть результы анализа остатков между предсказанными значениями зависимой переменной *IRISTYPE* и наблюдаемыми значениями. В качестве наблюдаемых значений использованы не имена классов цветков ириса, а их коды. Предсказанные значения вычислены посредством линейных уравнений регрессии. В рамке **Sample** (выборка) указывается, по какой выборке программа отобразит результаты анализа — анализируемой, кросс-проверочной или по полной выборке.

Кнопка **Classes** предназначена для выбора класса, для которого программа предоставит результаты анализа, по умолчанию будут выбраны все классы.

В рамке **Show Predicted and residuals values** (показать предсказанные значения и остатки) можно произвести соответствующие установки для вывода таблицы анализа.

Если установить флажок на *Exteded*, то в таблице будут отображены дополнительные статистики остатков, если установить флажок на *Spreadsheet for each dependent variable*, то таблицы будут введены отдельно для каждого значения зависимой переменной — сорта цветков ириса. В поле **Sort** можно выбрать принцип сортировки данных в таблице. Установите флажок на *Spreadsheet for each dependent variable* и нажмите кнопку **Predicted and residuals**, появятся таблицы с исходными, предсказанными значениями и остатками для каждого сорта (рис. 12.48).

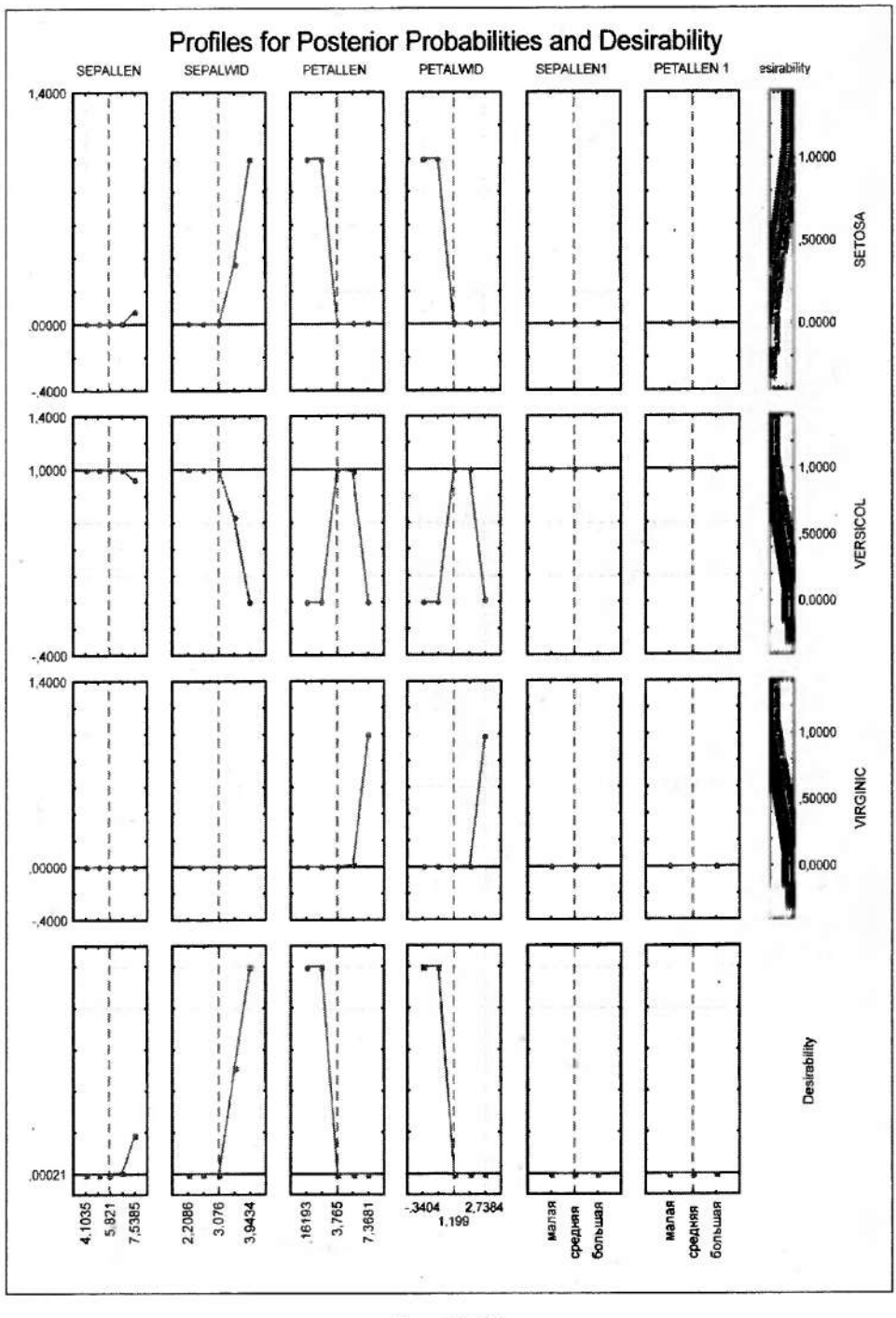

Profiles for Posterior Probabilities and Desirability
SEPALLEN
SEPALWID
PETALLEN
PETALWID
SEPALLEN1
PETALLEN 1
esirability
1,4000
,00000
-,4000
1,0000
,50000
0,0000
SETOSA
VERSICOL
VIRGINIC
Desirability
,00021
4,1035
5,821
7,5385
2,2086
3,076
3,9434
,16193
3,765
7,3681
-,3404
2,7384
1,199
малая
средняя
большая

Рис. 12.46

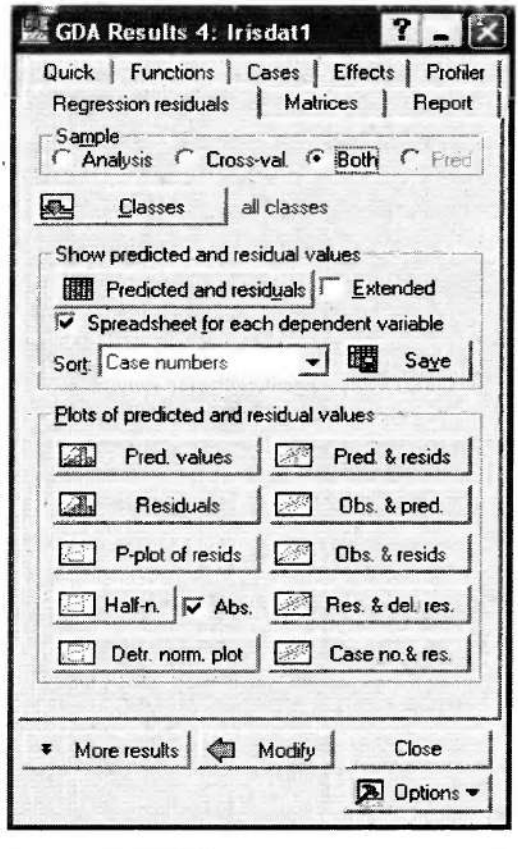


Рис. 12.47

Case name	Observed, Predicted, and Residual Values (I Sigma-restricted parameterization (Analysis and validation samples)			
	SETOSA Observed	SETOSA Predictd	SETOSA Resids	Sample Code
	1,000000	1,009334	-0,009334	0
	0,000000	-0,099486	0,099486	0
	0,000000	-0,031967	0,031967	0
	0,000000	0,033183	-0,033183	0
	0,000000	0,042188	-0,042188	0
	1,000000	0,999645	0,000355	0
	0,000000	0,136871	-0,136871	0

Рис. 12.48

При помощи кнопок в рамке **Plots of predicted and residuals values** можно построить различные графики результатов анализа. Нажмите, например, кнопку **Residuals**. Появятся гистограммы остатков всех сортов ириса. По виду гистограмм можно судить об адекватности уравнений регрессии, а значит, и о качестве дискриминации. Из графика видно (рис. 12.49), что закон распределения остатков напоминает нормальное распределение, поэтому можно говорить о достаточном уровне предсказания для сорта *SETOSA*.

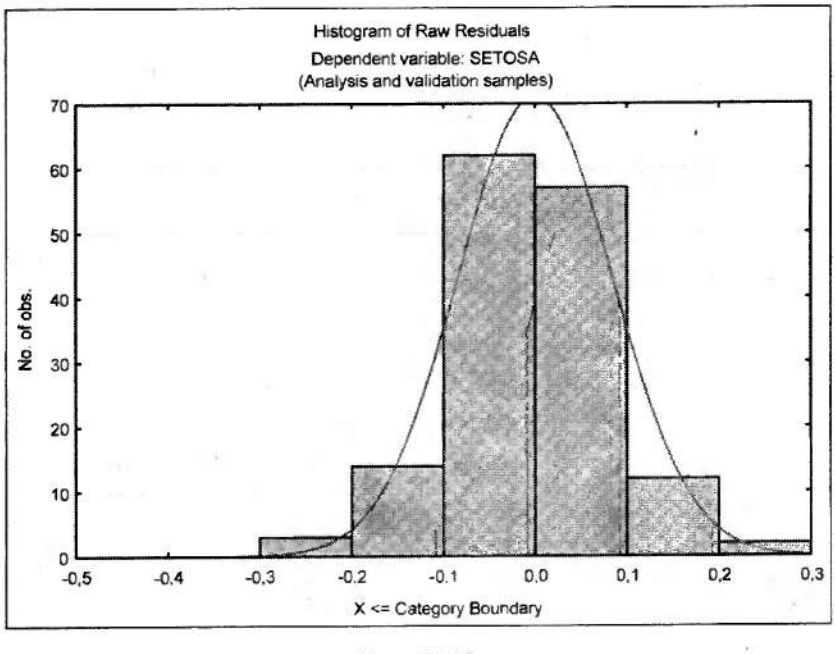


Рис. 12.49

Как и в модуле *DA*, в модуле *GDA* можно провести классификацию наблюдения, для которого класс неизвестен. Закройте все окна модуля *GDA* и вернитесь к файлу **IRISDAT1**. Чтобы не добавлять новое наблюдение, в строке, соответствующей наблюдению 150 в столбце *IRISTYPE* удалите название сорта цветка *VERSICOL* (рис.12.50). В окне **General Discriminant Analysis** выберите зависимую переменную, категориальные и непрерывные предикторы, как это было сделано ранее, но в окне **Cross-validation** либо не указывайте категориальную переменную-идентификатор *IDENTI*, либо в рамке **Status** выберите опцию **OFF** (отключить). Запустите процедуру анализа, и в окне **GDA Results** перейдите на вкладку **Cases**. В рамке **Sample** выберите опцию **Pred** (предсказать), в рамке **Case statistics to compute** — установите галочку против опции *Posterior probabilities* (апостериорные вероятности) и нажмите кнопку *Display statistics* (показать статистики). Появится таблица (рис.12.51) со значениями апостериорных вероятностей для наблюдения под номером 150.

	Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris							
	1 SEPAL	2 SEPAL1	3 SEPALW	4 PETAL	5 PETAL1	6 PETALW	7 IRISTYPE	8 IDENT
145	5,2	малая	3,5	1,5	малая	0,2	SETOSA	1
146	5,8	средняя	2,8	5,1	большая	2,4	VIRGINIC	1
147	6,7	большая	3,0	5,0	большая	1,7	VERSICOL	1
148	6,3	средняя	3,3	6,0	большая	2,5	VIRGINIC	1
149	5,3	малая	3,7	1,5	малая	0,2	SETOSA	1
150	5,0	малая	2,3	3,3	средняя	1,0	VERSICOL	1

Рис. 12.50

	Posterior Probabilities of Classifications (Irisd Prediction sample N = 1			
Case number	Observed Classif.	SETOSA prob.	VERSICOL prob.	VIRGINIC prob.
150		0,000000	1,000000	0,000000

Рис. 12.51

Из таблицы видно, что наибольшая вероятность, равная 1 соответствует сорту *VERSICOL*, т.е. программа, верно, определила сорт цветка.

Глава 13

Классификационный анализ без обучения

13.1. Кластерный анализ

Главное назначение кластерного анализа (от англ. *cluster* — гроздь, скопление) — разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. Методы кластерного анализа можно применять даже тогда, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству.

Техника кластеризации применяется в самых различных областях. Например, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям (таксономия — это распределение животных или растительных организмов по группам). В области психиатрии правильная диагностика кластеров симптомов таких болезней, как паранойя, шизофрения и т.д., является решающей для успешной терапии. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии предметов быта. Широкое применение нашел кластерный анализ в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать «горы» информации на пригодные для дальнейшей обработки группы, кластерный анализ оказывается весьма полезным и эффективным [6].

Отличием кластерного анализа от других методов классификации является отсутствие обучающей выборки (классификация без обучения) [9]. Большое достоинство кластерного анализа в том, что он дает возможность производить разбиение объектов не по одному параметру, а по ряду признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы. Это особенно важно, например, при прогнозировании конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m (m — целое) кластеров $Q_1, Q_2, ..., Q_m$ так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения. При этом объекты, принадлежащие одному и тому же кластеру, должны быть сходными, а объекты, принадлежащие разным кластерам, — разнородными.

Например, пусть G включает n стран, любая из которых характеризуется k параметрами: ВНП на душу населения (F_1); числом m автомашин на 1 тыс. чел. (F_2); душевым потреблением электроэнергии (F_3); душевым потреблением стали (F_4) и т.д. Тогда X_1 (вектор измерений) представляет собой набор указанных характеристик для первой страны, X_2 — для второй, X_3 — для третьей, ..., X_n — для n-й ($X_n = x_{n1}, x_{n2}, ..., x_{nk}$). Задача заключается в том, чтобы разбить страны на кластеры (группы однородности), например, по уровню развития.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонений:

$$W = \Sigma_j (X_j - \bar{X})^2 = \Sigma_j X_j^2 - (1/n)(\Sigma_j X_j)^2,$$

где X_j — вектор измерений j-го объекта; $\bar{X}$ — средний вектор измерений;

$$j = 1, ..., n.$$

Сходство между объектами G_i, G_j определим через понятие расстояния между векторами измерений X_i, X_j, так как интуитивно понятно, что чем меньше расстояние между объектами, тем они более схожи.

Напомним, что неотрицательная функция $d(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $d(X_i, X_j) \geq 0$ для всех X_i и X_j;

б) $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

в) $d(X_i, X_j) = d(X_j, X_i)$;

г) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$.

Значение $d(X_i, X_j)$ для X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между G_i и G_j соответственно выбранным характеристикам F_1, F_2, F_3, ..., F_k.

Перечислим часто употребляемые функции расстояний [16].

Evclidean distances (евклидова метрика) является наиболее популярной и вычисляется по формуле

$$d_e(X_i, X_j) = (\Sigma_k (x_{ik} - x_{jk})^2)^{1/2}.$$

Евклидово расстояние — геометрическое расстояние в многомерном пространстве. Обычно вычисляется по исходным, а не по стандартизованным данным. Поэтому на расстояния могут сильно влиять различия между осями, по координатам которых вычисляются эти расстояния. Так, если изменить единицу измерения одной из осей (например, сантиметры перевести в миллиметры), то изменится и исчисляемое расстояние, а значит, изменится и результат кластерного анализа.

Square Evclidean distances (квадрат евклидова расстояния) используют, если необходимо придать большие веса более отдаленным друг от друга объектам.

City-block Manhattan distances (манхэттенское расстояние городских кварталов) вычисляют по формуле

$$d_м(X_i, X_j) = \Sigma_k |x_{ik} - x_{jk}|.$$

Для этой меры влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат.

Chebychev distances metric (расстояние Чебышева) рассчитывают по формуле

$$d_ч(X_i, X_j) = max_k |x_{ik} - x_{jk}|$$

и применяют, когда желают определить два объекта как различные, если они различаются по какой-либо одной координате.

Иногда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это достигается с использованием *Power metric* (степенного расстояния Минковского), которое находят по формуле

$$d_r(X_i, X_j) = (\Sigma_k |x_{ik} - x_{jk}|^p)^{1/r}$$

Параметр p ответствен за постепенное взвешивание разностей по отдельным координатам, параметр r — за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра равны 2, то это расстояние совпадает с расстоянием Евклида.

Percent disagreement (процент несогласия) используется в тех случаях, когда данные являются категориальными. Это расстояние вычисляется по формуле

$$d_f(X_i, X_j) = (количество\ x_{ik} \neq x_{jk})/k.$$

Все приведенные расстояния пригодны, если объекты кластеризации можно представить как точки в *k-мерном* пространстве. При решении большого

количества задач из экономики или социологии объекты нельзя представить как точки в k-мерном пространстве. В этом случае целесообразно в качестве расстояния использовать *1 — Pearson r* (1 минус коэффициент корреляции Пирсона).

Понятием, противоположным расстоянию, является понятие сходства между объектами G_i и G_j.

Неотрицательная вещественная функция $S(X_i, X_j) = S_{ij}$ называется мерой сходства, если:

1) $0 \le S(X_i, X_j) < 1$ *для* $X_i \ne X_j$;
2) $S(X_i, X_i) = 1$;
3) $S(X_i, X_j) = S(X_j, X_i)$.

Алгоритмов кластерного анализа достаточно много. Все их можно подразделить на иерархические и неиерархические.

Иерархические (древовидные) процедуры — наиболее распространенные алгоритмы кластерного анализа по их реализации на ЭВМ. Различают агломеративные (от слова *agglomerate* — собирать) и итеративные дивизивные (от слова *division* — разделять) процедуры.

Принцип работы иерархических агломеративных (дивизивных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства). К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. На каждом шаге алгоритмы требуют вычисления матрицы расстояний, а следовательно, емкой машинной памяти и большого количества времени. В этой связи реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

Общий принцип работы агломеративного алгоритма следующий. На первом шаге каждое наблюдение G_i $(i = 1, 2, ..., n)$ рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и, с учетом принятого расстояния, по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Большинство программ, реализующих алгоритм иерархической классификации, предусматривают графическое представление классификации в виде дендрограммы.

В программе *STATISTICA* реализованы так называемые агломеративные методы минимальной дисперсии: *joining (tree clustering)* (древовидная кластеризация) и *two-way joining* (двухвходовая кластеризация), а также *k-means* (дивизивный метод *k-средних*).

В методе древовидной кластеризации предусмотрены различные правила иерархического объединения в кластеры [6, 16].

1. Правило *single Linkage* (одиночной связи). На первом шаге объединяются два наиболее близких объекта, т.е. имеющие максимальную меру сходства. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с одним из объектов кластера, т.е. для его включения в кластер требуется максимальное сходство лишь с одним членом кластера.

   Метод называют еще методом ближайшего соседа, так как расстояние между двумя кластерами определяется как расстояние между двумя наиболее близкими объектами в различных кластерах. Это правило «нанизывает» объекты для формирования кластеров. Недостатком данного метода является образование слишком больших продолговатых кластеров.
2. Правило *complete Linkage* (полных связей). Данный метод позволяет устранить недостаток, присущий методу одиночной связи. Суть правила в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше некоторого порогового значения S. В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения d. Таким образом, d определяет максимально допустимый диаметр подмножества, образующего кластер. Этот метод называют еще методом наиболее удаленных соседей, так как при достаточно большом пороговом значении d расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах.
3. Правило *unweighted pair-group average* (невзвешенного попарного среднего). В данном методе расстояние между двумя кластерами определяется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные группы, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.
4. *Weighted pair-group average* (взвешенное попарное среднее). Метод идентичен предыдущему, за исключением того, что при вычислении размеры соответствующих кластеров используются в качестве весовых коэффициентов. Желательно этот метод использовать, когда предполагаются неравные размеры кластеров.
5. *Unweighted pair-group centroid* (невзвешенный центроидный). Расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.
6. *Weighted pair-group centroid* (взвешенный центроидный). Идентичен предыдущему, за исключением того, что при вычислениях используют веса для учета разности между размерами кластеров. Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.
7. *Ward's method* (метод Уорда). В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая

есть не что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов отклонений. Этот метод направлен на объединение близко расположенных кластеров. Замечено, что метод Уорда приводит к образованию кластеров примерно равных размеров и имеющих форму гиперсфер.

Ранее мы рассмотрели методы кластеризации объектов (наблюдений), однако иногда кластеризация по переменным может привести к достаточно интересным результатам. Например, предположим, что собраны данные о различных показателях состояний пациентов, страдающих сердечными заболеваниями. Можно кластеризовать пациентов для определения групп однородности пациентов со сходными симптомами. В то же время можно кластеризовать показатели для определения групп однородности показателей, которые связаны со сходным физическим состоянием. В модуле «Кластерный анализ» также предусмотрена эффективная двухвходовая процедура (*two-way joining*), которая позволит кластеризовать сразу в двух направлениях — по наблюдениям и переменным.

Предположим, есть гипотезы относительно числа m кластеров (по переменным или наблюдениям). Тогда можно задать программе, создать ровно m кластеров так, чтобы они были настолько различны насколько это возможно. Именно для решения задач этого типа предназначен метод *k-means* (*k-средних*). Гипотеза может основываться на теоретических соображениях, результатах предшествующих исследований или догадке. Выполняя последовательное разбиение на различное число кластеров, можно сравнивать качество получаемых решений. Чаще всего начинают анализ, пытаясь разбить совокупность объектов на две группы, затем увеличивают их число до трех, четырех и т.д. Например, в предыдущем примере исследователь может предполагать, что пациенты попадают в основном в три различные категории. При помощи модуля он может убедиться — действительно ли это так. Программа начинает с m случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, чтобы минимизировать изменчивость внутри кластеров и максимизировать изменчивость между кластерами. Алгоритм случайным образом в пространстве назначает центры будущих кластеров. Затем вычисляет расстояние между центрами кластеров и каждым объектом, и объект приписывается к тому кластеру, к которому он ближе всего. Завершив приписывание, алгоритм вычисляет средние значения для каждого кластера. Этих средних будет столько, сколько используется переменных для проведения анализа, — k штук. Набор средних представляет собой координаты нового положения центра кластера. Алгоритм вновь вычисляет расстояние от каждого объекта до центров кластеров и приписывает объекты к ближайшему кластеру. Вновь вычисляются центры тяжести кластеров, и этот процесс повторяется до тех пор, пока центры тяжести не перестанут «мигрировать» в пространстве.

Наиболее известный метод представления матрицы расстояний или сходства основан на идее дендрограммы, или диаграммы дерева. Дендрограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний. С помощью дендрограммы можно графически или геометрически изобразить процедуру кластеризации при условии, что эта процедура оперирует только с элементами матрицы расстояний или сходства. Существует много способов построения дендрограмм. В дендрограмме объекты располагаются вертикально слева, результаты кластеризации — справа. Значения расстояний или сходства, отвечающие строению новых кластеров, изображаются над горизонтальной прямой поверх дендрограмм. На рис. 13.1 показан один из примеров дендрограммы.

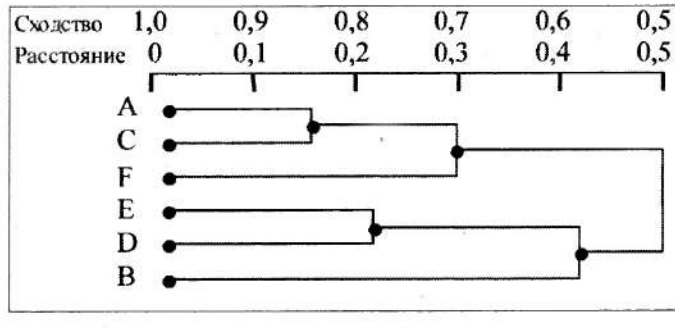


Рис. 13.1

Пример соответствует случаю шести объектов ($n = 6$) и k характеристик (признаков). Объекты A и C наиболее близки и поэтому объединяются в один кластер на уровне близости, равном 0,9. Объекты *D u E* объединяются при уровне 0,8. Теперь имеем 4 кластера: (*A, C*), (*F*), (*D, E*), (B). Далее образуются кластеры (*A, C, F*) и (*E, D, B*), соответствующие уровням близости, равным соответственно 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне 0,5.

Вид дендрограммы зависит от выбора меры сходства или расстояния между объектом и кластером и метода кластеризации.

Алгоритмы кластерного анализа имеют хорошую программную реализацию в пакете *STATISTICA*, которая позволяет решить задачи самой большой размерности.

13.2. Описание модуля *Cluster Analysis*

Для запуска модуля в меню **Statistics** надо щелкнуть по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выбрать команду **Cluster Analysis** (анализ кластера). Откроется стартовая панель модуля. На вкладке **Quick** находится список методов кластерного анализа, реализованных в программе *STATISTICA*. Это **Joining tree clustering** (древовидная кластеризация), **k-means clustering** (метод k-средних) и **Two-way joining** (двухвходовая кластеризация) Нажмите кнопку **Open Data**, появится окно **Select Spreadsheet** (выбрать

данные). Щелкните кнопкой **Files** и выберите в **Examples → Datasets** файл данных **Cars.sta** и нажмите кнопку **Open**. Открыть данные можно также через верхнее меню **File**, выбрав команду **Open**. В наблюдениях (строках таблицы) записаны названия (марки) автомобилей, в переменных (столбцах) параметры: *price* — цена, *ACCELERATION* — время в секундах, необходимое для разгона с места до скорости 60 миль в час, *BRAKING* — тормозной путь, *HANDLING* —наименьший радиус поворота, *MILEAGE* — расход горючего (количество миль, пройденных на одном галлоне бензина).

Необходимо разбить автомобили на несколько однородных групп, в которых автомобили мало отличаются друг от друга (существенно меньше, чем в совокупности). Сложность задачи в том, что надо сравнивать автомобили не по какому-то одному параметру (признаку), а по нескольким параметрам одновременно. Очевидно, разбив автомобили на группы, мы можем лучше в целом представить их совокупность с тем, чтобы затем более обоснованно принимать решение, например, при покупке или обмене одного автомобиля на другой.

В главной части стартовой панели высветите **k-means clustering** и нажмите **OK**, на экране появится диалоговое окно **k-means clustering** (рис. 13.2). Перейдите на вкладку **Advanced** и выберите переменные для анализа.

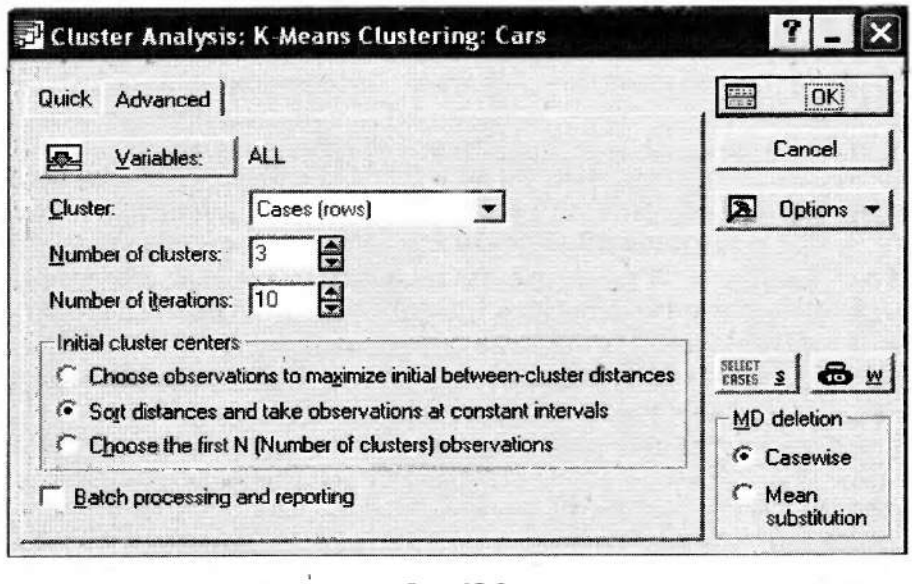


Рис. 13.2

Для этого нажмите кнопку **Variables** в левом верхнем углу экрана и откройте диалоговое окно **Select variables for analysis.** Для выбора всех параметров нажмите кнопку **Select All**, а затем — **OK**. Программа вернется в стартовое окно модуля.

На поле **Claster** надо выбрать объекты для кластеризации. Так как цель исследования — кластеризация автомобилей, которые являются в файле данных наблюдениями, выберите *Cases (rows)* (наблюдения (строки)).

В поле **Number of clasters** (число кластеров) нужно определить число групп, на которые мы хотим разбить автомобили. Запишите в это поле число 3.

В поле **Number of iterations** (число итераций) задается максимальное число итераций, используемых при построении классов. Задайте, например, число 10. Если необходимо провести кластеризацию не по всем объектам (автомобилям), воспользуйтесь кнопкой **Select cases**.

Группа опций *Initial claster centers* позволяет задать начальные центры кластеров: *Choose observations to maximize initial between-cluster distances* (выбрать наблюдения, максимизирующие начальные расстояния между кластерами); *Sort distances and take observations at constant intervals* (сортировать расстояния и выбрать наблюдения на постоянных интервалах); *Choose the first N (Namber of clasters) observations* (выбрать первые *N* (число кластеров) наблюдений). Выберите, например, вторую опцию и нажмите **OK**. Откроется окно результатов **k-means Clustering Results** (рис. 13.3).

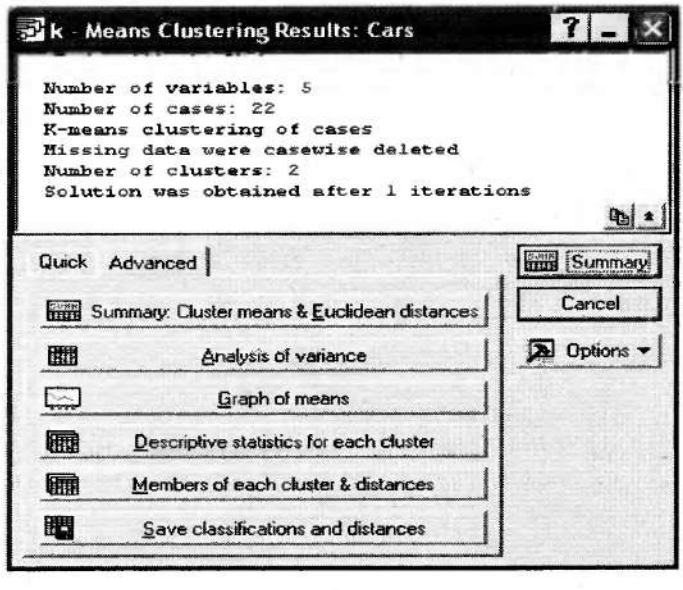


Рис. 13.3

В верхней информационной части окна представлены следующие данные:

- *Number of variables* (количество переменных);
- *Number of cases* (число наблюдений);
- *k-means clustering of cases* (метод *k-средних*);
- *Missing data were casewise deleted* (обработка пропущенных значений опущена);
- *Number of clusters* (число кластеров);
- *Solution was obtained after * iterations* (решение было найдено после* итераций).

Откройте вкладку **Advanced,** так как она содержит более подробную информацию о результатах анализа. Функциональное назначение кнопок открывшегося окна следующее.

Summary:Cluster means & Euclidean distances предназначена для вывода таблиц, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй — евклидовы расстояния и квадраты евклидовых расстояний между кластерами (рис. 13.4).

	Euclidean Distances between Clusters Distances below diagonal Squared distances above diagonal		
Cluster Number	No. 1	No. 2	No. 3
No. 1	0,000000	2,736755	0,916244
No. 2	1,654314	0,000000	2,231666
No. 3	0,957206	1,493876	0,000000

Рис. 13.4

Analysis of variance выводит таблицу дисперсионного анализа (рис. 13.5).

	Analysis of Variance (Cars)					
Variable	Between SS	df	Within SS	df	F	signif. p
PRICE	9,08159	2	11,91841	19	7,23881	0,004602
ACCELERATION	6,74790	2	14,25210	19	4,49794	0,025163
BRAKING	10,11892	2	10,88108	19	8,83457	0,001938
HANDLING	10,87750	2	10,12250	19	10,20857	0,000975
MILEAGE	7,99118	2	13,00882	19	5,83575	0,010573

Рис. 13.5

В таблице приведены значения межгрупповых *(Between SS)* и внутригрупповых *(Within SS)* дисперсий признаков. Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру и тем «качественнее» кластеризация. Параметры F и p также характеризуют вклад признака в разделение объектов на группы. Лучшей кластеризации соответствуют большие значения первого и меньшие значения второго параметра. Признаки с большими значениями p (например, больше 0,05) можно из процедуры кластеризации исключить.

Graph of means позволяет просмотреть средние значения для каждого кластера на линейном графике (рис. 13.6).

Descriptive statistics for each cluster выводит электронную таблицу с описательными статистиками для каждого кластера (среднее, дисперсия и т.д.).

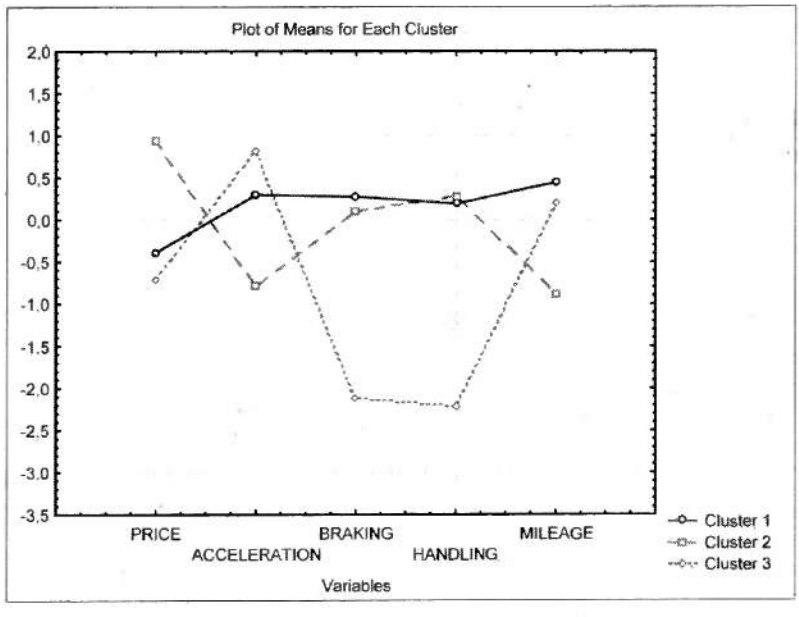


Рис. 13.6

Members of each clusters & distances предназначена для просмотра распределения объектов по кластерам. В таблице также будет указано расстояние от объекта до центра кластера.

Save classifications and distances сохраняет результаты классификации в файле *STATISTICA* для дальнейшего исследования. При этом в новом файле каждому наблюдению программой присваивается номер кластера, к которому он был отнесен при классификации. Из графиков, изображенных на рис. 13.6, видно, что в кластерах 1 и 2 средние параметров *BRAKING, HANDLING* незначительно отличаются друг от друга. Это свидетельствует о неудовлетворительном разбиении на группы. Возможно, это связано с тем, что эти характеристики слабо различимы для разных марок автомобилей. Поэкспериментируйте с составом переменных для кластерного анализа. Исключите, например, из рассмотрения переменные *BRAKING, HANDLING*. Как показывает рис. 13.7, расстояние между средними характеристик кластеров значительно увеличилось, также увеличилось общее расстояние между центрами кластеров (рис. 13.8), что свидетельствует о более успешной кластеризации.

Уменьшите число кластеров до 2, сохранив первоначальный состав характеристик автомобилей. Результаты кластеризации, приведенные на рис. 13.9, 13.10, также свидетельствуют о неуспешной кластеризации. Из этих данных следует, что переменные *BRAKING, HANDLING* вносят незначительный вклад при разбиении автомобилей на 2 кластера (малые значения *Between SS*, большие значения *Within SS* и соответственно малые значения *F* и большие значения *p*). Поэтому желательно из процедуры кластеризации их исключить. Таблица и график, изображенные на рис. 13.11,13.12, показывают, что разбиение автомобилей на две

однородные группы без учета параметров *BRAKING*, *HANDLING* также прошло успешно. По-видимому, это наиболее оптимальный вариант кластеризации из рассмотренных ранее.

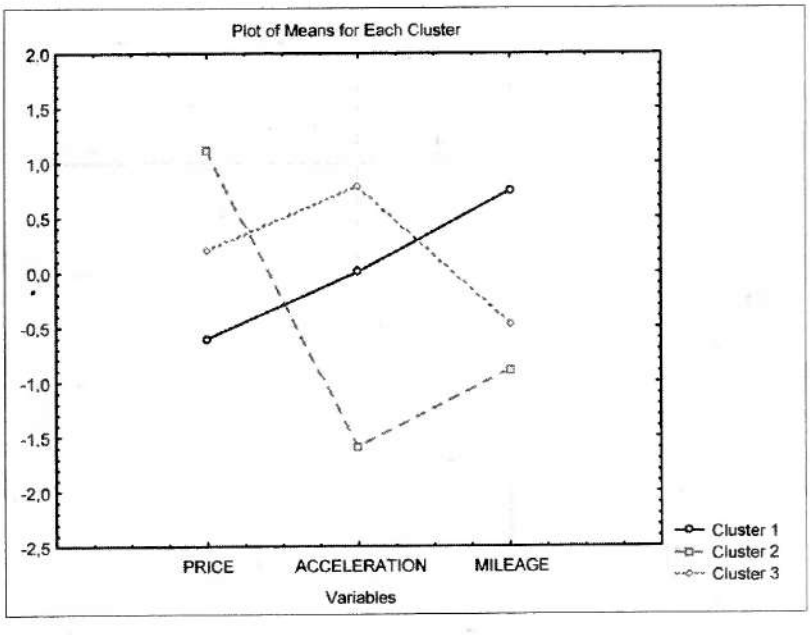


Рис. 13.7

Cluster Number	Euclidean Distances between C Distances below diagonal Squared distances above diag No. 1	No. 2	No. 3
No. 1	0,000000	0,938877	2,403320
No. 2	0,968957	0,000000	3,519362
No. 3	1,550264	1,875996	0,000000

Рис. 13.8

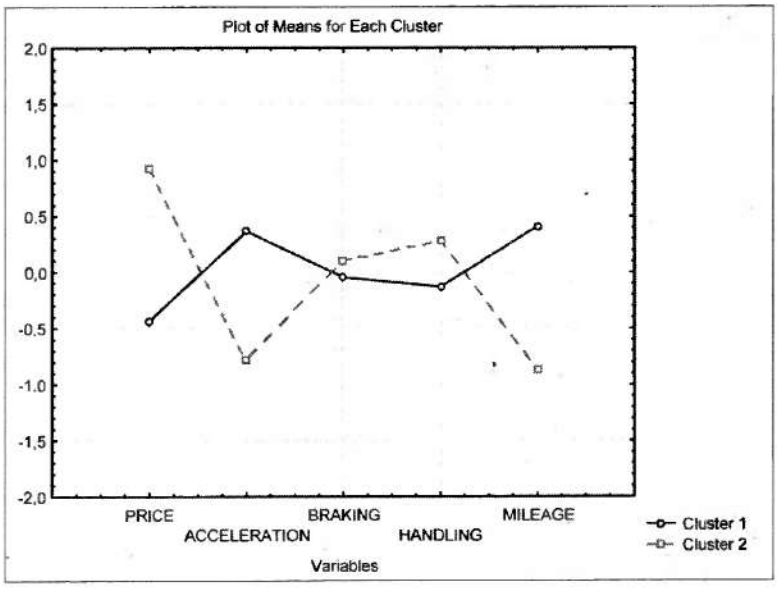


Рис. 13.9

Variable	Analysis of Variance (Cars)					
	Between SS	df	Within SS	df	F	signif. p
PRICE	8,911884	1	12,08812	20	14,74487	0,001023
ACCELERATION	6,283288	1	14,71671	20	8,53898	0,008426
BRAKING	0,101172	1	20,89883	20	0,09682	0,758898
HANDLING	0,806422	1	20,19358	20	0,79869	0,382113
MILEAGE	7,885540	1	13,11446	20	12,02572	0,002429

Рис. 13.10

Variable	Analysis of Variance (Cars)					
	Between SS	df	Within SS	df	F	signif. p
PRICE	8,911884	1	12,08812	20	14,74487	0,001023
ACCELERATION	6,283288	1	14,71671	20	8,53898	0,008426
MILEAGE	7,885540	1	13,11446	20	12,02572	0,002429

Рис. 13.11

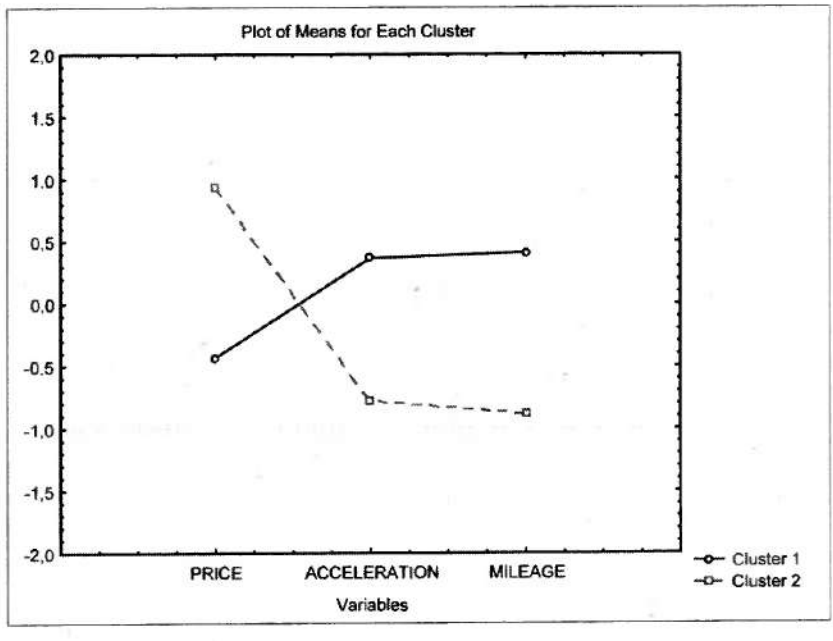


Рис. 13.12

Рассмотрим процедуру одновременной кластеризации по переменным (столбцам) и по случаям (строкам). В стартовой панели модуля выберите **Two-way joining**. Нажмите **OK.** Откроется диалоговое окно.

Нажмите кнопку **Variables** и укажите переменные для анализа, например, *Select All.* На вкладке **Advanced** имеется возможность выбрать пороговый параметр *Threshold Value* (значение порога). Пороговый параметр определяет, когда алгоритм рассматривает в матрице данных два числа как равные, а затем приписывает их к одному кластеру. Если эта величина слишком велика (по отношению к числам в матрице данных), то будет сформирован только один кластер; если она очень мала, то кластером будет являться каждая точка данных. Параметр может назначить пользователь — *User defined.* Но для большинства случаев рекомендуется величина по умолчанию — *Computed from data* (общее стандартное отклонение, деленное на 2). Опция *Batch processing and reporting* доступна при определенных установках в *Output Manager* (менеджере вывода). Нажмите **OK**. На экране появится окно результатов. В верхней информационной части окна указано число переменных; число наблюдений; пороговое значение; число полученных блоков разбиения; стандартное отклонение.

Следующие кнопки позволяют провести анализ результатов.

- **Summary: Two-way joining graph** (просмотреть графическое представление результатов).
- **Descriptive statistics for cases (row)** (описательные статистики для наблюдений).

- **Descriptive statistics for variables** (описательные статистики для переменных).
- **Reordered data matrix** (неупорядоченная матрица значений).

Нажмите кнопку **Summary: Two-way joining graph**. Появится цветной график результата кластеризации. Горизонтальная ось представляет параметры автомобилей, а вертикальная — марки. Одним цветом обозначены объекты, попавшие в один кластер. Так, например, *Mercedes* и *Audi* отнесены к одному блоку-кластеру по *PRICE; Olds, Corvett* и *Honda* объединены в один кластер по *MILAGE; Ford* и *Buick* — в один кластер по *BRAKING* и т.д.

Для кластеризации агломеративным методом древовидной кластеризации надо в стартовой панели высветить **Joining tree clustering** и нажать на **OK**. Появится диалоговое окно. Для переменных сделайте установку **Select All**. Так как кластеризуются автомобили, в поле **cluster** выберите пункт *Cases (row)*. В качестве исходных данных используется матрица значений (а не матрица расстояний), поэтому в поле **Input file** надо выбрать *Raw data*. В поле **Amalgamation (linkage) rule** (правило иерархического объединения) выберите правило объединения в кластеры, например *Single Linkage* (метод одиночной связи), и нажмите **OK**. Появится окно результатов. В верхней части окна записана информация: число переменных, число наблюдений, метод кластеризации, правило иерархического объединения, выбранная метрика (расстояние между объектами).

Кнопки в нижней части окна на вкладке **Advanced** предназначены для анализа результатов кластеризации:

- **Horizontal hierarchical tree plot** (горизонтальная древовидная диаграмма).
- **Vertical icicle plot** (вертикальная древовидная диаграмма).
- **Amalgamation schedule** (правило объединения в кластеры).
- **Graph of amalgamation schedule** (график порядка объединения).
- **Distance matrix** (матрица расстояний).
- **Descriptive statistics** (описательные статистики).

Нажмите **Horizontal hierarchical tree plot**. На рис. 13.13 показана горизонтальная древовидная диаграмма. Диаграмма начинается с каждого объекта в классе (в левой части диаграммы). Теперь представьте, что постепенно (очень малыми шагами) «ослабевает» критерий, показывающий, какие объекты являются уникальными, а какие нет. Другими словами, понижается порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате связывается все большее и большее число объектов и агрегируются (объединяются) все больше кластеров, состоящих из все сильнее различающихся элементов. На последнем шаге все объекты окончательно объединяются. На этих диаграммах горизонтальные оси представляют расстояние объединения (в вертикальных древовидных диаграммах вертикальные оси представляют расстояние объединения). Так, для каждого узла в графе (там, где формируется новый кластер) можно определить величину расстояния, для которого соответствующие элементы связываются в новый единственный кластер.

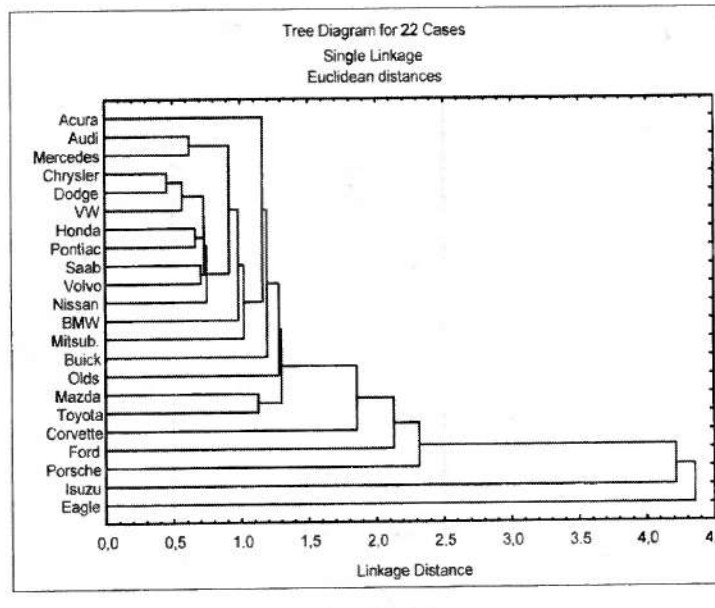


Рис. 13.13

Когда данные имеют ясную «структуру» в терминах кластеров объектов, сходных между собой, тогда эта структура может быть отражена в иерархическом дереве различными ветвями. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их.

13.3. Деревья классификации и их свойства

Деревья классификации — это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких независимых (предикторных) переменных. Чтобы понять основную идею данного метода, рассмотрим пример, приведенный в [16]. Представьте, что нужно придумать устройство, которое отсортирует коллекцию монет по их достоинству (например, 1, 2, 3 и 5 к.). Предположим, что какое-то измерение монет, например диаметр, известно и поэтому может быть использовано для построения иерархического устройства сортировки монет. Заставим монеты катиться по узкому желобу, в котором прорезана щель размером с однокопеечную монету. Если монета провалилась в щель, то это 1 к.; в противном случае она продолжает катиться дальше по желобу и натыкается на щель для двухкопеечной монеты; если она туда провалится, то это 2 к., если нет — значит, это 3 или 5 к., и т.д. Таким образом, построено дерево классификации. Решающее правило, реализованное в этом дереве классификации, позволяет эффективно рассортировать горсть монет.

В модуле **Classification Trees** программы *STATISTICA* с исчерпывающей полнотой реализованы методы построения бинарных деревьев классификации, основанных на ветвлении по одной предикторной переменной. Модуль также включает возмож-

ность построения деревьев классификации с ветвлением по значениям линейных комбинаций предикторных переменных, измеренных в интервальной шкале.

Цель построения дерева классификации заключается в предсказании значений категориальной зависимой переменной, и поэтому методы, реализованные в этом модуле, тесно связаны с более традиционными методами дискриминантного анализа, кластерного анализа, непараметрической статистики и нелинейного оценивания [16].

Способность деревьев классификации выполнять одномерное ветвление с предикторными переменными различных типов (категориальных, порядковых, интервальных) дает возможность анализировать вклады отдельных переменных в процедуру классификации.

Деревья классификации могут быть, а иногда и бывают, очень сложными. Однако использование специальных графических процедур позволяет упростить интерпретацию результатов даже для очень сложных деревьев. Возможность графического представления результатов и простота интерпретации во многом объясняют большую популярность деревьев классификации в прикладных областях, однако наиболее важные отличительные свойства деревьев классификации — их иерархичность и широкая применимость.

Широкая сфера применимости деревьев классификации делает их весьма привлекательным инструментом анализа данных, но не следует поэтому полагать, что его рекомендуется использовать вместо традиционных методов статистики. Напротив, если выполнены более строгие теоретические предположения, налагаемые традиционными методами, и выборочное распределение обладает некоторыми специальными свойствами (например, соответствие распределения переменных нормальному закону), то более результативным будет использование именно традиционных методов. Однако как метод разведочного анализа или как последнее средство, когда отказывают все традиционные методы, деревья классификации, по мнению многих исследователей, не знают себе равных.

Следует заметить, что если в модуле «Дискриминантный анализ» предусмотрена возможность классификации программой нового наблюдения, в модуле «Деревья классификации» такой возможности нет. Пользователь вынужден по полученным программой решающим правилам самостоятельно отнести новое наблюдение к тому или иному классу. Также программа не определяет вероятностных характеристик принадлежности каждого наблюдения к классам.

Деревья классификации широко используются в таких разнообразных прикладных областях, как медицина (диагностика), программирование (анализ структуры данных), ботаника (классификация) и психология (теория принятия решений). Они идеально приспособлены для графического представления, поэтому сделанные на их основе выводы гораздо легче интерпретировать, чем, если бы они были представлены только в числовой форме.

Реализованные в модуле **Classification Trees** методы дискриминантного одномерного ветвления по категориальным и порядковым предикторам и дискриминантного многомерного ветвления по линейным комбинациям порядковых предикторов представляют собой адаптацию соответствующих алгоритмов пакета *QUEST (Quick,*

Unbiased, Efficient Statistical Trees). *QUEST* — это программа деревьев классификации, разработанная *Loh и Shih* (1997), в которой используются улучшенные варианты метода рекурсивного квадратичного дискриминантного анализа и которая содержит ряд новых средств для повышения надежности и эффективности деревьев классификации.

Также в модуле **Classification Trees** имеется опция *Тип ветвления*, предоставляющая пользователю другой, концептуально более простой подход. Реализованный здесь алгоритм *Одномерного ветвления по методу CART* является адаптацией алгоритмов пакета *CART (Classification And Regression Trees)*. *CART* — это программа деревьев классификации, которая при построении дерева осуществляет полный перебор всех возможных вариантов одномерного ветвления.

Опции анализа *QUEST* и *CART* естественно дополняют друг друга. В случаях, когда имеется много предикторных переменных с большим числом уровней, поиск методом *CART* может оказаться довольно продолжительным. Кроме того, этот метод имеет склонность выбирать для ветвления те предикторные переменные, у которых больше уровней. Однако поскольку здесь производится полный перебор вариантов, есть гарантия, что будет найден вариант ветвления, дающий наилучшую классификацию (по отношению к обучающей выборке).

Метод *QUEST* — быстрый и несмещенный. Его преимущество в скорости перед методом *CART* становится особенно заметным, когда предикторные переменные имеют десятки уровней. Отсутствие у метода *QUEST* смещения в выборе переменных для ветвления также является его существенным преимуществом в случаях, когда одни предикторные переменные имеют мало уровней, а другие — много (предикторы со многими уровнями часто порождают «методы тыка», которые хорошо согласуются с данными, но дают плохую точность прогноза). Сочетание опций *QUEST* и *CART* позволяет полностью использовать всю гибкость аппарата деревьев классификации.

Метод деревья классификации хорош настолько, насколько удачным окажется выбор варианта анализа. Чтобы построить модель, дающую хороший прогноз, в любом случае нужно хорошо понимать природу взаимосвязей между предикторными и зависимыми переменными. Поэтому для успешного проведения анализа большую роль играют опыт, интуиция пользователя. Итак, методы анализа с помощью деревьев классификации можно охарактеризовать как набор иерархических, чрезвычайно гибких средств для предсказания пользователем принадлежности наблюдений (объектов) к определенному классу значений категориальной зависимой переменной по значениям одной или нескольких предикторных переменных.

13.4. Вычислительные методы. Модуль *Classification Trees*

Для решения задачи прогнозирования принадлежности объекта (случая) к определенному классу значений зависимой категориальной переменной по данным измерений одной или нескольких предикторных переменных было разработано большое число программ, реализующих метод деревьев классификации.

Процесс вычисления (построения) дерева классификации состоит из четырех основных этапов:

1. Выбор критерия точности прогноза.
2. Выбор вариантов ветвления.
3. Определение момента, когда дальнейшие ветвления следует прекратить.
4. Определение «подходящего размера» дерева.

Рассмотрим *этап 1*. Цель анализа с помощью деревьев классификации в конечном счете состоит в том, чтобы получить максимально точный прогноз. Наиболее точным прогнозом считается такой, который связан с наименьшей ценой. В большинстве приложений цена — это просто доля неправильно классифицированных наблюдений. Поэтому, как правило, самый лучший прогноз — такой, который дает наименьший процент неправильных классификаций.

В модуле **Classification Trees** предусмотрено два варианта задания цены неправильной классификации объектов *Misclassif. costs* (цена ошибок классифика-

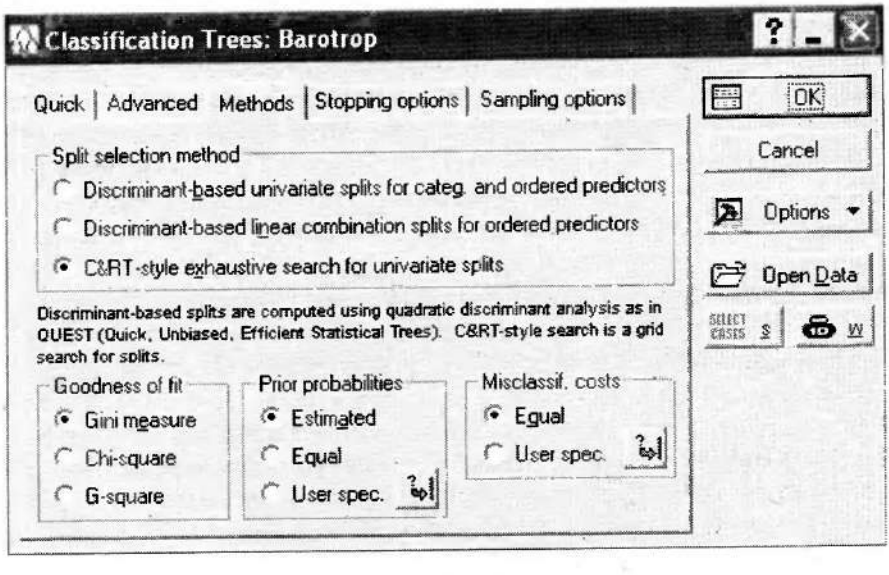


Рис. 13.14

ции), т.е. отнесения объекта, принадлежащего одному классу, к какому-то другому классу. Выборы этих вариантов задаются при помощи соответствующих опций в рамке **Misclassif. Costs** на вкладке **Methods** (методы) (рис. 13.14).

В первом варианте, когда цены берутся одинаковые *(Equal)*, все внедиагональные элементы матрицы цен ошибок классификации (прогнозируемые классы — по строкам, наблюдаемые классы — по столбцам) полагаются равными 1, и в выбранные значения *Prior probabilities* (априорных вероятностей) для классов зависимой переменной поправок не вносится.

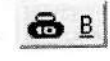
Бывает так, что по причинам, не связанным с размерами классов, для одних классов требуется более точный прогноз, чем для других. Например, гораздо

важнее выявить переносчиков инфекционного заболевания, постоянно контактирующих с другими людьми, чем тех же переносчиков, не имеющих постоянных контактов, — и это независимо от относительной численности тех и других. Если, например, принять, что избежать контактов с контактирующим переносчиком гораздо важнее, чем с неконтактирующим, то следует приписать ошибочной классификации контактирующего как неконтактирующего большую цену, чем ошибочной классификации неконтактирующего как контактирующего. Во втором варианте нужно выбрать для цен ошибок классификации опцию *User spec* (пользовательские). В результате этого на экране откроется таблица результатов, в которой указывают неотрицательные значения цены ошибок классификации для отдельных внедиагональных элементов матрицы цен ошибок классификации. Реально пользовательские цены ошибок классификации можно использовать для придания некоторым классам больших «весов», чем другим.

Априорные вероятности показывают, насколько мы, не зная ничего о значениях предикторных переменных, считаем вероятным, что объект будет принадлежать определенному классу. Можно заметить, что если априорные вероятности выбраны пропорциональными численности классов, а цена ошибки классификации — одинаковая для всех классов, то минимизация потерь в точности эквивалентна минимизации доли неправильно классифицированных наблюдений. Априорные вероятности выражают то, как, не располагая никакой априорной информацией о значениях предикторных переменных модели, оцениваем вероятность попадания объекта в тот или иной класс [6].

Выбор априорных вероятностей, используемых для минимизации потерь, очень сильно влияет на результаты классификации. Если различия между исходными частотами в данной задаче не считаются существенными или если мы знаем заранее, что классы содержат примерно одинаковое количество наблюдений, то тогда можно взять одинаковые априорные вероятности (*Equal*). В случаях, когда исходные частоты связаны с размерами классов (например, при работе со случайной выборкой), следует в качестве оценок для априорных вероятностей взять относительные размеры классов в выборке (*Estimated*). Наконец, если (на основании данных предыдущих исследований) располагаем какой-то информацией об исходных частотах, то априорные вероятности нужно выбирать с учетом этой информации (*User spec*). Например, априорная вероятность человека быть носителем рецессивного гена вдвое выше вероятности того, что этот ген имеет проявления. В любом случае, когда классу приписывается та или иная априорная вероятность, «учитывается» степень важности ошибки при классификации объектов этого класса. Минимизация потерь — это минимизация общего числа неправильно классифицированных наблюдений с априорными вероятностями, пропорциональными размерам классов (и ценами ошибки классификации, одинаковыми для всех классов), поскольку прогноз, чтобы давать меньший итоговый процент ошибок классификации, должен быть более точным на больших классах. Поэтому в модуле **Classification Trees** предусмотрено задание трех вариантов априорных вероятностей при помощи соответствующих опций в рамке **Prior probabilities** на вкладке **Methods** (методы) (рис.13.14):

- *Estimated* (оцениваемые), т.е. пропорциональные размерам классов зависимой переменной;
- *Equal* (одинаковые), т.е. одинаковые для всех классов зависимой переменной;
- User spec. (пользовательские). При выборе этой опции раскроется окно диалога, в котором надо задать априорную вероятность для каждого класса значений зависимой переменной. Если сумма всех введенных вероятностей не равна 1, программа *STATISTICA* автоматически внесет в них поправки, пропорциональные их величине.

Использование весов для весовой переменной в качестве множителей наблюдений для агрегированных данных также имеет отношение к минимизации цены классификации. Однако вместо того, чтобы использовать веса наблюдений для агрегированных данных, можно ввести подходящие априорные вероятности и/или цены ошибки классификации и получить те же самые результаты, не тратя времени на обработку множества наблюдений, имеющих одинаковые значения всех переменных. В модуле «Деревья классификации» веса наблюдений трактуются непосредственно как множители наблюдений. Доля неправильных классификаций на агрегированных данных с использованием весов наблюдений будет такой же, как при анализе того же набора данных, в котором наблюдения продублированы соответствующее число раз. Задать веса наблюдений можно при помощи кнопки, расположенной справа-внизу окна **Classification Trees**  .

За исключением простейших случаев, взаимосвязи между априорными вероятностями, ценами ошибок классификации и весами наблюдений являются довольно сложными. Однако если минимизация цены соответствует минимизации доли неправильных классификаций, все эти обстоятельства можно не принимать во внимание.

2. *Второй этап* анализа заключается в том, чтобы выбрать способ ветвления по значениям предикторных переменных. В соответствии с иерархической природой деревьев классификации такие ветвления производятся последовательно, начиная с корневой вершины, затем переходят к вершинам потомкам, пока дальнейшее ветвление не прекратится и «неразветвленные» вершины потомки окажутся терминальными. Терминальные вершины (или листья) — это узлы дерева, начиная с которых никакие решения больше не принимаются. На рисунках терминальные вершины показываются программой красными пунктирными линиями, а остальные — так называемые решающие вершины, или вершины ветвления, — сплошными черными линиями. Началом дерева считается самая верхняя решающая вершина, которую иногда также называют корнем дерева.

Различные методы выбора типа ветвления реализованы как опции в рамке **Split selection methods** (выбор типа ветвления) на вкладке **Methods** (рис.13.14) [6]. Первый метод — *Discriminant-based univariate split for categ. and ordered predictors* (дискриминантное одномерное ветвление) — можно использовать для предикторных переменных, измеренных в различных шкалах: номинальной, порядковой, интервальной, в шкале отношений. Второй метод — *Discriminant-based*

linear combination split for ordered predictors (дискриминантное многомерное ветвление по линейной комбинации) требует, чтобы предикторы были измерены как минимум в интервальной шкале. В обоих методах ветвления строятся с помощью квадратичного дискриминантного анализа, как это делается при построении деревьев классификации по методу *QUEST*. Третий тип ветвления – это *C&RT-style exhaustive search for univariate splits* (полный перебор вариантов одномерного ветвления методом *CART*). Его, как и первый метод, можно использовать для всех типов предикторных переменных. В отличие от дискриминантных методов ветвления, в этом методе, для того чтобы найти наилучший вариант ветвления, производится последовательный перебор всех возможных комбинаций уровней предикторных переменных.

Рассмотрим более подробно перечисленные методы.

Если выбрана опция *Discriminant-based univariate split for categ. and ordered predictors*, то прежде всего нужно решать вопрос, какую из вершин дерева, построенного к данному моменту, следует расщепить на данном шаге и какую из предикторных переменных при этом использовать. Для каждой вершины вычисляются *p-уровни* для проверки значимости зависимостей между принадлежностью объектов к классам и уровням каждой из предикторных переменных. В случае категориальных предикторов (шкалы – номинальная, порядковая) p вычисляются для проверки критерия χ^2 гипотезы о независимости принадлежности к классам от уровня категориального предиктора в данном узле дерева. В случае некатегориальных предикторов (шкалы – интервальная, отношений) p вычисляются для анализа *ANOVA* взаимосвязи классовой принадлежности и значений предиктора в данном узле. Если, наименьший из вычисленных p оказался меньше p Бонферони для множественных 0,05-сравнений (принимается по умолчанию) или иного порогового значения, установленного пользователем, то для разветвления этого узла выбирается та предикторная переменная, которой соответствует наименьшее значение p. Если среди p не оказалось ни одного, меньшего чем заданное пороговое значение, то p вычисляются по статистическим критериям, устойчивым к виду распределения, например *F-Левена*. В случае некатегориальных предикторов для построения двух относящихся к данной вершине классов применяется алгоритм кластеризации *k-средних* (k=2). При этом находятся корни квадратного уравнения, характеризующего различие средних значений по классам предиктора, и для каждого из корней вычисляются значения порога ветвления. Выбирается вариант ветвления, для которого значение порога ветвления ближе к среднему по классу. В случае категориального предиктора создаются фиктивные переменные, представляющие уровни этого предиктора, а затем с помощью метода сингулярного разложения фиктивные переменные преобразуются в совокупность независимых предикторов. Затем применяется описанный алгоритм для некатегориальных предикторов, после чего полученное ветвление «проецируется обратно» в уровни исходной категориальной переменной и трактуется как различие между двумя множествами уровней этой переменной. Описанные процедуры довольно сложны, однако они позволяют уменьшить смещение при выборе ветвления, которое характерно для полного перебора деревьев с одномерным

ветвлением по методу *CART*. Смещение имеет место в сторону выбора переменных с большим числом уровней ветвления, и при интерпретации результатов оно может исказить относительную значимость влияния предикторов на значения зависимой переменной.

При выборе опции *Discriminant-based linear combination split for ordered predictors* способ использования непрерывных некатегориальных переменных, участвующих в линейной комбинации, очень похож на тот, который применялся в предыдущем методе для категориальных переменных. С помощью сингулярного разложения непрерывные предикторы преобразуются в новый набор независимых предикторов. Затем применяются процедуры создания классов и поиска ветвления, ближайшего к среднему по классу, после чего результаты «проецируются назад» в исходные непрерывные предикторы и представляются как одномерное ветвление линейной комбинации предикторных переменных.

Третья опция — *C&RT-style exhaustive search for univariate splits* предполагает перебор всех возможных вариантов ветвления по каждой предикторной переменной. Далее находится тот из них, который дает наибольший рост для критерия согласия (или, что-то же самое, наибольшее уменьшение отсутствия согласия). Для категориальной предикторной переменной, принимающей в данном узле k значений, имеется ровно $2^{(k-1)} - 1$ вариантов разбиения множества ее значений на две части. Для некотегориального предиктора, имеющего в данном узле k различных уровней, имеется $k - 1$ точек, разделяющих разные уровни. Таким образом, количество различных вариантов ветвления, которые необходимо просмотреть, будет очень большим, если в задаче много предикторов, у них много уровней значений и в дереве много терминальных вершин.

После выбора этой опции активируется рамка **Goodness of fit** (рис. 13.14), где в качестве критерия согласия может быть выбрана одна из трех возможных мер: *Gini measure* (мера Джини), *Chi-square* (χ^2) и *G-square* (G^2). Критерии согласия используются для выбора наилучшего из всех возможных вариантов ветвления.

Мера Джини однородности вершины принимает нулевое значение, когда в данной вершине имеется всего один класс. Если используются априорные вероятности, оцененные по размерам классов или исходя из одинаковой цены ошибок классификации, то мера Джини вычисляется как сумма всех попарных произведений относительных размеров классов, представленных в данной вершине; ее значение будет максимальным, когда размеры всех классов одинаковы.

Мера χ^2 — это мера χ^2 Бартлета.

Мера G^2 — есть мера максимума правдоподобия χ^2, используемая в моделировании структурными уравнениями.

3. *Третий этап* анализа заключается в выборе момента, когда следует прекратить дальнейшие ветвления. Метод деревья классификации обладает тем свойством, что если не установлено ограничение на число ветвлений, то можно прийти к «чистой» классификации, когда каждая терминальная вершина содержит только один класс наблюдений (объектов). Как правило, при анализе с помощью деревьев классификации данные о классификации зависимой переменной или уровни значений предикторных переменных содержат ошибки измерений или со-

ставляющую белого шума. Поэтому было бы нереально пытаться продолжать сортировку до тех пор, пока каждая терминальная вершина не станет «чистой».

В модуле на вкладке **Stopping options** (параметры остановки, рис. 13.15) в рамке **Stopping rule** (правило остановки) реализованы три варианта остановки: *Prune on misclassification error* (отсечение по ошибке классификации), *Prune on deviance* (отсечение по вариации), и *Fact-style direct stopping* (прямая остановка по методу *FACT*).

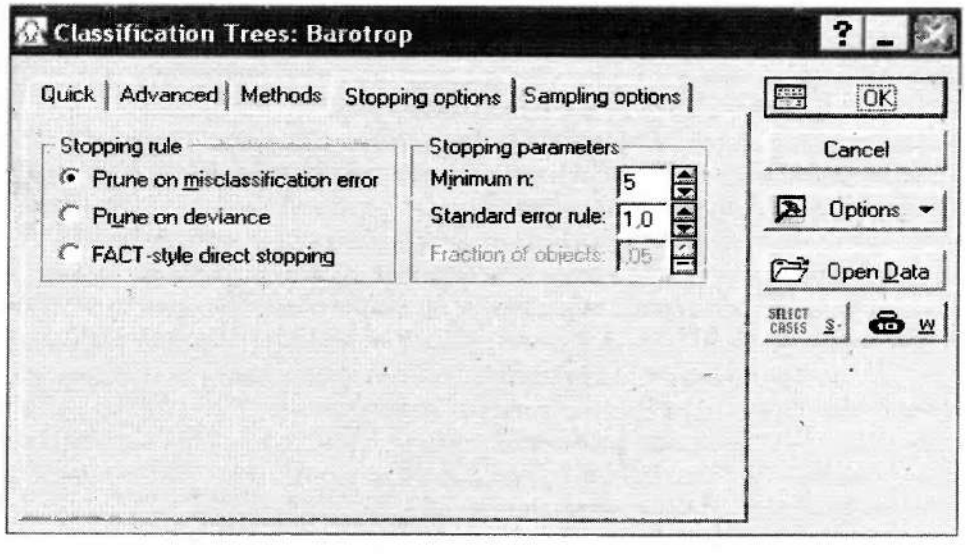


Рис. 13.15

Первые два правила отсечения (см. ниже вторая стратегия выбора размера дерева — метод автоматического построения дерева) в некоторой степени похожи на пошаговое обратное продвижение из модуля Дискриминантный анализ. «Ветви» последовательно «отсекаются» от полного дерева классификации подобно тому, как при пошаговом обратном продвижении в модуле «Дискриминантный анализ» предикторы последовательно исключаются из уравнений прогноза. Затем дерево классификации «подходящего размера» выбирается из «усеченных» деревьев с помощью специального правила стандартной ошибки.

При выборе этих двух правил в рамке **Stopping parameters** (параметры остановки) активизированы опции *Minimum n* (минимальное *n*) и *Standard error rule* (правило стандартной ошибки).

В правиле прямая остановка по методу *FACT* используется совершенно иной подход. Здесь полное дерево классификации, содержащее все возможные ветвления, рассматривается как имеющее «подходящий размер». При данной опции надо определить момент прекращения ветвлений, задавая значение в поле ввода **Fraction of objects** (доля неклассифицированных объектов). Ветвление по предикторным переменным продолжается до тех пор, пока каждая терминальная вершина дерева классификации или не станет «чистой» (т.е. не будет содержать

неправильно классифицированных объектов-наблюдений), или количество отнесенных к этой вершине объектов из одного или нескольких классов не станет меньше заданной доли от общей численности соответствующего класса (классов).

Опции *Stopping parameters* позволяют управлять временем остановки ветвлений и выбором усеченного дерева «подходящего размера». Для этого предусмотрено 3 параметра: *Minimum n, Standard error rule* и *Fraction of objects*.

Minimum n. Если в качестве правила остановки был выбран метод *Prune on misclassification error*, значение, которое пользователь задает в поле ввода **Minimum n**, используется для управления моментом, когда прекращается выбор ветвлений и начинается отсечение. Ветвление по предикторным переменным продолжается до тех пор, пока все терминальные вершины дерева классификации не станут «чистыми» (т.е. не будут содержать неправильно классифицированных наблюдений) или будут содержать не более чем заданное минимальное число неправильно классифицированных наблюдений. Усечение дерева начинается, когда все терминальные вершины будут удовлетворять этому критерию.

Standard error rule. Если в качестве правила остановки был выбран метод *Prune on misclassification error*, значение, которое пользователь задает в поле ввода **Standard error rule**, используется для выбора дерева классификации «подходящего размера» из последовательности усекаемых деревьев. Правило стандартной ошибки действует следующим образом.

В последовательности деревьев ищется усеченное дерево с наименьшей ценой кросс-проверки. Соответствующее значение обозначим через *Min. CV*, а стандартную ошибку цены кросс-проверки для этого дерева — через *Min. SE*. Затем в качестве дерева «подходящего размера» в последовательности деревьев возьмем то, у которого наименьшее число терминальных вершин с ценой кросс-проверки, не превышающей *Min. CV* плюс значение поля *Standard error rule*, умноженное на *Min. SE*. Малые (близкие к нулю) значения параметра *Standard error rule* приводят, как правило, к тому, что выбранное дерево «подходящего размера» оказывается лишь чуть-чуть «проще» (по числу терминальных вершин), чем дерево с наименьшей ценой кросс-проверки. Большие же значения параметра *Standard error rule* приводят, как правило, к тому, что выбранное дерево «подходящего размера» оказывается значительно «проще» (по числу терминальных вершин), чем дерево с наименьшей ценой кросс-проверки. Таким образом, отсечение по цене/сложности находит усеченное дерево с наименьшим числом терминальных вершин, имеющее цену кросс-проверки, не превышающую *Min. CV* плюс значение поля **Standard error rule**, умноженное на *Min. SE*.

Fraction of objects. Если в качестве правила остановки выбрано *FACT-style direct stopping*, значение параметра, введенного в поле **Fraction of objects,** используется для управления выбором дерева классификации «подходящего размера». Ветвление по предикторным переменным продолжается до тех пор, пока все терминальные вершины дерева классификации не станут «чистыми» или пока количество объектов из прогнозируемого класса для данной вершины не станет меньше заданной доли от общего числа объектов этого класса. Минимальное количество наблюдений для каждого класса значений зависимой переменной, вычисленное

исходя из этой доли, можно увидеть в таблице результатов при помощи кнопки **Class minimum objects** (минимум объектов в классе), которая доступна через окно диалога **Classification Trees-Results**, на вкладке **Predicted classes** (предсказанные классы.).

4. С определением момента, когда дальнейшие ветвления следует прекратить, непосредственно связан *четвертый этап* — определение «подходящих размеров» дерева. Очевидно, что чем больше размерность дерева классификации, тем точнее прогноз. Но сложнее интерпретация результатов и решающие правила, поэтому пользователю труднее сделать прогноз о принадлежности к классу нового наблюдения. Можно высказать ряд общих соображений о том, что следует считать «подходящими размерами» для дерева классификации. Дерево классификации должно быть достаточно сложным для того, чтобы учитывать имеющуюся информацию, и в то же время оно должно быть как можно более простым для возможности интерпретировать результаты. Дерево должно уметь использовать ту информацию, которая улучшает точность прогноза, и игнорировать ту, которая прогноза не улучшает.

Одна из возможных стратегий выбора размера дерева состоит в том, чтобы наращивать его до нужного размера, который определяется самим пользователем на основе уже имеющихся данных, диагностических сообщений системы, выданных на предыдущих этапах анализа, или, на крайний случай, интуиции.

Другая стратегия связана с использованием хорошо структурированного и документированного набора процедур для выбора «подходящего размера» дерева, разработанных Бриманом (*Breiman*). Благодаря опциям модуля **Classification Trees** можно использовать любую (или одновременно обе) из двух различных стратегий выбора дерева «подходящего размера» среди всех возможных деревьев.

Начнем с описания первой стратегии, в которой пользователь сам устанавливает размеры дерева, до которых оно может расти. В этом варианте в качестве правила остановки надо выбрать опцию *Fact-style direct stopping*, а затем задать при помощи опции *Fraction of objects* долю неклассифицированных, которая позволяет дереву расти до нужного размера.

В программе предусмотрено 3 способа оценки того, насколько удачно выбран пользователем размер дерева — 3 варианта кросс-проверки для построенного дерева классификации:

а) кросс-проверка на тестовой выборке — наиболее предпочтительный вариант кросс-проверки. В этом варианте кросс-проверки дерево классификации строится по исходной — обучающей выборке, а его способность к прогнозированию проверяется путем предсказания классовой принадлежности элементов тестовой выборки. Если значение цены на тестовой выборке окажется больше, чем на обучающей выборке, то это свидетельствует о плохом результате кросс-проверки. Возможно, в этом случае следует поискать дерево другого размера, которое бы лучше выдерживало кросс-проверку. Цена — это доля неправильно классифицированных наблюдений при условии, что были использованы оцениваемые априорные вероятности, а цены ошибок классификации были взяты одинаковыми. Тестовая и обучающая выборки могут быть образованы из двух независимых

наборов данных, или, если в нашем распоряжении имеется большая обучающая выборка, мы можем случайным образом отобрать часть (например, треть или половину) наблюдений и использовать ее в качестве тестовой выборки. В модуле **Classification Trees** кросс-проверка на тестовой выборке выполняется с помощью введения специальной переменной — идентификатора *(sample identifier)* выборки, содержащей код выборки (обучающая или тестовая), к которой относится данный объект.

Можно открыть таблицу результатов **Misclassification matrix** на вкладке **cross-validation** окна **Classification Trees Results** в рамке **Test sample**, содержащую матрицу ошибок классификации на тестовой выборке, чтобы узнать, какое количество объектов каждого класса (столбцы) ошибочно отнесено к другому классу (строки). Эта таблица результатов содержит также цену кросс-проверки и стандартное отклонение цены кросс-проверки, вычисленные по результатам классификации на тестовой выборке. Другая таблица результатов — **Predicted classes** (предсказанные классы) для каждого объекта (наблюдения) из тестовой выборки показывает его номер (или имя объекта, если оно используется), его класс, его прогнозируемый класс и терминальную вершину дерева классификации, к которой приписан этот объект;

б) *v-кратная* кросс-проверка. Этот вид кроєс-проверки целесообразно использовать в случаях, когда в нашем распоряжении нет отдельной тестовой выборки, а обучающее множество слишком мало для того, чтобы из него выделять тестовую выборку. Параметры кросс-проверки задаются на вкладке **Sampling options** (опции выбора, рис. 13.16). Положительное целое число, которое вводится в поле **Seed for random number generator** (начальное значение датчика случайных чисел), является начальным значением для датчика случайных чисел, порождающего заданное число случайных выборок из обучающей выборки. Задаваемое пользователем значение v *(v-fold cross validation)* (*v-кратной* кросс-проверки) по умолчанию равно 3, определяет число случайных выборок — по возможности одинакового объема, которые формируются из обучающей выборки. Дерево классификации нужного размера строится v раз, причем каждый раз поочередно одна из выборок не используется в его построении, но затем используется как тестовая выборка для кросс-проверки. Таким образом, каждая подвыборка $v-1$ раз участвует в обучающей выборке и ровно один раз служит тестовой выборкой. Цены кросс-проверки, вычисленные для всех v тестовых выборок, затем усредняются, и в результате получается *v-кратная* оценка для цены кросс-проверки, которая вместе со своей стандартной ошибкой доступна в таблице результатов **Sequential tree** (последовательность деревьев). Значение, заданное в поле ввода **p-level for split variable selection** (*p-уровень*), используется в процессе ветвления, если выбран *Discriminant-based* (дискриминантный тип ветвления). Он определяет, по какому критерию выбирается переменная для ветвления: по значимости *F-Левена* (статистический критерий в дисперсионном анализе, устойчивый к виду распределения) или по значимости стандартного одномерного *F*;

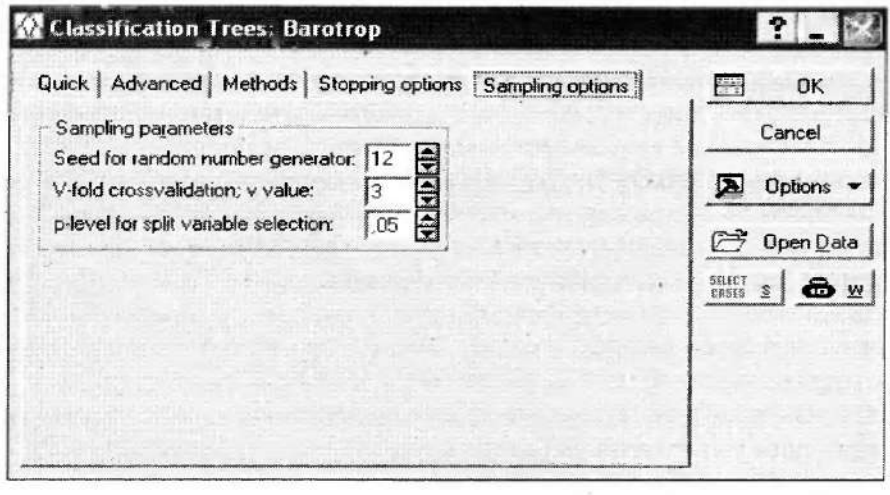


Рис. 13.16

в) глобальная кросс-проверка. В этом варианте производится заданное число итераций, причем всякий раз часть обучающей выборки (равная единице, деленной на заданное целое число) оставляется в стороне, а затем по очереди каждая из отложенных частей используется как тестовая выборка для кросс-проверки построенного дерева классификации. По умолчанию число итераций равно 3. Этот вариант кросс-проверки, вероятно, уступает методу *v-кратной* кросс-проверки в случае, если была выбрана опция *Fact-style direct stopping*, однако он может оказаться очень полезным для проверки методов автоматического построения дерева. Для того чтобы осуществить глобальную кросс-проверку, надо в окне **Classification Trees Results** на вкладке **cross-validation** в рамке **Learning sample** (обучающая выборка) в поле **v-vold for GCV** (число выборок для ГКП) задать число итераций и нажать кнопку **Perform global CV** (выполнить кросс-проверку, рис. 13.17).

Вторая из возможных стратегий выбора «подходящего размера» для дерева — метод автоматического построения дерева Бримана, который реализован кросс-проверочным отсечением либо по минимальной цене-сложности, либо по минимальному отклонению-сложности. Единственное различие между этими двумя опциями — способ измерения ошибки прогноза. При первой опции используется функция потерь, равная доли неправильно классифицированных объектов при оцениваемых априорных вероятностях и одинаковых ценах ошибок классификации. При второй опции используется мера, основанная на принципе максимума правдоподобия и называемая отклонением.

Для того чтобы в модуле **Classification Trees** выполнить кросс-проверочное отсечение по минимальной цене-сложности, нужно выбрать опцию *Prune on misclassification error* в качестве правила остановки. Кросспроверочное отсечение по минимальному отклонению-сложности выполняется, если в качестве правила остановки выбрана опция *Prune on deviance*. Эти опции диалога (рис. 13.15) были подробно описаны ранее. Функция цены, которая требуется

для кросс-проверочного отсечения по минимальной цене-сложности, вычисляется по мере построения дерева, начиная с ветвления в корневой вершине, пока дерево не достигнет максимально допустимого размера, определяемого величиной *Minimum n*.

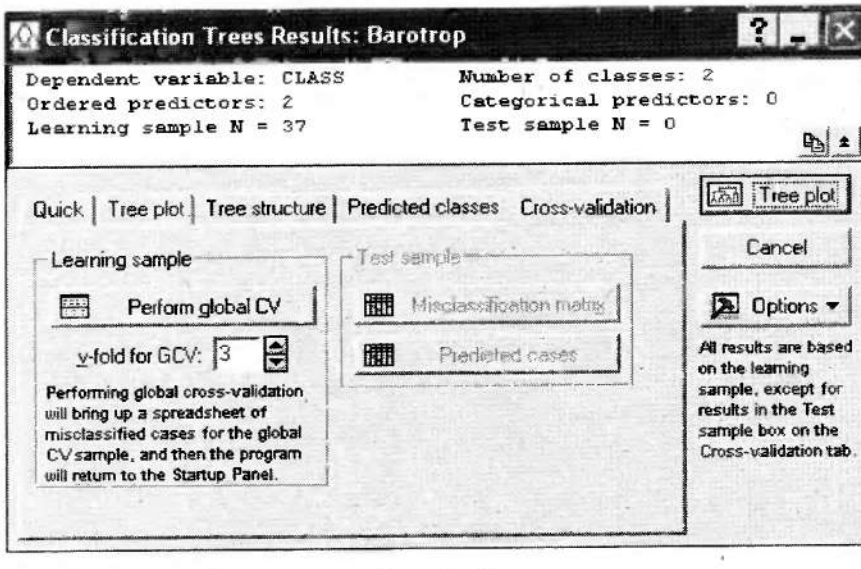


Рис. 13.17

Цена для обучающей выборки пересчитывается при каждом новом ветвлении дерева, так что в результате получается, вообще говоря, убывающая последовательность цен (это отражает улучшение качества классификации). Цена обучающей выборки называется ценой обучения, чтобы отличать ее от цены кросс-проверки. Это необходимо делать, поскольку *v-кратная* кросс-проверка также производится при каждом новом ветвлении дерева [6]. В качестве значения цены для корневой вершины следует использовать оценку цены кросс-проверки из *v-кратной* кросс-проверки. Размер дерева можно определить как число терминальных вершин, потому что для бинарных деревьев при каждом новом ветвлении размер дерева увеличивается на единицу. Введем так называемый параметр сложности. Положим его сначала равным нулю, и для каждого дерева (начиная с исходного, представленного одной вершиной) будем вычислять функцию, равную цене дерева плюс значение параметра сложности, умноженное на размер дерева. Станем теперь постепенно увеличивать значение параметра сложности, пока значение этой функции для максимального дерева не превысит ее значения для какого-либо из деревьев меньшего размера, построенных на предыдущих шагах. Примем это меньшее дерево за новое максимальное дерево и будем дальше увеличивать значение параметра сложности, пока значение функции для этого дерева не станет больше ее значения для какого-то еще меньшего дерева. Будем продолжать этот процесс до тех пор, пока дерево с единственной корневой вершиной не станет максимальным. В этом алгоритме используется так называемая

штрафная функция (параметр сложности). Она представляет собой линейную комбинацию цены, которая в общем случае убывает с ростом дерева, и размера дерева, который линейно растет. По мере того как значение параметра сложности увеличивается, большие по размеру деревья получают все больший штраф за свою сложность, пока не будет достигнуто пороговое значение, при котором более высокая цена меньшего дерева будет перевешиваться сложностью большего дерева. Последовательность максимальных деревьев, которая получается в процессе выполнения этого алгоритма, обладает рядом замечательных свойств. Они являются вложенными, поскольку при последовательном усечении каждое дерево содержит все вершины следующего (меньшего) дерева в последовательности. Поначалу при переходе от очередного дерева к последующему отсекается, как правило, большое число вершин, однако по мере приближения к корневой вершине на каждом шаге будет отсекаться все меньше вершин. Деревья последовательности усекаются оптимально в том смысле, что каждое дерево в последовательности имеет наименьшую цену среди всех деревьев такого же размера.

Выберем теперь из последовательности оптимально усеченных деревьев дерево «подходящего размера». Естественным критерием здесь является цена кросс-проверки. Не будет никакой ошибки, если в качестве дерева «подходящего размера» выберем то, которое дает наименьшую цену кросс-проверки, однако часто оказывается, что есть еще несколько деревьев с ценой кросс-проверки, близкой к минимальной. Бриман и другие высказывают разумное предложение, что в качестве дерева «подходящего размера» нужно брать наименьшее (наименее сложное) из тех, чьи цены кросс-проверки несущественно отличаются от минимальной. Авторы предложили правило «*1 × SE*» — в качестве дерева «подходящего размера» нужно брать наименьшее дерево из тех, чьи цены кросс-проверки не превосходят минимальной цены кросс-проверки плюс умноженная на единицу стандартная ошибка цены кросс-проверки для дерева с минимальной ценой кросс-проверки.

SE-правило модуля «Деревья классификации» позволяет кроме единицы (значение по умолчанию) выбирать и другие значения для множителя. Так, если взять его значение равным 0, то в результате в качестве дерева «подходящего размера» мы получим дерево с наименьшей ценой кросс-проверки. Значения, большие 1, дадут в качестве дерева «подходящего размера» дерево, значительно меньшее по размеру, чем дерево с наименьшей ценой кросс-проверки. Существенное преимущество «автоматического» выбора дерева состоит в том, что оно позволяет избежать как «недо-», так и «пересогласованности» с данными. Процедура «автоматического» выбора дерева направлена на то, чтобы выбирать наиболее простое (наименьшее по размеру) дерево с ценой кросс-проверки, близкой к минимальной, и тем самым избегать потери точности прогноза.

Кросс-проверочное отсечение по минимальной цене-сложности и последующий выбор дерева «подходящего размера» — действительно «автоматические» процедуры. Алгоритм самостоятельно принимает все решения, необходимые для выбора дерева «подходящего размера», за исключением разве что выбора множителя в *SE-правиле*. В связи с этим возникает вопрос о том, насколько хорошо воспроизводятся

результаты, т.е. не может ли получиться так, что при повторении этого процесса «автоматического выбора» будут строиться деревья, сильно отличающиеся друг от друга по размеру. Именно здесь очень полезной может оказаться глобальная кросс-проверка. При глобальной кросс-проверке все этапы анализа повторяются заданное число раз (по умолчанию — 3), и при этом часть наблюдений используется как тестовая выборка для кросс-проверки полученного дерева классификации. Если средняя цена тестовых выборок, которая называется ценой глобальной кросс-проверки, превышает цену кросс-проверки выбранного дерева или если стандартная ошибка цены глобальной кросс-проверки превышает стандартную ошибку цены кросс-проверки для выбранного дерева, то это свидетельствует о том, что процедура «автоматического» выбора дерева вместо устойчивого выбора дерева с минимальным оцененным значением цены дает недопустимо большой разброс результатов.

13.5. Примеры анализа модулем *Classification Trees*

Пример 1. Из библиотеки **Example** откройте файл данных **Barotrop**. В файле приведены данные о координатах — долготе (*LONGITUDE*) и широте (*LATITUDE*) для 37 циклонов, достигающих силы урагана, по двум классификациям (*CLASS*) циклонов — бароклинным (*BARO*) и тропическим (*TROP*).

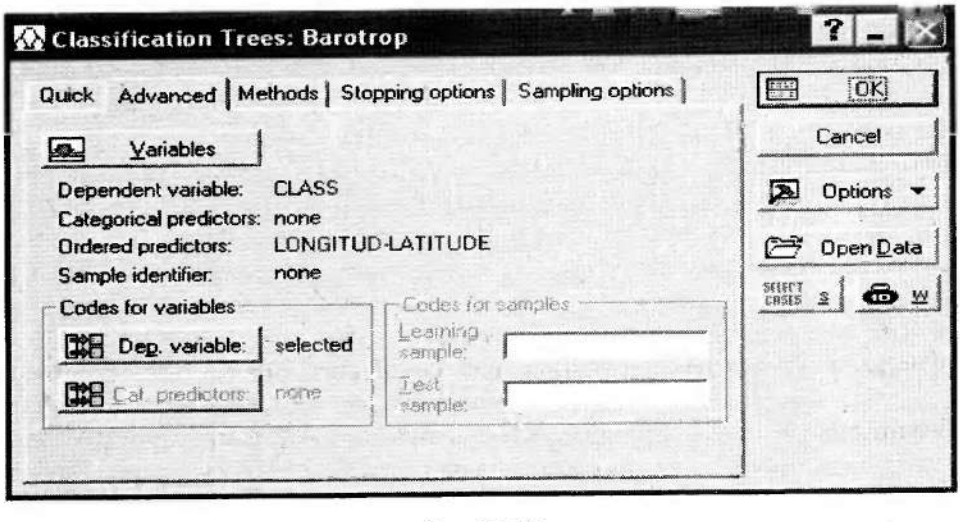


Рис. 13.18

В верхнем меню **Statistics** щелкните по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выберите команду **Classification Trees**. Откроется стартовая панель модуля (рис. 13.18), в котором на вкладке **Advanced** нажмите кнопку**Variables**, во вновь открывшемся окне **Select the dependent...** в поле **Dependent variable** высветите переменную *CLASS*, в поле

Ordered predictors — *LONGITUDE, LATITUDE.* Файл не содержит категориальных предикторов *(Categorical predictors)* и нет переменной-идентификатора *(Sample Identifier)*, поэтому в соответствующих полях не высвечиваем переменные. Нажмите **OK**. Программа вернется в предыдущее окно, щелкнув по **Dep. Variables**, задайте коды зависимой переменной — *BARO, TROP.*

Для выбора метода ветвления перейдите на вкладку **Methods** и выберите в рамке **Split selection methods** тип ветвления — *C&RT-style exhaustive search for univariate splits*, в рамке **Misclassif. costs** — цену неправильной классификации *Equal*, в рамке **Prior probabilities** — априорные вероятности *Estimated*, в рамке **Goodness of fit** — критерий согласия *Gini measure* (рис. 13.15).

Для выбора момента, когда следует прекратить ветвление, перейдите на вкладку **Stoping options** и в рамке **Stopping rule** выберите в качестве правила остановки — отсечение по ошибке классификации *(Prune on misclassification error)*, а в рамке **Stopping parameters** — минимальное число неправильно классифицированных наблюдений в терминальной вершине *(Minimum n)* = 5 (рис.13.15).

Для выбора дерева классификации «подходящего размера» из последовательности усеченных деревьев в рамке **Stopping parameters** задайте значение параметра «правило стандартной ошибки» *(Standard error rule)*, равное 1. Напомним, что сделанные установки для выбора размера дерева классификации соответствуют второй из возможных стратегий выбора «подходящего размера» для дерева — методу автоматического построения дерева Бримана. Щелкните по **OK.** Откроется окно **Classification Trees Results** (рис. 13.19).

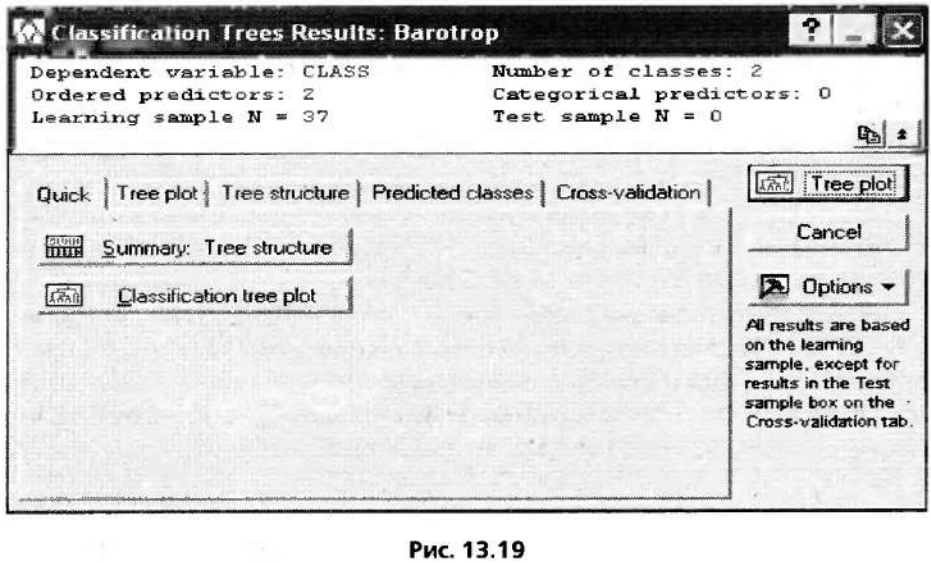


Рис. 13.19

В информационной части окна указано название зависимой переменной, количество интервальных переменных (2), объем обучающей выборки (37), объем тестовой выборки (0). На вкладке *Quick* представлены две кнопки — **Summary. Tree structure** и **Classification tree plot.**

Нажмите последовательно на каждую из них. На рис. 13.20 в таблице представлены номера узлов (вершин) (*Node*); номера дочерних вершин (*Child nodes*) на левой и правой ветвях (*Left, Right branch*); исходное количество объектов (*Observed*) в классах; предсказанные классы (*Predicted classes);* условия ветвления (*Split conditions).*

	Tree Structure (Barotrop) Child nodes, observed class n's, predicted class, and split condition for each node						
Node	Left branch	Right branch	n in cls BARO	n in cls TROP	Predict. class	Split constant	Split variable
1	2	3	19	18	BARO	-67,7500	LONGITUD
2	4	5	9	18	TROP	-62,5000	LONGITUD
3			10	0	BARO		
4			9	0	BARO		
5			0	18	TROP		

Рис. 13.20

Так, например, из таблицы следует, что левая ветвь содержит два узла под номерами 2 и 4, правая — два узла под номерами 3, 5. Далее, из строки 1 таблицы следует, что в первой вершине все ураганы — 19 *BARO* и 18 *TROP* классифицированы (предсказаны) как *BARO*. Из вершины 1 выходят две ветви (правая и левая) с соответствующими вершинами 2 и 3.

Условие разделения ураганов по вершинам 2 и 3 следующее: если значение переменной *LONGITUDE* <= (меньше либо равно) 67,75, то ураганы классифицируются как *TROP*, в противном случае — как *BARO*. Из строк 2 и 3 следует, что по данному условию 9 ураганов *BARO* и 18 ураганов *TROP* классифицированы как *TROP*, а 10 ураганов *BARO* классифицированы как *BARO*. Из строк 4 и 5 следует, что при помощи правила *LONGITUDE* <= 62,5, 9 ураганов *BARO* и 18 ураганов *TROP* классифицированы соответственно как *BARO* и *TROP*.

Интерпретация результатов таблицы значительно упрощается, если воспользоваться графом дерева классификации, изображенным на рис. 13.21.

Красным пунктиром (на мониторе) обозначены терминальные вершины (листья) дерева (3–5), дальнейшее ветвление из которых невозможно. Заметим, что все имеющиеся на рисунке метки и легенды являются пользовательским текстом, поэтому их можно редактировать, перемещать или удалять. Внутри больших прямоугольников, изображающих вершины дерева, нарисованы гистограммы, высота которых соответствует количеству ураганов определенного класса в данной вершине.

Пользователь по своему усмотрению может выбрать те или иные параметры графа дерева. Перейдите на вкладку **Tree plot**. Откроется окно, изображенное на рис. 13.22.

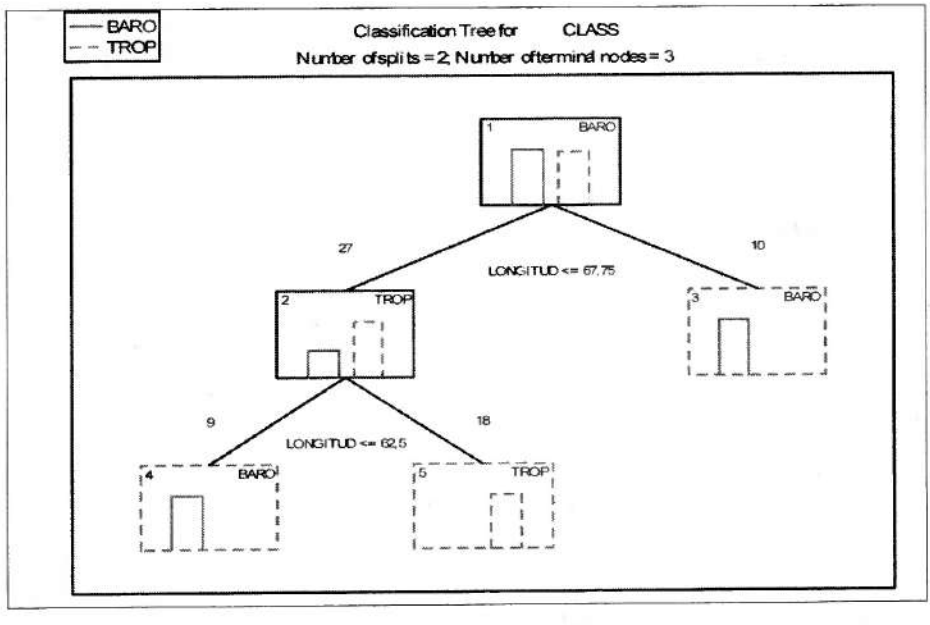


Рис. 13.21

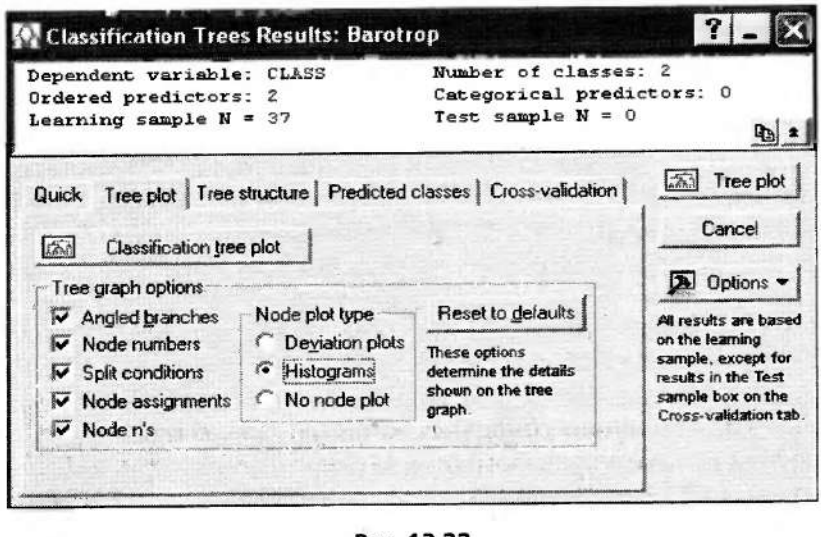


Рис. 13.22

Рассмотрим основные установки этого окна:

- если на вкладке **Tree plot** в рамке **Tree graph options** (опции графа дерева классификации) отмечена галочкой позиция *Angled branches* (диагональные ветви), то граф дерева изображается с диагональными ветвями. В противном случае линии ребер (ветвей) графа проводятся по горизонталям и вертикалям;

- если обозначена опция *Node number* (номера вершин), то около каждой вершины графа дерева пишется номер вершины в качестве метки;
- если обозначена опция *Split conditions* (условия ветвления), то около каждой нетерминальной вершины пишется соответствующее ей условие ветвления. В случае одномерного ветвления по порядковой предикторной переменной условие записывается так: имя переменной, знак «меньше или равно» и взятая с обратным знаком константа функции ветвления. В случае одномерного ветвления по категориальной предикторной переменной условие записывается так: имя переменной, знак равенства и категория (или категории) предикторной переменной, удовлетворяющая условию ветвления. В случае ветвления по линейной комбинации порядковых предикторных переменных в условии записывается следующее: *F(0)* (сокращенное обозначение линейной комбинации предикторных переменных), затем знак «меньше или равно» и взятая с обратным знаком константа функции ветвления. Для всех других видов ветвлений в условии ветвления указывается, что все наблюдения (объекты), удовлетворяющие условию ветвления, направляются в левую дочернюю вершину; те же наблюдения (объекты), которые не удовлетворяют условию ветвления, направляются в правую дочернюю вершину;
- если обозначена опция *Node assignments* (приписанные классы), то граф дерева изображается с метками классов, приписанных вершинам, которые указывают прогнозируемый класс для объектов этой вершины;
- если обозначена опция *Node n's* (*n* вершины), то около каждой ветви графа дерева в виде пользовательского текста выводится информация о количестве наблюдений (объектов), направленных по этой ветви;
- если в рамке *Node plot type* (тип диаграммы для вершин дерева классификации) обозначена опция *Deviation plots* (диаграммы отклонений), то внутри каждой вершины графа дерева рисуется диаграмма, аналогичная двумерной столбчатой диаграмме. Когда все наблюдения, попавшие в вершину, расклассифицированы правильно, столбик, соответствующий прогнозируемому классу, будет направлен вверх, а столбики остальных классов — вниз, так что удачную классификацию легко можно распознать визуально;
- если обозначена опция *Histograms*, то внутри каждой вершины графа дерева рисуется диаграмма, аналогичная двумерной гистограмме. На этой диаграмме количество объектов определенного класса в данной вершине изображается столбиком соответствующей высоты. Когда все наблюдения, попавшие в эту вершину, расклассифицированы правильно, столбик, соответствующий прогнозируемому классу для этой вершины, будет высоким, а столбики остальных классов — маленькими;
- если обозначена опция *No node plot* (вершины без диаграмм), то граф дерева изображается без диаграмм вершин.

Кнопка **Reset to defaults** (установки по умолчанию) восстанавливает значения опций графа дерева, принятых по умолчанию.

Следует проявлять аккуратность, если в выбранных переменных есть пропущенные данные, так как программа произведет построчное удаление пропущенных значений, т.е. из дальнейшего анализа будут исключены все наблюдения, у которых отсутствует значение хотя бы одной из переменных, выбранных для анализа. Результаты, которые будут таким образом получены для переменных, не содержащих пропусков, будут использовать не всю доступную информацию (а именно: не будут учитываться наблюдения, в которых какая-то другая переменная имеет пропущенное значение).

На вкладке **Tree structure** расположены следующие опции (рис. 13.23).

Classification tree structure — таблица результатов, в которой выведена вся информация о графе дерева.

Child nodes — таблица результатов, где для каждой вершины ветвления в выбранном дереве указаны соответствующая левая и правая дочерние вершины, к которым относятся объекты (наблюдения), если они соответственно удовлетворяют или не удовлетворяют условию ветвления в данной вершине. Для терминальных вершин дерева дочерние вершины не указываются.

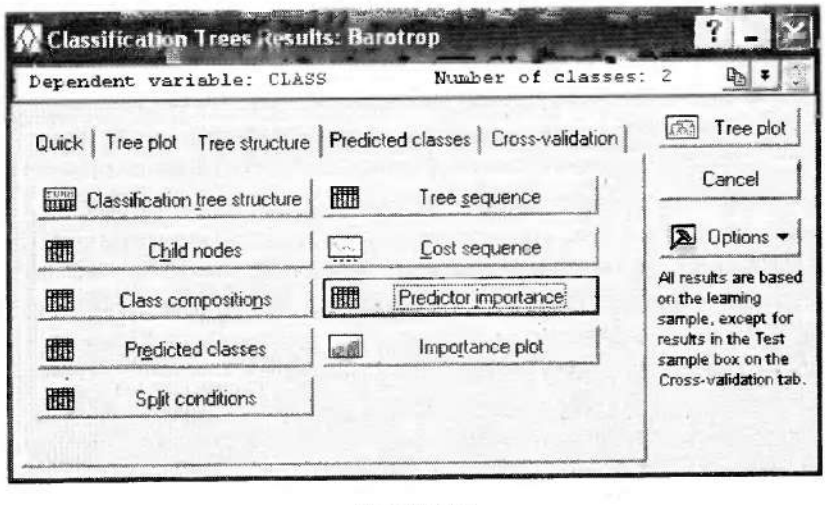


Рис. 13.23

Class compositions (структура классов) — таблица результатов, в которой для каждой вершины выбранного дерева приводится та же информация, что и в таблице результатов *Child nodes*, и дополнительно — данные о числе объектов каждого класса, отнесенных к этой вершине.

Predicted classes — таблица результатов, в которой для каждой вершины выбранного дерева приводится та же информация, что и в таблице результатов *Class compositions*, и дополнительно (появляется новый столбец — *Predicted classes*) — название класса, к которому приписываются объекты (наблюдения), попавшие в эту вершину.

Split conditions (условия ветвления) — таблица результатов, в которой для каждой вершины ветвления в выбранном дереве указано соответствующее ей условие

ветвления. Если в качестве типа ветвления было выбрано дискриминантное одномерное ветвление или полный перебор деревьев с одномерным ветвлением по методу *CART*, то условия ветвления выдаются как *Split constant* (постоянная ветвления) и *Split variable* (переменная ветвления) (рис. 13.24). Если переменная ветвления является категориальным предиктором, то будет выдана *Split category* (категория ветвления), а если переменная ветвления является порядковым предиктором, будет выдан *Split coefficient* (коэффициент ветвления). В случае если в качестве типа ветвления было выбрано дискриминантное многомерное ветвление по линейным комбинациям, для каждой из порядковых предикторных переменных будут выданы *Split constant* и *Split coefficient*. Для терминальных вершин дерева никакой информации об условиях ветвления не выдается.

	Split conditions (Barotrop Split condition for each node	
Node	Split constant	Split variable
1	-67,7500	LONGITUD
2	-62,5000	LONGITUD
3		
4		
5		

Рис. 13.24

Tree sequence (последовательность деревьев) — таблица результатов для последовательности деревьев. Если в качестве правила остановки было выбрано отсечение по ошибке классификации или по вариации, то в таблице результатов выводятся *Terminal nodes* (терминальные вершины); *CV cost* (цена кросс-проверки); *Std. error* (ее стандартная ошибка); *Resub. Cost* (цена обучения); *Node complexity* (сложность каждого из усеченных деревьев) (рис. 13.25, 13.26).

	Tree Sequence (Barotrop) Statistics for successive trees Selected tree denoted by *				
Tree number	Terminal nodes	CV cost	Std. error	Resub. cost	Node complxty
*1	3	0,000000	0,000000	0,000000	0,000000
2	1	0,486486	0,082169	0,486486	0,243243

Рис. 13.25

	Tree Sequence (Barotrop) Statistics for successive trees Selected tree denoted by *				
Tree number	Terminal nodes	CV cost	Std. error	Resub. cost	Node complxty
*1	3	0,000000	0,000000	0,00000	0,00000
2	1	0,486486	0,082169	51,26586	25,63293

Рис. 13.26

Tree number	Tree Sequence (Barotrop) Statistics for successive trees Selected tree denoted by *				
	Terminal nodes	CV cost	Std. error	Resub. cost	Node complxty
*1	3	0,00	0,00	0,00	0,00

Рис. 13.27

Звездочкой помечено дерево, которое было признано деревом «подходящего размера». Если в качестве правила остановки была выбрана прямая остановка по методу *FACT*, то вся указанная информация выводится для выбранного дерева «подходящего размера» (рис. 13.27).

Cost sequence (график последовательности цен) – раскрывается окно с графиком последовательности цен. Если было выбрано правило остановки отсечением по ошибке классификации или по вариации, на графике изображаются цена кросс-проверки и цена обучения для каждого дерева из последовательности усеченных деревьев. Выбранное усеченное дерево «подходящего размера» помечено звездочкой (рис. 13.28).

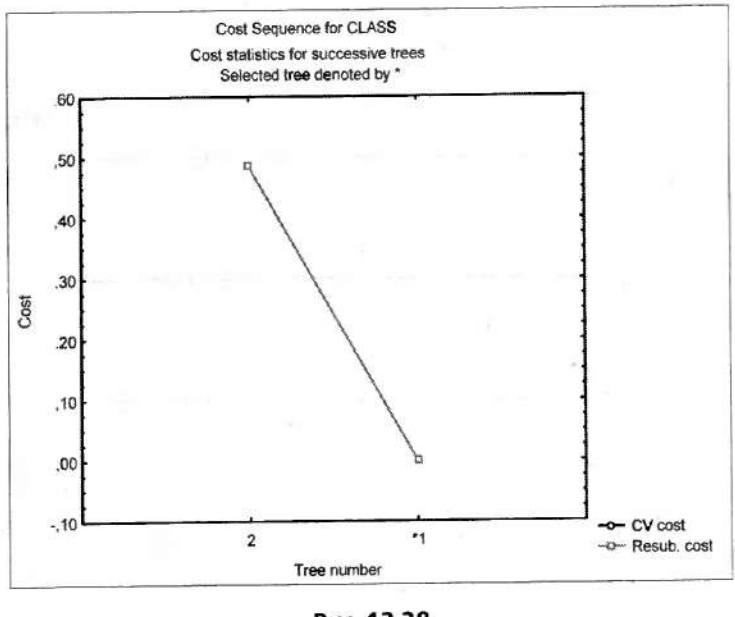


Рис. 13.28

Если в качестве правила остановки была выбрана прямая остановка по методу *FACT*, то эта информация выводится для выбранного дерева «подходящего размера».

Predictor importance (значимость предикторов) доступна, когда выбрано дискриминантное одномерное ветвление или полный перебор деревьев с одномерным ветвлением по методу *CART*. При нажатии этой кнопки раскрывается таблица результатов, в которой для каждой из анализируемых предикторных переменных устанавливается ранг ее значимости по 100-балльной шкале (рис. 13.29). Из данной таблицы и решающих правил на графе дерева классификации (рис. 13.21) следует, что долгота циклонов (*LONGITUDE*) является значимой предикторной переменной, определяющей принадлежность циклонов к классам *BARO* и *TROP*.

Variable	Predictor V Based on 0=low imp Ranking
LONGITUD	100
LATITUDE	2

Рис. 13.29

Importance plot (диаграмма значимости предикторов) изображает столбчатую диаграмму значимости предикторов, где для каждой предикторной переменной показан ее ранг значимости по 100-балльной шкале.

На вкладке **Predicted classes** расположены следующие кнопки (рис. 13.30).

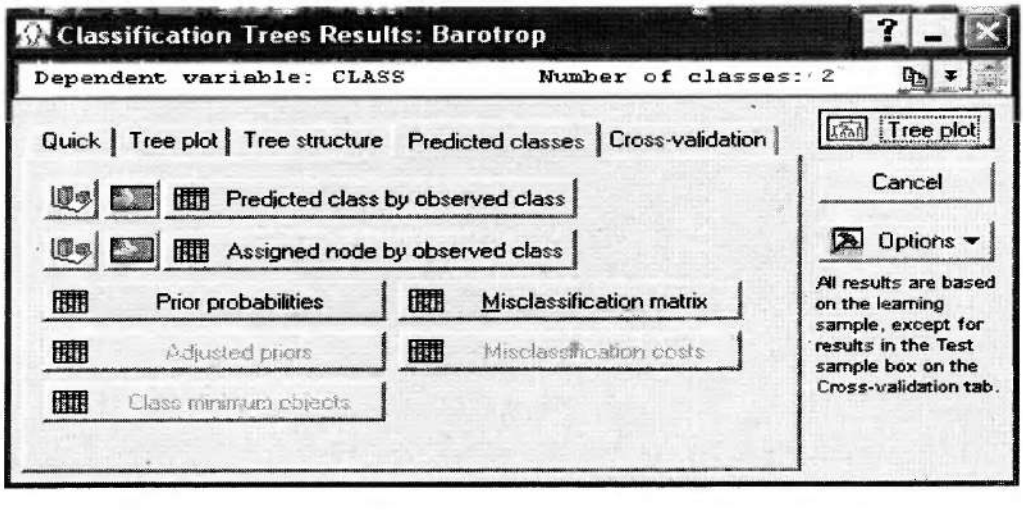


Рис. 13.30

Predicted class by observed class (предсказанные и наблюдаемые объекты в классах). В этой таблице результатов (рис. 13.31) выводится вся информация о том, сколько объектов каждого из наблюдаемых классов значений зависимой переменной отнесено по результатам классификации к тому или иному классу. Также в ней выдается объем обучающей выборки. Исходные классы соответствуют столбцам матрицы, предсказанные классы — строкам. При нажатии на имеющиеся здесь же кнопки (трехмерная гистограмма) или (дискретная карта линий уровня) вся эта информация представляется графически в цветном изображении, и если в задаче много классов, то легче бывает увидеть ошибки классификации (рис. 13.32).

	Predicted Class x Observed Class n's (Barotrop) Predicted (row) x observed (column) matrix Learning sample N = 37	
Class	Class BARO	Class TROP
BARO	19	0
TROP	0	18

Рис. 13.31

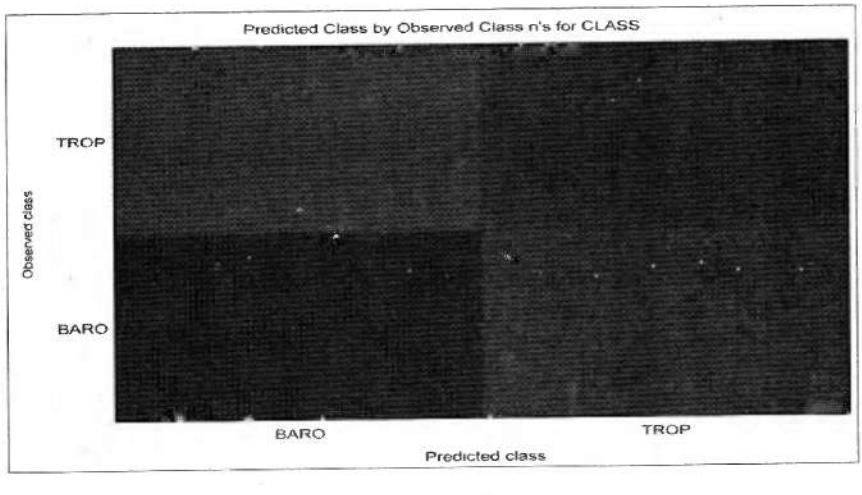


Рис. 13.32

Assigned node by observed class (обозначенные вершины, исходные классы). В таблице выводится вся информация о том, сколько объектов каждого из наблюдаемых классов попало в каждую терминальную вершину. Также в ней выдается объем обучающей выборки. Исходные классы соответствуют столбцам матрицы, предсказанные классы — строкам.

Prior probabilities (априорные вероятности) — таблица результатов априорных вероятностей. Вверху в поле сообщений выдается информация о том, как выбирались априорные вероятности: оценивались по выборке, одинаковые

или пользовательские. Здесь же выдается объем обучающей выборки. В первом столбце таблицы приводятся априорные вероятности для каждого класса зависимой переменной, во втором столбце — число элементов в каждом классе (рис. 13.33).

Adjusted prior (скорректированные априорные вероятности). Кнопка доступна, если ранее была выбрана опция пользовательские цены ошибок классификации. При нажатии этой кнопки раскрывается таблица результатов. В ней выдаются априорные вероятности для каждого класса значений зависимой переменной, скорректированные с учетом пользовательских цен ошибок классификации.

Misclassification matrix (ошибки классификации) — в этой таблице выводится информация о том, сколько объектов каждого из наблюдаемых классов по результатам классификации было ошибочно отнесено к другому классу. Кроме того, в поле сообщений выдается объем обучающей выборки. Исходные классы соответствуют столбцам матрицы, предсказанные классы — строкам (рис. 13.34).

	Class Prior Probat Priors are estimate Learning sample N	
Class	Prior probs.	Class n
BARO	0,513514	19
TROP	0,486486	18

Рис. 13.33

	Learning Samp Predicted (row) Learning samp	
Class	Class BARO	Class TROP
BARO		0
TROP	0	

Рис. 13.34

Misclassification costs (цена ошибок классификации). Кнопка доступна, если ранее в рамке **Misclassif. Costs** на вкладке **Methods** (методы) (рис. 13.14) была выбрана опция пользовательские цены ошибок классификации. Для всех классов значений зависимой переменной (по столбцам) в таблице результатов выводится заданная пользователем цена неправильной классификации объекта определенного класса как объекта другого класса (по строкам).

Class minimum objects (минимум объектов в классе). Кнопка доступна, если в качестве правила остановки была выбрана прямая остановка по методу *FACT*. В поле сообщений таблицы выводится значение доли неклассифицированных, введенное ранее в соответствующее поле ввода стартовой панели. Далее, для каждого класса значений зависимой переменной выдается минимальное n остановки. Минимальное n остановки — это ближайшее целое число к размеру класса, умноженному на долю неклассифицированных и на отношение априорной вероятности данного класса к наименьшей из априорных вероятностей всех классов. Если при этом была выбрана опция пользовательские цены ошибок классификации, то используются скорректированные априорные вероятности.

Для выполнения глобальной кросс-проверки перейдите на вкладку **Cross-validation** (кросс-проверка, рис. 13.35). При выборе опции *v-fold GCV* можно задать число выборок (кратность) кросс-проверки, которое будет применяться для оценки ошибки глобальной кросс-проверки.

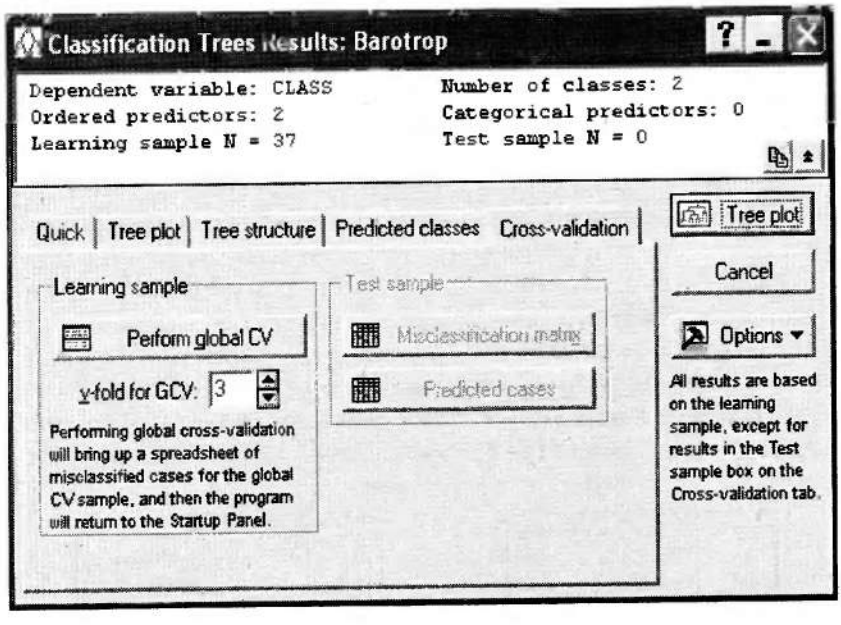


Рис. 13.35

Кнопка **Misclassification matrix** (матрица ошибок классификации) на тестовой выборке доступна, если для анализа была выбрана тестовая выборка. При нажатии этой кнопки раскрывается таблица результатов, в которой выводится информация о том, сколько объектов каждого из наблюдаемых классов по результатам классификации было ошибочно отнесено к другому классу. Исходные классы соответствуют столбцам матрицы, предсказанные классы — строкам. В поле сообщений выводятся *Global CV cost* (цена кросс-проверки) и *s.d.Global CV cost* (стандартное отклонение) цены кросс-проверки, вычисленные по данным классификации на тестовой выборке.

Кнопка **Predicted cases** (предсказанные классы) для тестовой выборки также доступна, если для анализа была выбрана тестовая выборка. Для каждого объекта из тестовой выборки в таблице результатов выдаются его номер (или имя объекта, если оно используется), наблюдаемый класс, прогнозируемый класс и терминальная вершина, к которой этот объект был отнесен. Кроме того, для каждого объекта указаны категории (для категориальных предикторов) или значения (для порядковых предикторов) всех предикторных переменных.

Нажмите кнопку **Perform global CV** (выполнить глобальную кросс-проверку). Запустится процедура кросс-проверки и откроется окно (рис. 13.36). Нажмите кнопку **Global CV Misclassification matrix** (матрица ошибок классификации глобальной кросс-проверки), появится таблица ошибок классификации глобальной кросс-проверки (рис. 13.37). Из данной таблицы следует, что при глобальной кросс-проверке один ураган *BARO* неверно классифицирован как ураган *TROP*, а все ураганы *TROP* классифицированы верно. При этом *Global CV cost* (цена глобальной кросс-проверки) составила 0,02703, *s.d.Global CV cost* (стандартное от-

клонение) цены — 0,02666 и эти величины незначительно отличаютс от цены кросс-проверки (0,0) и ее стандартной ошибки (0,0).

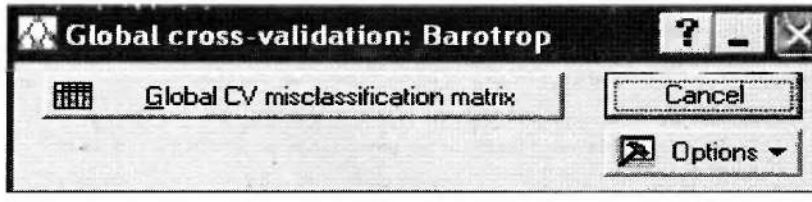


Рис. 13.36

Global CV Sample Misclassification Matrix (Barotrop)
Predicted (row) x observed (column) matrix
Global CV cost = ,02703; s.d. CV cost = ,02666

Class	Class BARO	Class TROP
BARO		1
TROP	0	

Рис. 13.37

Таким образом, из приведенных результатов можно сделать вывод, что для данного файла данных успешно завершилась процедура классификации методом **Classification Trees**. Получено решающее правило, состоящее из двух этапов, которое позволит произвольный циклон классифицировать как ураган *BARO* или *TROP* по значениям долготы.

На первом этапе проверяем справедливость неравенства

$$долгота \leq 67{,}75.$$

Если неравенство выполняется, то циклон является ураганом *BARO*. Если неравенство несправедливо, то циклон может быть ураганом *BARO* или *TROP*.

На втором этапе для циклонов, неклассифицированных на предыдущем этапе как *BARO*, проверяем справедливость неравенства

$$долгота \leq 62{,}5.$$

Если неравенство выполняется, то циклон является ураганом *TROP*. Если неравенство не выполняется, то циклон является ураганом *BARO*.

Пример 2. Рассмотрим работу модуля на примере с использованием метода одномерного дискриминантного ветвления по категориальным и порядковым предикторам. Из библиотеки **Example** откройте файл данных **Boston2**, в котором приведены данные о жилищном строительстве в Бостоне. Цена участка под застройку категориальная, классифицируется как *LOW* (низкая), *MEDIUM* (средняя) или *HIGH* (высокая) в зависимости от значений зависимой переменной *PRICE*. Имеется один категориальный предиктор — *CAT1* и 12 порядковых предикторов — *ORD1–ORD12*. В качестве тестовой выборки используем копию обу-

чающей выборки. Переменная-идентификатор выборки *SAMPLE* имеет код 1 для *LEARNING* (обучающей) и 2 — для *TEST* (тестовой) выборок. Число наблюдений (участков), т.е. объем обучающей выборки, равно 506.

Щелкните по кнопке **Variables** (переменные) и выберите переменную *PRICE* в качестве зависимой переменной, переменную *CAT1* — в качестве категориального предиктора, переменные *ORD1–ORD12* в качестве порядковых предикторов и переменную *SAMPLE* — в качестве идентификатора выборки. На вкладке **Methods** окна **Classification Trees** в рамке **Split selection methods** выберите *Discriminant-based univariate split for categ. and ordered predictors* (дискриминантное одномерное ветвление), в рамке **Prior probabilities** — опцию *Equal*, в рамке **Misclassif. Costs** — также опцию *Equal* (рис. 13.14). На вкладке **Sampling options** задайте *V-fold crossvalidation* (число случайных выборок), равное 10. На остальных вкладках все опции установлены по умолчанию.

Нажмите **OK**. В открывшемся окне **Classification trees results** перейдите на вкладку **Tree structure** и нажмите кнопку **Tree sequence**, появится таблица последовательности деревьев (рис. 13.38). Как видно из этой таблицы, полученное дерево имеет цену кросс-проверки *(CV cost)*, равную 0,262128, и ее стандартную ошибку *(Std. error)* — 0,019424, цену для обучающей выборки *(Resub. cost)* — 0,240978 и «сглаженную» сложность вершины *(Node complxty)* — 0,004269. Деревья с минимальной ценой кросс-проверки (деревья с номерами 23 и 24) имеют цену кросс-проверки, равную 0,248204, со стандартной ошибкой 0,019080, а выбранное дерево является наиболее простым среди тех деревьев, чья цена кросс-проверки не превосходит 0,248204 + 0,019080 = 0,267284. Так, деревья 23, 24 имеют соответственно 25 и 19 терминальных вершин, а дерево 29 — всего 8.

Если теперь нажать на кнопку **Cost sequence** (последовательность цен), то в графическом виде будут представлены последовательности цен обучения и цен кросс-проверки (рис.13.39). Как видно из графика, цена обучения *(Resub. cost)* заметно уменьшается с увеличением размера дерева. В то же время цена кросс-проверки *(CV cost)* с ростом размера дерева быстро достигает минимума, а затем — для очень больших размеров дерева — начинает расти. Выбранное дерево «подходящего размера» располагается близко к точке перегиба этой кривой, т.е. близко к той точке, где первоначальное резкое уменьшение цены кросс-проверки начинает сходить на нет.

Процедура «автоматического» отбора дерева выбрала относительно простое (т.е. маленькое) дерево с ценой кросс-проверки, близкой к минимуму. Тем самым мы избежали потери точности прогноза, которая имела бы место в случае, если бы в качестве дерева «подходящего размера» было выбрано слишком маленькое или слишком большое дерево.

Перейдите на вкладку **Cross — validation**, в рамке **Test sample** нажмите кнопку **Misclassification matrix** (ошибки классификации на тестовой выборке), появится таблица результатов (рис. 13.40).

Tree number	Tree Sequence (Boston2) Statistics for successive trees Selected tree denoted by * Terminal nodes	CV cost	Std. error	Resub. cost	Node complxty
1	87	0,292544	0,019607	0,111515	0,000000
2	86	0,292544	0,019607	0,111527	0,000012
3	84	0,292613	0,019657	0,111596	0,000035
4	83	0,282979	0,019586	0,111665	0,000069
5	81	0,279126	0,019549	0,111828	0,000081
6	80	0,279126	0,019549	0,111966	0,000138
7	77	0,279126	0,019549	0,113893	0,000642
8	74	0,279126	0,019549	0,115901	0,000669
...	...	...	...	...	...
21	29	0,253764	0,019096	0,174906	0,002048
22	28	0,249980	0,019064	0,176971	0,002065
23	25	0,248204	0,019080	0,184967	0,002665
24	19	0,248204	0,019080	0,200971	0,002667
25	17	0,251988	0,019123	0,206751	0,002890
26	15	0,255980	0,019244	0,212775	0,003012
27	14	0,258090	0,019336	0,216629	0,003854
28	9	0,261944	0,019404	0,236709	0,004016
*29	8	0,262128	0,019424	0,240978	0,004269
30	7	0,280326	0,019936	0,248478	0,007499
31	5	0,286131	0,020057	0,264867	0,008195

Рис. 13.38

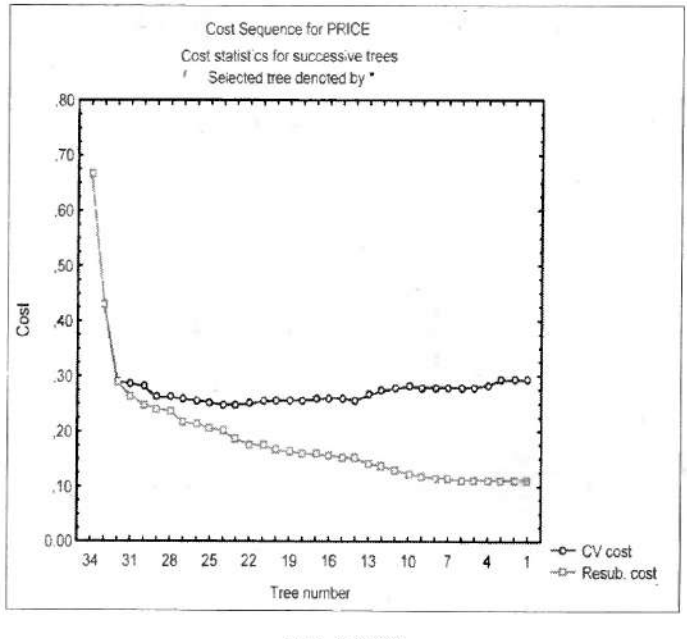


Рис. 13.39

Class	Test Sample Misclassific Predicted (row) x observ CV cost = ,24098; s.d. C		
	Class LOW	Class MEDIUM	Class HIGH
LOW		14	2
MEDIUM	48		29
HIGH	1	28	

Рис. 13.40

В этой таблице для каждого наблюдаемого класса в тестовой выборке показано число наблюдений, которые были ошибочно отнесены к одному из двух других классов (48 наблюдений класса *MEDIUM* были ошибочно отнесены к классу *LOW*, а 14 наблюдений класса *LOW* были ошибочно отнесены к классу *MEDIUM* и т.д.). Можно заметить, что большинство ошибок связано с определением наблюдений среднего класса. В поле сообщений этой таблицы результатов выдаются цена кросс-проверки *CV cost* (0,2409) и ее стандартное отклонение *s.d. CV cost* (0,0189), рассчитанные по ошибкам классификации на тестовой выборке. Поскольку тестовая выборка является точной копией обучающей, цена кросс-проверки на тестовой выборке совпадает с ценой для обучающей выборки *(Resub. cost)*.

Теперь на этой же вкладке введите в поле **v-fold GCV** (число выборок), используемых для глобальной кросс-проверки, значение 10 и нажмите кнопку **Perform global CV** (выполнить глобальную кросс-проверку). На экран будет выдана таблица результатов *Global CV Sample Misclassification matrix* (матрица ошибок классификации глобальной кросс-проверки, рис. 13.41). Обратите внимание, что количество ошибок классификации, расположенных в соответствующих ячейках матрицы на рис. 13.40, 13.41, достаточно близки (за исключением значений 28 и 42). Значения цены глобальной кросс-проверки (0,2814) и ее стандартного отклонения (0,01973) также близки к значениям цены кросс-проверки (0,2621) и ее стандартной ошибки (0,0194) для выбранного дерева. Это значит, что процедура «автоматического» отбора дерева смогла выбрать дерево с ошибкой, близкой к минимальной.

Class	Global CV Sample Misclassification Matrix (Boston2) Predicted (row) x observed (column) matrix Global CV cost = ,2814; s.d. CV cost = ,01973		
	Class LOW	Class MEDIUM	Class HIGH
LOW		22	3
MEDIUM	43		30
HIGH	3	42	

Рис. 13.41

Перейдите на вкладку **Tree plot** и нажмите кнопку **Classification tree plot**, появится граф дерева классификации (рис. 13.42). На этом графе видно, что выбранное дерево «подходящего размера» имеет 8 терминальных вершин, 7 ветвлений и 14 ветвей, поэтому полностью его понять и интерпретировать довольно сложно. Еще сложнее при помощи полученных решающих правил классифицировать новое наблюдение (новый участок). Но можно отметить, что при отнесении наблюдений к определенному классу (*HIGH, MEDIUM, LOW*) значимыми предикторами являются такие предикторы, как *ORD12* и *ORD5*.

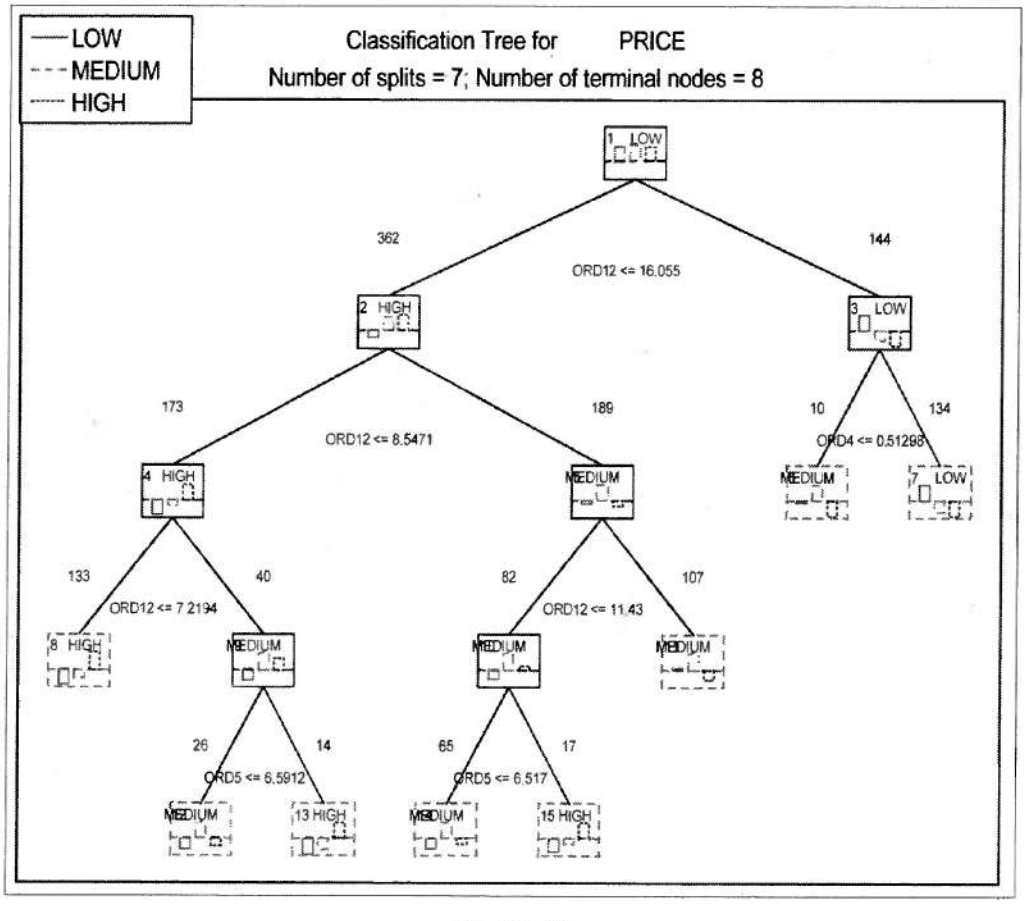


Рис. 13.42

Если открыть таблицу результатов, нажав на кнопку **Predictor importance** (значимость предикторов) на вкладке **Tree structure**, то можно убедиться в том, что эти предикторы действительно являются значимыми, так как их ранг 100 и 73 для *ORD12* и *ORD5* соответственно (рис. 13.43). Можно также определить, какая терминальная вершина дерева классифицирует большинство наблюдений с высокой, средней и низкой ценой. Для этого нажмите кнопку **Assigned node**

by observed class (обозначенные вершины, исходные классы). Появится таблица (рис. 13.44), в которой каждой терминальной вершине (строке) соответствует количество участков с высокой, средней и низкой ценой (столбцы). Из этой таблицы видно, что большинство наблюдений с высокой (*HIGH*) ценой классифицирует терминальная вершина с номером 8 (111 участков), со средней ценой (*MEDIUM*) — терминальная вершина с номером 11 (64 участка), с низкой (*LOW*) ценой — терминальная вершина с номером 7 (118 участков). Та же самая информация, изображенная в виде цветовой гаммы (дискретная карта линий уровня), станет доступной, если нажать на кнопку  .

Variable	Predictor V Based on u 0=low impo Ranking
CAT1	11
ORD1	22
ORD2	18
ORD3	31
ORD4	37
ORD5	73
ORD6	38
ORD7	17
ORD8	18
ORD9	32
ORD10	40
ORD11	31
ORD12	100

Рис. 13.43

Node	Terminal Node x Observed Terminal node (row) x obse Learning sample N = 506 Class LOW	Class MEDIUM	Class HIGH
6	3	7	0
7	118	14	2
8	0	22	111
11	34	64	9
12	3	18	5
13	0	2	12
14	8	42	15
15	1	4	12

Рис. 13.44

Глава 14

Методы редукции данных

14.1. Факторный анализ

Главными целями факторного анализа являются сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных [6]. Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации переменных.

Сокращение достигается путем выделения скрытых общих факторов, объясняющих связи между наблюдаемыми признаками (переменными) объекта, т.е. вместо исходного набора переменных появится возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

Взаимосвязи между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная путем подгонки линия регрессии дает графическое представление зависимости. Если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме, то такая переменная будет включать наиболее существенные черты обеих переменных. Итак, произошло сокращение числа переменных — две заменили одной. Причем новый фактор (переменная) является линейной комбинацией двух исходных. Приведенный пример, в котором

две коррелированные переменные объединены в один фактор, показывает главную идею факторного анализа.

В основном процедура выделения факторов подобна вращению, максимизирующему дисперсию исходного пространства переменных. Например, на диаграмме рассеяния можно рассматривать линию регрессии как ось X, повернув ее так, что она совпадает с прямой регрессии. Этот тип вращения называется вращением, максимизирующим дисперсию (варимакс), так как цель вращения заключается в максимизации изменчивости новой переменной (фактора) и минимизации разброса исходных переменных. Если пример с двумя переменными распространить на большее число переменных, то вычисления становятся сложнее, однако основной принцип представления двух или более зависимых переменных одним фактором остается в силе.

Число наблюдаемых объектов может быть большим и взаимосвязи между ними чрезвычайно сложными. Однако наблюдая объект, выдвигаем гипотезу, что существует небольшое число факторов, которые влияют на измеряемые параметры. Естественно желание выделить как можно меньшее число скрытых общих факторов и чтобы выделенные факторы как можно точнее приближали наблюдаемые параметры, описывали связи между ними.

Выделяемые таким образом факторы называют общими, так как они воздействуют на все признаки (параметры) объекта, а не на какой-то один признак или группу признаков. Эти факторы являются гипотетическими, скрытыми, их нельзя измерить непосредственно, однако существуют статистические методы их выделения.

Рассмотрим модель факторного анализа [9]. Пусть задана система переменных $X_1, X_2, ..., X_n$. Например, X_1 — производительность труда, X_2 — фондоотдача ..., X_n — себестоимость. Значения переменных или признаков $X_1, X_2, ..., X_n$ известны для каждого из N предприятий (объектов). Представим исходную информацию в виде матрицы $X = x_{ji}$ размерности $(n \times N)$. Каждая строка состоит из значений одного показателя для каждого из N объектов исследования. Предполагается, что каждый элемент этой матрицы x_{ji} является результатом воздействия некоторого числа m гипотетических общих факторов и одного характерного фактора

$$x_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + ... + a_{jr}f_{ri} + ... + a_{jm}f_{mi} + d_j v_{ji}, \quad (14.1)$$

где a_{jr} — весовой коэффициент j-й переменной на r-м общем факторе или нагрузка j-й переменной на r-м общем факторе; f_{ri} — значение r-го общего фактора на i-м объекте исследования; d_j — нагрузка или весовой коэффициент j-й переменной на j-м характерном факторе; v_{ji} — значение j-го характерного фактора на i-м объекте исследования; $j = 1, ..., n; i = 1, ..., N; r = 1, ... m; m << n$.

Так как массив данных $X = x_{ji}$ представляет величины различной размерности, то для того чтобы перейти к безразмерным величинам, пронормируем элементы массива.

$$y_{ji} = (x_{ji} - \overline{X}_j)/S_j, \quad (14.2)$$

где $\overline{X}_j$ — выборочное среднее j-й переменной (признака); S_j — выборочная дисперсия j-й переменной. После этих преобразований получим

$$y_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + ... + a_{jr}f_{ri} + ... + a_{jm}f_{mi} + d_j v_{ji}, \quad (14.3)$$

где a_{jm} — неизвестные коэффициенты, называемые факторными нагрузками; $d_j v_{ji}$ называется остатком (невязкой), или остаточным специфическим фактором. Задача состоит в том, чтобы оценить a_{jm} некоторым оптимальным образом.

Обычно в моделях факторного анализа предполагаются выполненными следующие предположения [2]:

- x_{ji} имеют многомерное нормальное распределение;
- общие факторы f_{ri} являются либо некоррелированными случайными величинами с дисперсией 1, либо неизвестными случайными параметрами;
- остатки (остаточные факторы) v_{ji} имеют нормальное распределение, не коррелированны между собой и не зависят от общих факторов.

Если в качестве критерия оптимальности используют минимум расхождения между ковариационной матрицей исходных признаков и той, которая получается после оценивания факторных нагрузок (мера «расхождения» двух матриц, в данном случае есть просто евклидова норма их разности), то приходят к методу главных компонент.

Если критерием оптимальности является максимальная близость исходных корреляций признаков к тем, которые получены в модели после оценивания нагрузок, то говорят о методах анализа главных факторов.

Правая часть выражения (14.3) линейна и напоминает выражение для регрессионного анализа. Однако здесь есть большая разница. В регрессионном анализе система переменных предполагается измеряемой непосредственно (например, взяты из отчетной документации предприятий). Однако в факторном анализе общие и характерные факторы являются гипотетическими (неизвестными). Их нужно оценить методами математической статистики и линейной алгебры.

14.2. Описание модуля *Factor Analysis*

В меню **Statistics** щелкните по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выберите команду **Factor Analysis** (анализ факторов). Откроется стартовая панель модуля. Рассмотрим все его компоненты и опишем некоторые из них. В поле **Input File** (файл входных данных) надо указать тип исходного файла, с которым предстоит работать. В модуле возможны следующие типы исходных данных:

- *Correlation Matrix* (корреляционная матрица);
- *Raw Data* (исходные данные).

Выберите, например, *Raw Data*. Это обычный файл данных, где по строкам записаны значения переменных. В правом нижнем углу окна, за всеми функциональными кнопками находится поле *MD deletion* (обработка пропущенных значений). В этом поле необходимо задать один из способов, которым будут обрабатываться при анализе пропущенные значения (незаполненные ячейки):

- *Casewise* (способ исключения пропущенных случаев);
- *Pairwise* (парный способ исключения пропущенных значений);
- *Mean Substitution* (подстановка среднего вместо пропущенных значений).

Способ *Casewise* состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки (наблюдения), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. Итак, в таблице остаются только те наблюдения, в которых нет ни одного пропуска.

В способе *Pairwise* игнорируются пропущенные наблюдения не для всех переменных, а лишь для выбранной пары. Все наблюдения, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляционной матрицы, когда последовательно рассматриваются все пары переменных.

Способ *Mean Substitution* предполагает при выполнении анализа заполнение пустых клеток средними значениями.

Очевидно, в способе *Pairwise* остается больше наблюдений для обработки, чем в способе *Casewise*. Тонкость, однако, состоит в том, что в способе *Pairwise* оценки различных коэффициентов корреляции строятся по различному числу наблюдений. Выберите, например, способ *Casewise*.

Дальнейшее рассмотрение требует работы уже с конкретными данными, поэтому следующим действием откройте файл, содержащий исходные данные для анализа (если он еще не открыт). В качестве примера рассмотрите имеющийся в программе *STATISTICA* файл **Factor.sta** из библиотеки **Examples**. Об этом файле шла речь при изучении модуля **Canonical Analysis**. Теперь, когда есть данные для анализа, выбран способ обработки пропущенных значений, перейдем к выбору переменных, для которых будем проводить факторный анализ.

Для того чтобы сделать это, задействуйте кнопку **Variables.** Появится окно выбора переменных **Select the variables for the factor analysis** (выбрать переменные для факторного анализа). Кнопка **Select All** (выбрать все) позволяет выбрать все переменные сразу.

Щелкните в стартовом окне модуля кнопкой **OK**. Программа начнет анализ выбранных переменных, появится окно **Define Method of Factor Extraction** (определить метод выделения факторов). В информационной части окна (рис. 14.1) сообщается, что пропущенные значения обработаны методом *Casewise*. Обработано 100 случаев и100 случаев принято для дальнейших вычислений. Корреляционная матрица вычислена для 10 переменных. Нижняя часть текущего диалогового окна состоит из трех вкладок. Выделите вкладку **Descriptives,** так как факторный анализ надо начинать с вычисления корреляционной матрицы. Ее анализ позволит оценить степень коррелированности переменных между собой. И если эта степень окажется высокой,

то данные переменные можно объединять в один фактор. А процедура вычисления корреляционной матрицы доступна именно из этого окна.

Кнопка **Review corelations, means, standard deviations** предназначена для построения корреляционной матрицы, вычисления средних, стандартных отклонений.

Кнопка **Compute multiple regression analyses** осуществляет запуск процедуры множественного регрессионного анализа.

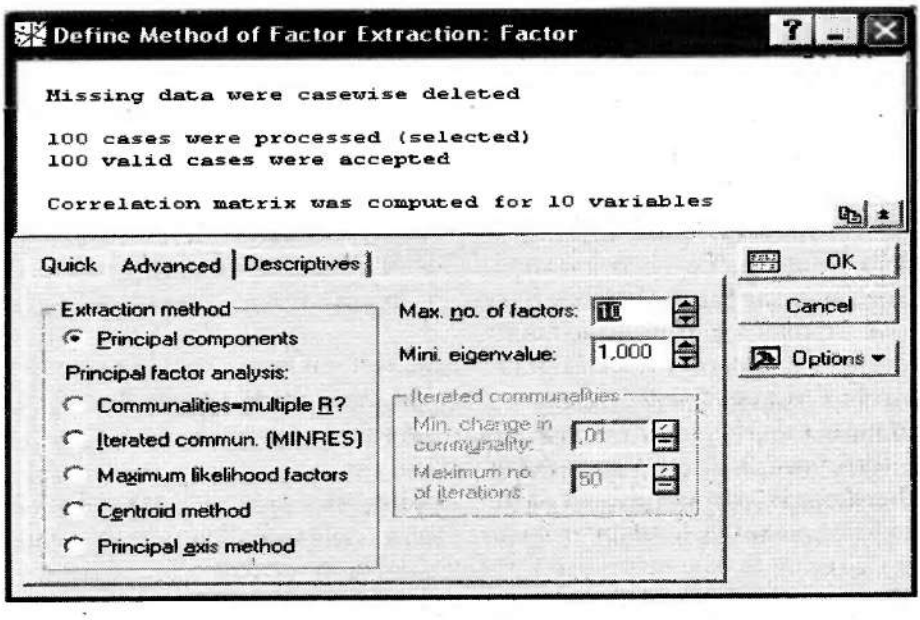


Рис. 14.1

Нажмите кнопку **Review corelations, means, standard deviations.** Откроется окно **Review Descriptive Statistics** (обзор описательных статистик). на вкладке **Quick (Advanced)** нажмите кнопку **Correlations**. На рис. 14.2 изображен фрагмент корреляционной матрицы, из которого видно, что коэффициенты корреляции переменных *WORK* с переменными *HOME* имеют малые значения, в то время как с другими группами переменных принимают большие значения. Этот факт отразится на результатах последующих этапов факторного анализа.

Нажмите кнопку **Cancel** и вернитесь в исходное окно **Define Method of Factor Extraction.** Выделите вкладку **Advanced.** на этой вкладке имеются следующие поля:

- **Maximum no. of factors** (максимальное число факторов);
- **Minimum eigenvalue** (минимальное собственное значение).

В поле **Minimum eigenvalue** устанавливается минимальное собственное значение, т.е. если собственные значения окажутся меньше, чем установленный здесь минимум, то они игнорируются.

Variable	Correlations (Factor) Casewise deletion of MD N=100							
	WORK1	WORK2	WORK3	HOBBY1	HOBBY2	HOME1	HOME2	HOME3
WORK1	1,00	0,65	0,65	0,60	0,52	0,14	0,15	0,14
WORK2	0,65	1,00	0,73	0,69	0,70	0,14	0,18	0,24
WORK3	0,65	0,73	1,00	0,64	0,63	0,16	0,24	0,25
HOBBY1	0,60	0,69	0,64	1,00	0,80	0,54	0,63	0,58
HOBBY 2	0,52	0,70	0,63	0,80	1,00	0,51	0,50	0,48
HOME1	0,14	0,14	0,16	0,54	0,51	1,00	0,66	0,59
HOME2	0,15	0,18	0,24	0,63	0,50	0,66	1,00	0,73
HOME3	0,14	0,24	0,25	0,58	0,48	0,59	0,73	1,00
MISCEL1	0,61	0,71	0,70	0,90	0,81	0,50	0,64	0,59
MISCEL2	0,55	0,68	0,67	0,84	0,76	0,42	0,59	0,52

Рис. 14.2

В поле **Maximum no. of factors** пользователь устанавливает количество факторов, которые необходимо выделить для анализируемых данных. Можно установить любое значение, не превышающее количество переменных, но не любой полученный таким образом результат окажется правильным. Для того чтобы получить интерпретируемый результат, на практике используют несколько полезных критериев.

В методе главных компонент [6] по умолчанию предполагается, что дисперсии всех переменных равны 1. Тогда общая дисперсия равна общему числу переменных (для нашего примера — 10). Это означает, что наибольшая изменчивость, которая потенциально может быть выделена, равна 10. Максимально возможное число выделяемых факторов равно числу переменных. Каждому фактору соответствует дисперсия, объясненная этим фактором. Дисперсии, соответствующие факторам, называются собственными значениями.

Для просмотра собственных значений факторов в окне **Define Method of Factor Extraction** произведите следующие установки параметров: *Maximum no. of factors = 10* и *Minimum eigenvalue = 0*. Далее нажмите **OK**. В открывшемся окне **Factor Analysis Results** нажмите кнопку **Eigenvalues**, появится таблица с собственными числами (рис. 14.3).

Во втором столбце таблицы приведены дисперсии выделенных факторов — собственные числа. В третьем столбце для каждого фактора приводится процент от общей дисперсии (в данном примере она равна 10). Как видно, первый фактор объясняет 61% общей дисперсии, второй фактор — 18% и т.д. Четвертый столбец содержит накопленную или кумулятивную дисперсию. Как только получена информация о том, сколько дисперсии выделил каждый фактор, можно перейти к вопросу, сколько факторов следует оставить.

Критерий Кайзера. Сначала можете отобрать только факторы с собственными значениями, большими 1. По существу это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером и является, вероятно, наиболее широко используемым. В приведенном примере на основе данного критерия выделяются только два фактора, так как остальные не подходят под условие, наложенное на собственные значения.

	Eigenvalues (Factor) Extraction: Principal components			
Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6,118369	61,18369	6,11837	61,1837
2	1,800682	18,00682	7,91905	79,1905
3	0,472888	4,72888	8,39194	83,9194
4	0,407996	4,07996	8,79993	87,9993
5	0,317222	3,17222	9,11716	91,1716
6	0,293300	2,93300	9,41046	94,1046
7	0,195808	1,95808	9,60626	96,0626
8	0,170431	1,70431	9,77670	97,7670
9	0,137970	1,37970	9,91467	99,1467
10	0,085334	0,85334	10,00000	100,0000

Рис. 14.3

Критерий каменистой осыпи. Критерий является графическим методом, впервые предложенным Кэттелем.

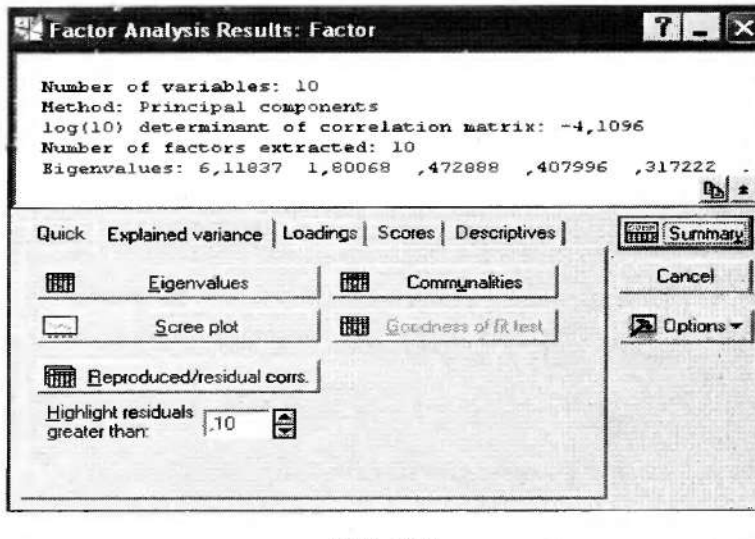


Рис. 14.4

Надо изобразить собственные значения, представленные в таблице в виде графика. Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. на вкладке **Explained variance** нажмите кнопку **Scree plot** (рис. 14.4).

Из построенного графика (рис. 14.5) видно, что в соответствии с этим критерием можно пытаться выделить 2 или 3 фактора.

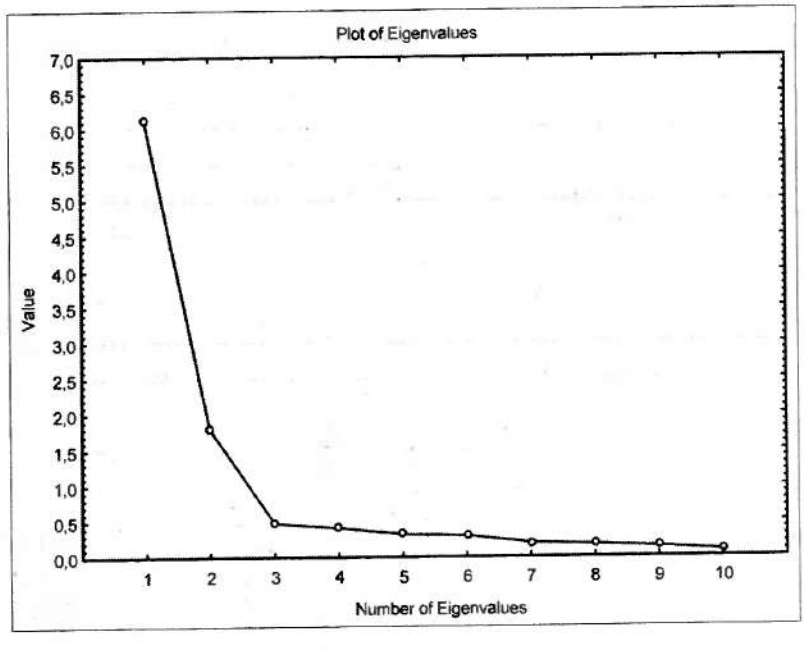


Рис. 14.5

Различные методы выделения факторов расположены на вкладке **Advanced** окна **Define Method of Factor Extraction** и объединены в группу опций под заголовком **Extraction method** (метод выделения). Как говорилось в математическом анонсе, в зависимости от критерия оптимальности возможен анализ либо методом *Principal components* (методом главных компонент), либо одним из методов, объединенных в группу *Principal factor analysis* (анализ главных факторов).

В группе *Principal factor analysis* предусмотрены следующие методы:

- *Communalities = multiple R**2* (общности равны квадрату коэффициента множественной корреляции);
- *Iterated Communalities (MINRES)* (итеративные общности или минимальные остатки);
- *Maximum likelihood factors* (максимальное правдоподобие);
- *Centroid method* (центроидный метод);
- *Principal axis method* (метод главных осей).

Выберите опцию *Principal components*. Чтобы лучше понять основные моменты факторного анализа, предположите, что неизвестны критерии определения числа факторов, и поэтому начните анализ с максимального числа факторов. Сохраните значения максимального числа факторов — 10 и минимального собственного значения — 0 (если собственное значение не будет установлено в 0, то количество выделенных факторов не будет равняться 10).

Щелкните кнопкой **OK**, и на экране появится уже знакомое окно **Factor Analysis Results**. В верхней информационной части окна указаны:

- *Number of variables* (число анализируемых переменных);
- *Method* (метод анализа);
- *log(10) determination of correlation matrix* (десятичный логарифм детерминанта корреляционной матрицы);
- *Number of factor extraction* (число выделенных факторов);
- *Eigenvalues* (собственные значения). В нижней части окна находятся функциональные кнопки, позволяющие всесторонне численно и графически просмотреть результаты анализа.

Нажмите кнопку **Summary. Factor loadings** (итоги, факторные нагрузки). на рис. 14.6 приведен фрагмент таблицы с факторными нагрузками — корреляциями между переменными и выделенными факторами.

	Factor Loadings (Unrotated) (Factor) Extraction: Principal components (Marked loadings are > ,700000)				
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
WORK_1	-0,652601	0,514217	0,301687	0,439108	-0,0137
WORK_2	-0,756976	0,494770	-0,078826	-0,211795	-0,0908
WORK_3	-0,745706	0,456680	-0,104749	0,030826	-0,2049
HOBBY_1	-0,941630	-0,021835	0,012653	0,001861	0,1206
HOBBY_2	-0,875615	0,051643	0,099675	-0,324541	-0,0158
HOME_1	-0,576062	-0,604977	0,490999	-0,114927	-0,1125
HOME_2	-0,671289	-0,617962	-0,125776	0,159963	0,2250
HOME_3	-0,641532	-0,573925	-0,268572	0,152709	-0,3625
MISCEL_1	-0,951516	0,013513	-0,050164	0,026706	0,0767

Рис. 14.6

Из таблицы видно, что первому и второму факторам (*Factor 1, Factor 2*) соответствуют бульшие значения коэффициентов корреляции, чем остальным факторам. Причем с увеличением номера фактора значения коэффициентов корреляции стремительно уменьшаются. При правильно выбранном количестве факторов таблицы факторных нагрузок должны выявлять закономерности, проявляющиеся в следующем. Факторные нагрузки должны объединять переменные в группы, для которых коэффициенты корреляции с факторами принимают бульшие значения по одной группе и меньшие значения по другой.

Из сказанного следует нецелесообразность рассмотрения всех десяти факторов. Воспользуйтесь результатами этой таблицы, критерием Кэттеля, критерием Кайзера и назначьте число факторов — 2.

Из фрагмента таблицы результатов, приведенного на рис. 14.7, видно, что есть некоторая закономерность в значении факторных нагрузок, а именно группе переменных *WORK* соответствуют бульшие значения коэффициентов корреляции с фактором 1, чем с фактором 2. Аналогичные данные получим для групп переменных *HOBBY* и *MISCEL*. Но в такой форме выявленные закономерности трудно проинтерпретировать.

Чтобы получить интерпретируемое решение, надо применить повороты осей, которые достигаются вращением факторов. Как уже говорилось, если пространство общих факторов найдено, то с помощью поворота системы координат в принципе можно получить бесчисленное множество решений. Конечно, такое количество решений — абсурд. Важно найти интерпретируемое решение. Программа предлагает несколько способов вращения [6]:

- *Varimax raw* (варимакс исходных);
- *Varimax normalized* (варимакс нормализованных);
- *Biquartimax raw* (биквартимакс исходных);
- *Biquartimax normalized* (биквартимакс нормализованных);
- *Quartimax raw* (квартимакс исходных);
- *Quartimax normalized* (квартимакс нормализованных);
- *Equamax raw* (эквимакс исходных);
- *Equamax normalized* (эквимакс нормализованных).

Метод варимакс предназначен для максимизации дисперсий квадратов исходных факторных нагрузок по переменным для каждого фактора, что эквивалентно максимизации дисперсий в столбцах матрицы квадратов исходных факторных нагрузок.

Целью метода биквартимакс является одновременная максимизация суммы дисперсий квадратов исходных факторных нагрузок по факторам и максимизация суммы дисперсий квадратов исходных фактоных нагрузок по переменным. Это эквивалентно одновременной максимизации дисперсий в строках и столбцах матрицы квадратов исходных факторных нагрузок.

Метод квартимакс означает максимизацию дисперсий квадратов факторных нагрузок по факторам для каждой переменной, что эквивалентно максимизации дисперсий в строках матрицы квадратов исходных факторных нагрузок.

Метод эквимакс можно рассматривать как взвешенную смесь вращения по методам варимакс и квартимакс, что эквивалентно одновременной максимизации дисперсий в строках и столбцах матрицы квадратов исходных факторных нагрузок. Однако в отличие от вращения по методу биквартимакс относительный вес, назначенный критерию варимакс при вращении, равен количеству факторов, деленному на 2.

Дополнительный термин *normalized* (нормализованные) в названии методов указывает на то, что факторные нагрузки в процедуре нормализуются, т.е. делятся на корень квадратный из соответствующей общности. Термин *raw* (исходные) показывает, что вращаемые нагрузки не нормализованы.

В поле **Factor rotation** окна **Factor Analysis Results** на вкладке **Quick** выберите метод поворота осей, например *Varimax raw*, и щелкните по **Summary**. Из фрагмента таблицы факторных нагрузок (рис. 14.8) следует, что *Factor 1* имеет высокие факторные нагрузки по переменным *WORK* и низкие по переменным *HOME*, а *Factor 2* — наоборот: низкие по переменным *WORK* и высокие по переменным *HOME*. При этом факторные нагрузки, соответствующие переменным групп *HOBBY* и *MISCEL*, принимают промежуточные значения. Это и означает, что выделенные два фактора наилучшим образом характеризуют данные.

Выявление и интерпретация закономерностей в таблицах факторных нагрузок — достаточно трудоемкий процесс. Процедура значительно упрощается, если использовать графическое представление факторных нагрузок. Нажмите кнопку **Plot of factor loadings** (двумерный график нагрузок). График, представленный на рис. 14.9, иллюстрирует соотношение между факторами и группами переменных. Видно, что группа переменных *WORK* занимает на плоскости крайнее левое верхнее положение, а группа переменных *HOME* — крайнее правое нижнее положение. Следовательно, *Factor 1* отвечает за удовлетворение, получаемое на работе, а *Factor 2* измеряет удовлетворенность домашней жизнью. Поэтому можно сделать вывод, что общая удовлетворенность исследуемой группы людей, в основном, определяется двумя факторами — удовлетворенностью работой и удовлетворенностью домом.

Variable	Factor Loadings (Unrotate Extraction: Principal comp (Marked loadings are > ,7 Factor 1	Factor 2
WORK_1	-0,652601	0,514217
WORK_2	-0,756976	0,494770
WORK_3	-0,745706	0,456680
HOBBY_1	-0,941630	-0,021835
HOBBY_2	-0,875615	0,051643
HOME_1	-0,576062	-0,604977
HOME_2	-0,671289	-0,617962
HOME_3	-0,641532	-0,573925
MISCEL_1	-0,951516	0,013513
MISCEL_2	-0,900333	0,048154
Expl.Var	6,118369	1,800682
Prp.Totl	0,611837	0,180068

Рис. 14.7

Variable	Factor Loadings (Varima Extraction: Principal com (Marked loadings are > , Factor 1	Factor 2
WORK_1	0,830623	-0,019320
WORK_2	0,902408	0,058905
WORK_3	0,870524	0,082595
HOBBY_1	0,739857	0,582885
HOBBY_2	0,731191	0,484489
HOME_1	0,097371	0,829676
HOME_2	0,165722	0,897242
HOME_3	0,168370	0,844159
MISCEL_1	0,768988	0,560555
MISCEL_2	0,748861	0,502121
Expl.Var	4,561544	3,357507
Prp.Totl	0,456154	0,335751

Рис. 14.8

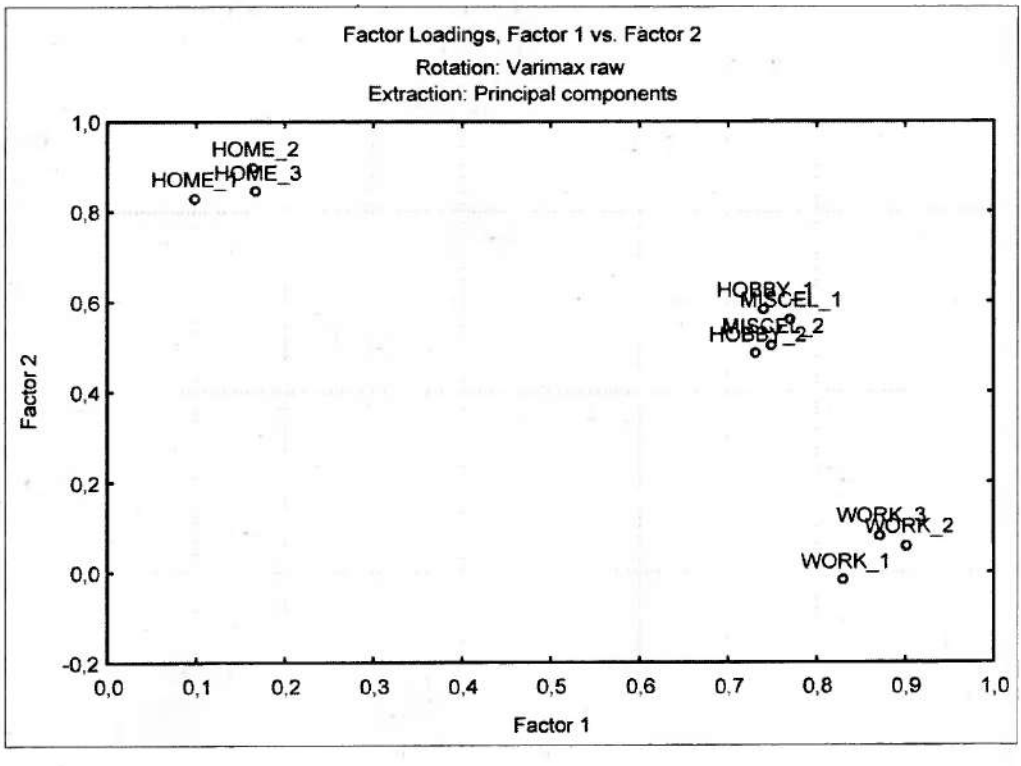


Рис. 14.9

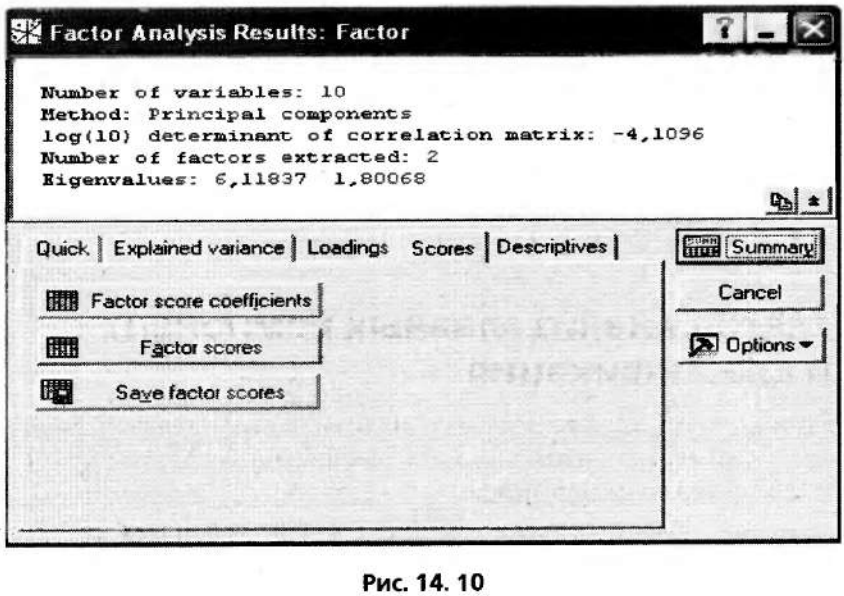


Рис. 14. 10

В диалоге **Factor Analysis Results** перейдите на вкладку **Scores** (рис. 14.10). Нажмите кнопку **Factor Score coefficients**, откроется таблица с коэффициентами линейных уравнений регрессий (рис. 14.11), по которым программа посчитает значения факторов для каждого наблюдения (респондентов).

Нажмите кнопку **Factor Scores**, появится таблица (рис. 14.12), в которой отображены значения факторов для каждого респондента. По этим значениям можно судить об отношении респондентов к *Factor 1* и *Factor 2*. Положительное значение фактора соответствует позитивному отношению респондента, а отрицательное — негативному.

Variable	Factor Score Coefficients (F Rotation: Varimax raw Extraction: Principal compor Factor 1	Factor 2
WORK_1	0,256768	-0,164304
WORK_2	0,263925	-0,145425
WORK_3	0,249750	-0,129616
HOBBY_1	0,115785	0,102111
HOBBY_2	0,131660	0,063002
HOME_1	-0,126453	0,325194
HOME_2	-0,118337	0,340306
HOME_3	-0,107542	0,317830
MISCEL_1	0,128865	0,087384
MISCEL_2	0,133727	0,066977

Рис. 14.11

Case	Factor Scores (Factor) Rotation: Varimax raw Extraction: Principal comp Factor 1	Factor 2
1	0,77326	-0,59909
2	-1,95924	-0,42839
3	-1,31803	-0,13560
4	0,17915	-0,70837
5	0,08277	-1,64135
6	-1,42460	0,42254
7	-0,19411	-0,39425
8	0,95212	-1,13020
9	0,03346	-0,20582
10	-0,70690	-0,41079
11	-0,18579	-1,75809
12	0,23559	1,19109

Рис. 14.12

Величина положительного фактора соответствует силе предпочтения данного фактора (для отрицательного — наоборот). Таким образом, процедура редукции данных позволила выделить два значимых фактора — *Factor 1* и *Factor 2* и сократить число переменных с 10 до 2.

14.3. Метод анализ главных компонент и классификация

На практике часто возникает задача анализа данных большой размерности. Метод анализ главных компонент и классификация позволяет решить эту задачу и служит для достижения двух целей:

- уменьшение общего числа переменных (редукция данных) для того, чтобы получить *«главные»* и *«некоррелирующие»* переменные;
- классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Данный метод имеет сходство с факторным анализом в постановочной части решаемых задач, но имеет ряд существенных отличий:

- при анализе главных компонент не используются итеративные методы для извлечения факторов;
- наряду с активными переменными и наблюдениями, используемыми для извлечения главных компонент, можно задать вспомогательные переменные и/или наблюдения; затем вспомогательные переменные и наблюдения проектируются на факторное пространство, вычисленное на основе активных переменных и наблюдений;
- перечисленные возможности позволяют использовать метод как мощное средство для классификации одновременно переменных и наблюдений.

Решение основной задачи метода анализ главных компонент и классификация достигается созданием векторного пространства латентных (скрытых) переменных (факторов) с размерностью меньше исходной (исходная размерность определяется числом переменных для анализа в исходных данных) [6].

Предположим, необходимо выбрать объект (например, автомобиль) по двум критериями например, мощность двигателя и стоимость. Если значения этих двух критериев принять за координаты точек на плоскости, соответствующих различным автомобилям, то получим диаграмму рассеяния, которая покажет, что можно построить линию, проходящую через большинство точек и, в частности, через центр облака точек. В этом случае линия регрессии будет представлять два свойства автомобилей и, следовательно, может использоваться для выбора автомобиля. Тем не менее если принять во внимание и другие технические параметры автомобиля, например время разгона до 100 км/ч, то обычная парная регрессия переменных не поможет в принятии решения, так как она уже не будет представлять все три свойства автомобиля. Таким образом, становится ясным, что раз число переменных больше двух, то регрессия двух переменных уже не подходит для нашей задачи. Для случая с несколькими переменными требуется что-то, что является общим для всех переменных и может быть использовано как «значение» вида объектов. Если выразить геометрически, то это должна быть линия или линии (оси факторов), которые проходят через центр облака точек многомерного пространства. Анализ главных компонент является тем методом, который может сделать это. Новые факторные оси построены в пространстве меньшей размерности, на них можно спроектировать пространство переменных анализа.

Математически вычисление факторов в основном состоит в диагонализации симметричной матрицы: матрицы корреляций или ковариаций в зависимости от того, нужно ли данные стандартизировать или центрировать относительно средних значений. В обоих случаях результатом будет новый набор некоррелированных переменных (главных компонент), которые являются линейными комбинациями первоначальных переменных. Число переменных становится меньше,

и внутренняя дисперсия данных стремится к максимально возможному значению. Фактически в этом случае создается новое пространство — факторное, на которое можно спроектировать переменные и наблюдения, затем можно классифицировать на категории.

Главные компоненты — это прямые линии, которые наилучшим образом соответствуют облакам точек в векторных пространствах переменных и наблюдений, согласно критерию наименьших квадратов. По критерию наименьших квадратов главные компоненты (факторы) получаются как результат максимальной суммы квадратов ортогональных проекций. Следовательно, строится векторное подпространство меньшей размерности, которое заменяет первоначальное векторное пространство. Хотя фактор извлекается так, чтобы максимально объяснить разброс данных, редко удается сделать это полностью. Поэтому извлекается еще один фактор и т.д. По крайней мере число факторов, извлекаемых таким образом, никогда не превысит число переменных анализа.

В программе *STATISTICA* метод анализ главных компонент реализован для векторных пространств переменных и наблюдений. Если предположить, что исходная таблица данных состоит из n строк (наблюдений или объектов) и p переменных (признаков, характеризующих объекты), то для данных, содержащихся в p переменных, проводится анализ в *n-мерном* пространстве, заданном p переменными, а для данных, содержащихся в n наблюдениях, проводится анализ в *p-мерном* пространстве, заданном n наблюдениями. Так, метод предусматривает нахождение методом наименьших квадратов в векторном пространстве переменных R_n, ортогональных векторов — факторные оси, которые используются для вычисления факторных координат точек — переменных из пространства R_n. Проектирование переменных из пространства R_n на факторное пространство F_q, созданное набором факторов, помогает обнаруживать скрытые различия между переменными. Аналогично метод предусматривает нахождение методом наименьших квадратов в векторном пространстве наблюдений R_p, ортогональных векторов — факторные оси, которые используются для вычисления факторных координат точек — наблюдений из пространства R_p. Проектирование наблюдений из пространства R_p на факторное пространство G_q, созданное набором факторов, помогает обнаруживать скрытую структуру данных.

В соответствии с идеологией метода главных компонент можно разделить переменные на две группы: переменные анализа или активные переменные и вспомогательные переменные. Оба набора переменных относятся к одним и тем же данным и, следовательно, коррелируют между собой. Главные компоненты (факторы) будут вычислены только по переменным анализа (активным переменным). Вспомогательные переменные можно затем спроектировать на подпространство факторов, чтобы сделать выводы об этих переменных, даже если они не участвовали непосредственно в вычислениях. То есть вспомогательные переменные используются только для интерпретации результатов. Заметим, что такое разделение переменных необязательно и должно исходить из существа задачи.

Аналогично наблюдения можно разделить на вспомогательные и активные наблюдения для анализа. Это может быть сделано с помощью группирующей переменной, с использованием одного из ее значений в качестве кода для задания наблюдений анализа. Остальные наблюдения будут считаться вспомогательными наблюдениями. При этом только основные наблюдения будут участвовать в вычислениях главных компонент. Вспомогательные наблюдения позже проектируются на векторное подпространство, образованное факторами, которые были вычислены на основе переменных анализа и основных наблюдений. Выводы на основе вычисленных факторов применимы и к вспомогательным наблюдениям, даже если они не участвовали в наблюдениях.

Как было замечено, метод главных компонент позволяет вычислять главные компоненты с помощью матрицы корреляций или матрицы ковариаций. При реализации метода на вычисляемые факторы будут влиять различия вариабельности (изменчивости) активных переменных. Следовательно, анализ будет успешным, только если такие различия представляют интерес для проводимых исследований. В большинстве случаев эти различия несущественны просто потому, что они связаны с измерениями в различных шкалах. Например, два различных типа измерений температуры по Цельсию и Фаренгейту могут использоваться в двух переменных. Очевидно, что учет этих различий в анализе приведет к отрицательным результатам. В этом случае рекомендуется преобразовать данные, чтобы исключить различие в масштабах. Так как эти измерения произведены в шкале интервалов (связь между температурой по Фаренгейту и Цельсию имеет вид линейной зависимости — $F = (5/9)C + 42$) и измерения отличаются точкой начала отсчета (62) и масштабом (5/9), данные надо преобразовать, а именно: центрировать относительно средних и масштабировать стандартными отклонениями, т.е. надо выбрать матрицу корреляций для вычисления главных компонент. Если измерения отличаются только точкой начала отсчета, данные нужно центрировать только относительно их средних, по этой причине главные компоненты необходимо вычислять через матрицу ковариаций. Очевидно, если в таблице исходных данных присутствуют разнотипные переменные (например, вес, длина, температура) или дисперсии однотипных переменных существенно отличаются, то для вычисления главных компонент надо выбрать корреляционную матрицу.

Собственные значения матрицы ковариаций или корреляций переменных анализа играют важную роль в вычислении главных компонент. Дополнительно к определению факторных координат переменных и наблюдений они предоставляют информацию о дисперсии, которую можно проанализировать по числу факторов. Эта информация может быть в дальнейшем использована для определения порядка, на который вы можете уменьшить размеры пространства первоначальных переменных и наблюдений без потери данных. На основе собственных значений построены различные критерии для определения оптимального числа факторов. Так как сумма собственных значений равна числу «активных» переменных и среднее собственных значений равно 1, то общий критерий состоит в том, чтобы начать с собственных значений, которые больше 1.

Один из важных вопросов, на который дается ответ в методе главных компонент, является ли число главных компонент оптимальным, т.е. могли бы они (главные компоненты) идеально представить весь набор точек (переменных и наблюдений). Так как каждое собственное значение матрицы корреляций или ковариаций является показателем объясненной дисперсии каждой главной компоненты, то процент общей дисперсии (объясненной) можно приписать к данному числу факторов. Этот процент называют «качеством отображения», и он является важной мерой дисперсии, вычисляемой по данному набору главных компонент.

14.4. Описание модуля *Principal Components & Classification Analysis*

Из библиотеки **Example → Datasets** откройте файл данных **Activities**, в котором приведены различные характеристики образа жизни для 28 групп людей. В качестве активных переменных используем 7 видов социальной активности: *WORK* (работа), *TRANSPORT* (транспорт), *CHILDREN* (дети), *HOUSEHOLD* (домашний быт), *SHOPPING* (покупки), *PERSONAL CARE* (личное время), *MEAL* (еда), которым посвящают время представители каждой из 28 групп. Показателем является общее время, посвященное данному виду активности представителями группы в часах. При анализе пропущенные данные замените на соответствующие средние. В качестве вспомогательных переменных укажите 3 переменные: *SLEEP* (сон), *TV* (телевизор) и *LEISURE* (досуг). Для того чтобы проиллюстрировать способ задания основных и вспомогательных наблюдений, в файл данных добавлена дополнительная группирующая переменная *GENDER* (пол), принимающая значения *MALE* (мужчины), *FEMALE* (женщины). Это означает, что одна часть групп состоит из женщин, другая — из мужчин. Для присвоения меток точкам на графиках добавлена переменная *GEO.REGION* (регион) [6].

Цель данного анализа — изучение взаимосвязей между различными показателями социальной активности, чтобы выявить скрытые факторы, которые упростили бы процесс классификации изучаемых групп населения. Для достижения этой цели необходимо определить факторные оси в пространстве меньшей размерности, на которые можно спроектировать пространство переменных анализа, а также сделать возможной визуализацию этих групп, т.е. нанести результаты на карту полученного пространства.

В верхнем меню **Statistics** щелкните по **Multivariate Exploratory Techniques** и выберите команду **Principal Components & Classification Analysis**. Откроется стартовая панель модуля (рис. 14.13), в котором на вкладке **Advanced** нажмите кнопку **Variables**. В открывшемся окне **Select the variables...** в поле **Variables for analysis** (переменные для анализа) высветите переменные *WORK* (работа) — *MEAL* (еда), в поле **Supplimentary variables** (вспомогательные переменные) высветите переменные *SLEEP* — *LEASURE*, в поле **Active cases variable** (переменные с основными наблюдениями) — *GENDER*, в поле **Grouping variable** — *GEO.REGION* (рис. 14.14).

Нажмите кнопку **OK**. В открывшемся окне **Principal Components and Classification Analysis** в поле **Code for active cases** выберите значение группирующей переменной *FEMALE* в качестве кода для основных наблюдений.

При помощи контекстного меню, щелкнув правой кнопкой мыши последовательно на именах переменных и просмотрев в *Statistics of Block Data → Block columns → Means, SD's*, легко проверить, что дисперсии и средние переменных в файле исходных данных значительно отличаются. Поэтому в рамке **Analysis based on** (анализ основан на) выберите опцию *Correlations* для проведения анализа на основе корреляционной матрицы. Далее в рамке **Compute variances** (вычисление дисперсии) выберите опцию *as SS/N-1*, в рамке **MD deletion** (удаление пропущенных данных) — опцию *Mean Substitution* (замена средним) и нажмите кнопку **OK**.

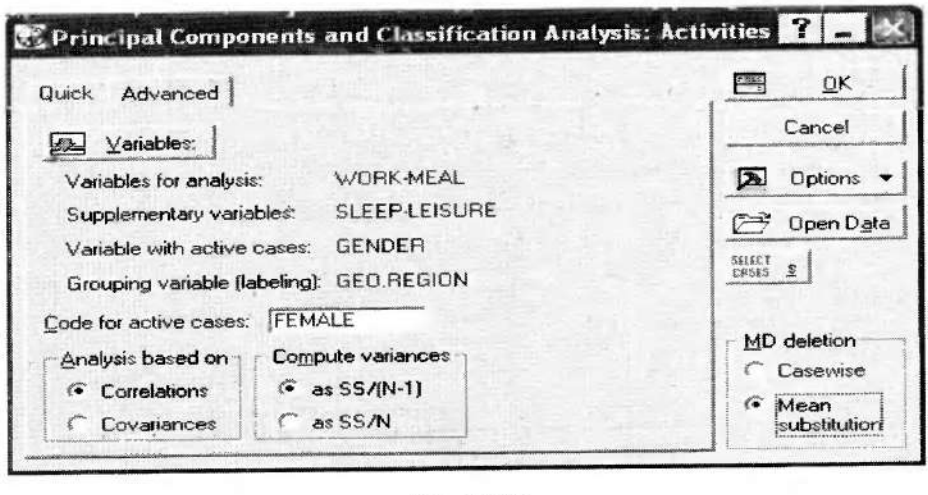


Рис. 14.13

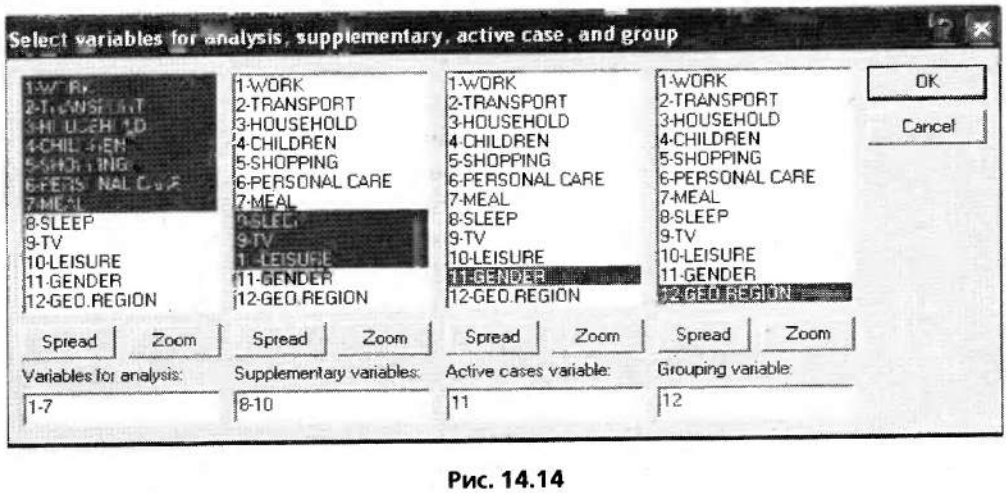


Рис. 14.14

В появившемся окне результатов анализа **Principal Components and Classification Analysis Results** в информационной части указано количество основных и вспомогательных переменных и наблюдений (рис. 14.15).

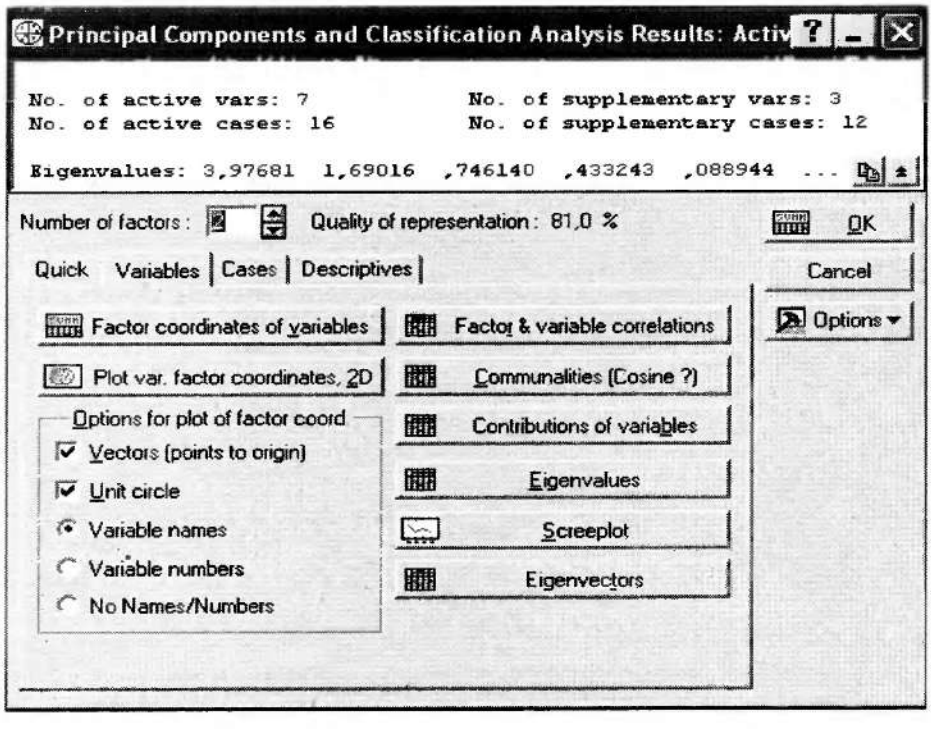


Рис. 14.15

Нажмите кнопку **Screeplot,** программа построит график каменистой осыпи, на котором в виде кусочно-линейной функции изображены собственные значения (рис. 14.16). По критерию Кэттеля (см.§ 14.2) надо определить собственное значение, начиная с которого «горка» теряет свою кривизну, т.е. убывание собственных значений максимально замедляется. Число выделяемых факторов должно быть равно номеру этого собственного значения. Из графика видно, что такими собственными значениями являются значения 2 или 3. Поэтому число выделяемых факторов может быть равно 2 или 3. В поле **Number of factors** установите число факторов, равным 2. При этом качество представления (*Quality of representation)* поменяет свое значение со 100% на 81%.

Нажмите кнопку **Eigenvalues** (собственные значения), чтобы построить таблицу собственных значений (рис. 14.17). В этой таблице для каждого собственного значения также приведен процент объясненной дисперсии (*Total variance*), кумулятивное собственное значение (*Cumulative Eigenvalue*) и кумулятивный процент (*Cumulative %*) объясненной дисперсии. Собственные значения представлены в порядке убывания, отражая тем самым степень важности соответствующих выделенных факторов для объяснения вариации исходных данных. Так,

фактор, соответствующий максимальному собственному значению (3,976814), описывает приблизительно 56,8% общей вариации. Второй фактор для значения (1,690162) описывает 25,77% общей вариации и т.д.

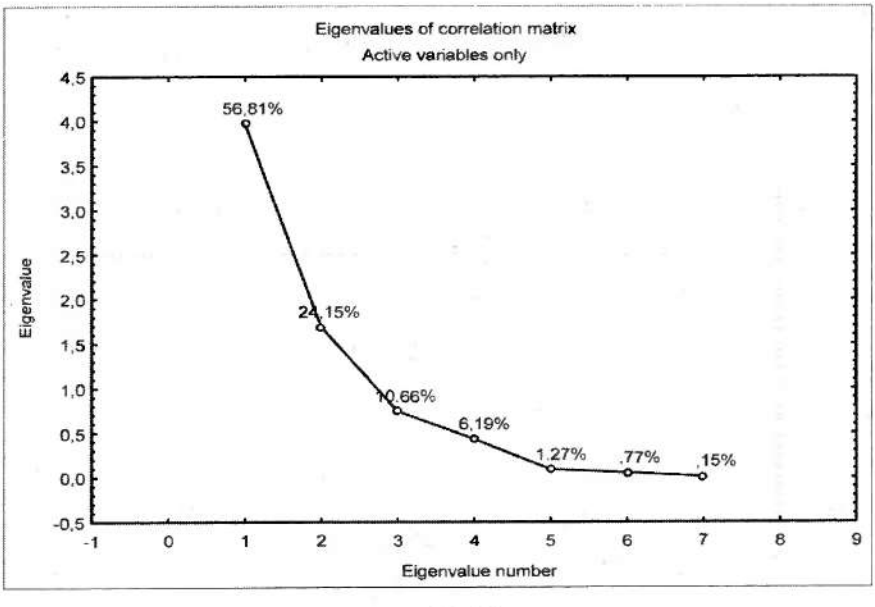


Рис. 14.16

Value number	Eigenvalues of correlation matrix, and related statist Active variables only			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,976814	56,81163	3,976814	56,8116
2	1,690162	24,14518	5,666976	80,9568
3	0,746140	10,65914	6,413116	91,6159
4	0,433243	6,18918	6,846359	97,8051
5	0,088944	1,27063	6,935303	99,0758
6	0,054063	0,77233	6,989366	99,8481
7	0,010634	0,15191	7,000000	100,0000

Рис. 14.17

Когда анализируются корреляционные матрицы, сумма собственных значений равна числу (активных) переменных, для которых выделены (рассчитаны) факторы, при этом «среднее ожидаемое» собственное значение равно 1. На практике применяют различные критерии для правильного выбора размерности факторного пространства. Наиболее простой из них — оставить только те факторы, собственные значения которых больше 1. В данном примере только два первых

собственных значения больше 1 и они объясняют приблизительно 82% общей вариации. Таким образом, значения собственных чисел подтвердили правильность выбора числа выделяемых факторов — 2.

Нажмите кнопку **Factor coordinates of variables** (факторные координаты переменных), чтобы получить таблицу координат исходных факторов в пространстве новых выделенных факторов (рис. 14.18).

Variable	Factor coordinates of tl Active and Supplement *Supplementary variabl	
	Factor 1	Factor 2
WORK	-0,941018	0,275054
TRANSPORT	-0,851971	-0,185457
HOUSEHOLD	0,912134	0,036525
CHILDREN	0,779245	-0,354216
SHOPPING	0,326204	-0,917236
PERSONAL CARE	-0,536329	-0,685359
MEAL	0,729504	0,377189
*SLEEP	0,590196	0,318393
*TV	0,280880	-0,568769
*LEISURE	0,476076	-0,318265

Рис. 14. 18

В терминологии факторного анализа факторные координаты в методе главных компонент также называют «факторными нагрузками». С точки зрения математики, главный компонент — это линейная комбинация переменных, которые сильно коррелируют с ним. В дальнейшем подразумевается, что факторные координаты переменной — это корреляции между переменной и факторными осями.

Следовательно, интерпретация главных компонент должна быть сделана в терминах корреляции, т.е. нужно выделить те переменные (наблюдения), которые имеют наибольшие (абсолютные) значения факторных координат для данных факторов. Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Другими словами, чем больше величина факторной координаты переменной, тем лучше переменные показывают структуру, представленную этим фактором. Например, фактор с высокими факторными нагрузками для трех измерений размеров человека, таких, как вес, рост и окружность грудной клетки, может рассматриваться как представляющий «размер» (т.е. абстракция трех переменных) человека.

Координаты отображаются как для активных переменных, так и для вспомогательных. Как видно из таблицы, первая факторная ось, соответствующая собственному значению 3.976, наиболее сильно коррелирует с переменными *WORK*, *TRANSPORT*, (сильные отрицательные корреляции), *PERSONAL CARE* (умеренные отрицательные корреляции), *MEAL, SLEEP* (умеренная положительная корреляция),

HOUSEHOLD и *CHILDREN* (сильные положительные корреляции). Поэтому можно субъективно обозначить первую выделенную факторную ось как социальную активность, связанную с работой, домом и детьми. Вторую же ось, соответствующую собственному значению 1,69, можно обозначить как социальную активность, связанную с такими видами деятельности, как покупки, личное время, телевидение (сильные и умеренные корреляции с *SHOPPING*, *TV*, *PERSONAL CARE*).

Аналогичная кнопка находится на вкладке **Cases**. Факторные координаты наблюдений (**Factor coordinates of cases**) — это не корреляции, как в случае с переменными. Наблюдения с большими координатами лучше показывают структуру, представленную фактором.

Процессу интерпретации факторов помогают графики факторных координат переменных и наблюдений. Нажмите на вкладке **Variables** кнопку **Plot var. factor coordinates. 2D**, чтобы построить соответствующий график для выделенных факторов (рис. 14.19). Как видно из рисунка, все переменные изображены в виде точек на единичном круге, так как корреляции (координаты точек) наблюдений с факторными осями принимают значения из интервала [0, 1].

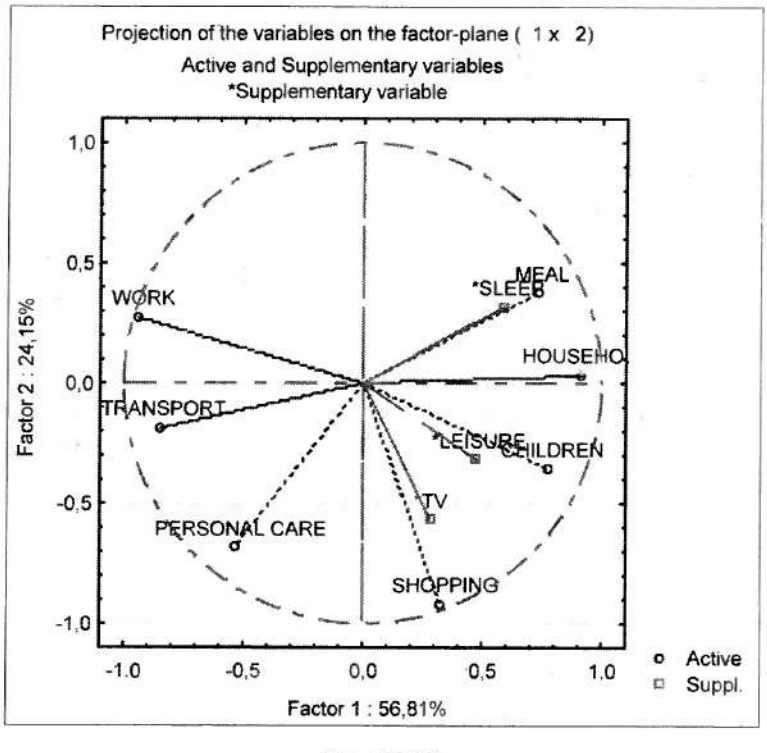


Рис. 14.19

Горизонтальная ось системы координат соответствует фактору 1 (*Factor 1*), а вертикальная — фактору 2 (*Factor 2*). В зависимости от знаков координат точки расположены в соответствующих квадрантах плоскости. Основные и вспомога-

тельные переменные изображены (на мониторе) соответственно кружочком синего цвета и прямоугольником красного цвета. Этот круг является визуальным индикатором того, насколько хорошо каждая переменная воспроизводится текущим набором выделенных факторов — чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат.

Нажмите кнопку **Contributions of variables** (вклад переменных), появится таблица (рис. 14.20) с вкладами основных переменных. Вклад переменной — это относительный вклад переменной в дисперсию факторной оси.

Variable	Variable contributions	
	Factor 1	Factor 2
WORK	0,222669	0,044762
TRANSPORT	0,182522	0,020350
HOUSEHOLD	0,209210	0,000789
CHILDREN	0,152691	0,074235
SHOPPING	0,026757	0,497776
PERSONAL CARE	0,072331	0,277913
MEAL	0,133820	0,084176

Рис. 14.20

Значения этой статистики используются для отсеивания переменных, перед тем как они рассматриваются на основе факторных координат, т.е. корреляций для интерпретации факторных осей. Естественно, те переменные должны быть кандидатами для дальнейшей проверки, вклад (относительный) которых в дисперсию оси фактора больше. Обратите внимание, что значения вкладов «пропорциональны» факторным нагрузкам.

Аналогичная кнопка находится на вкладке **Cases**. Как и в случае переменных, вклады основных наблюдений *Contributions of cases* также являются их относительными вкладами в дисперсию факторной оси. Следовательно, вклад наблюдения — это мера важности наблюдения в качестве определителя факторной оси. Больший вклад наблюдения «утяжеляет» его в факторе. Следовательно, при переходе к интерпретации главных компонент сначала рассматриваются наблюдения с большим вкладом.

Нажмите кнопку **Communalities [Cosine 2]**. Программа построит таблицу общностей переменных (рис. 14.21).

Variable	Communalities. based Active and Supplemen *Supplementary variab	
	From 1 factor	From 2 factors
WORK	0,885515	0,961170
TRANSPORT	0,725854	0,760248
HOUSEHOLD	0,831988	0,833322
CHILDREN	0,607222	0,732691
SHOPPING	0,106409	0,947731
PERSONAL CARE	0,287648	0,757366
MEAL	0,532177	0,674448
*SLEEP	0,348331	0,449705
*TV	0,078893	0,402391
*LEISURE	0,226649	0,327941

Рис. 14.21

Общность — это доля объясненной дисперсии, которая характеризует степень общности переменной (наблюдения) с другими переменными (наблюдениями) по заданному числу факторов. Геометрически это квадрат косинуса угла, образованного радиус вектором переменной (наблюдением) и факторной осью. На вкладке **Cases** этой кнопке соответствует кнопка с названием **Cosine 2**. В таблице, которая откроется после нажатия этой кнопки (рис.14.22), также представлена дополнительная информация о принадлежности наблюдения к основным или вспомогательным наблюдениям. Каждому наблюдению также будет поставлено в соответствие значение группирующей переменной *GEO REGION*.

На вкладке **Cases** щелкните по кнопке **Plot cases factor coordinates,2D** (график наблюдений в факторном пространстве). Появится график (рис. 14.23), на котором изображаются как основные (*FEMALES*) наблюдения, которые использовались при расчете факторов (кружочки синего цвета), так и вспомогательные (*MALES*) наблюдения (квадратики красного цвета).

Case	Cosine squares, based on correlations (Activities) Active cases variable: GENDER Labelling variable: C Code for active cases: FEMALE Suppl. case values			
	Factor 1	Factor 2	GENDER	GEO.REGION
EMU	0,75243	0,01208	MALE	WEST
EWU	0,68324	0,250961	FEMALE	WEST
UWU	0,21258	0,59802	FEMALE	WEST
MMU	0,70912	0,03662	MALE	WEST
MWU	0,20083	0,63840	FEMALE	WEST
SMU	0,65769	0,10023	MALE	WEST
SWU	0,65025	0,22902	FEMALE	WEST
EMW	0,36550	0,49443	MALE	WEST
EWW	0,24993	0,60944	FEMALE	WEST
UWW	0,82170	0,05346	FEMALE	WEST
MMW	0,33514	0,52367	MALE	WEST
MWW	0,62314	0,29149	FEMALE	WEST
SMW	0,28958	0,44568	MALE	WEST
SWW	0,08866	0,45641	FEMALE	WEST
EME	0,76586	0,11727	MALE	EAST
EWE	0,72313	0,03600	FEMALE	EAST
UWE	0,81228	0,02979	FEMALE	EAST
MME	0,27368	0,39035	MALE	EAST
MWE	0,09616	0,06685	FEMALE	EAST
SME	0,73617	0,03525	MALE	EAST
SWE	0,85001	0,02442	FEMALE	EAST
EMY	0,74743	0,13018	MALE	EAST
EWY	0,50735	0,17308	FEMALE	EAST
UWY	0,77247	0,00109	FEMALE	EAST
MMY	0,74464	0,12000	MALE	EAST
MWY	0,39633	0,13218	FEMALE	EAST
SMY	0,73128	0,00397	MALE	EAST
SWY	0,83523	0,05268	FEMALE	EAST

Рис. 14.22

Обратите внимание, что основные и вспомогательные наблюдения сгруппированы в разных областях плоскости, т.е. они объединены в группы однородности — кластеры. При этом кластер с вспомогательными мужскими группами расположен в центральной и нижней части второго квадранта, т.е. имеет отрицательные значения координат по первой, горизонтальной оси и положительные значения координат по второй, вертикальной оси (кроме одного наблюдения). Напомним, что горизонтальная факторная ось была нами интерпретирована как социальная активность, связанная с работой, домом и детьми. При этом отрицательную часть оси определяют переменные *WORK*, *TRANSPORT* (рис. 14.19).

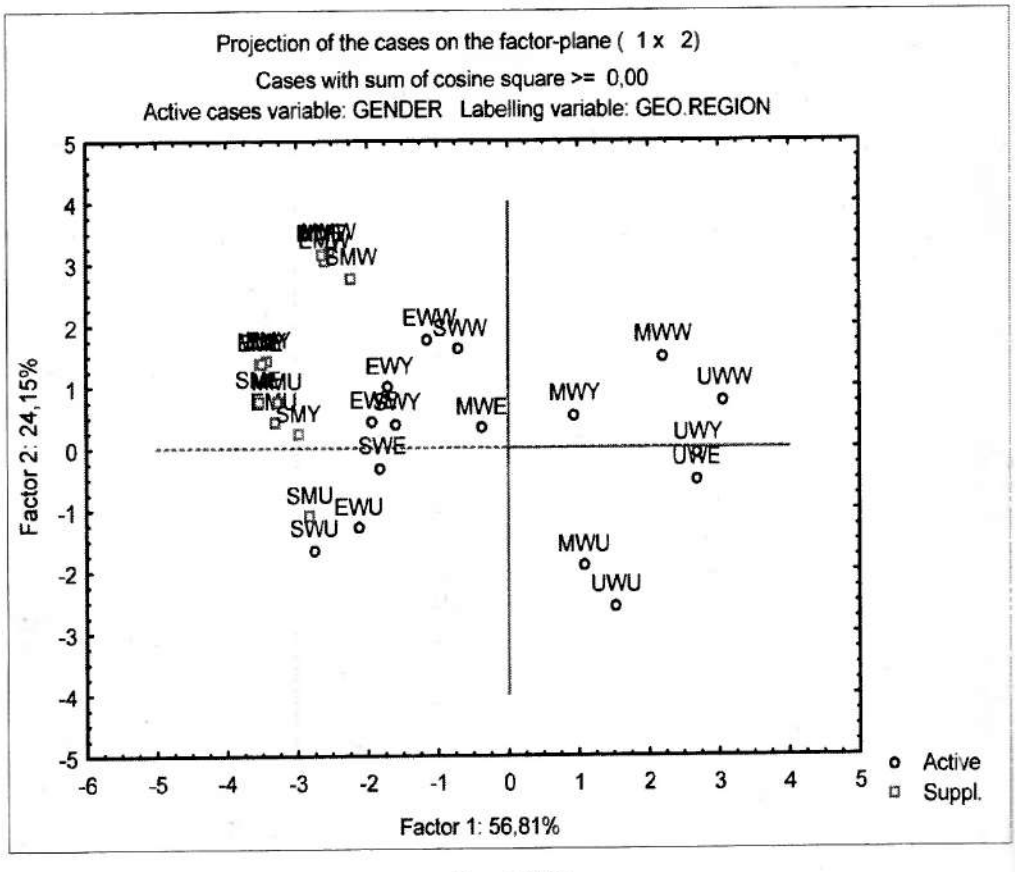


Рис. 14.23

Из графика видно, что мужские группы объединены в области переменной *WORK*. Это означает, что социальная активность мужских групп в основном сконцентрирована в работе. Проявления социальной активности женских групп носят более разносторонний характер – они сгруппированы более или менее равномерно на всей плоскости: в области переменных *PERSONAL CARE, CHILDREN, HOUSEHOLD, SHOPPING, MEAL, SLEEP, LEISURE.*

Глава 15

Методы анализа и упрощения геометрической структуры данных

15.1. Многомерное шкалирование

Многомерное шкалирование можно рассматривать как альтернативу факторному анализу [16], в котором достигается сокращение числа переменных, путем выделения латентных (непосредственно не наблюдаемых) факторов, объясняющих связи между наблюдаемыми переменными. Цель многомерного шкалирования — поиск и интерпретация латентных переменных, дающих возможность пользователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков. Далее, как правило, будем говорить лишь о сходствах объектов, имея в виду, что на практике это могут быть расстояния или степени связи между ними. В факторном анализе сходства между переменными выражаются с помощью матрицы коэффициентов корреляций. В многомерном шкалировании в качестве исходных данных можно использовать произвольный тип матрицы сходства объектов: расстояния, корреляции и т.д.

Несмотря на то что имеется много сходства в характере исследуемых вопросов, методы многомерное шкалирование и факторный анализ имеют ряд существенных отличий. Так, факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. Многомерное шкалирование не накладывает таких ограничений, оно может быть применимо, если задана матрица попарных сходств объектов. В терминах различий получаемых результатов факторный анализ стремится извлечь больше факторов — латентных переменных по сравнению с многомерным шкалированием. Поэтому многомерное шкалирование часто приводит к проще интерпретируемым решениям. Однако более существенно то, что метод многомерное шкалирование можно применять к любым типам расстояний или сходств, в то время как факторный анализ требует, чтобы в качестве исходных данных была использована корреляционная матрица переменных или по файлу исходных данных сначала была вычислена матрица корреляций.

Суммируя сказанное, можно сказать, что многомерное шкалирование потенциально применимо к более широкому классу исследовательских задач. Причем модуль «Многомерное шкалирование» является единственным в системе *STATISTICA* статистическим модулем, который не может использовать на входе необработанные исходные данные.

Основное предположение многомерного шкалирования заключается в том, что существует некоторое метрическое пространство существенных базовых характеристик, которые неявно и послужили основой для полученных эмпирических данных о близости между парами объектов. Следовательно, объекты можно представить как точки в этом пространстве. Предполагают также, что более близким (по исходной матрице) объектам соответствуют меньшие расстояния в пространстве базовых характеристик. Поэтому, многомерное шкалирование — это совокупность методов анализа эмпирических данных о близости объектов, с помощью которых определяется размерность пространства существенных для данной содержательной задачи характеристик измеряемых объектов и конструируется конфигурация точек (объектов) в этом пространстве. Это пространство («многомерная шкала») аналогично обычно используемым шкалам в том смысле, что значениям существенных характеристик измеряемых объектов соответствуют определенные позиции на осях пространства. Таким образом, на входе всех алгоритмов многомерного шкалирования используется матрица, элемент которой на пересечении ее i-й строки и j-го столбца содержит сведения о попарном сходстве анализируемых объектов (объекта i и объекта j). На выходе алгоритма многомерного шкалирования получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во вспомогательных шкалах, связанных с латентными переменными, откуда и название многомерное шкалирование), причем размерность нового пространства признаков существенно меньше размерности исходного (за это собственно и идет борьба).

Логику многомерного шкалирования можно проиллюстрировать на следующем простом примере. Предположим, что имеется матрица попарных расстояний

(т.е. сходства некоторых признаков) между некоторыми городами. Анализируя матрицу, надо расположить точки с координатами городов в двумерном пространстве (на плоскости), максимально сохранив реальные расстояния между ними. Полученное размещение точек на плоскости впоследствии можно использовать в качестве приближенной географической карты. В общем случае многомерное шкалирование позволяет таким образом расположить объекты (города в нашем примере) в пространстве некоторой небольшой размерности (в данном случае она равна двум), чтобы достаточно адекватно воспроизвести наблюдаемые расстояния между ними. В результате можно измерить эти расстояния в терминах найденных латентных переменных. Так, в нашем примере можно объяснить расстояния в терминах пары географических координат Север/Юг и Восток/Запад.

Данные в исходной матрице близости (сходстве) объектов могут быть получены различными способами. Вообще говоря, многомерное шкалирование ориентируется на экспертные оценки близости объектов, когда респонденту предъявляют пары объектов и он должен упорядочить их по степени внутреннего сходства, которое иногда оценивается в баллах. Также данные — матрица расстояний между наблюдениями — могут быть получены каким-либо модулем *STATISTICA*, например, **Cluster Analysis — Joining tree clustering** (кластерный анализ — иерархическая классификация).

Как было замечено, для анализа в модуле «Многомерное шкалирование» используются файлы, содержащие матрицу корреляции, сходства, различия или расстояний между объектами. Однако чтобы программа *STATISTICA* распознавала файл как матричный, он должен удовлетворять следующим условиям:

- число строк равно числу столбцов плюс 4;
- основная матрица должна быть квадратной, а имена наблюдений должны быть такими же, как и имена переменных;
- последние четыре строки должны содержать имена наблюдений и следующие статистики:
  - *Means* (средние). В этой строке приводится среднее каждой переменной; для матриц сходства и различия ее можно оставить пустой (т.е. ничего не вводить);
  - *Std.Dev* (стандартное отклонение). В этой строке приводится стандартное отклонение каждой переменной. Для матриц сходства и различия ее можно оставить пустой (т.е. ничего не вводить);
  - *No.Cases* (число наблюдений). Здесь необходимо ввести число наблюдений, из которых составляется матрица;
  - *Matrix*. Здесь необходимо ввести тип матричного файла: 1 — корреляции; 2 — сходства; 3 — различия и 4 — ковариации.

15.2. Вычислительные методы Многомерного шкалирования

Многомерное шкалирование — это не просто определенная процедура, а скорее способ наиболее эффективного размещения объектов, приближенно сохраняющий наблюдаемые между ними расстояния. Другими словами, многомерное шкалирование размещает объекты в пространстве заданной размерности и проверяет, насколько точно полученная конфигурация сохраняет расстояния между объектами. При размещении объектов программа максимизирует критерий согласия, характеризующий качество подгонки модели к данным, называемый стрессом. Модуль «Многомерное шкалирование» как бы погружает анализируемые объекты (о которых известна лишь матрица попарных расстояний в исходном пространстве признаков) в некоторое, специальным образом подобранное пространство латентных признаков, имеющее небольшую размерность (в программе *STATISTICA* — от 1 до 9; чаще всего 2...5).

Мерой, наиболее часто используемой для оценки качества подгонки модели (отображения), измеряемого по степени воспроизведения исходной матрицы сходств, является величина критерия согласия — стресса φ, который для текущей конфигурации определяется так [16]:

$$\varphi = \sum_i \sum_j \left(d_{ij} - f\left(\delta_{ij}\right)\right)^2 \qquad (15.1)$$

Здесь d_{ij} — воспроизведенные расстояния в пространстве заданной размерности, $f(\delta_{ij})$ — неметрическое монотонное преобразование исходных данных (расстояний) δ_{ij}. Таким образом, модуль «Многомерное шкалирование» реализует неметрический подход к шкалированию, т.е. воспроизводит не количественные меры сходств объектов, а лишь их относительный порядок. Обычно используется одна из нескольких похожих мер сходства. Тем не менее большинство из них сводится к вычислению суммы квадратов отклонений наблюдаемых расстояний δ_{ij} (либо их некоторого монотонного преобразования $f(\delta_{ij})$) и воспроизведенными расстояниями d_{ij}. Таким образом, чем меньше значение стресса, тем лучше матрица исходных расстояний согласуется с матрицей результирующих расстояний.

Вообще говоря, чем больше размерность пространства, используемого для воспроизведения расстояний, тем лучше «согласие» воспроизведенной матрицы с исходной (меньше значение стресса). Если взять размерность пространства равной числу переменных, то возможно абсолютно точное воспроизведение исходной матрицы расстояний. Однако нашей целью является упрощение решаемой задачи с тем, чтобы объяснить матрицу сходства (расстояний) в терминах лишь нескольких важнейших факторов (латентных переменных или вспомогательных шкал).

Вернемся к примеру с расстояниями между городами. Если получена двумерная карта, намного проще представить себе расположение городов и планировать передвижение между ними, чем, если бы имелась только матрица попарных расстояний. Рассмотрим, почему уменьшение числа факторов (или вспомогательных

шкал) может приводить к ухудшению представления исходной матрицы. Обозначим буквами *A, B, C* и *D, E, F* две тройки городов. Соответствующие им точки и попарные расстояния между ними показаны в двух табличках (матрицах).

	A	*B*	*C*
A	*0*		
B	*90*	*0*	
C	*90*	*90*	*0*

	D	*E*	*F*
D	*0*		
E	*90*	*0*	
F	*180*	*90*	*0*

Первой матрице соответствует случай, когда все города удалены друг от друга в точности на 90 км, а второй — когда города *D* и *F* удалены на 180 км.

Можно ли три точки, соответствующие городам (объектам), расположить в одномерном пространстве (на прямой)? Действительно, три точки, соответствующие городам *D, E* и *F*, могут быть расположены на прямой линии:

$$D - 90\ км - E - 90\ км - F.$$

D удален на 90 км от города *E*, и *E* — на 90 км от *F*, а город *D* — на 90 + 90 = 180 км от *F*. Если попытаться проделать то же самое с городами *A, B* и *C*, то видно, что соответствующие им точки уже нельзя разместить на прямой с сохранением исходной структуры расстояний. Однако эти точки можно расположить на плоскости, например, в виде треугольника:

A

90 км *90 км*

B *90 км* *C*

Располагая эти три точки так, можно в точности воспроизвести все расстояния между ними. Без лишних деталей этот пример показывает, как конкретная матрица расстояний (сходств) связана с числом искомых латентных переменных (размерностью результирующего пространства). Конечно, реальные данные никогда не являются такими точными и содержат случайный шум, т.е. случайную изменчивость, влияющую на различие между воспроизведенной и исходной матрицей.

Обычно для выбора размерности пространства, в котором будет воспроизводиться наблюдаемая матрица, используют график зависимости стресса от размерности (график каменистой осыпи). Этот критерий впервые был предложен Кэттелом в контексте решения задачи снижения размерности в факторном анализе. Кэттел предложил найти такую абсциссу на графике, в которой график стресса начинает визуально сглаживаться в направлении правой, пологой его части, таким образом, уменьшение стресса максимально замедляется. Образно говоря, линия на рисунке напоминает скалистый обрыв, а черные точки на графике напоминают камни, которые ранее упали вниз. Таким образом, внизу наблюдается как бы каменистая осыпь из таких точек.

Вторым критерием для решения вопроса о размерности с целью интерпретации является ясность полученной конфигурации точек. Иногда, как в нашем примере с городами, результирующие координаты легко интерпретируются. В других случаях точки на графике могут образовывать ту или иную разновидность случайного облака, и не существует непосредственного способа для интерпретации латентных переменных. В последнем случае следует постараться немного увеличить число координатных осей и рассмотреть получаемые в результате конфигурации. Чаще всего получаемые решения проще удается проинтерпретировать. Однако если точки на графике не следуют какому-либо образцу, а также если график стресса не показывает какого-либо явного изгиба (и не похож на край обрыва), то данные, скорее всего, являются случайным шумом.

Ввиду специфики задач, решаемых многомерным шкалированием, этот метод получил широкое распространение в маркетинговых исследованиях. Важнейшая задача, которую решает данный метод, — это построение пространственной карты восприятия товаров (услуг), предлагаемых на рынке. Это не только позволяет выявить значимые характеристики, влияющие на потребительские предпочтения, но также предоставляет возможность графически представить результаты и существенно облегчить интерпретацию данных.

Многомерное шкалирование весьма популярно в психологическом исследовании восприятия личности. Анализируются сходства между определенными чертами характера с целью выявления основополагающих личностных качеств.

В последнее время многомерное шкалирование активно используется для решения задач, связанных с сегментацией рынка. Цель построения пространственной карты — нахождения «пустот» на рынке, которые фирма может «заполнить» своим товаром. Разбив пространство на гомогенные пространства и выявив характеристики группы респондентов, демонстрирующих сходство предпочтений, можно составить представление о том, каким должен быть продукт, ориентированный на данную группу потребителей, а также о правильной ориентации рекламной кампании существующих продуктов.

15.3. Описание модуля *Multidimensional Scaling*

Сначала создадим матричный файл. Для этого воспользуемся файлом **Cars** из **Examples**. Данный файл нами подробно был описан в разделе Кластерный анализ. Произведите последовательность действий по следующей схеме: **Open → Examples → Datasets → Cars → Statistics → Multivariate exploratory Techniques → Cluster Analysis → Joining tree clustering.**

В открывшемся окне **Cluster Analysis** нажмите кнопку **Variables** и выберите все переменные для анализа, в поле **Cluster** укажите *Cases (rows)*, в поле **Distance measure** — *Euclidean distance* и нажмите **OK**. В открывшемся окне **Joining Results Advanced** нажмите кнопку **Matrix**. Программа автоматически сохранит матричный файл различий, в котором на пересечении строки и столбца приведено евклидово расстояние (различие) между автомобилями, марки которых прописаны в названии строки и столбца. Файл можете сохранить под любым именем (напри-

мер, **Cars1)** с расширением *smx*. На рис. 15.1 приведен фрагмент файла. Глядя на эту матрицу, трудно понять, насколько близки те или иные марки автомобилей в общей совокупности. Поэтому представляет интерес рассмотрение этих данных на общем плане, например, на плоскости в естественно интерпретируемых шкалах. Эту задачу можно решить в модуле **Multidimensional Scaling.**

	Cars 1					
	1 Acura	2 Audi	3 BMW	4 Buick	5 Corvette	6 Chrysler
Acura	0,00	3,15	2,81	2,77	4,06	2,39
Audi	3,15	0,00	1,20	2,25	2,45	1,58
BMW	2,81	1,20	0,00	2,83	1,86	1,44
Buick	2,77	2,25	2,83	0,00	4,40	1,69
Corvette	4,06	2,45	1,86	4,40	0,00	3,08
Chrysler	2,39	1,58	1,44	1,69	3,08	0,00

Рис. 15.1

Для запуска модуля **Multidimensional Scaling в** верхнем меню **Statistics** щелкните по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выберите команду **Multidimensional Scaling**. Откроется стартовая панель модуля. Нажмите кнопку **Variables** и откройте диалоговое окно **Select variables for analysis**. Для выбора всех переменных нажмите кнопку **Select All,** а затем — на **OK.** Программа вернется в стартовое окно модуля.

В поле **Number of dimensions** можно выбрать размерность пространства. По умолчанию она равна 2. Доступное максимальное значение размерности равно 9 или же числу переменных (объектов) минус один, в зависимости от того, что меньше. Отметим, что, выполняя последовательно несколько анализов и принимая установки по умолчанию, пользователь может последовательно оценить решения для пространств размерности n, n-1, n-2 и т.д. Например, сначала можно специфицировать размерность 5, и после возвращения к диалоговому окну **Results** программа по умолчанию вычислит решение для размерности 4 и т.д. Мы будем рассматривать двухмерное пространство.

Рассмотрим вкладку **Opions**, которая позволяет выбрать первоначальную конфигурацию *(Starting configuration)* для проведения текущего анализа. В рамке **Starting configuration** находятся следующие опции (рис. 15.2):

- *Standard Guttman-Lingoes* (стандартная Гутмана-Лингоуса). Эта процедура выполняет анализ главных компонент и в большинстве ситуаций предоставляет адекватную начальную конфигурацию, необходимую для работы итерационного алгоритма подгонки модели. По умолчанию начальной конфигурацией является стандартная конфигурация Гутмана-Лингоуса;

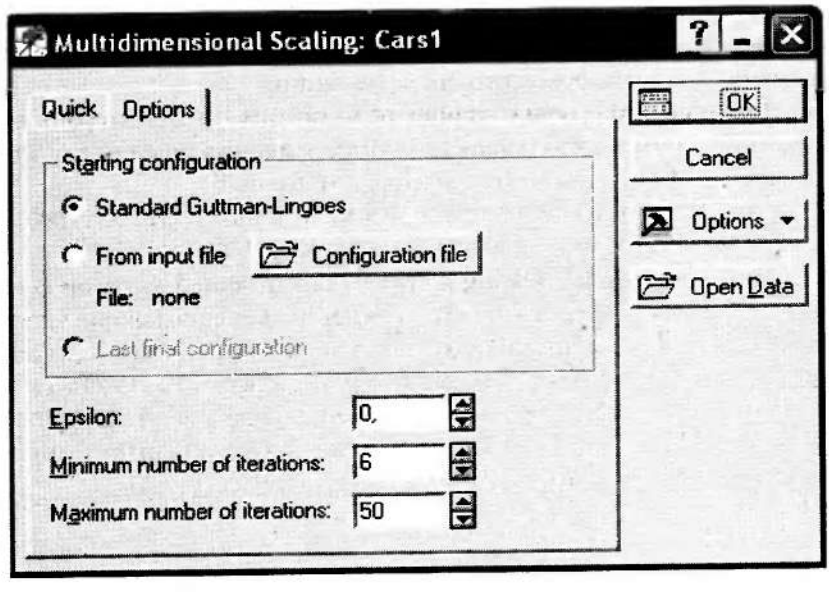


Рис. 15.2

- *From input file* (из файла). Начальная конфигурация строится из обычного файла с координатами, сохраненного в формате *STATISTICA*. Этот файл может быть создан с помощью модуля «Управление данными» или получен в качестве файла конфигурации в одном из предыдущих анализов по методу многомерного шкалирования. Этот файл начальной конфигурации должен содержать столько же случаев, сколько и текущий матричный файл. Кроме того, число переменных в нем должно быть не меньше выбранной для текущего анализа размерности;
- *Last final configuration* (последняя полученная конфигурация). После возвращения к стартовой панели из окна **Results**, по умолчанию, начальной конфигурацией является последняя полученная. В этом случае в качестве начальной конфигурации будет использоваться решение, полученное для предыдущей модели. Так, предположим, что только найдено трехмерное решение для матрицы сходства и программа вернулась к стартовой панели модуля. Если следующим шагом пользователь хотел бы найти двухмерное решение (не меняя установки **Starting configuration** в комбинированном поле), программа автоматически использует предыдущую полученную конфигурацию для первых двух координатных осей в качестве начальной для последующего анализа. Конечно, если были выбраны новые переменные (кнопка **Variables**) или пользователь увеличил размерность, то опция *Last final configuration* не будет доступна.

В поле **Epsilon** (Эпсилон) задается наименьшее расстояние, рассматриваемое программой как существенное или значимое. Все расстояния меньше указанного в этом поле значения будут рассматриваться программой как нулевые. Таким образом, чем меньше эта величина, тем большее число итераций потребуется и тем

более точное решение будет получено. Изменяйте этот параметр, только если итерационная процедура подгонки модели не может сойтись (т.е. невозможно получить решение) даже после большого числа итераций.

В поле **Minimum** и **Maximum number of iterations** (число итераций) можно указать минимальное и максимальное допустимое число итераций для алгоритма подгонки модели. Минимальное число итераций равно 6.

Оставьте это окно без изменений и нажмите кнопку **OK**. Откроется окно **Parameter Estimation** (оценивание параметров, рис. 15.3).

Модуль **Multidimensional Scaling** является реализацией методов неметрического многомерного шкалирования. В качестве начальной конфигурации программа вычисляет главные компоненты для матрицы сходства/различия. Затем программа запустит процесс итерационного поиска решения по методу наискорейшего спуска. Цель этих итераций — минимизировать нестандартизованный стресс — φ и коэффициент отчуждения Гутмана — k. Нестандартизованное значение стресса вычисляется, как было замечено, по формуле (15.1).

Коэффициент отчуждения k определяется как

$$k = \sqrt{1 - \left(\sum_{ij} d_{ij}\delta_{ij}\right)^2 / \sum_{ij} d_{ij}^2 \sum_{ij} \delta_{ij}^2} \qquad (15.2)$$

В общем случае программа *STATISTICA* старается минимизировать разницу между воспроизведенными расстояниями и монотонным преобразованием исходных расстояний. Другими словами, программа пытается воспроизвести расстояния с сохранением порядка их величин (отсюда и название неметрическое многомерное шкалирование).

Заметим, что при использовании метода наискорейшего спуска оцениваемые значения вычисляются с помощью перестановки ранговых образов Гутмана. При этом количество итераций, необходимых для их вычисления, записывается в первый столбец *s*. Полученные значения уточняются на втором этапе. Для этого после каждой итерации наискорейшего спуска программа выполняет до пяти итераций по методу преобразования с помощью монотонной регрессии — второй столбец *t*. Эта процедура минимизирует стандартизованный стресс S:

$$S = \sqrt{\sum_{ij} \left(d_{ij} - f(\delta_{ij})\right)^2 / \sum_{ij} d_{ij}^2} \qquad (15.3)$$

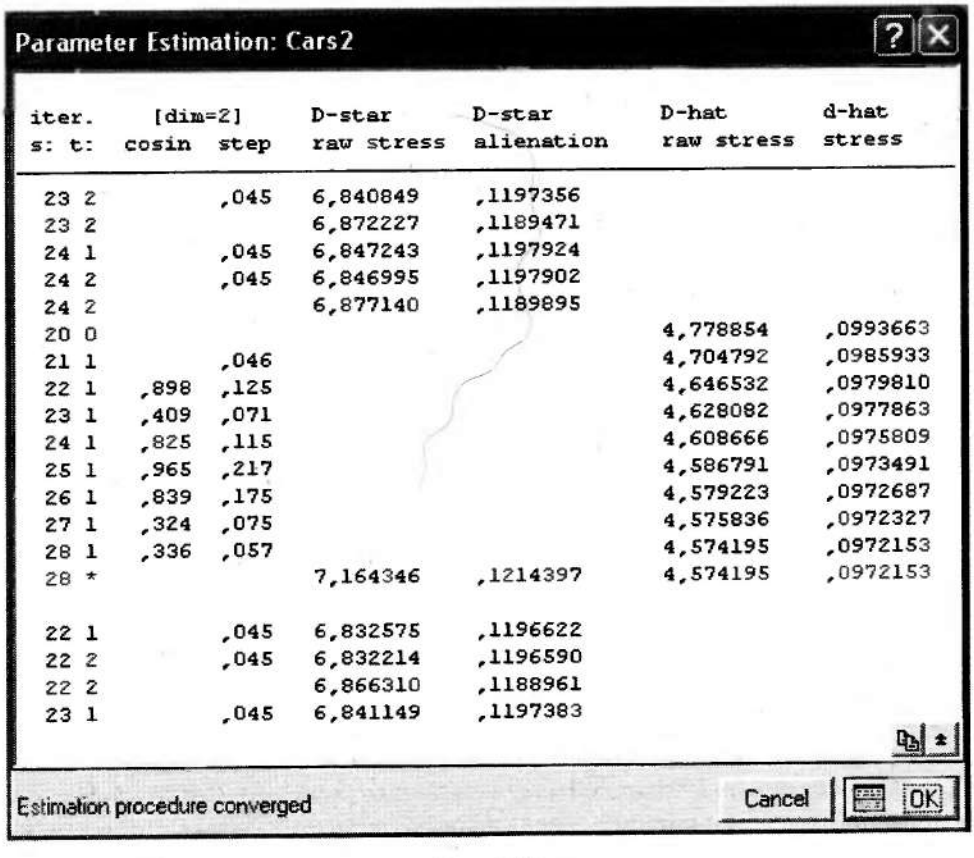


Рис. 15.3

Перед получением окончательной конфигурации программа выполнит несколько таких итераций.

D-star (*D* со звездочкой) вычисляются с помощью процедуры перестановки ранговых образов Гутмана. В общем случае эта процедура пытается воспроизвести порядок следования рангов расстояний в исходной матрице расстояний.

D-hat (*D* с крышечкой) вычисляются методом преобразования с помощью монотонной регрессии. При использовании этого метода программа пытается подобрать монотонное преобразование (регрессию), наиболее точно воспроизводящее исходные расстояния .

В итоге программа определит лучшую двухмерную конфигурацию. Она будет выделена синим цветом и * (28*).

Далее нажмите кнопку **ОК**. Появится окно **Results** (рис. 15.4).

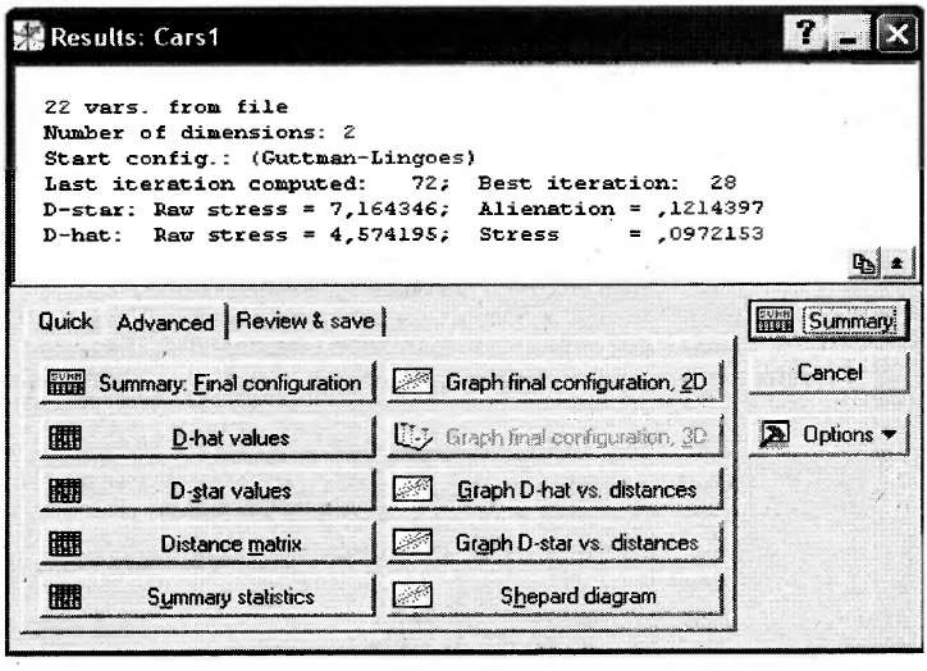


Рис. 15.4

В информационном поле окна приведены оценки параметров наилучшей конфигурации. На вкладке **Advanced** нажмите кнопку **Summary statistics** (итоговые статистики). Появится таблица, содержащая воспроизведенные расстояния (первый столбец *Distance*) и соответствующие им значения *D-hat и D-star* (рис. 15.5).

	Final Configuration (Cars1) D-star: Raw stress = 7,164346; D-hat: Raw stress = 4,574195;		
	Distance	D-star	D-hat
D(7, 6)	0,135718	0,082796	0,114886
D(21, 7)	0,104274	0,104274	0,114886
D(13, 2)	0,136754	0,127910	0,114886
D(17,10)	0,082796	0,135718	0,114886
D(22,19)	0,213889	0,136754	0,194637
D(21,10)	0,232409	0,138712	0,194637
D(22,10)	0,138712	0,149570	0,194637
D(15,10)	0,281578	0,150076	0,194637
D(17, 6)	0,203418	0,164541	0,194637
D(21, 6)	0,164541	0,203418	0,194637

Рис. 15.5

Элементы таблицы результатов будут отсортированы по исходным значениям *D-hat, D-star*, а для обозначения элементов используются их матричные индексы: *D(X, Y)*, где *X* — соответствующая строка в матрице входов, и *Y* — соответствующая колонка. Например, элемент *D*(7,6) соответствует элементу в 7-й строке и 6-м столбце матрицы расстояний, т.е. в нашем случае это расстояние между *Dodge* и *Chrysler*. Причем чем ближе значения *Distance, D-hat, D-star*, тем точнее определено расстояние относительно входных данных. Таким образом, можно сказать, что в нашем примере *D* (21,7) определено с большей степенью точности, а *D*(7,6) — с меньшей. Но в общем для двухмерного пространства, расстояния воспроизведены с приемлемой точностью.

Нажмите **Summary: Final configuration** (окончательная конфигурация). Откроется таблица результатов с координатами окончательной конфигурации (рис. 15.6).

	Final Configuration (Cars1) D-star: Raw stress = 7,164346; Alienation = ,1214397 D-hat: Raw stress = 4,574195; Stress = ,0972153	
	DIM. 1	DIM. 2
Acura	0,41051	-0,781206
Audi	-0,22895	0,238138
BMW	-0,38318	-0,146188
Buick	0,64401	0,083318
Corvette	-1,10836	-0,059648
Chrysler	0,06941	-0,096415

Рис. 15.6

Кнопка **D-hat values** открывает таблицу результатов с преобразованными входными значениями, вычисленными с помощью монотонной регрессии (рис. 15.7).

	D-hat 'distances' in Final Configuration (Cars1) D-star: Raw stress = 7,164346; Alienation = ,1214397 D-hat: Raw stress = 4,574195; Stress = ,0972153					
	Acura	Audi	BMW	Buick	Corvette	Chrysler
Acura	0,00000(	1,19656!	1,05729'	0,92410:	1,75717'	0,92109'
Audi	1,19656!	0,00000(	0,37049!	0,81890:	0,92410:	0,49774:
BMW	1,05729'	0,37049!	0,00000(	1,05729'	0,61561-	0,37049!
Buick	0,92410:	0,81890:	1,05729'	0,00000(	1,97410:	0,51666!
Corvette	1,75717'	0,92410:	0,61561-	1,97410:	0,00000(	1,19215!
Chrysler	0,92109'	0,49774:	0,37049!	0,51666!	1,19215!	0,00000(

Рис. 15.7

Кнопка **D-star values** открывает таблицу результатов с преобразованными входными значениями, вычисленными с помощью процедуры ранговых образов Гутмана (рис. 15.8).

	D-star 'distances' in Final Configuration (Cars1) D-star: Raw stress = 7,164346; Alienation = ,1214397 D-hat: Raw stress = 4,574195; Stress = ,0972153					
	Acura	Audi	BMW	Buick	Corvette	Chrysler
Acura	0,000000	1,276303	1,069499	1,034792	1,758188	0,893657
Audi	1,276303	0,000000	0,341673	0,775278	0,933454	0,482803
BMW	1,069499	0,341673	0,000000	1,113682	0,603111	0,414120
Buick	1,034792	0,775278	1,113682	0,000000	1,961975	0,566715
Corvette	1,758188	0,933454	0,603111	1,961975	0,000000	1,203314
Chrysler	0,893657	0,482803	0,414120	0,566715	1,203314	0,000000

Рис. 15.8

Кнопка **Distance matrix** открывает таблицу с матрицей расстояний, воспроизведенных для конфигурации точек в пространстве заданной размерности (рис. 15.9).

	Distances in Final Configuration (Cars1) D-star: Raw stress = 7,164346; Alienation = ,1214397 D-hat: Raw stress = 4,574195; Stress = ,0972153					
	ACURA	AUDI	BMW	BUICK	CORVETTE	CHRYSLER
ACURA	0,000000	1,203314	1,016462	0,895502	1,681548	0,765040
AUDI	1,203314	0,000000	0,414120	0,886576	0,928462	0,448266
BMW	1,016462	0,414120	0,000000	1,052518	0,730321	0,455323
BUICK	0,895502	0,886576	1,052518	0,000000	1,758188	0,602050
CORVETTE	1,681548	0,928462	0,730321	1,758188	0,000000	1,178343
CHRYSLER	0,765040	0,448266	0,455323	0,602050	1,178343	0,000000

Рис. 15.9

Теперь рассмотрим графическую интерпретацию данных.

Shepard diagram. После выбора этой опции программа построит диаграмму рассеяния, на которой изображена зависимость воспроизведенных расстояний от исходных расстояний (рис. 15.10). Расстояния отсортированы в порядке возрастания. Такая диаграмма рассеяния называется диаграммой Шепарда. По оси ординат Y откладываются воспроизведенные расстояния, а по оси абсцисс X — истинные расстояния между объектами. На этом графике также строится график ступенчатой функции. Ее линия представляет так называемые величины D с крышечкой, т.е. результат монотонного преобразования $f(\delta_{ij})$ исходных данных.

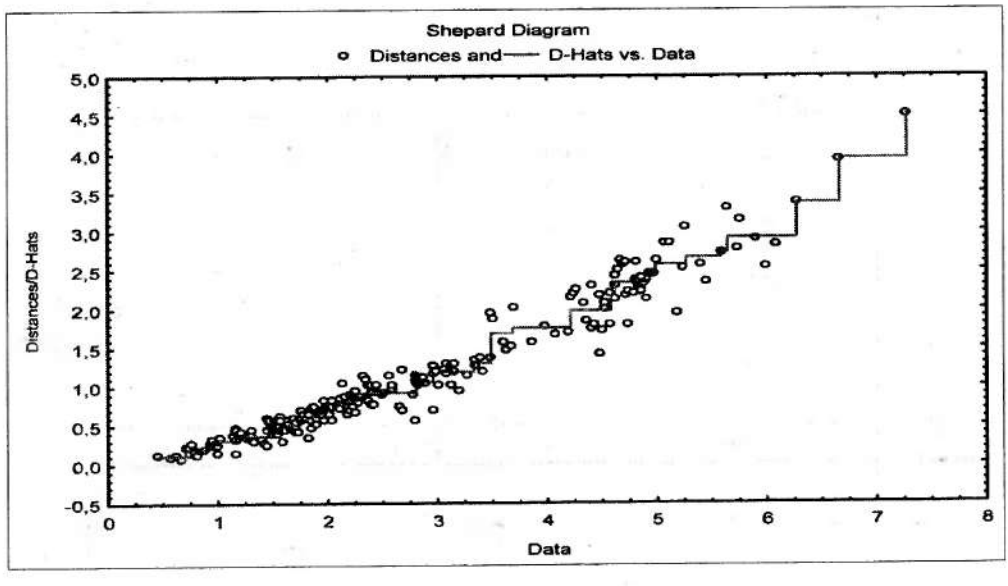


Рис. 15.10

Таким образом, эта ступенчатая функция будет либо монотонно возрастающей, либо монотонно убывающей. Чем лучше согласие ступенчатой функции с точками данных на диаграмме рассеяния, тем лучше согласие с моделью, т.е. тем лучше будет воспроизведение расстояний (а именно — точнее будет сохранен порядок следования их величин) в пространстве заданной размерности.

Если бы все воспроизведенные результирующие расстояния легли на эту ступенчатую линию, то ранги наблюдаемых расстояний (сходств) были бы в точности воспроизведены полученным решением (пространственной моделью). Отклонения от этой линии показывают на ухудшение качества согласия (т.е. качества подгонки модели).

Диаграмма Шепарда, представленная на рисунке, показывает достаточно незначительные отклонения от графика ступенчатой функции, что свидетельствует о хорошем качестве подгонки модели. По крайней мере, можно утверждать, что преобразование исходной матрицы расстояний размерностью 26×26 в матрицу координат объектов 26×2 произошло без существенной потери информации.

Graph final configuration 2D. Этот график позволяет отобразить окончательную конфигурацию объектов на плоскости. Нажмите кнопку **Graph final configuration, 2D** в окне **Results**, на вкладке **Advanced.** Появится окно, в котором надо назначить оси диаграммы рассеяния. Выберите, например, в колонке **First(X) — Dimession 1**, а в колонке **Second(Y) — Dimession 2**. После чего нажмите **OK**, появится двухмерный график (рис. 15.11).

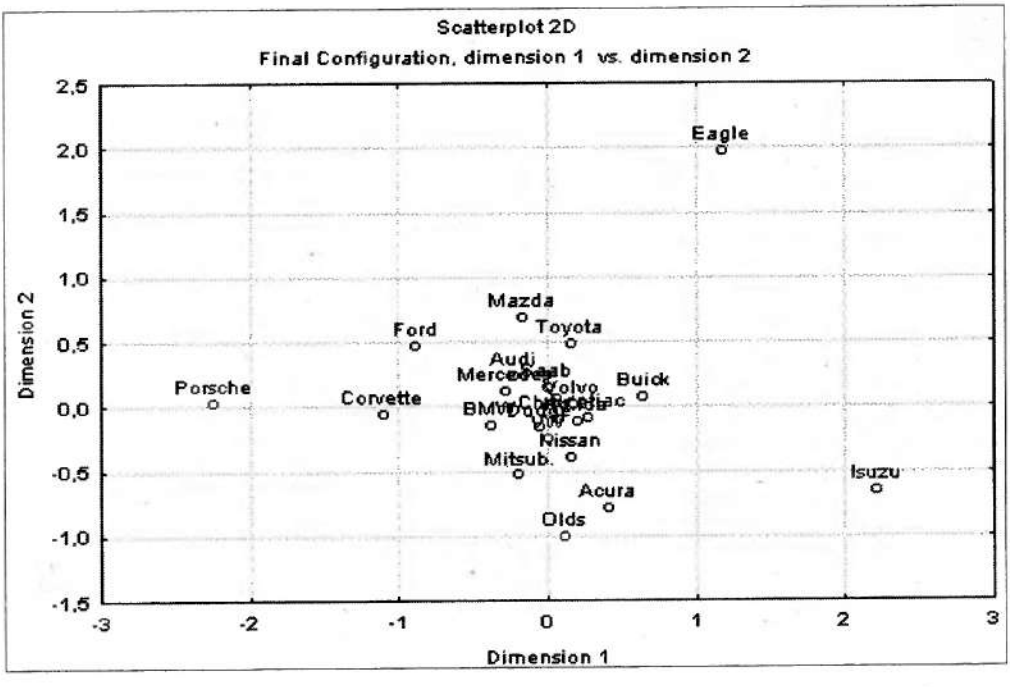


Рис. 15.11

На графике отчетливо видны марки автомобилей, «близких» друг другу по определенным характеристикам (*VOLVO*, *BMW*, *MERCEDES*, *NISSAN* и т.д.), на большем расстоянии от них находятся *OLDS*, *ACURA*, *CORVETTE*, *FORD*, *MAZDA*, и на значительном расстоянии от них расположены *PORSCHE*, *EAGLE*, *ISUZU*. Теперь труднообозримая матрица представлена на плоскости и легко «читается». Этот график является тем, за что идет борьба в многомерном шкалировании, — набор многомерных данных представлен ясно и отчетливо.

Graph D-hat vs. Distances. С помощью этого графика (рис. 15.12) можно визуализировать зависимость преобразованных значений входных данных (*D* с крышечкой) от преобразованных расстояний. Чем плотнее точки на графике группируются вокруг диагональной линии, тем лучше согласие с данными для выбранной модели.

Graph D-star vs. Distances. С помощью этого графика (рис. 15.13) можно визуализировать зависимость преобразованных значений входных данных (*D* со звездочкой) от преобразованных расстояний. Чем плотнее точки на графике группируются вокруг диагональной линии, тем лучше согласие с данными для выбранной модели.

Graph final configuration, 3D. Этот график позволяет отобразить окончательную конфигурацию объектов в трехмерном пространстве. В нашем примере соответствующая опция сейчас недоступна, так как рассматривали двухмерное пространство. Чтобы ее активизировать в окне **Results**, нажмите кнопку **Cancel**. После этого программа вернется в окно **Multidimensional Scaling**. Обратите внимание: теперь

в поле **Number of dimensions** стоит размерность 1, а не 2, как первоначально выбирали.

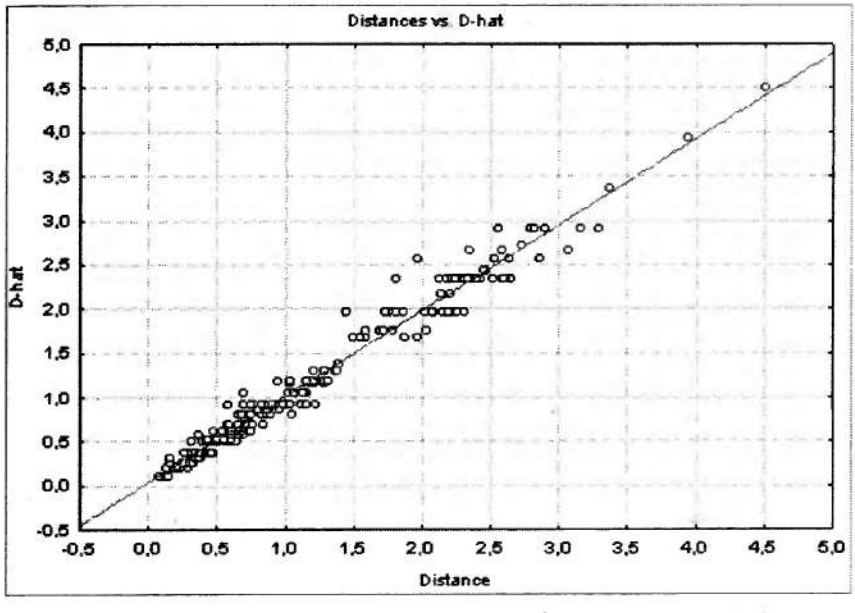


Рис. 15.12

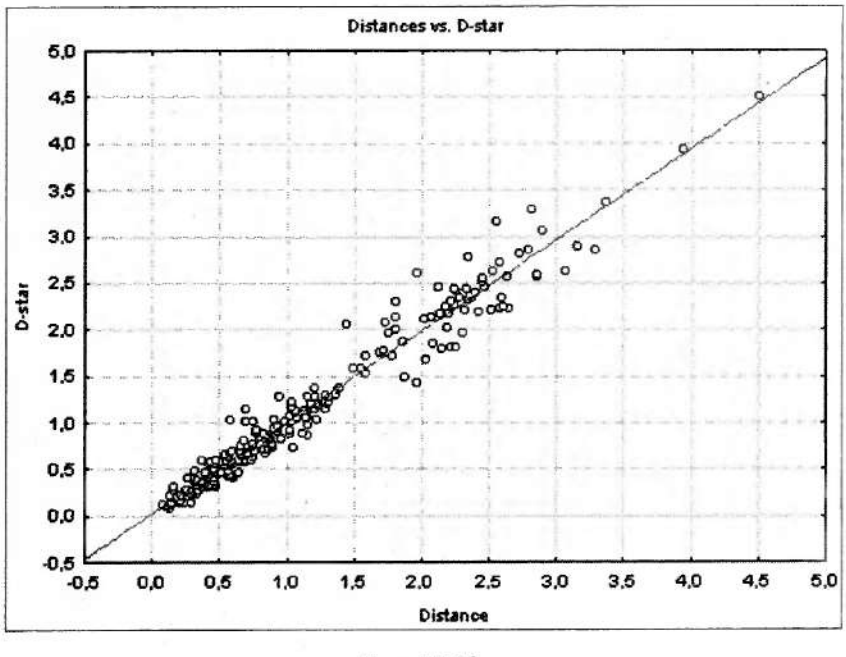


Рис. 15.13

Как упоминалось ранее, программа уменьшает размерность пространства при повторном обращении. Итак, для того чтобы посмотреть трехмерный график, надо в этом поле поставить цифру 3 и нажать **ОК**. После чего программа вернется в окно **Results**. Нажмите кнопку **Graph final configuration, 3D**, появится окно, в котором надо назначить оси диаграммы рассеяния. Поставьте в соответствие осям координат переменные и нажмите **ОК**. Появится график (рис. 15.14), из которого видно, что окончательная конфигурация объектов в трехмерном пространстве не противоречит конфигурации в двухмерном. Марки автомобилей *Porsche, Eagle, Isuzu* так же, как и в двухмерном случае, расположены на предпочтительном расстоянии от остальных марок автомобилей.

Рассмотрим кнопки на вкладке **Review&Save**. На этой вкладке (рис. 15.15) можно еще раз просмотреть и сохранить *Distance matrix* — таблицу результатов с матрицей расстояний, воспроизведенных для конфигурации точек в пространстве заданной размерности; *Start configuration* — матрицу данных начальной конфигурации; *Final configuration* — координаты точек окончательной конфигурации в стандартных файлах *STATISTICA*. Эти файлы могут быть использованы в других модулях *STATISTICA*.

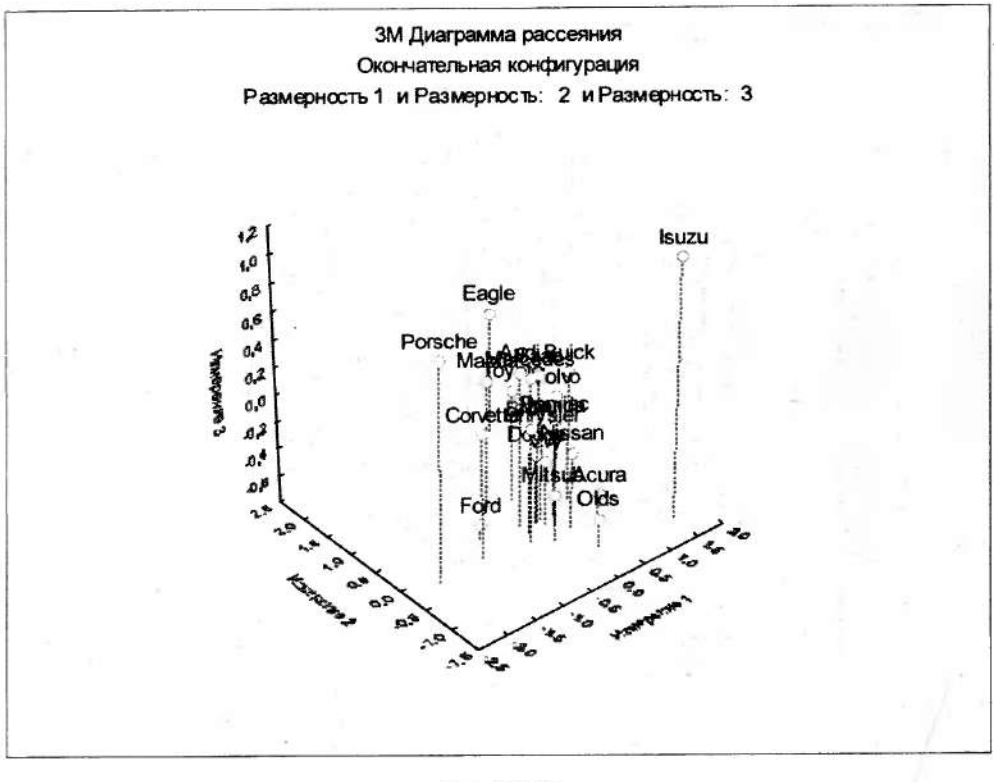


Рис. 15.14

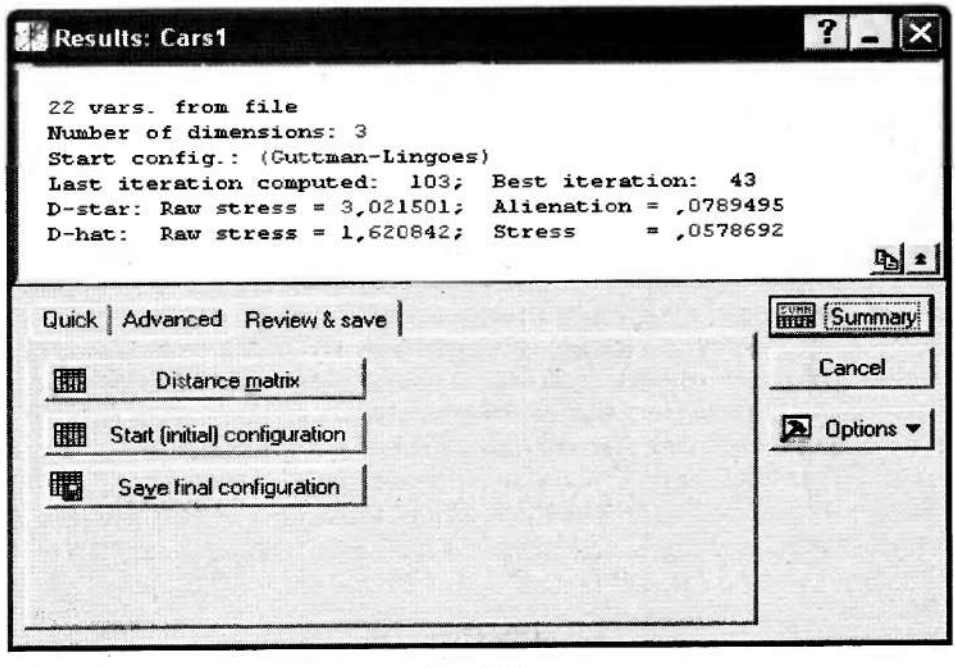


Рис. 15.15

Например, можно выбрать сохраненную конфигурацию в качестве начальной для последующего анализа. Также с помощью модуля «Управление данными» можно объединить файл с другим файлом, содержащим дополнительные переменные.

Чтобы проверить правильность выбора размерности пространства, как отмечалось ранее, используется критерий каменистой осыпи. Для построения графика кривой каменистой осыпи проведите последовательные вычисления *D-star:raw stress* для размерностей от 6 до 1. Запишите полученные данные в таблицу (рис. 15.16) и постройте линейный график — *Line Plot* (рис. 15.17).

	Каменистая осыпь
	1 **D-star: raw stress**
Dim1	30,04894
Dim2	6,274416
Dim3	2,995366
Dim4	0,6334
Dim5	0,0000086
Dim6	0,0000073

Рис. 15.16

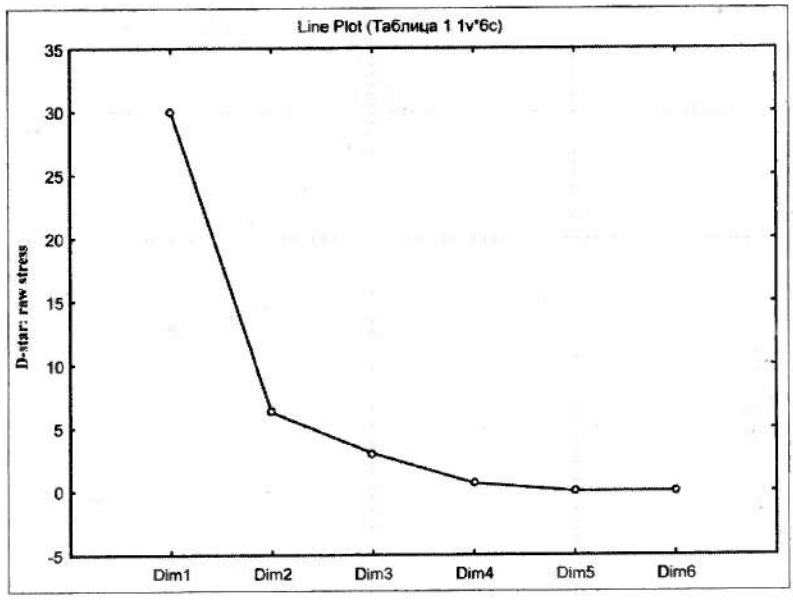


Рис. 15.17

Из графика согласно критерию каменистой осыпи (критерия Кэттела) следует, что для воспроизведения расстояний между марками автомобилей необходимо выбрать двухмерное пространство, так как в точке с абсциссой *Dim2* максимально замедляется уменьшение стресса.

15.4. Анализ соответствий

Анализ соответствий содержит описательные и разведочные методы анализа двухвходовых и многовходовых таблиц [16]. Эти методы позволяют исследовать структуру и взаимосвязь группирующих переменных, включенных в таблицу. Одной из наиболее общих разновидностей многовходовых таблиц являются частотные таблицы сопряженности. В классическом анализе соответствий частоты в таблице сопряженности стандартизуются таким образом, чтобы сумма наблюдений во всех ячейках была равна 1.

Практическое применение анализ соответствий может найти в таких сферах, как маркетинг, управление, социология, политика и т.д. К примеру, в маркетинге можно при выборе целевой аудитории новых маркетинговых программ с помощью анализа соответствий исследовать структуру группирования покупателей по личным предпочтениям. В политике результаты опросов общественного мнения населения могут быть представлены в виде таблиц сопряженности. Например, значения в ячейках этой таблицы могут обозначать количество голосов, отданных данному кандидату определенной социальной группой населения. Всестороннее исследование таблиц такого типа может быть проведено как раз с помощью анализа соответствий. В управлении может быть произведен анализ

распределения доходов между различными социальными группами. В социологии позволяет исследовать структуру группирующих переменных, включенных в двухвходовые и многовходовые таблицы. А также путем анализа соответствия можно исследовать предпочтительность того или иного программного продукта в зависимости от конкретных к нему требований.

Одна из целей анализа соответствий — представление содержимого таблицы относительных частот в виде расстояний между отдельными строками и/или столбцами таблицы в пространстве, возможно, более низкой размерности. Каким образом это достигается, лучше всего показать на простом примере.

Допустим, что собраны данные о пристрастии к курению сотрудников некоторой компании. Это простая двухвходовая таблица (рис. 15.18).

Сотрудники	Категории курящих				Общее к-во
Группа сотрудников	(1) Некурящие	(2) Слабо	(3) Средне	(4) Сильно	Всего по строке
(1) Старшие менеджеры	4	2	3	2	11
(2) Младшие менеджеры	4	3	7	4	18
(3) Старшие сотрудники	25	10	12	4	51
(4) Младшие сотрудники	18	24	33	13	88
(5) Секретари	10	6	7	2	25
Всего по столбцу	61	45	62	25	193

Рис. 15.18

Можно считать, что 4 числа в каждой строке данной таблицы являются координатами 4-мерного пространства и значит можно вычислить расстояния (евклидовы) между 5 точками (строками) этого 4-мерного пространства. Расстояния между данными точками в 4-мерном пространстве отражают всю информацию о сходствах между строками в том смысле, что чем меньше расстояние, тем больше сходство между категориями курящих. Теперь предположим, что возможно найти пространство меньшей размерности для представления точек-строк, которое сохраняет всю или почти всю информацию о различиях между строками. Например, представить всю информацию о сходстве между строками (категории работников) в виде 1, 2 или 3-мерного графика. Это может и не быть практически полезным для маленьких таблиц, но сильно выиграет представление и интерпретация очень больших таблиц (в которых, например, записаны предпочтения для 10 потребительских товаров 100 групп респондентов) в результате упрощения, полученного путем применения методов анализа соответствий. Например, представить 10 потребительских товаров в двумерном пространстве.

Рассмотрим вычислительные аспекты работы программы. Вычисляются относительные частоты для введенной таблицы, так что сумма всех элементов таблицы будет равна 1 (каждый элемент делится на 193 — общее число наблюдений). Полученная нормированная таблица показывает, как распределена единичная масса по ячейкам.

В терминологии анализа соответствий суммы по строкам и столбцам в матрице относительных частот называются массой строки и столбца соответственно.

Инерция определяется как значение статистики χ^2 (Хи-квадрат) Пирсона для двухвходовой таблицы, деленное на общее количество наблюдений (193 в примере).

Критерий χ^2 Пирсона — это наиболее простой критерий проверки значимости связи между двумя категоризованными переменными. Критерий Пирсона основывается на том, что в двухвходовой таблице ожидаемые частоты при гипотезе «между переменными нет зависимости» можно вычислить непосредственно [10]. Это непараметрический критерий, его применение никак не связано с распределением табулированных переменных. Рассмотрим двухмерную таблицу сопряженности $n(i,j)$; $i = 1, 2, ..., r, j = 1, 2, ..., s$, состоящую из r строк и s столбцов.

Обозначим

$$n(i) = n(i,1) + n(i,2) + ... + n(i,s), i = 1, 2, ..., r,$$
$$n(j) = n(1,j) + n(2,j) + ... + n(r,j), j = 1, 2, ..., s, n = \Sigma\, n(i,j),$$

где $n(i)$, $n(j)$ называются маргинальными частотами, так как они располагаются по краям таблицы. В терминах анализа соответствий маргинальные частоты называются *профилями*.

Ожидаемой частотой $\underline{n}(i,j)$, соответствующей наблюдаемой частоте $n(i,j)$, называется произведение маргинальных частот, соответствующих строке i и столбцу j, деленное на общее число наблюдений n, т.е. $\underline{n}(i,j) = n(i) \times n(j)/n$.

Статистика χ^2 Пирсона вычисляется по формуле

$$\chi^2\ Пирсона = \Sigma\ (n(i,j) - \underline{n}(i,j))^2 / n(i,j).$$

Эта статистика замечательна тем, что при достаточно большом числе наблюдений ее распределение можно приблизить распределением χ^2 и, значит, вычислить приближенный *p-уровень* критерия.

Если строки и столбцы таблицы полностью независимы друг от друга, то элементы таблицы могут быть воспроизведены (через ожидаемые частоты) при помощи профилей строк и столбцов. Любое отклонение от ожидаемых величин (ожидаемых при гипотезе о полной независимости переменных по строкам и столбцам) будет давать вклад в совокупную статистику χ^2 Пирсона. Поэтому, чем более зависимы строки и столбцы, тем большее значение принимает статистика Пирсона и меньшее *p-уровень* критерия. Таким образом, анализ соответствий можно рассматривать как метод декомпозиции статистики χ^2 Пирсона для двухвходовых таблиц (инерция = χ^2/число наблюдений) с целью определения пространства наименьшей размерности, позволяющего представить отклонения от ожидаемых величин. Это напоминает задачу факторного анализа, где осуществляется декомпозиция совокупной вариации, так чтобы снижение размерности переменных приводило к наименьшим потерям в матрице ковариаций исходных переменных.

Очевидно, что не меньший интерес могут вызывать суммарные величины по столбцам, в этом случае можно представить точки-столбцы в пространстве меньшей размерности, которое удовлетворительно воспроизводит сходство (и расстояния) между относительными частотами для столбцов таблицы. В действительности возможно одновременное отображение на одном графике точек-столбцов и точек-строк, представляющее всю имеющуюся информацию, содержащуюся в двухвходовой таблице.

Вычисления проводятся над следующими матрицами:

- P обозначает матрицу относительных частот, т.е. каждый элемент P вычисляется как соответствующая частота из таблицы ввода, деленная на сумму всех элементов таблицы ($n(i,j) / n$);
- r обозначает вектор сумм элементов строк матрицы P;
- c обозначает вектор сумм элементов столбцов матрицы P;
- D_r обозначает диагональную матрицу, элементы главной диагонали D_r равны соответствующим суммам элементов строк P;
- D_c обозначает диагональную матрицу, элементы главной диагонали D_c равны соответствующим суммам элементов столбцов P.

Вычисление координат строк и столбцов в пространстве меньшей размерности базируется на обобщенном сингулярном разложении матрицы P, которое имеет вид $P = A D_u B$, так что

$$A D_r^{-1} A = B D_c^{-1} B = I,$$

где A — матрица левосторонних обобщенных сингулярных векторов, B — матрица правосторонних обобщенных сингулярных векторов, D_u — диагональная матрица, диагональные элементы которой равны обобщенным сингулярным числам, D_r^{-1} — матрица, обратная к D_r, D_c^{-1} — матрица, обратная к D_c^{-1} и I — единичная матрица (диагональная матрица, все диагональные элементы которой равны 1).

Вычисление координат для точек-строк и точек-столбцов зависит от выбора способа стандартизации. При различных способах стандартизации сохраняется расположение точек-строк и точек-столбцов относительно друг друга.

При канонической стандартизации координаты строк вычисляются по формуле

$$F = D_r^{-1} A D_u^S,$$

а координаты столбцов по формуле

$$G = D_c^{-1} B D_u^S.$$

При стандартизации по профилям строк координаты строк вычисляются по матрице профилей строк $R = D_r^{-1} P$, а именно: координаты строк вычисляются по формуле

$$F = D_r^{-1} A Du,$$

а координаты столбцов по формуле

$$G = D_c^{-1} B.$$

Данная опция удобна, когда вы заинтересованы в интерпретации расстояний между точками-строками, координаты столбцов не должны при этом рассматриваться.

При стандартизации по профилям столбцов координаты столбцов вычисляются по матрице профилей столбцов, а именно: координаты столбцов вычисляются по формуле

$$F = D_c^{-1}BD_u,$$

а стандартные координаты строк по формуле

$$G = D_r^{-1}A.$$

Данная опция удобна, когда вы заинтересованы в интерпретации расстояний между точками-столбцами, координаты строк не должны при этом рассматриваться.

Продолжим рассмотрение вычислительных аспектов анализа соответствий в процессе изучения модуля **Correspondence Analysis.**

15.5. Описание модуля *Correspondence Analysis*

Рассмотрим последовательность шагов при работе с модулем **Correspondence Analysis**. Из библиотеки **Example** откройте файл данных **Smoking** (рис. 15.19). В файле приведены группы сотрудников: *SR.MANAGERS* (старшие менеджеры), *JR.MANAGERS* (младшие менеджеры), *SR.EMPLOYEES* (старшие сотрудники), *JR.EMPLOYEES* (младшие сотрудники), *SECRETARIES* (секретари); категории курящих: *NONE* (нет), *LIGHT* (слабо), *MEDIUM* (средне), *HEAVY* (сильно) и соответствующие им частоты.

	Simple correspondence analysis e			
	1 NONE	2 LIGHT	3 MEDIUM	4 HEAVY
SR.MANAGERS	4	2	3	2
JR.MANAGERS	4	3	7	4
SR.EMPLOYEES	25	10	12	4
JR.EMPLOYEES	18	24	33	13
SECRETARIES	10	6	7	2

Рис. 15.19

В верхнем меню **Statistics** щелкните по **Multivariate Exploratory Techniques** (многомерные исследовательские методы) и выберите команду **Correspondence Analysis**. Откроется стартовая панель модуля, в котором по умолчанию открыта вкладка стандартного (обычного) анализа соответствий **Correspondence Analysis** (рис. 15.20).

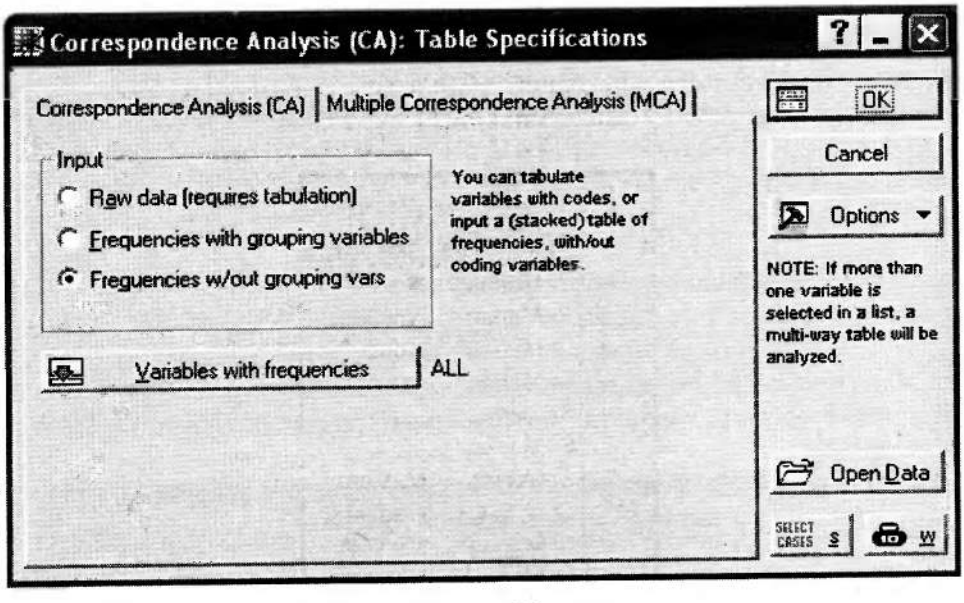


Рис. 15.20

В поле **Input** надо определить характер вводимых данных.

Опция *Raw data (requires tabulation)* (исходные данные, требуется табуляция) означает [6], что программа ожидает на входе категоризованные переменные с кодами, однозначно определяющими, к какой категории принадлежит каждый объект, например, данные могут быть в виде таблицы, являющейся фрагментом файла **Smoking3** из библиотеки **Example** (рис. 15.21). Другими словами, исходные данные представлены в виде таблицы, строки которой соответствуют наблюдениям (например, сотрудники компании), а столбцы — переменным, характеризующим наблюдения. Но среди этих переменных должны быть минимум две категоризованные переменные.

Нажмите кнопку **Row and column variable(s)** (переменные для строк и столбцов), откроется окно **Select Coding Variables Specifying the Table** (выберите кодирующие переменные для строк и столбцов). Укажите переменные, соответствующие строкам и столбцам таблицы (рис. 15.22).

Далее нажмите кнопку **Codes for grouping variables** (коды для группирующих переменных) и выберите коды переменных для анализа. Щелкните по **OK**, программа построит двухвходовую таблицу частот и откроется окно **Correspondence Analysis Results** (результаты анализа соответствий).

Опция *Frequencies with grouping variables* (частоты с группирующими переменными) означает, что программа ожидает в качестве входных данных категоризованные переменные с кодами, однозначно определяющими принадлежность к той или иной категории. Но дополнительно должна быть переменная, содержащая частоты или какие-либо другие величины, опреде ляющие меру соответствия для категорий рассматриваемых группирующих переменных. Например, файл данных может иметь вид, изображенный на рис. 15.23. При табуляции программа

в ячейке, соответствующей кодам категоризованных переменных проставит частоту, которая стоит в матрице исходных данных в строке с именами соответствующих кодов. Например, в исходной таблице переменные могут быть закодированы так, как это сделано в файле **Smoking2** (рис. 15.23).

	1 EMPLOYEE	2 SMOKING
1	Sr.Manag	None
2	Sr.Manag	None
3	Sr.Manag	None
4	Sr.Manag	None
5	Sr.Manag	Light
6	Sr.Manag	Light
7	Sr.Manag	Medium
8	Sr.Manag	Medium
9	Sr.Manag	Medium
10	Sr.Manag	Heavy
11	Sr.Manag	Heavy
12	Jr.Manag	None
13	Jr.Manag	None
14	Jr.Manag	None
15	Jr.Manag	None
16	Jr.Manag	Light
17	Jr.Manag	Light

Рис. 15.21

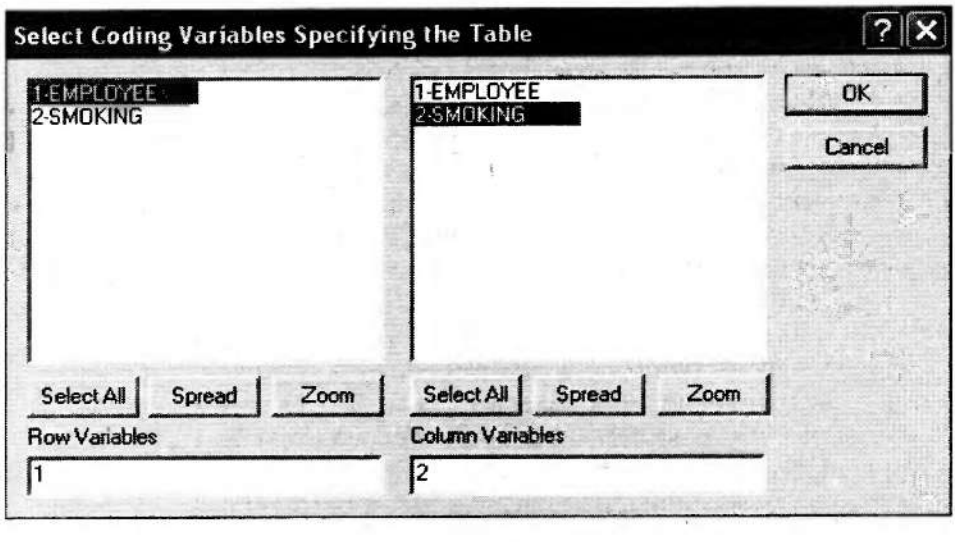


Рис. 15.22

При выборе этой опции активными являются кнопки **Row and column variable(s)** (переменные для строк и столбцов) **Variable with frequencies/counts** (переменная с частотами/количествами совпадений), **Codes for grouping variables** (коды для группирующих переменных). Если в таблице имеется несколько ссылок на одну и ту же ячейку, то соответствующие величины «частотной» переменной будут складываться, и полученная сумма будет приписана соответствующей ячейке таблицы. Для запуска процедуры **Correspondence Analysis** надо последовательно нажать на три кнопки: **Row and column variable(s)**, **Variable with frequencies/counts, Codes for grouping variables**.

Опция *Frequencies w/out grouping vars* (частоты без группирующих переменных) означает, что программа ожидает, что выбранные переменные (и объекты) содержат только частоты (или некоторые другие меры соответствия). Например, файл данных **Smoking.sta** организован подобным образом. Нажмите кнопку **Variables with frequencies** (переменные с частотами), откроется окно (рис. 15.24) **Select Variables with Frequencies (Counts)** (выберите переменные с частотами), в котором надо указать переменные, содержащие некоторую меру соответствия, сходства, неупорядоченности, связи и т.д. Например, нажмите кнопку **Select All** (выбрать все) и щелкните **OK**. Откроется окно **Correspondence Analysis Results** (рис. 15.25) на вкладке **Quick**. Из информационной части окна следует, что зависимость между группами сотрудников и категориями курящих присутствует (*p-уровень* критерия значительно меньше 1), но слабая (*p-уровень* больше чем 0,05).

	1 EMPLOYEE	2 SMOKING	3 FREQUENCIES
1	Sr.Manag	None	4
2	Sr.Manag	Light	2
3	Sr.Manag	Medium	3
4	Sr.Manag	Heavy	2
5	Jr.Manag	None	4
6	Jr.Manag	Light	3
7	Jr.Manag	Medium	7
8	Jr.Manag	Heavy	4
9	Sr.Empl	None	25
10	Sr.Empl	Light	10
11	Sr.Empl	Medium	12
12	Sr.Empl	Heavy	4
13	Jr.Empl	None	18
14	Jr.Empl	Light	24
15	Jr.Empl	Medium	33
16	Jr.Empl	Heavy	13
17	Secretar	None	10
18	Secretar	Light	6
19	Secretar	Medium	7
20	Secretar	Heavy	2

Рис. 15.23

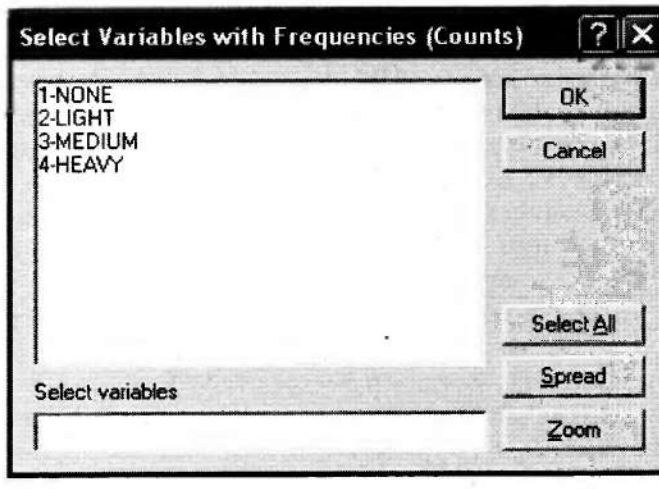


Рис. 15.24

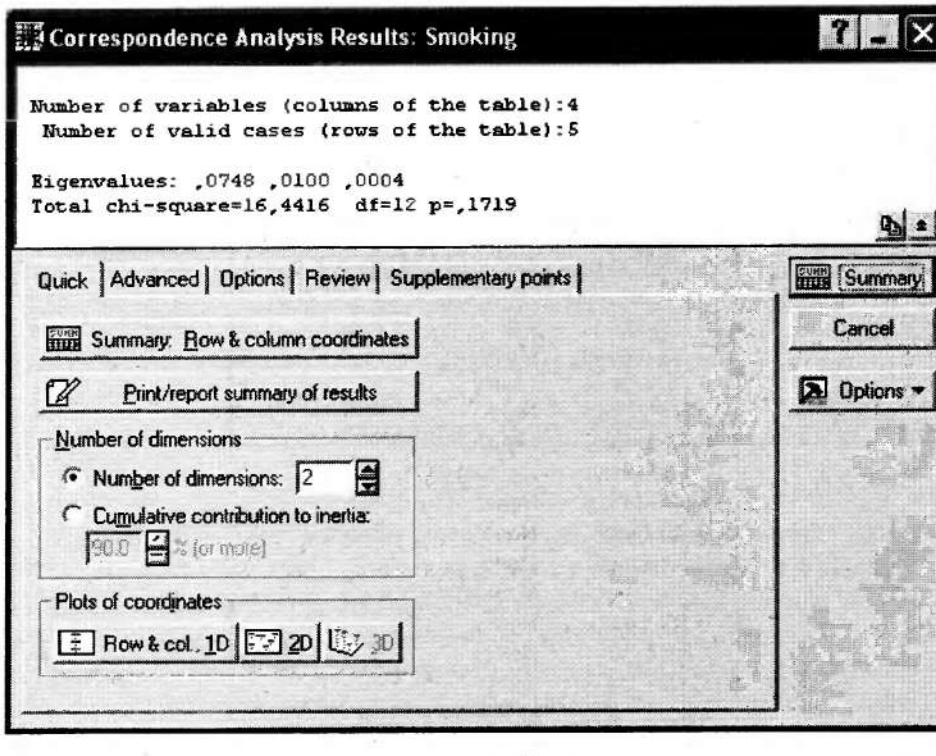


Рис. 15.25

Нажмите кнопку **Summary. Row & column coordinates** (координаты строк и столбцов). Появятся две таблицы результатов (одна для точек-строк, другая для точек-столбцов) с координатами строк и столбцов, а также многочисленные

статистики, которые позволяют оценить адекватность выбора размерности пространства для данного решения (рис. 15.26—15.27). Если имеются дополнительные точки-строки и/или точки-столбцы, то информация об их координатах вместе с соответствующими статистиками также выводится в окно результатов.

В столбцах 2, 3 приведены координаты строк (столбцов) в пространстве заданной размерности. Метод вычисления координат и их интерпретация зависят от выбранного метода стандартизации в рамке **Standardization** (стандартизация) на вкладке **Options** (опции).

	Row Coordinates and Contributions to Inertia (Smoking) Input Table (Rows x Columns): 5 x 4 Standardization: Row and column profiles							
Row Name	Row Number	Coordin. Dim.1	Coordin Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine Dim.1
SR.MANAGERS	1	-0,07	0,19	0,06	0,89	0,03	0,00	0,09
JR.MANAGERS	2	0,26	0,24	0,09	0,99	0,14	0,08	0,53
SR.EMPLOYEES	3	-0,38	0,01	0,26	1,00	0,45	0,51	1,00
JR.EMPLOYEES	4	0,23	-0,06	0,46	1,00	0,31	0,33	0,94
SECRETARIES	5	-0,20	-0,08	0,13	1,00	0,07	0,07	0,87

Рис. 15.26

	Column Coordinates and Contributions to Inertia (Smoking) Input Table (Rows x Columns): 5 x 4 Standardization: Row and column profiles								
Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine Dim.1	Ine Di
NONE	1	-0,39	0,03	0,32	1,00	0,58	0,65	0,99	
LIGHT	2	0,10	-0,14	0,23	0,98	0,08	0,03	0,33	
MEDIUM	3	0,20	-0,01	0,32	0,98	0,15	0,17	0,98	
HEAVY	4	0,29	0,20	0,13	0,99	0,19	0,15	0,68	

Рис. 15.27

Столбец *Mass* (масса) содержит соответствующие суммы по строкам или столбцам для таблицы относительных частот.

Столбец *Quality* (качество) содержит информацию о качестве представления соответствующей точки-строки или точки-столбца в координатной системе соответствующей размерности, которая была выбрана пользователем. Качество точки определено как отношение квадрата расстояния от данной точки до начала координат в пространстве сниженной размерности к квадрату расстояния от данной точки до начала координат в пространстве максимальной размерности (напомним, что метрикой здесь является метрика χ^2 Пирсона). Качество изменяется в пределах от 0 до 1. Чем ближе значение к 1, тем качество выше. При увеличении размерности пространства качество возрастает. Мера качества, вычисляемая в программе *STATISTICA*, не зависит от выбранного метода стандартизации. Низкое качество

означает, что данные строка или столбец недостаточно хорошо представлены в пространстве с заданным числом измерений — размерностью пространства.

Качество точки представляет долю вклада данной точки в величину общей инерции (χ^2) в пространстве заданной размерности. Однако эта величина не показывает, насколько соответствующая точка в действительности обусловливает общую инерцию (величину χ^2).

Столбец *Relative Inertia* (относительная инерция) представляет долю общей инерции, которая приходится на данную точку, и не зависит от размерности, которую выбрал пользователь.

Столбец *Inertia Dim* (относительная инерция каждого измерения) содержит относительный вклад соответствующей точки в инерцию, приходящуюся на рассматриваемое измерение — ось координат. Таким образом, данная величина отражена в отчете для каждой точки-строки или точки-столбца и для каждого измерения.

Столбец *Cosine 2* (косинус 2) содержит качество для каждой точки, обусловленное соответствующим измерением. Сумма этих величин для некоторой рассматриваемой строки по всем измерениям равна качеству данной точки-строки. Эта величина может также интерпретироваться как корреляция данной точки с соответствующим измерением. Термин *Cosine 2* связан с геометрической интерпретацией, так как это квадрат косинуса угла между вектором с координатами данной точки и рассматриваемой осью.

Кнопка **Print/report summary of results** (печать/вывод в *отчет*) предназначена для занесения информации, расположенной в информационном поле, и для отображения таблиц результатов анализа в отчете. Если текущие установки **Output Manager** (диспетчер вывода) не определяют вывод в отчет, при нажатии этой кнопки появится диалоговое окно с предложением изменить параметры **Output Manager**.

Рамка **Number of dimensions** (размерность) позволяет определить число измерений, для которых будут вычисляться координаты строк и столбцов. При установке в одноименном поле определяется фиксированное число измерений в поле редактирования.

При выборе опции *Cumulative contribution to inertia* (кумулятивный вклад в инерцию) программа определит минимальное число необходимых измерений по кумулятивной величине инерции, которая не должна превышать кумулятивный процент, задаваемый пользователем в поле ввода под данной опцией. Отметим, что максимальное число собственных значений, которые можно получить для двухвходовой таблицы, равно *min(r,s)-1*. Если выбрать максимально возможное число измерений, то можно полностью воспроизвести информацию, содержащуюся в таблице ввода.

Опции в рамке **Plots of coordinates** (графики координат) позволяют отобразить координаты строк и столбцов совместно с координатами дополнительных точек на *1D*, *2D* или *3D* графике. Можно использовать средство *Brushing* (кисть) для удаления некоторых точек графика (например, оставив только те точки, которые четко отделены от остальных). Ориентация осей произвольна, они могут

поворачиваться, например, на 180°. Можно быстро осуществить данный «поворот шкал», щелкнув правой кнопкой мыши на графике. Помимо имеющихся здесь графиков существуют и другие графики, доступные на вкладке **Advanced** (расширенный). Если нажать на любую из кнопок *1D*, *2D* или *3D*, то соответствующий график будет выведен на экран для всех измерений, пар измерений или триплетов измерений согласно установкам в рамке **Number of dimensions** (размерность).

При установках в поле **Number of dimensions,** равным 1 и 2, нажмите последовательно кнопки *1D* и *2D*. Программа отобразит координаты обоих факторов — *Группы сотрудников* и *Категории курящих* на одномерной (рис. 15.28) и двумерной (рис. 15.29) диаграммах.

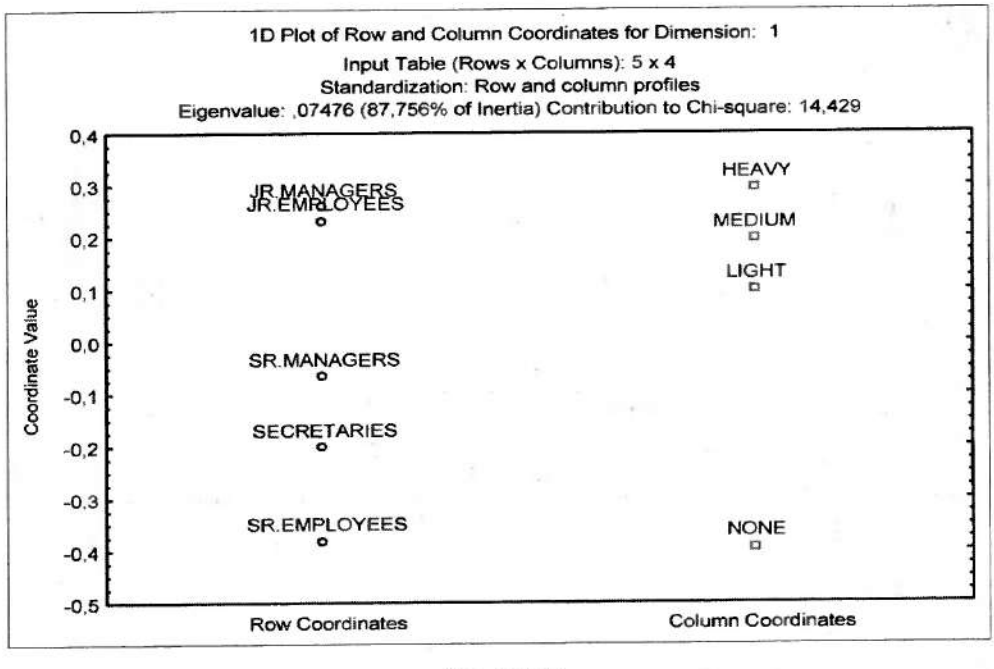


Рис. 15.28

На графиках одновременно отображены оба фактора в виде точек в пространствах меньшей размерности (1 и 2), которые максимально точно воспроизводят сходство и расстояния между относительными частотами для строк столбцов таблицы исходных данных.

Преимущество двухмерного пространства заключается в том, что строки, отображаемые в виде близких точек, близки друг к другу и по относительным частотам.

Перейдите на вкладку **Review** и нажмите кнопку **Row percentages**, программа построит таблицу относительных частот в процентах по строке (сумма элементов строки = 100%). Из рис. 15.28 видно, что, рассматривая положение точек по первой оси, *SR.EMPLOYEES* и *SECRETARIES* относительно близки по координатам. Такое же сходство этих групп сотрудников просматривается и из строк таблицы

относительных частот, изображенной на рис. 15.30. Наибольшее различие присутствует между *SR.EMPLOYEES* и *JR.MANAGERS*.

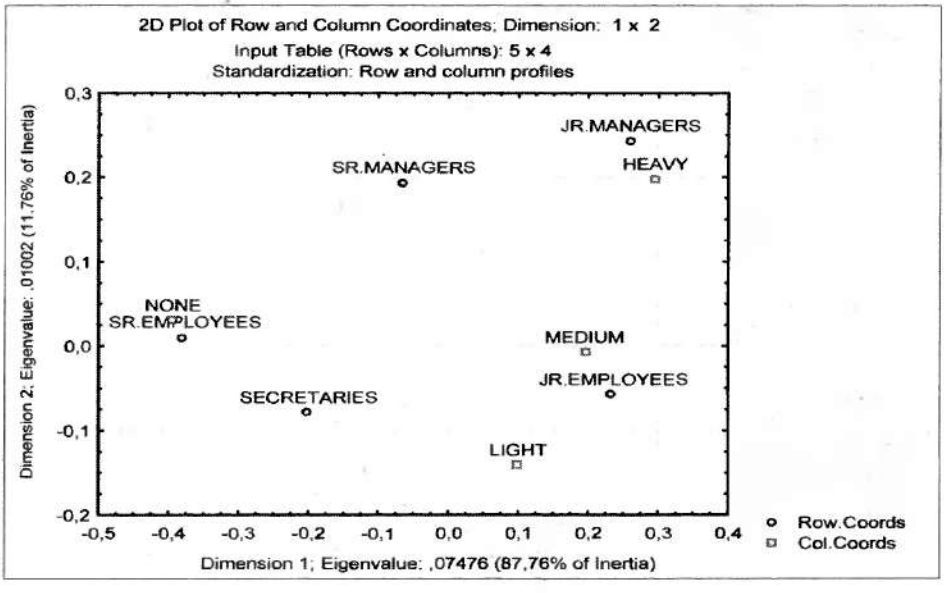


Рис. 15.29

	Percentages of Row Totals (Smoking) Input Table (Rows x Columns): 5 x 4 Total Inertia=,08519 Chi2=16,442 df=12 p=,17190				
	NONE	LIGHT	MEDIUM	HEAVY	Total
SR.MANAGERS	36,36364	18,18182	27,27273	18,18182	100,0000
JR.MANAGERS	22,22222	16,66667	38,88889	22,22222	100,0000
SR.EMPLOYEES	49,01961	19,60784	23,52941	7,84314	100,0000
JR.EMPLOYEES	20,45455	27,27273	37,50000	14,77273	100,0000
SECRETARIES	40,00000	24,00000	28,00000	8,00000	100,0000

Рис. 15.30

Самое главное — правильный графический анализ результатов [10]. Обычно горизонтальная ось соответствует максимальной инерции (87,76%). Вертикальная ось — минимальной инерции (11,76%). Рядом со значениями инерции указаны собственные значения. Пересечение двух осей — это центр тяжести наблюдаемых точек (точка с координатами 0,0), соответствующий средним профилям. Если точки принадлежат одному и тому же типу (являются либо строками, либо столбцами), то чем меньше расстояние между ними, тем теснее связь. Для того чтобы установить связь между точками разного типа (между строками и столбцами), следует рассмотреть углы между ними с вершиной в центре тяжести. Общая схема визуальной оценки степени зависимости такова:

- выберите две произвольные точки разного типа (например, *SR.EMPLOYEES* и *NONE*);
- соедините их отрезками прямых с центром тяжести;
- если образованный угол острый, то строка и столбец положительно коррелированны друг с другом (в нашем случае это означает, что старшим сотрудникам соответствует больший процент некурящих);
- если образованный угол тупой, то корреляция между переменными отрицательная;
- если угол прямой, то корреляция отсутствует.

Из приведенных правил следует, что сходство между *SR.EMPLOYEES* и *SECRETARIES* можно объяснить положительной корреляцией с переменной *NONE*. Из таблицы (рис. 15.30) видно, что этим категориям сотрудников соответствуют наибольшие проценты некурящих (соответственно 49 и 40%). Между *SR.MANAGERS* и *MEDIUM* корреляция практически равна нулю (этой категории сотрудников соответствует небольшой процент средне курящих — 27%). Между *JR.EMPLOYEES* и *NONE* корреляция отрицательная, что также подтверждается данными из таблицы на рис. 15.30: этой категории сотрудников соответствует минимальный процент некурящих — 20%. Схема визуальной оценки степени зависимости между строками и столбцами хорошо работает при сильной зависимости между группирующими переменными (*p-уровень* критерия χ^2 меньше чем 0,05).

На вкладке **Options** (рис. 15.31) можно выбрать метод стандартизации.

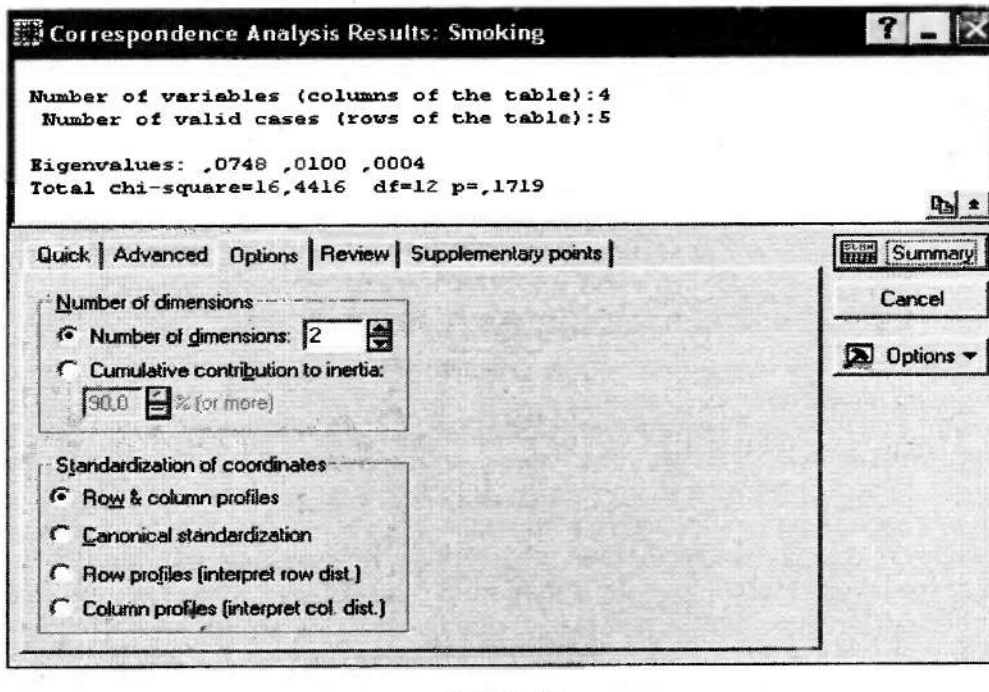


Рис. 15.31

Необходимость в стандартизации вызвана следующими соображениями. Во введении термин «расстояние» также использовался для обозначения различий между строками и столбцами матрицы относительных частот, которые в свою очередь представлялись в пространстве меньшей размерности в результате использования методов анализа соответствий. В действительности расстояния, представленные в виде координат в пространстве соответствующей размерности, — это не просто евклидовы расстояния, вычисленные по относительным частотам столбцов и строк, а некоторые взвешенные расстояния. Процедура подбора весов устроена таким образом, чтобы в пространстве более низкой размерности метрикой являлась бы метрика χ^2 Пирсона. При этом, если сравниваете точки-строки, то выбираете стандартизацию профилей строк или стандартизацию профилей строк и столбцов; если сравниваете точки-столбцы, то выбираете стандартизацию профилей столбцов или стандартизацию профилей строк и столбцов. Метод стандартизации можно выбрать в рамке **Standardization of coordinates:** *Row & column profiles* (профили строк и столбцов), *Canonical Standardization* (каноническая стандартизация), *Row profiles* (профили строк), *Column profiles* (профили столбцов).

На вкладке **Advanced** (рис. 15.32) нажмите кнопку **Eigenvalues** (собственные значения). Появится таблица (рис. 15.33) с сингулярными значениями, собственными числами и соответствующими статистиками. В последнем столбце приведено разложение статистики χ^2 Пирсона по собственным значениям.

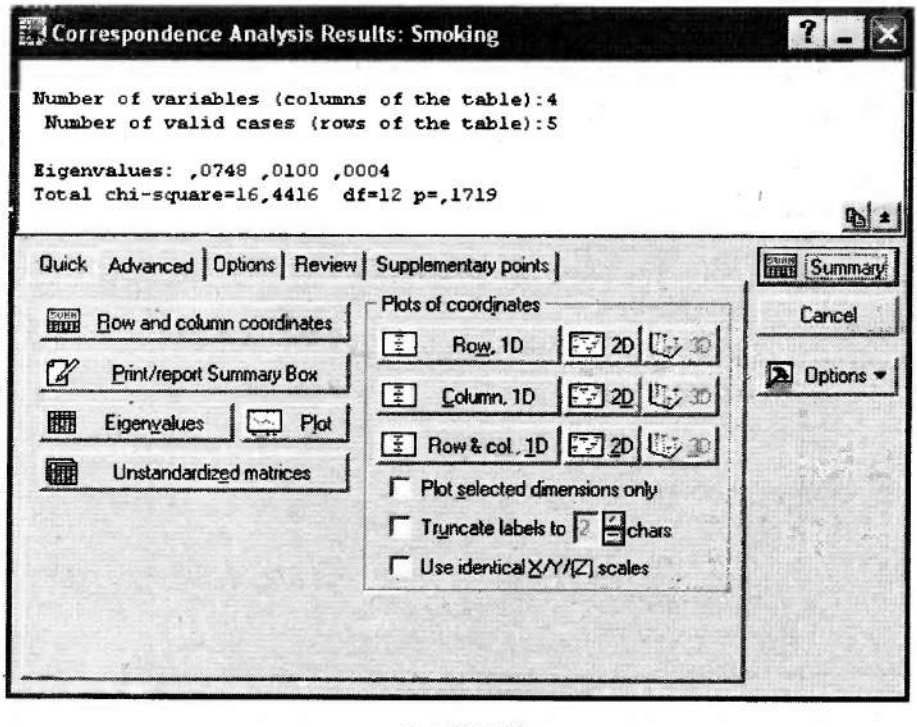


Рис. 15.32

Number of Dims.	Eigenvalues and Inertia for all Dimensions (Smoking) Input Table (Rows x Columns): 5 x 4 Total Inertia=,08519 Chi?=16,442 df=12 p=,17190 Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,273421	0,074759	87,75587	87,7559	14,42851
2	0,100086	0,010017	11,75865	99,5145	1,93332
3	0,020337	0,000414	0,48547	100,0000	0,07982

Рис. 15.33

Из таблицы видно, что одна размерность объясняет 87,76% инерции, а это значит, что для рассматриваемой двухвходовой таблицы значения относительных частот, которые восстанавливаются по одной размерности, дают вклад в величину статистики χ^2 (и, следовательно, инерции) в размере 87,76% от первоначального. Две размерности позволяют объяснить 99,51% значения χ^2 Пирсона. Заметим, что измерения (оси координат) выбираются так, чтобы расстояния между строками или столбцами были максимальны и увеличение размерности приводило бы к уменьшению доли совокупной величины χ^2 (и, следовательно, инерции), приходящейся на каждую дополнительную размерность. Это как раз можно наблюдать на рис. 15.33. При увеличении размерности значение χ^2 уменьшается.

Нажмите кнопку **Plot**, программа построит график собственных значений (рис. 15.34).

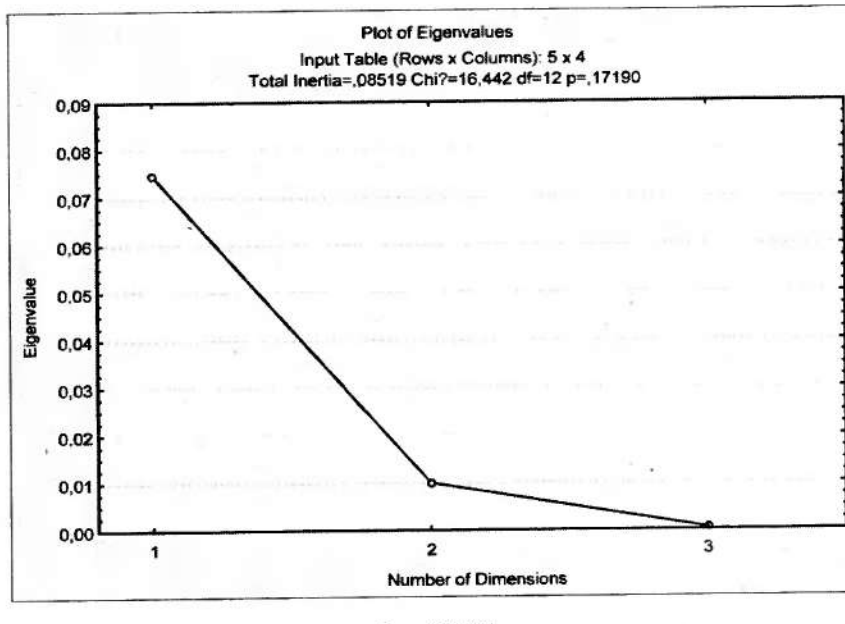


Рис. 15.34

Если нажать кнопку **Unstandardized matrices**, то появится таблица результатов с относительными частотами (т.е. величины входной таблицы, поделенные на сумму всех величин) и таблицы правых и левых обобщенных сингулярных векторов. Данная опция полезна, если необходимо обрабатывать матрицы средствами *STATISTICA Visual Basic*, например, для применения нестандартных методов стандартизации координат строк и столбцов.

На вкладке **Review** (просмотр) можно просмотреть входную таблицу и таблицы, вычисляемые по данной входной таблице (рис. 15.35). Рассмотрим функциональное назначение кнопок.

Observed frequencies (наблюдаемые частоты). Данная кнопка выводит таблицу результатов с входной таблицей (например, частоты в двухвходовой матрице).

Row percentages (проценты по строке). Кнопка выводит таблицу частот, преобразованную к виду процентов по строкам, т.е. если просуммировать величины по всем строкам, то результат будет равен 100%. Координаты для точек-строк вычисляются по данной таблице, если была выбрана стандартизация по *профилям* строк или *профилям* строк и столбцов.

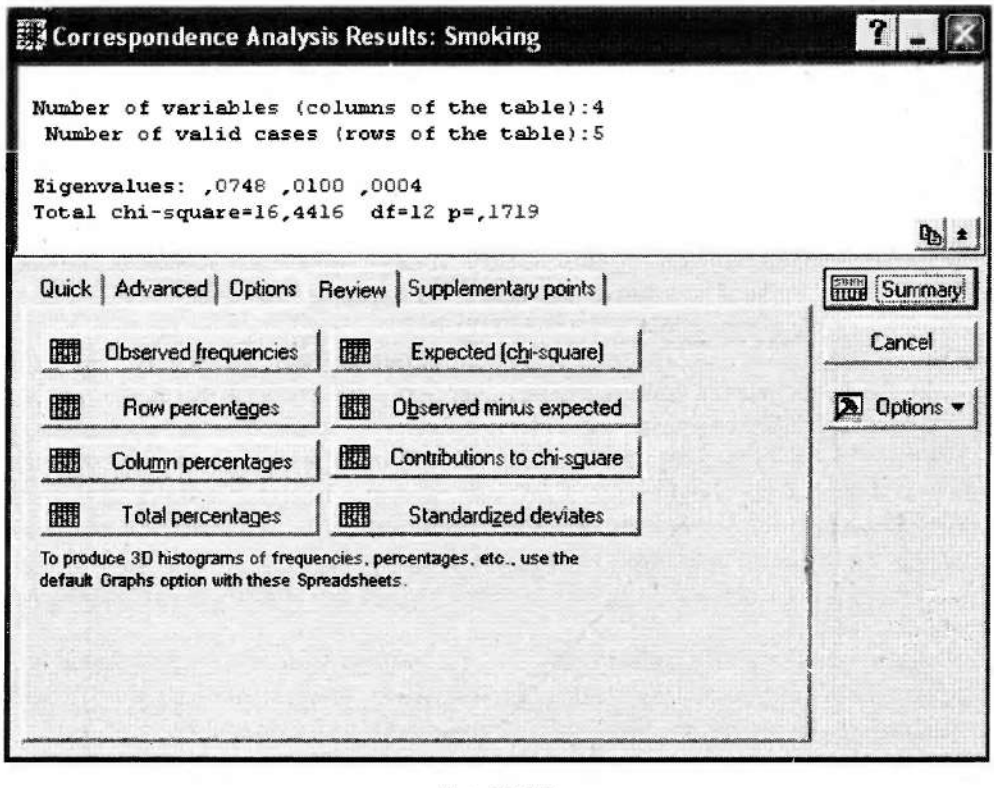


Рис. 15.35

Column percentages (проценты по столбцу). Кнопка выводит на экран таблицу частот, преобразованную к виду процентов по столбцам, т.е. если просуммировать величины по всем столбцам, то результат будет равен 100%. Координаты

для точек-столбцов вычисляются по данной таблице, если была выбрана стандартизация по *профилям* столбцов или *профилям* строк и столбцов.

Total percentages (проценты от общего числа). Кнопка выводит на экран таблицу частот, преобразованную к виду процентов от суммы всех элементов рассматриваемой входной матрицы. Таким образом, если просуммировать величины по всем столбцам и строкам, то результат будет равен 100%.

Expected (chi-square) (ожидаемые χ^2). Кнопка выводит на экран таблицу результатов с ожидаемыми частотами, вычисление которых базируется на гипотезе о том, что переменные по строкам и столбцам не зависят друг от друга. Используется стандартная формула для вычисления ожидаемых частот, базирующаяся на маргинальных суммах входной частотной таблицы.

Observed minus expected (наблюдаемые минус ожидаемые). Кнопка выводит на экран таблицу разностей между наблюдаемыми и ожидаемыми величинами.

Contributions to chi-square (вклад χ^2). Кнопка выводит на экран величины вклада каждой ячейки в совокупную величину χ^2 (данная опция будет работать, только если входная таблица содержит частоты). Сумма всех величин равна общей величине χ^2.

Standardized deviates (стандартизованные отклонения). Величины стандартизованного отклонения вычисляются как квадратный корень от соответствующих величин вклада в χ^2.

Дополнительную помощь в интерпретации результатов может оказать включение дополнительных точек-строк или столбцов, которые на первоначальном этапе не участвовали в анализе. Вкладка **Supplementary point** (дополнительные точки-столбцы и/или точки-строки) позволяет задать дополнительные точки. Результаты для данных точек будут автоматически включены в *отчет*, если нажать кнопку **Row and column coordinates** на вкладке **Advanced** или вывести на экран графики точек-строк и точек-столбцов.

Если на стартовой панели модуля **Correspondence Analysis** выделить вкладку **Multiple Correspondence Analysis** (многомерный анализ соответствий), откроется окно (рис. 15.36) с одноименным названием.

Многомерный анализ соответствий можно рассматривать как обобщение анализа соответствий на случай многовходовой таблицы частот. В ранее рассмотренном примере анализировали двухвходовую таблицу частот, соответствующую категориям курящих и сотрудников фирмы. Если добавить еще, по крайне мере, одну категорию, например, *Пол*, то придем к задаче многомерного анализа соответствий.

Многомерный анализ соответствий — это анализ соответствий на бинарной (индикаторной) матрице, где объекты расположены по строкам, а группирующие переменные — по столбцам. Фрагмент такой матрицы приведен на рис. 15.37.

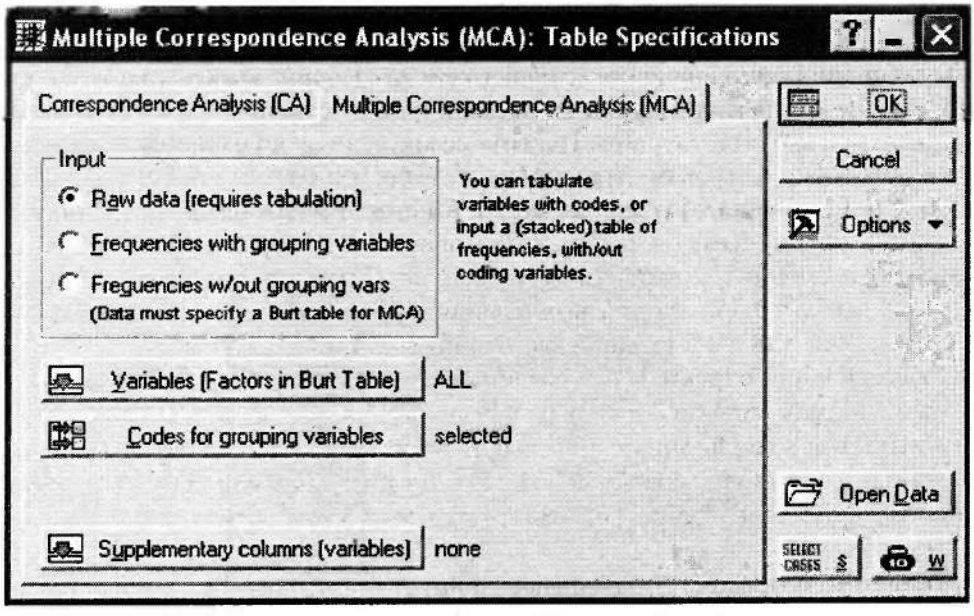


Рис. 15.36

	Группа сотрудников					Отношение к курению			
Номер наблюдения	*Старший менеджер*	*Младший менеджер*	*Старший сотрудник*	*Младший сотрудник*	*Секретарь*	*Некурящий*	*Слабо*	*Средне*	*Сильно*
1	1	0	0	0	0	1	0	0	0
2	0	1	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...

Рис. 15.37

Если объект принадлежит некоторой категории, то элемент на пересечении соответствующей строки и столбца равен 1, в противном случае — 0. Например, объект 1 представляет *Старшего менеджера*, который принадлежит категории *Некурящие*. Так как многомерный анализ соответствий предполагает случай более двух переменных, то бинарная матрица может дополнительно включать переменные *Мужчина* и *Женщина*, которые аналогично кодируются 0 или 1. Таким образом, окончательный результат может представлять взаимосвязи между переменными *Пол*, *Склонность к курению*, *Занимаемая должность* (*Группа сотрудников*). Реальные вычисления в многомерном анализе соответствий не используют индикаторную матрицу (которая может быть очень большой, если рассматривается много объектов и переменных). Для вычислений используется матричное произ-

ведение транспонированной и исходной бинарной матрицы — матрица Берта. Данная квадратная матрица табулирует связи между всеми имеющимися категориям. Модуль **Correspondence Analysis** позволяет использовать матрицу Берта в расчетах без каких-либо преобразований.

Программа *STATISTICA* также может создавать матрицу Берта по переменным, заданным обычным способом, в виде группирующих переменных. На рис. 15.38 приведен фрагмент файла данных трех группирующих переменных.

	Correspondence analysis example.		
	1 EMPLOYEE	2 SMOKING	3 GENDER
1	Sr.Manag	None	Male
2	Sr.Manag	None	Female
3	Sr.Manag	None	Male
4	Sr.Manag	None	Male
5	Sr.Manag	Light	Male
6	Sr.Manag	Light	Male
7	Sr.Manag	Medium	Female
8	Sr.Manag	Medium	Male
9	Sr.Manag	Medium	Male
10	Sr.Manag	Heavy	Male
11	Sr.Manag	Heavy	Male
12	Jr.Manag	None	Female
13	Jr.Manag	None	Female
14	Jr.Manag	None	Female

Рис. 15.38

Составленная программой матрица Берта состоит из девяти блоков (рис.15.39): подматрицы кросстабуляции переменной *MANAGERS* с переменной *MANAGERS*; подматрицы кросстабуляции переменной *MANAGERS* с переменной *SMOKING*; подматрицы кросстабуляции переменной *MANAGERS* с переменной *GENDER*; подматрицы кросстабуляции переменной *SMOKING* с переменной *MANAGERS*; подматрицы кросстабуляции переменной *SMOKING* с переменной *SMOKING*; подматрицы кросстабуляции переменной *SMOKING* с переменной *GENDER*; подматрицы кросстабуляции переменной *GENDER* с переменной *MANAGERS*; подматрицы кросстабуляции переменной *GENDER* с переменной *SMOKING*; подматрицы кросстабуляции переменной *GENDER* с переменной *GENDER*.

Заметим, что матрица Берта симметрична, а суммы диагональных элементов в каждом блоке, представляющем кросстабуляцию некоторой переменной с собой, равны общему числу объектов (193).

	1 EMP:Sr.Man	2 EMP:Jr.Man	3 EMP:Sr.Emp	4 EMP:Jr.Emp	5 EMP:Sec	6 SM:Non	7 SM:Lig	8 SM:Med	9 SM:Heav	10 GEN:Mal	11 GEN:Fem
EMPLOYEE:Sr.Manag	11	0	0	0	0	4	2	3	2	9	2
EMPLOYEE:Jr.Manag	0	18	0	0	0	4	3	7	4	10	8
EMPLOYEE:Sr.Empl	0	0	51	0	0	25	10	12	4	34	17
EMPLOYEE:Jr.Empl	0	0	0	88	0	18	24	33	13	55	33
EMPLOYEE:Secretar	0	0	0	0	25	10	6	7	2	18	7
SMOKING:None	4	4	25	18	10	61	0	0	0	39	22
SMOKING:Light	2	3	10	24	6	0	45	0	0	29	16
SMOKING:Medium	3	7	12	33	7	0	0	62	0	42	20
SMOKING:Heavy	2	4	4	13	2	0	0	0	25	16	9
GENDER:Male	9	10	34	55	18	39	29	42	16	126	0
GENDER:Female	2	8	17	33	7	22	16	20	9	0	67

Рис. 15.39

Если в диалоговом окне **Multiple Correspondence Analysis** в рамке **Input** (входные данные) установлена опция *Raw data (requires tabulation)* (исходные данные (требуется табуляция)) вкладки **Multiple Correspondence Analysis** стартовой панели модуля «Анализ соответствий», то *STATISTICA* ожидает на входе (категоризованные) группирующие переменные с кодами, однозначно определяющими, к какой категории принадлежит каждый объект. Вид файла **Smoking4** будет идентичен файлу, изображенному на рис. 15.21, но появятся дополнительные столбцы, например, столбец *Gender* (пол) с кодами *Male* (мужчина), *Female* (женщина) (рис. 15.38). Затем *STATISTICA* вычислит соответствующие переменные для входной таблицы. При выборе этой опции активными будут кнопки **Variables (Factor in But Table)** (переменные (факторы в таблице Берта)), **Codes for grouping variables**, **Supplementary columns (variables)** (добавить столбцы (переменные)).

Если при помощи этих кнопок произвести соответствующие установки и нажать на **OK**, будет запущена процедура **Multiple Correspondence Analysis**. Анализ и интерпретация результатов аналогичны одномерному анализу соответствий. Для просмотра матрицы Берта в окне результатов модуля «Многомерный анализ соответствий» перейдите на вкладку **Review** (рис. 15.35) и нажмите кнопку **Observed frequencies** (исходные частоты).

Если выбрана опция *Frequencies with grouping variables* (частоты с группирующими переменными), то *STATISTICA* ожидает в качестве входных данных категоризованные переменные с кодами, однозначно определяющими принадлежность той или иной категории. Дополнительно программа ожидает на входе переменную, содержащую частоты или какие-либо другие величины, определяющие меру

соответствия для категорий, рассматриваемых группирующих переменных. Вид файла будет идентичен файлу, изображенному на рис. 15.23, но также появятся дополнительные столбцы, например, столбец *Gender* с кодами *Male*, *Female*. Изменится структура кнопок внизу диалога. Активными будут кнопки **Variables (Factor in But Table)**, **Codes for grouping variables**, **Variable with frequencies counts** (переменная с частотами совпадений) и **Supplementary columns (variables)**.

Если выбрана опция *Frequencies w/out grouping variables* (частоты без группирующих переменных), то программа ожидает, что выбранные переменные (и объекты) содержат только частоты (или некоторые другие меры соответствия). Структура такого файла соответствует файлу **Smoking5**, изображенному на рис. 15.39. По сути, это матрица Берта, которую строит программа при выборе двух предыдущих опций *Raw data (requires tabulation)*, *Frequencies with grouping variables*. При этом будет активной одна кнопка **Variable with frequencies.** После выбора переменных для анализа станет активной кнопка **Specify structure of table** (задайте структуру таблицы). После нажатия этой кнопки откроется окно **Specify the dimensions of the...** (задайте структуру таблицы) (рис. 15.40).

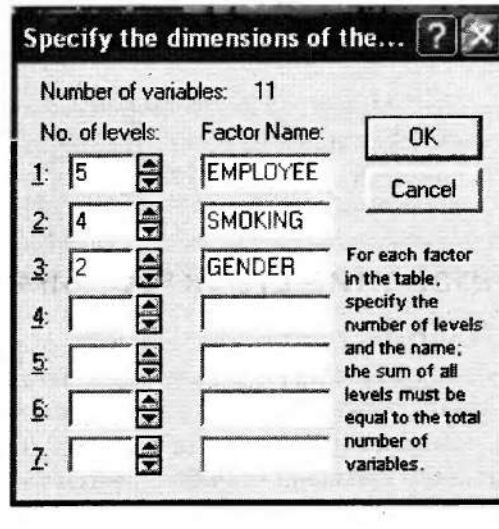


Рис. 15.40

Здесь в поле **No. Of levels** надо указать количество категорий каждой переменной, а в поле **Factor Name** — имена переменных (факторов). Заметим, что суммарное число категорий должно быть равно общему числу наблюдений в исходной матрице. Дальнейший анализ и интерпретация результатов аналогичны раннее рассмотренному одномерному анализу соответствий.

Глава 16

Причинное моделирование

16.1. Моделирование структурными уравнениями

Наметившийся в последнее время прогресс в области многомерного статистического анализа и анализа корреляционных структур, объединенный с новейшими вычислительными алгоритмами, послужил отправной точкой для создания новой, но уже получившей признание техники моделирования структурными уравнениями (*SEPATH*) [16]. Эта необычайно мощная техника многомерного анализа включает методы из различных областей статистики, множественная регрессия и факторный анализ получили здесь естественное развитие и объединение.

Объектом моделирования структурными уравнениями являются сложные системы, внутренняя структура которых не известна («черный ящик»). Наблюдая параметры системы при помощи *SEPATH*, можно исследовать ее структуру, установить причинно-следственные взаимосвязи между элементами системы.

Рассмотрим основные задачи, для решения которых используются структурные уравнения.

1. Причинное моделирование или анализ путей, при проведении которого предполагается, что между переменными имеются причинные взаимосвязи. Возможна проверка гипотез и подгонка параметров причинной модели, описываемой линейными уравнениями. Причинные модели могут включать явные или латентные (неявные) переменные, или и те и другие.
2. Подтверждающий факторный анализ, используемый как развитие обычного факторного анализа для проверки определенных гипотез о структуре факторных нагрузок и корреляций между факторами.
3. Факторный анализ второго порядка, являющийся модификацией факторного анализа, при проведении которого для получения факторов второго порядка анализируется корреляционная матрица общих факторов.
4. Регрессионные модели, являющиеся модификацией многомерного линейного регрессионного анализа, в котором коэффициенты регрессии могут быть зафиксированы равными друг другу или каким-нибудь заданным значениям.
5. Моделирование ковариационной структуры, которое позволяет проверить гипотезу о том, что матрица ковариации имеет определенный вид. Например, с помощью этой процедуры можно проверить гипотезу о равенстве дисперсий у всех переменных.
6. Моделирование корреляционной структуры, которое позволяет проверить гипотезу о том, что матрица корреляции имеет определенный вид. Классическим примером является гипотеза о том, что матрица корреляции имеет циклическую структуру.
7. Модели структуры средних, которые позволяют исследовать структуру средних, например, одновременно с анализом дисперсий и ковариаций.

Структурные уравнения, включающие только линейные связи между явными и латентными переменными, могут быть изображены в виде диаграмм путей. Поэтому даже начинающий пользователь может провести сложный анализ с минимальными затратами времени на обучение [16].

Постановка задачи структурного моделирования выглядит следующим образом [2]. Пусть имеются переменные, для которых известны статистические моменты, например, матрица выборочных коэффициентов корреляции или ковариации. Такие переменные называются явными. Они могут быть характеристиками сложной системы. Реальные связи между наблюдаемыми явными переменными могут быть достаточно сложными, однако предполагаем, что имеется некоторое число скрытых переменных, которые с известной степенью точности объясняют структуру этих связей. Таким образом, с помощью латентных переменных строится модель связей между явными и неявными переменными. В некоторых задачах латентные переменные можно рассматривать как причины, а явные — как следствия, такие модели называются причинными. Допускается, что скрытые переменные, в свою очередь, могут быть связаны между собой. Структура связей допускается достаточно сложной, однако тип ее постулируется — это связи, описываемые линейными уравнениями. Какие-то параметры линейных моделей известны, какие-то нет, и являются свободными параметрами.

Обозначим неизвестные параметры $a_1, a_2, ..., a_k$, a матрицу выборочных коэффициентов корреляции или ковариации — через S. Пересчитаем эту матрицу формально с помощью модели. Получим новую матрицу S', являющуюся функцией от $a_1, a_2, ..., a_k$. Пусть $d(S, S')$ — некоторая функция, измеряющая различие двух матриц. Задача состоит в том, чтобы построить оценки неизвестных параметров, дающие минимум $d(S, S')$. Различные функции d соответствуют различным методам оценивания.

Основная идея моделирования структурными уравнениями состоит в том, что можно проверить, связаны ли переменные Y и X линейной зависимостью $Y = aX$, анализируя их дисперсии и ковариации. Эта идея основана на простом свойстве среднего и дисперсии: если умножить каждое число на некоторую константу k, среднее значение также умножится на k, при этом стандартное отклонение умножится на модуль k.

Например, рассмотрим набор из трех чисел 1, 2, 3. Эти числа имеют среднее, равное 2, и стандартное отклонение, равное 1. Если умножить все три числа на 4, то легко посчитать, что среднее значение будет равно 8, стандартное отклонение — 4, а дисперсия — 16. Таким образом, если есть наборы чисел X и Y, связанные зависимостью $Y = 4X$, то дисперсия Y должна быть в 16 раз больше, чем дисперсия X. Поэтому можно проверить гипотезу о том, что Y и X связаны уравнением $Y = 4X$, сравнением дисперсий переменных Y и X.

Эта идея может быть различными способами обобщена на несколько переменных, связанных системой линейных уравнений. При этом правила преобразований становятся более громоздкими, вычисления более сложными, но основной смысл остается прежним — можно проверить, связаны ли переменные линейной зависимостью, изучая их дисперсии и ковариации.

Заметим, что если бы были известны наблюдаемые значения Y и X, то найти значение параметра k можно было бы по методу наименьших квадратов. Но в структурном моделировании обе переменные или одна из них могут быть латентными, т.е. с неизвестными значениями.

Процесс моделирования структурными уравнениями состоит из 5 этапов.

1. Формирование модели — предварительное описание способов, которыми предположительно явные и скрытые переменные связаны между собой (вначале это делается графически, на языке диаграмм путей, затем переводится на язык системы).
2. С помощью некоторых правил программа перерабатывает модель, сформулированную на языке системы (*PATH*), в модель для дисперсий и ковариаций переменных. Программа определяет, какие значения дисперсий и ковариаций переменных получаются в текущей модели на основании входных данных. Модель записывается в файл модели с расширений **.cmd*.
3. Проверка программой, насколько хорошо полученные дисперсии и ковариации удовлетворяют предложенной модели.
4. Сообщение программы пользователю о полученных результатах статистических испытаний, а также вывод оценок параметров и стандартных

ошибок для численных коэффициентов в линейных уравнениях вместе с большим количеством дополнительной диагностической информации.

5. На основании этой информации, пользователь решает, хорошо ли текущая модель согласуется с исходными данными. Если качество подгонки неудовлетворительно, возвращаются к первому шагу и переструктурируют модель. Постепенно изменяя модель, добиваются приемлемой степени ее адекватности исходным данным.

Основные этапы процесса структурного моделирования на рис. 16.1 показаны в виде диаграммы.

Рис. 16.1

Для задания структурных связей между переменными *STATISTICA* использует командный язык *PATH*, который по своим возможностям похож на диаграммы путей. Только для простых систем можно описывать связи и переменные сразу на языке *PATH*. Сложные системы, по крайней мере, в главных своих элементах, вначале изображаются графически. Для этого служат диаграммы путей. Так же, как в программировании используются блок-схемы, прежде чем написать программу, так и в *SEPATH* используются диаграммы путей. Диаграммы путей можно построить на бумаге либо непосредственно на экране, используя графические возможности системы или какого-нибудь текстового редактора. Они изображают переменные, связанные линиями, которые служат для отображения причинных

связей. Каждая связь или путь включает в себя две переменные (заключенные в прямоугольник или овал), соединенные стрелками или дугами. Диаграммы удобнее всего использовать в качестве инструмента для указания, какие переменные вызывают изменения в других переменных.

Рассмотрим классическое линейное регрессионное уравнение

$$Y = aX + E.$$

Его можно изобразить в виде диаграммы путей (рис. 16.2).

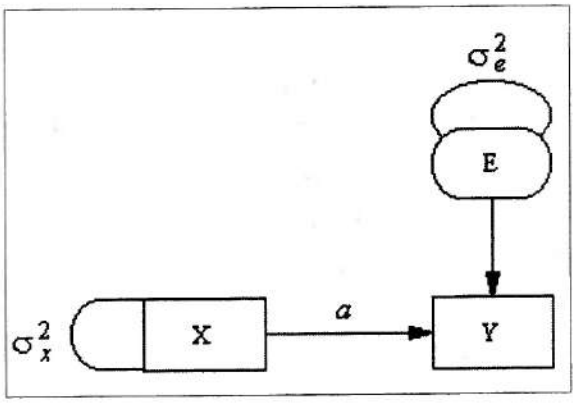


Рис. 16.2

Как видно, все переменные в системе уравнений размещены на диаграмме в прямоугольниках или овалах. Каждое уравнение отображается на диаграмме следующим путем: все независимые переменные (переменные в правой части уравнения) имеют стрелки, указывающие на зависимые переменные. Весовые коэффициенты располагаются вблизи соответствующих стрелок.

Заметим, что кроме представления линейных зависимостей в виде стрелок, диаграмма также содержит некоторые другие выражения. Во-первых, дисперсия независимых переменных, которая должна быть задана для проверки модели структурных связей, показана на диаграмме с использованием дуг. Во-вторых, некоторые переменные изображены в овальных, а не в прямоугольных рамках. Явные переменные (т.е. переменные, которые можно измерить непосредственно) на диаграммах изображаются внутри прямоугольников. Латентные переменные (т.е. те, которые нельзя непосредственно измерить, например, факторы в факторном анализе или остатки в регрессионном) изображаются внутри овалов или окружностей. Например, переменная E на диаграмме может рассматриваться как остаток линейной регрессии, когда значение Y предсказывается по значению X. Такие остатки не наблюдаются непосредственно, но могут быть вычислены по известным значениям Y и X (если известны значения параметров, в нашем случае это весовой коэффициент a).

Хотя диаграммы путей могут использоваться для отражения причинных связей в наборе переменных, они не предполагают реального наличия таких связей.

Диаграммы путей часто используются для простого и изоморфного представления системы линейных уравнений. Поэтому они могут выражать линейные связи вне зависимости от того, имеются ли на самом деле описанные причинные связи. Следовательно, хотя мы интерпретируем диаграмму на рисунке как «*X* влияет на *Y*», диаграмма также может обозначать графическое представление линейного регрессионного соотношения между *X* и *Y* (без причинной связи между ними).

Следует помнить, что невозможно идеальное соответствие модели и данных по нескольким причинам. Структурные модели с линейными зависимостями, как и все математические модели, являются только приближениями реальных явлений. Далее, природные зависимости, как правило, далеки от линейных. Поэтому истинные зависимости между переменными, скорее всего, нелинейны. Более того, истинность многих статистических предположений, накладываемых на проверяемую модель, остается под большим вопросом. Поэтому в прикладных исследованиях вместо вопроса «Идеально ли модель согласуется с данными?» должен волновать вопрос «Согласуется ли она достаточно хорошо, чтобы быть полезной для практического использования и разумного объяснения структуры наблюдаемых данных?».

Далее, надо знать, что идеальное соответствие модели данным не всегда означает, что модель верна. Невозможно доказать, что модель верна — «умение доказывать правильность модели эквивалентно умению предсказывать будущее» [16]. Так, если определенная причинная модель верна, то она согласуется с наблюдаемыми данными. Но модель, согласующаяся с данными, не обязательно верна, так как, возможно, существует другая модель, которая ничуть не хуже согласуется с теми же данными.

Как уже говорилось, переменные могут быть связаны нелинейно. Возможна и линейная связь, которая не зависит от того, что мы выбрали в качестве причины в модели. Древнее изречение «наблюдаемая зависимость не означает причинной зависимости» остается верным даже для сложной и многомерной корреляции. Причинное моделирование позволяет исследовать только то, насколько данные отличаются от соответствующих выводов причинной модели (а именно от предполагаемой ковариационной структуры). Если система линейных уравнений, соответствующая диаграмме путей, хорошо согласуется с данными, это позволяет оставить модель для дальнейшего анализа или использования, но не доказывает ее истинность.

Правила построения диаграмм путей основаны на следующих соображениях [6].

Диаграммы путей состоят из переменных, связанных дугами и стрелками, представляющими, соответственно, направленные и ненаправленные связи между переменными. Эти переменные должны быть либо эндогенными, либо экзогенными.

Эндогенная переменная (внутрисистемная) — это переменная, которая входит в качестве зависимой переменной хотя бы в одно линейное уравнение структурной модели. Эндогенные переменные на диаграммах путей легко отличить от остальных переменных, так как на них указывает как минимум одна стрелка.

Экзогенная переменная (внесистемная) — это переменная, которая не входит в качестве зависимой переменной ни в одно уравнение структуной модели. Экзо-

генные переменные легко отличить на диаграммах путей от остальных переменных, так как на них не указывает ни одна стрелка.

Как было замечено, переменные также могут быть либо явными, либо неявными (латентными). Следовательно, любая переменная относится к одной из четырех категорий: явной эндогенной (*manifest endogenous*), явной экзогенной (*manifest exogenous*), латентной эндогенной (*latent endogenous*) и латентной экзогенной (*latent exogenous*). Правила состоят из 9 пунктов.

1. Явные переменные всегда изображаются в прямоугольниках (или квадратах), а латентные переменные — всегда внутри овала или окружности.
2. Каждая направленная связь представляется непосредственно с помощью стрелки между двумя соответствующими переменными.
3. Ненаправленные связи не обязательно должны явно отображаться на диаграмме (см. далее правило 9 по поводу неявного представления ненаправленных связей).
4. Ненаправленные связи, явно отображаемые на диаграмме, обозначаются в виде дуги от переменной к самой себе или к другой переменной.
5. Эндогенные переменные не могут соединяться с другими перемеными с помощью дуг.
6. Номера свободных параметров выводятся в виде чисел, размещенных на, около или немного выше середины дуги или стрелки.
7. Фиксированное значение для дуги или стрелки всегда приводится в виде числа с плавающей точкой. Это число обычно располагается на, около или немного выше середины дуги или стрелки.
8. Диаграммы, относящиеся к разным вероятностным пространствам, отделяются разграничительной линией и словами *Группа 1* (для первого пространства), *Группа 2* и т.д. в каждой области диаграммы.
9. Для всех экзогенных переменных должны быть явно или неявно указаны с помощью фиксированных значений или свободных параметров их дисперсии и ковариации.

Если эти ковариации или дисперсии выражены неявно, выполняются следующие правила:

а) для латентных экзогенных переменных дисперсии, не имеющие явного выражения на диаграмме, предполагаются фиксированными и равными 1.0, а ковариации, не имеющие явного описания, предполагаются равными 0;

б) для явных экзогенных переменных дисперсии и ковариации, не имеющие явного представления на диаграмме, полагаются свободными параметрами, каждый из которых имеет различный порядковый номер. Эти номера параметров не совпадают ни с какими номерами параметров, явно употребляемыми на диаграмме.

Заметим, путевые диаграммы, изображенные в виде графов в *STATISTICA*, не воспринимаются программой в автоматическом режиме, а переводятся на язык *PATH1* пользователем вручную при помощи процедуры мастер путей. Поэтому

перечисленные правила составлены с целью введения некоего стандарта на построение путевых диаграмм. Это позволит: сделать менее громоздким изображение и описание путевой диаграммы, а также однозначно и правильно интерпретировать диаграммы, построенные и опубликованные разными авторами.

Путевые диаграммы на языке *PATH1* уже доступны для компьютерного анализа. Существуют простые правила, задающие соответствия между представлением модели с помощью диаграммы и представлением на языке *PATH*.

1. Каждая стрелка или дуга представляется на отдельной строке.
2. Пробелы игнорируются.
3. Имена явных переменных представляются полным именем, заключенным в квадратные скобки.
4. Имя [*CONSTANT*] резервируется для обозначения переменной с дисперсией 0 и средним 1.
5. Имена скрытых переменных записываются в круглых скобках.
6. Прямые связи (стрелки) представляются записью:

   VNAME1–<#1>{<#2>}–> VNAME2,

   где *VNAME1* и *VNAME2* — имена явных или скрытых переменных; <#1> — номер параметра, т.е. целое число между 1 и 30 000; этот номер требуется, если путь имеет свободный параметр, который оценивается системой; <#2> — начальное значение, с которого начинается приближение при оценивании свободного параметра.
7. Непрямые связи (дуги) представляются в форме

   VNAME1–<#1 >{<#2>}–VNAME2,

   где элементы записи имеют тот же смысл, что в пункте 6.
8. Различные группы обозначаются с помощью предложения вида

   GROUP<#>,

   где <#> — число групп. Все *PATH1* файлы начинаются с предложения *GROUP1*. Группы должны быть составлены по порядку, начиная с первой. Перед началом записи команд новой группы необходимо закончить записи, относящиеся к предыдущей группе, командой *ENDGROUP* — конец группы.
9. Пустые строки в командном файле, а также любые строки, начинающиеся знаком *, рассматриваются как комментарии. Каждый элемент диаграммы путей имеет соответствующий элемент в языке *PATH1*. Это соответствие легко установить, язык *PATH1* построен максимально близко к диаграммам путей. Можно было бы, глядя на диаграмму, писать программу на языке *PATH1*.

В *STATISTICA* имеется удобное средство, позволяющее писать программу в диалоговом режиме. Это средство называется **Path Wizards** (мастер путей),

с его помощью шаг за шагом определяется модель, задаются типы переменных и устанавливаются связи между ними.

16.2. Стартовое окно модуля *SEPATH*

Важно отметить, что вычислительные процедуры в модуле **SEPATH** реализованы в предположении нормальности наблюдаемых случайных величин.

Для запуска модуля **SEPATH** в меню **Statistics** щелкните по **Advanced Linear/Nonlinear Models** и выберите команду **Structural Equation Modeling.** Стартовое окно модуля будет иметь вид, изображенный на рис. 16.3. На вкладке **Advanced** доступны опции, с помощью которых можно открывать и сохранять модели, задавать параметры и группирующую переменную, а также доступны конструктор путей и мастер путей. Рассмотрим функциональное назначение кнопок на этой вкладке.

Path tool (конструктор путей). Данная кнопка вызывает конструктор путей, удобное средство для быстрого и безошибочного ввода и редактирования программ анализа путей.

Path wizards (мастер путей). Эта кнопка открывает окно **Wizards SEPATH**, в котором можно выбрать один из стандартных мастеров путей. Выбранный мастер шаг за шагом переведет диаграмму пути на язык *PATH1*.

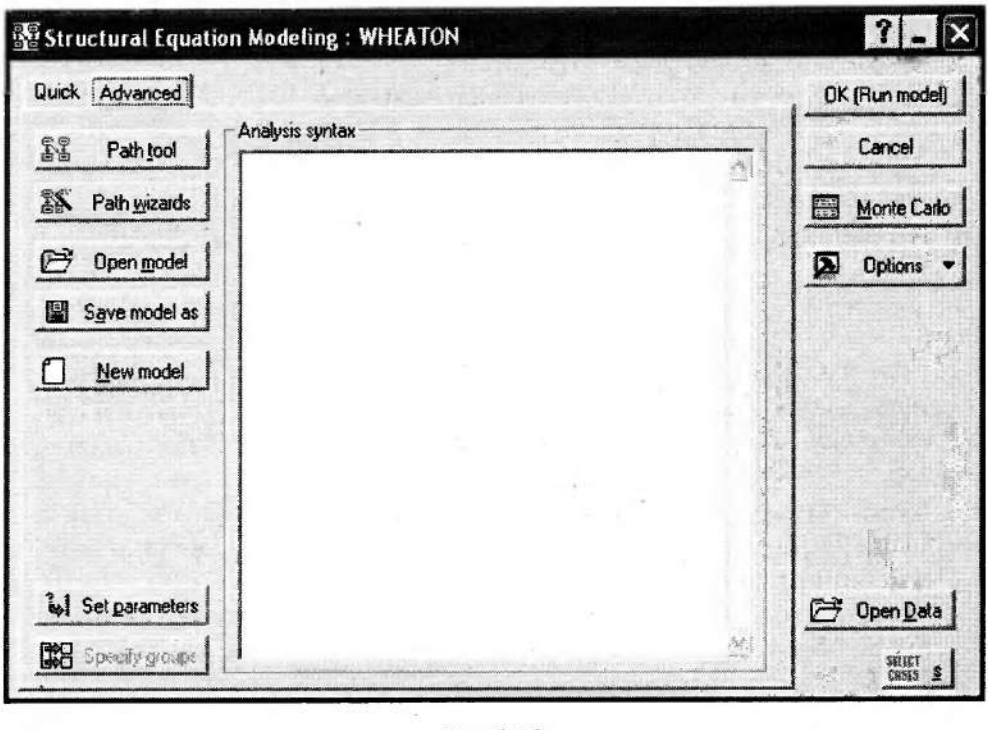


Рис. 16.3

Analysis Syntax (редактирование текста). Это окно можно использовать для просмотра и редактирования текста текущего файла модели. Указанная функция часто используется для изменения формата текста с целью улучшения его читаемости, а также для добавления поясняющих комментариев.

Open model (открыть модель). Эта кнопка открывает стандартное окно **Open file**, в котором можно выбрать ранее созданный и сохраненный файл с расширением **.cmd* для дальнейшего анализа и/или модификации.

Save model (сохранить модель). Данная кнопка позволяет сохранить текущую модель в текстовом формате **.cmd*.

New model (создать модель). Для того чтобы удалить текущую модель (показанную в поле **Analysis Syntax**) и определить новую, надо нажать эту кнопку.

Set parameters (параметры). Указанная кнопка отображает диалоговое окно **Analysis parameters**, в котором можно уточнить параметры анализа.

Specify groups (задать группы). Если используемый файл данных содержит матрицы ковариаций или корреляций между группами, одна или несколько из этих групп могут быть выбраны для анализа. Эта кнопка открывает окно **Specify grouping**. Если файл данных содержит необработанные исходные данные, после нажатия кнопки откроется окно **Specify grouping variables** (задайте группирующую переменную), в котором можно задать используемую группирующую переменную и коды.

В качестве файла данных рассмотрим матричный файл **Wheaton** (рис. 16.4) из **Examples → Datasets → Sepath**. Данные являются ковариационной матрицей, вычисленной по выборке объема 923. Каждая запись в исходной выборке соответствует одному человеку. Необходимо построить регрессионную модель, связывающую социоэкономический статус человека со степенью его отчужденности от общества, по данным, полученным для каждого человека в 1967 г. и в 1971 г. В модуле **SEPATH** исходные данные могут иметь структуру обычной таблицы в формате *STATISTICA*, программа самостоятельно преобразует в матричный файл с ковариациями.

	SEPATH Electronic Manual Example #1					
	1 ANOMIA67	2 POWLES67	3 ANOMIA71	4 POWLES71	5 EDUCATN	6 SEINDEX
ANOMIA67	11,834	6,947	6,819	4,783	-3,839	-21,899
POWLES67	6,947	9,364	5,091	5,028	-3,889	-18,831
ANOMIA71	6,819	5,091	12,532	7,495	-3,841	-21,748
POWLES71	4,783	5,028	7,495	9,986	-3,625	-18,875
EDUCATN	-3,839	-3,889	-3,841	-3,625	9,61	35,522
SEINDEX	-21,899	-18,831	-21,748	-18,775	35,522	450,288
Means						
Std Dev						
No Cases	932					
Matrix	4					

Рис. 16.4

В файле данных приняты следующие обозначения: *ANOMIA67*, *ANOMIA71* – аномальность в поведении; *EDUCATN* – образование; *POWLESS67*, *POWLESS71* – отсутствие способностей; *SEINDEX* – общительность. Все приведенные переменные, характеризующие различные параметры человека, – это наблюдаемые (явные) переменные, так как их значения приведены в файле данных. Но в модели должны присутствовать еще латентные переменные: *SES* – социоэкономический статус; *AL67*, *AL71* – отчужденность от общества. Явные переменные *EDUCATN*, *SEINDEX*, *ANOMIA67*, *ANOMIA71*, *POWLESS67*, *POWLESS71* относятся к эндогенным. Латентную переменную *SES* можно считать экзогенной, а латентные переменные *AL67*, *AL71* – эндогенными.

16.3. Построение диаграммы путей

Задача состоит в том, чтобы составить модель, оценить неизвестные параметры модели и далее определить, насколько адекватно модель описывает ковариационную структуру данных. Сначала воспользуемся графическими возможностями программы *STATISTICA* и построим диаграммы путей. Диаграммы путей пользователь строит, исходя из своего понимания моделируемой сложной системы, и процедура эта не формализована, т.е. для одной и той же системы могут быть построены различные диаграммы. Проблема в том, чтобы построить (или выбрать из построенных) диаграмму, наиболее адекватно описывающую систему.

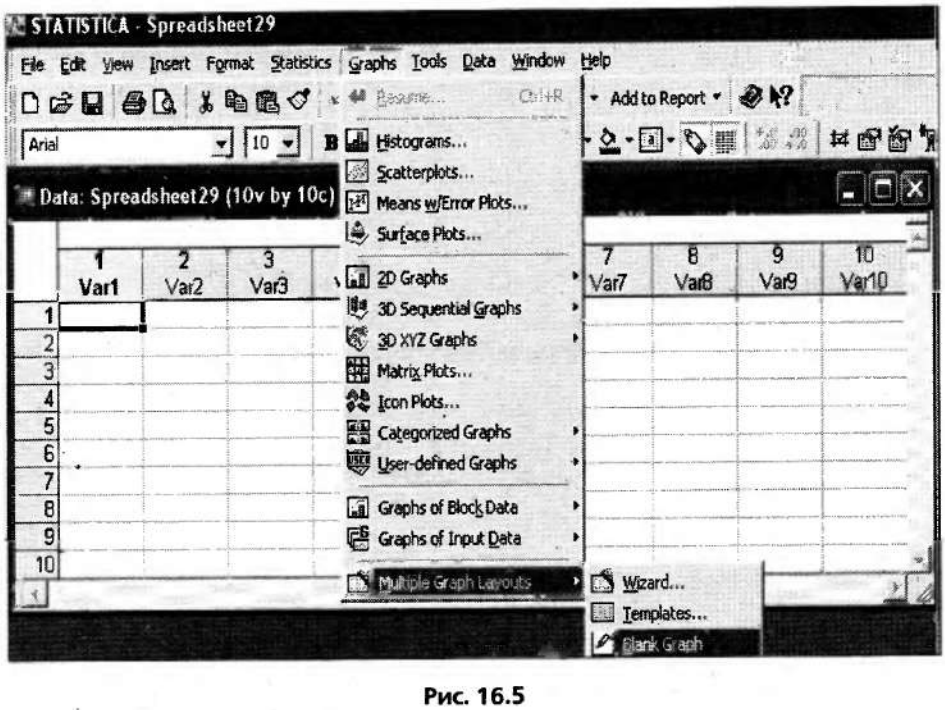


Рис. 16.5

Рассмотрим процедуру построения диаграмм в *STATISTICA*. Вряд ли нам удастся построить диаграммы лучшие, чем предложены в [2, 6]. Щелкните по кнопке **Graphs** в верхнем меню. В появившемся окне (рис.16.5), выберите команду **Multiple Graph Layouts** (множественные слои графика), а затем выберите **Blank Graph** (пустой график).

На экране появится графическое окно **Graph1: New Graph**, в нем можно создать самую сложную диаграмму, используя горизонтальную панель с кнопками, позволяющими рисовать любые фигуры: прямоугольники, стрелки, окружности, дуги и т.д. Строить такие диаграммы можно и в текстовых редакторах, например, в *Word*. Нарисуйте в *STATISTICA* диаграммы, изображенные на рис. 16.6–16.8.

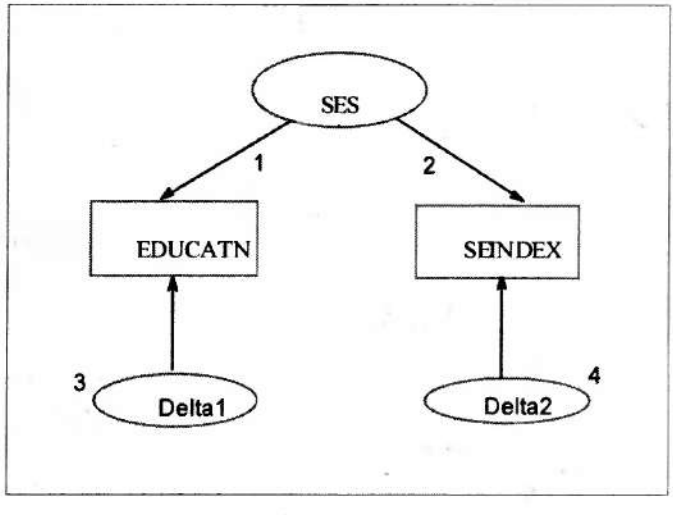


Рис. 16.6

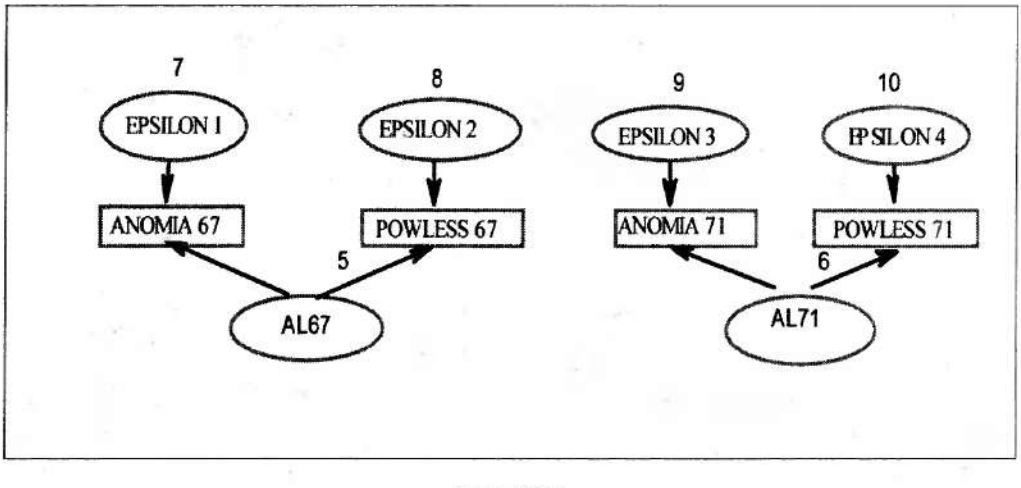


Рис. 16.7

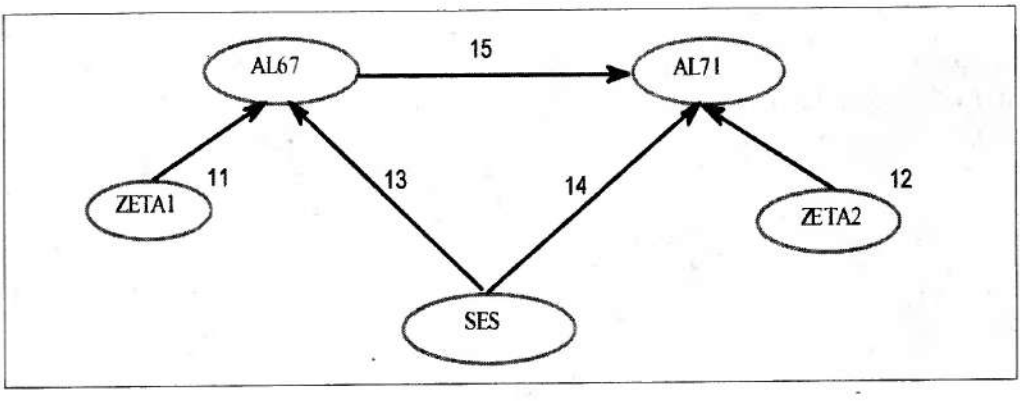


Рис. 16.8

На рис. 16.6 изображена однофакторная модель с одним общим социоэкономическим фактором SES и двумя явными переменными *EDUCATN* и *SEINDEX*. Эта модель соответствует двум регрессионным уравнениям

$$EDUCATN = a_1 SES + Delta1$$

и

$$SEINDEX = a_2 SES + Delta2,$$

где a_1 и a_2 — неизвестные коэффициенты (свободные параметры); *Delta1*, *Delta2* — остаточные переменные.

На рис. 16.7 изображена двухфакторная модель с общими факторами *AL67*, *AL71*:

$$ANOMIA67 = AL67 + EPSILON1,\ POWLESS67 = a_5 AL67 + EPSILON2,$$

$$ANOMIA71 = AL71 + EPSILON3,\ POWLESS71 = a_6 AL71 + EPSILON4.$$

На рис. 16.8 изображены регрессионные зависимости между фактором SES и факторами *AL67*, *AL71*:

$$AL67 = a_{13} SES + ZETA1 \text{ и } AL71 = a_{14} SES + a_{15} AL67 + ZETA2,$$

где a_{13}, a_{14}, a_{15} — неизвестные коэффициенты, *ZETA1*, *ZETA2* — ошибки, дисперсии которых также являются свободными параметрами.

На рис. 16.9 изображена общая модель, включающая три предыдущие модели.

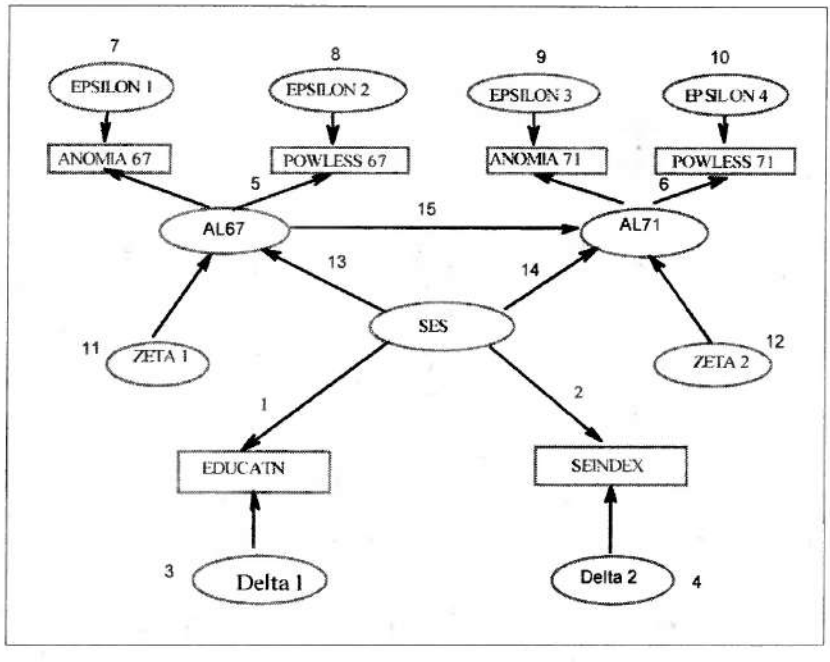


Рис. 16.9

Рассмотрим соображения, по которым общая модель была представлена как совокупность трех более простых моделей. При построении диаграммы сложной системы желательно мысленно разделить ее на части меньшего размера. Как и многие другие модели структурных уравнений, наша модель может быть представлена в виде классической модели *LISREL Karl Joreskog* [6], состоящей из трех меньших моделей, две из которых являются общими факторами моделей и обычно называются моделями измерений (рис. 16.7), и одной модели множественной регрессии, называемой структурной моделью (рис. 16.8). В этом примере нас в первую очередь интересует регрессионная зависимость между социоэкономическим статусом (*SES*) и отчуждением личности (*AL*) в два различных момента. Как это часто случается с данными в социологии, наблюдаемые переменные, используемые для оценивания социоэкономического статуса и отчужденности, имеют различную степень надежности. Следовательно, корреляции между наблюдаемыми переменными ослаблены возможной недостоверностью данных, поэтому полученная регрессионная зависимость между ними может оказаться ошибочной. Для того чтобы справиться с этой проблемой, в моделях типа *LISREL* постулируется регрессионная зависимость между латентными переменными, которые содержат ошибки измерений в виде общих факторов для наблюдаемых переменных. Таким образом, имеются две модели измерений, одна факторная модель для латентных экзогенных переменных — на рис. 16.6, а другая для латентных эндогенных переменных — на рис. 16.7. На рис. 16.9 модели измерений расположены в верхней и нижней частях диаграммы, а структурная модель находится в центре диаграммы.

Рассмотрим смысл некоторых числовых обозначений на диаграммах. Числа 1 и 2, расположенные у стрелок из *SES* к *EDUCATN* и *SEINDEX*, представляют номера нагрузок (свободные параметры, коэффициенты a_1, a_2) фактора *SES* на эти переменные, которые вычисляются программой. Дуги с расположенными рядом с ними числами 3 и 4 соответствуют дисперсиям переменных остатков *DELTA1* и *DELTA2*. Эти числа представляют свободные параметры, которые должны быть также оценены программой. Стрелки из *AL67* в *ANOMIA67* и из *AL71* в *ANOMIA71* не имеют рядом расположенных целых чисел. Это означает, что нагрузки (коэффициенты при *AL67, AL71*) равны 1. В процессе реализации процедуры **Path wizards** программа сама определяет, какие параметры модели (коэффициенты при переменных) являются свободными, а какие нет. Поэтому вид диаграммы путей всегда должен быть уточнен после реализации модуля **SEPATH.**

16.4. Мастер путей — *Path Wizards*

После построения диаграмм путей необходимо перевести графическую модель на язык *PATH1*. С учетом сложности модели проделаем это последовательно, шаг за шагом, используя диаграммы на рис. 16.6–16.9.

Щелкните по кнопке **Path wizards** (мастер путей). В появившемся окне **SEPATH Wizard – Select Wizard** (рис. 16.10) представлены две опции: непосредственно *Structural Modeling* (структурное моделирование) и *Confirmatory factor analysis* (подтверждающий факторный анализ).

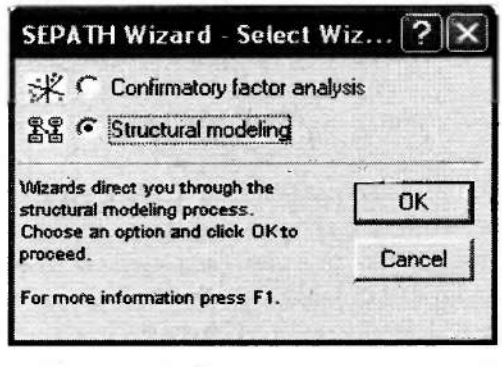


Рис. 16.10

Модели подтверждающего факторного анализа очень похожи на обычные модели с общими факторами, за исключением того, что некоторые нагрузки факторов и корреляции между факторами могут быть определены равными друг другу или нулю. Эти условия на структуру факторов легко проверить с помощью специальной записи гипотез о структуре факторов. Мастер факторного анализа позволяет определить практически любую подтверждающую факторную модель всего за несколько секунд, если эта модель имеет 8 или меньше общих факторов. Большинство реальных моделей попадает в эту категорию. Определение моделей

большего размера также не составляет особого труда — для этого следует воспользоваться конструктором путей.

Выберите опцию *Structural Modeling* и щелкните **OK**. Откроется окно **Structural Modeling — Exogenous Variables**, в котором надо определить экзогенные переменные. В поле **Exogenous variables** напишите *SES*. Это экзогенный общий фактор, определяющий социоэкономический статус человека. Нажмите кнопку **Vars** (переменные) и определите явные переменные, связанные с экзогенным фактором *SES — EDUCATN, SEINDEX* (рис. 16.11). Щелкните далее **OK**. Программа возвратится в окно **Structural Modeling — Exogenous Variables**. В поле **Base name for residual variables** (основное имя для остаточных переменных) предложит по умолчанию имя для остаточных переменных *Delta* (поэтому это имя было использовано в путевой диаграмме на рис. 16.9). При желании можно изменить это имя, для этого надо войти в соответствующую строку и написать другое имя.

В рамке **Residual Vars** (остаточные переменные) выберите опцию *Uncorrelated* (некоррелированные). Из путевой диаграммы следует, что ранее была выбрана модель с некоррелированными остаточными переменными, так как остаточные факторы *DeltaI, Delta2* не коррелированны между собой (между ними нет никакого пути). Если выбрать **Correlated**, то будет рассматриваться модель с коррелированными остаточными переменными. То же относится к рамке **Factors** (факторы) — выберите опцию *Uncorrelated* (рис. 16.11).

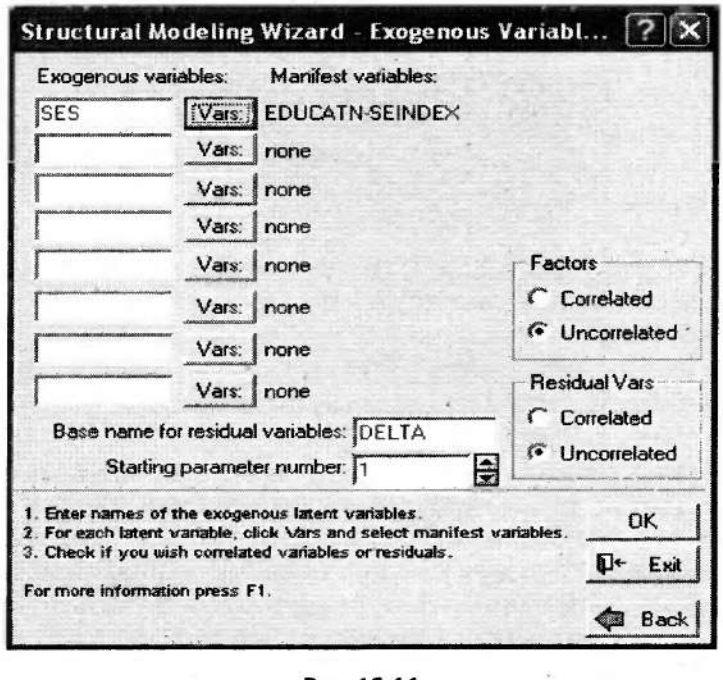


Рис. 16.11

Определите теперь эндогенные переменные. Щелкните кнопкой **OK** в окне **Define Exogenous Variables**. В автоматически появившемся окне **Structural Modeling – Endogenous variables** (определить эндогенные переменные) произведите выбор эндогенных переменных в соответствии с рис. 16.7, где дана двухфакторная модель с двумя эндогенными факторами *AL67, AL71*. Прежде всего запишите в 1-е поле **Endogenous variables** фактор *AL67*. Далее щелкните кнопкой **Vars** и в появившемся окне выберите переменные *ANOMIA87, POWLESS67*. Проделайте то же для переменной *AL71* (рис. 16.12).

В поле **Base name for residual variables** автоматически будет прописано имя для остаточных переменных – *EPSILON* (это имя также было использовано в путевой диаграмме), в поле **Base name for disturbances** (основное имя для дисперсии) – *ZETA*. Нажмите **OK**. Появится окно (рис. 16.13) **Define Structural Equation Paths** (определить пути структурных уравнений). Воспользуйтесь диаграммой на рис. 16.8. На этой диаграмме представлена регрессионная модель, связывающая факторы *SES, AL67, AL71*. Переведем эту модель на язык *PATH1*.

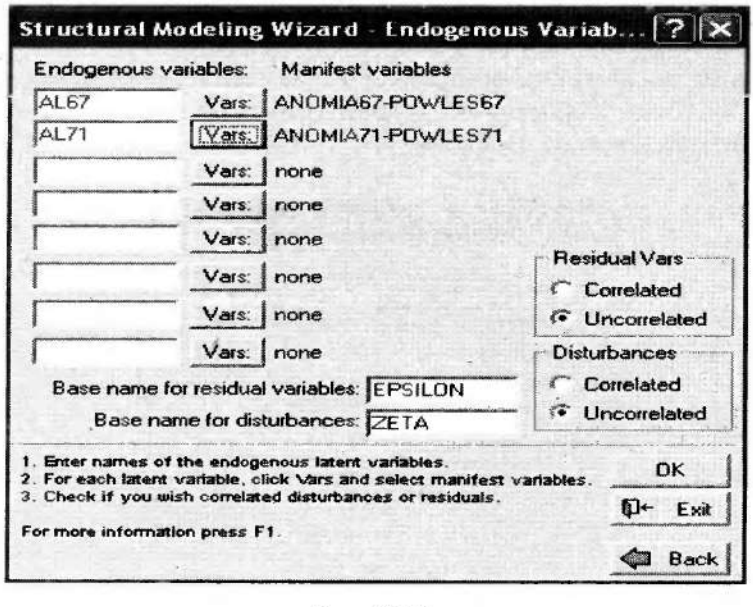


Рис. 16.12

Вначале в списке *From*: (из) высветите переменную *SES*, в списке *To* (в) выделите переменную *AL67*. Щелкните кнопкой *Add* (добавить). В списке *Paths* (пути) увидите запись (*SES*)-13->(*AL67*).

Таким же образом установите последовательно связи между факторами *SES* и *AL71*, между *AL67* и *AL71*.

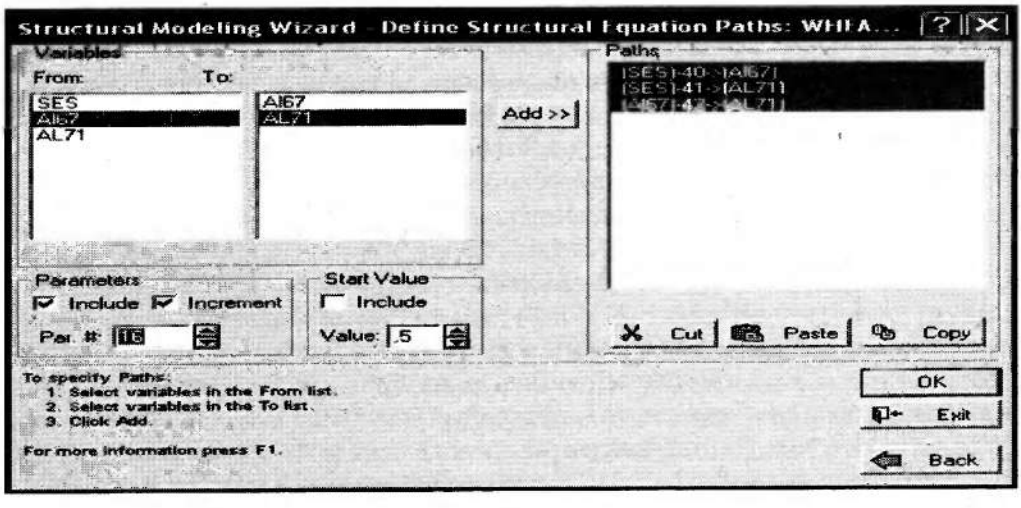


Рис. 16.13

Щелкните **OK**, в появившемся окне (рис. 16.14) выберите опцию *Append this model to existing program* (присоединить эту модель к существующей) и вновь щелкните **OK**. Модель определена, и программа вернулась в стартовую панель модуля **Structural Equation Modeling** (рис. 16.15).

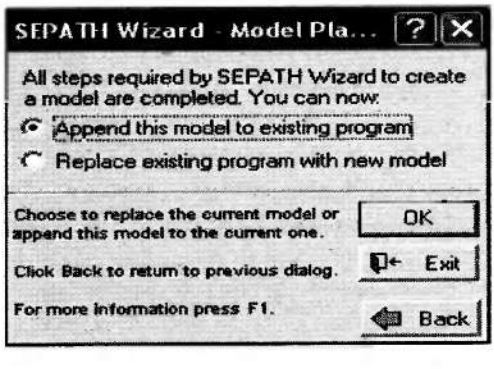


Рис. 16.14

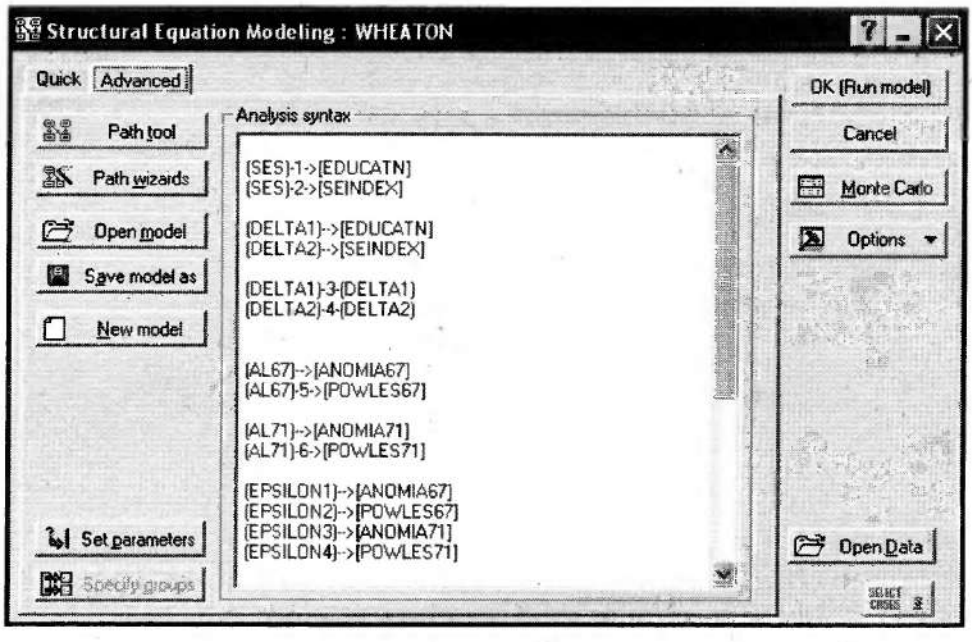


Рис. 16.15

Обратите внимание на то, что программа самостоятельно определила, какие параметры модели (нагрузки) свободны, а какие нет. Так, в строке [*AL67*]> [*ANOMIA67*] между фактором *AL67* и переменной *ANOMIA67* отсутствует номер свободного параметра, т.е. коэффициент при *AL67* равен 1, а в строке [*AL67*]–5> [*POWLES67*] присутствует номер 5, что соответствует свободному параметру a_5, в последствии оцениваемому программой.

Для того чтобы изменить, добавить новые связи на языке *PATH1*, нажмите кнопку **Path tool** (конструктур путей), откроется окно **Path Construction Tool** (рис. 16.16).

Таким образом, модель с языка диаграмм переведена на язык *PATH1*. Если в стартовой панели модуля нажать кнопку **Set Parameters** (установить параметры), то на экране появится окно **Analysis Parameters** (параметры анализа) (рис. 16.17).

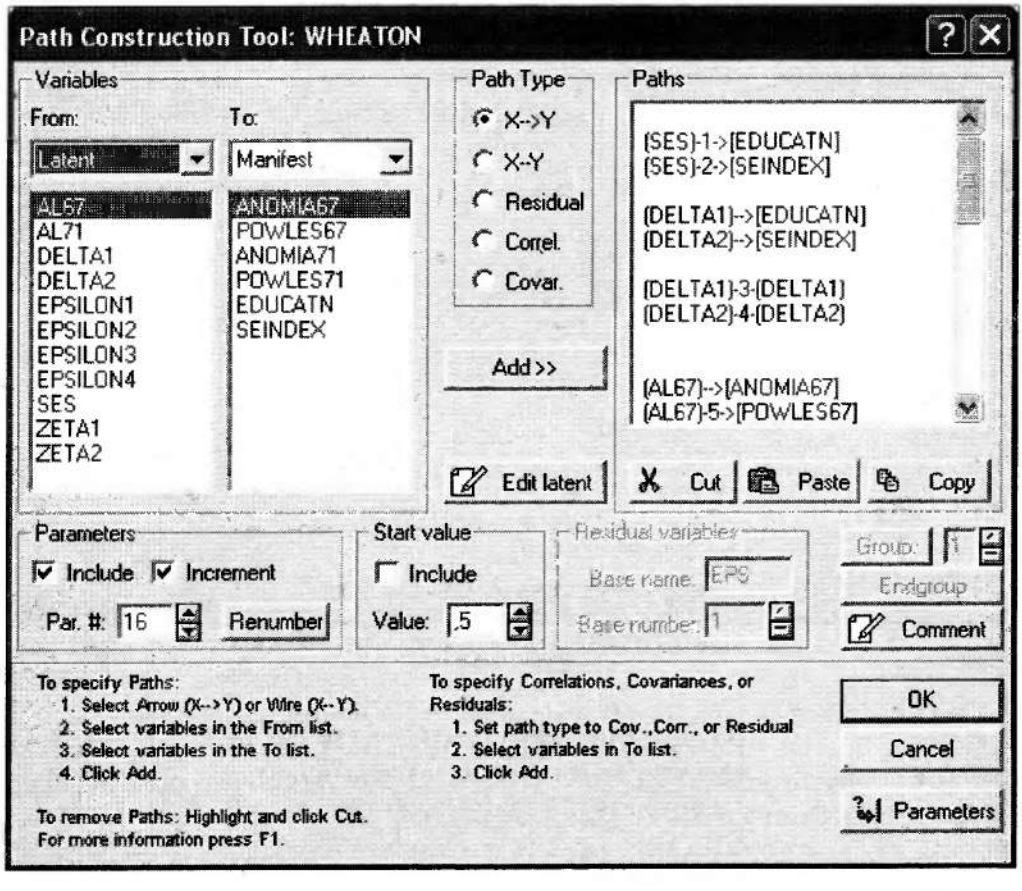


Рис. 16.16

Рассмотрим назначение основных параметров диалога.

1. Рамка **Data to analyze** (данные для анализа).

Covariance's (ковариации). При выборе этой опции анализируется ковариационная матрица входных переменных. Если файл содержит необработанные исходные данные, то **SEPATH** проанализирует их и вычислит соответствующую матрицу ковариаций. Распределение элементов ковариационной матрицы отличается от распределения элементов корреляционной матрицы. Это видно по диагональным элементам: у ковариационной матрицы они являются дисперсиями переменных, а у корреляционной они всегда равны 1.

Correlations (корреляции). Если выбрана эта опция, то **SEPATH** вычисляет корреляционную матрицу входных данных и проводит ее анализ. **SEPATH** может оценить новую полностью стандартизованную модель, в которой все переменные, явные и латентные, стандартизованы и имеют единичную дисперсию, также позволяя оценить стандартные ошибки коэффициентов путей.

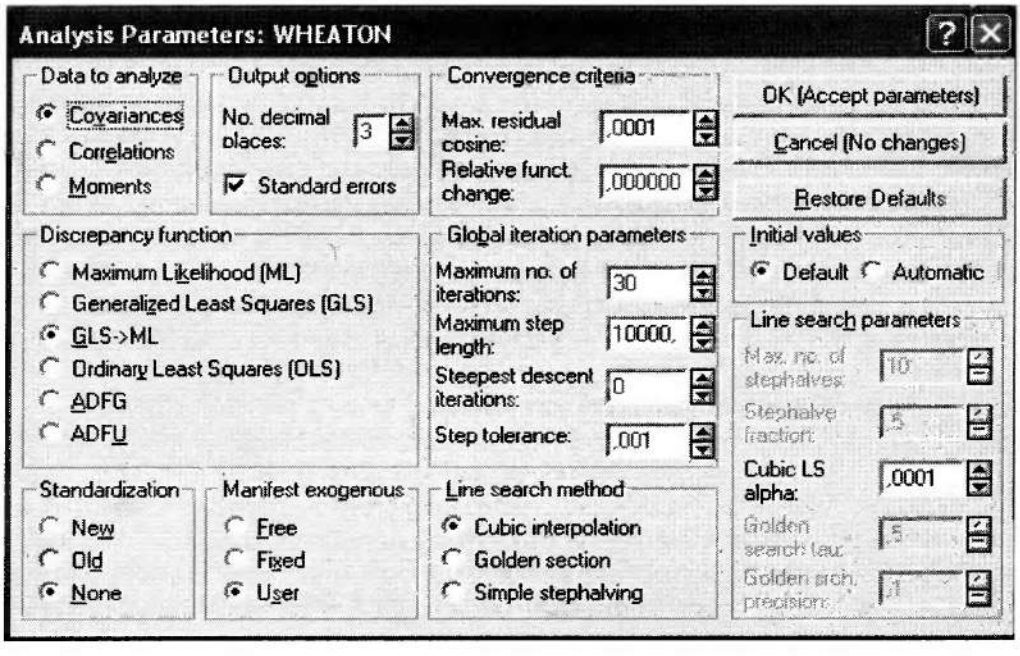


Рис. 16.17

Moments (моменты). При такой установке **SEPATH** анализирует расширенную матрицу произведений моментов вместо ковариационной матрицы. Эта опция выбирается, если анализируются модели, включающие свободные члены или структурированные средние.

2. Рамка **Discrepancy function** (функция несогласия). Здесь выбирается, какая функция или функции несогласия будут минимизироваться в процессе оценивания параметров.

Maximum Likelihood (ML) (максимум правдоподобия). В этом случае производится оценивание по методу максимума правдоподобия *Wishart* (Уишарта), если анализируются корреляции и ковариации, и по обычному методу максимума правдоподобия, если анализируются моменты.

Generalized Least Squares (GLS) (обобщенный метод наименьших квадратов) позволяет провести оценивание по обобщенному методу наименьших квадратов.

GLS → ML. Устанавливается по умолчанию. *SEPATH* выполняет 5 итераций с использованием метода наименьших квадратов вне зависимости от заданного значения максимального числа итераций в поле рамки **Global iteration parameters**, а после этого использует оценивание по методу максимума правдоподобия.

Ordinary Least Squares (OLS) (метод наименьших квадратов). Выполняется оценивание с помощью обычного метода наименьших квадратов.

ADFG (асимптотически свободное от распределения оценивание с помощью Граммиана). *SEPATH* производит асимптотически свободное от распределения оценивание, которое не требует многомерной нормальности. В этом случае на

матрицу весов накладывается условие Грамма (т.е. симметричность и положительная полуопределенность). Перед проведением *ADFG* оценивания *SEPATH* выполняет *GLS*.

ADFU (асимптотически свободное от распределения несмешанное оценивание). *SEPATH* выполняет асимптотически свободное от распределения несмешанное оценивание, которое также не предполагает многомерной нормальности. В этом случае эмпирическая матрица весов является несмещенной оценкой для истинной матрицы весов. Оцениваемая матрица может не удовлетворять условиям Грамма. В этом случае матрица весов не имеет обратной, поэтому *ADFU* оценивание невыполнимо и необходимо повторить оценивание с использованием *ADFG* опции. *SEPATH* выполняет сначала оценивание по *GLS*, а затем *ADFG* оценивание.

3. Рамка **Standardization** (стандартизация).

New (новая). В случае выбора этой опции все латентные переменные решения, зависимые и независимые, имеют дисперсию 1 . В отличие от старого метода стандартизации, в этом случае также вычисляются стандартные ошибки. Использование этой опции вместе с опцией *Correlations* в рамке **Data to analyze** позволяет оценить полностью стандартизованную модель путей, в которой все переменные, явные и латентные, имеют единичные дисперсии. В этой модели стандартные ошибки могут быть оценены для всех параметров.

Old (старая). Некоторые программы выполняют вычисления после завершения оценивания, преобразуя полученное решение в стандартизированную форму. Этот метод позволяет находить стандартизованное решение быстрее, так как не проверяется выполнение ограничивающих условий при каждой итерации, но при этом могут быть вычислены стандартные ошибки.

None (нет). Вычисляется нестандартизованное решение.

4. Рамка **Manifest exogenous** (явные экзогенные).

Free (свободные). Если выбрана эта опция, дисперсии и ковариации между явными экзогенными переменными рассматриваются как свободные параметры и добавляются к модели, хотя их значения явно не указываются. Начальные значения принимаются равными наблюдаемым значениям дисперсии и ковариации. Если выбрана эта опция, любая попытка задать дисперсии и ковариации явных экзогенных переменных приведет к сообщению об ошибке.

Fixed (фиксированные). В этом случае дисперсии ковариации явных экзогенных переменных фиксированы в процессе итерации (и равны выборочным значениям), а по окончании итерации рассматриваются, как если бы они были свободными параметрами. Этот подход неприменим, если в рассматриваемой модели имеется несколько явных экзогенных переменных и несколько свободных параметров, которые сильно замедляют процесс оценивания и делают его результаты малодостоверными. Полученные при этом результаты следует перепроверить с использованием опции *Free*, чтобы гарантировать точность оценок. Если выбран этот метод, любая попытка задать дисперсии и ковариации явных экзогенных переменных приведет к сообщению об ошибке.

User (определяемые пользователем). В данном случае пользователь должен описать дисперсии и ковариации для всех явных экзогенных переменных с использованием стандартного синтаксиса *PATH1*. Это описание может быть создано за несколько секунд с использованием опции *Covar* в рамке **Path Type** (тип пути) в окне **Path Construction Tool** (рис.16.16).

5. Рамка **Convergence criteria** (критерий сходимости).

Max. residual cosine (максимум косинуса остатков), значения этого критерия малы при стабилизировавшихся значениях параметров, оно может быть изменено в поле ввода справа от названия критерия, по умолчанию оно равно 0,0001, и это значение срабатывает практически во всех ситуациях.

Relative funct. Change (критерий относительного приращения функции). Значение этого критерия должно быть очень маленьким.

6. Рамка **Global iteration parameters** (общие параметры итерации). Рекомендуется не менять значения параметров итерации или изменять их незначительно.

Maximum no. of interations (максимальное число итераций). В это поле вводится максимально допустимое число итераций (по умолчанию равно 30), при достижении которого процесс остановится, программа при этом выдаст сообщение о превышении установленного максимального числа итераций. Минимальное число итераций равно 0, при этом программа просто посчитает функцию несогласия и оцененную матрицу ковариаций. Максимальное число итераций не должно превышать 1000.

Maximum step length (максимальный размер шага). В этом поле задается максимально допустимое значение длины вектора приращения.

Steepest descent interations (число итераций наискорейшего спуска). В данное поле вводится число итераций с наикратчайшим спуском, используемое при стандартном процессе итераций.

Step tolerance (минимальная толерантность значимого параметра). В этом поле вводится значение допустимости, при котором параметр временно исключается из итеративного процесса. Значение допустимости вычисляется как разность между единицей и квадратом множественной корреляции между выбранным параметром и остальными параметрами. Если параметр в процессе итерации становится сильно избыточным по отношению к другим параметрам, приближенный Гессиан, используемый при проведении итераций по методу Гаусса-Ньютона, становится нестабильным. Очень низкое значение допустимости означает, что данный параметр никогда не будет удален из процесса итерации. Следовательно, процесс итерации может аварийно завершиться, если приближение Гессиана окажется почти сингулярным. Высокое значение допустимости означает, что параметры не будут изменяться, если они заметно коррелированы с другими параметрами, в этом случае итерация не достигнет точки решения.

7. Рамка **Line search method** (метод линейного поиска). После выбора направления, в котором будет происходить изменение параметров на текущем шаге итерации, программа выбирает длину вектора приращения, т.е. проблема минимизации

зации сводится к оцениванию одной вместо n неизвестных. Для выбора длины шага может быть использован один из следующих трех методов.

1) *Cubic interpolation* (кубическая интерполяция) отличается быстротой и относительной устойчивостью; отлично работает в подавляющем большинстве случаев;

2) *Golden section* (золотое сечение) позволяет точно решить задачу линейной минимизации на каждом шаге итерации; часто сходится на несколько итераций быстрее, чем метод кубической интерполяции, но работает дольше, поскольку на каждом шаге итерации требует оценивания большего числа функций;

Simple stephalving (деление пополам) — самый быстрый из рассмотренных методов, но во многих случаях, когда два предыдущих метода сходятся к минимуму, деление пополам пропускает точку минимума и расходится.

8. Рамка **Initial values** (начальные значения).

Default (по умолчанию) использует начальное значение 0,5 для всех свободных параметров, кроме дисперсий и ковариаций (или корреляций) явных экзогенных переменных. Начальные значения этих параметров принимаются равными наблюдаемым значениям в исходной выборке. Если указана эта опция, вы можете также указать конкретное начальное значение любого свободного параметра, включив желаемое значение в скобках в тексте программы сразу после номера параметра. Так, например, следующая строка устанавливает начальное значение — 0,47 для параметра с номером 5.

```
(F1)-5{- 0,47->[X2]
```

Automatic (автоматически). В этом случае начальные значения свободных параметров получаются автоматически, с использованием модификации метода.

9. Рамка **Line search parameters** (параметры линейного поиска).

Max. No. of Stephalves (максимальное число делений пополам). Устанавливает максимальное число делений пополам, производимое на каждом шаге итерации при проведении для линейного поиска метода деления пополам.

Stephalves Fraction (коффициент деления пополам). Это параметр, который устанавливает отношение, в котором происходит деление отрезка при переходе от текущего шага итерации к следующему, используя деление пополам.

Cubic LS alpha (МНК альфа кубической интерполяции) контролирует минимальную величину, на которую должно уменьшиться значение функции несогласия, чтобы новый шаг итерации был приемлем при использовании метода **Cubic LS alpha** (по умолчанию, равно 0,0001). Позволяет рассматривать как допустимое почти любое уменьшение функции несогласования.

Golden search tau (тау золотого сечения) контролирует диапазон, в котором происходит поиск по методу *Golden search*.

Golden srch. Precision (точность золотого сечения) определяет точность, с которой происходит оценивание по методу *Golden search*.

10. Рамка **Output options** (опции вывода).

No. decimal blaces (число десятичных знаков) задает число десятичных знаков, выводимых по умолчанию при распечатке результатов в текстовом окне результатов и в таблицах результатов.

Standard errors (стандартные ошибки). Если установлен флажок в этом поле, программа автоматически будет выводить оценки для стандартных ошибок всех параметров в выходном тексте *PATHI* и таблице с результатами оценивания параметров для модели. Однако если для оценивания модели был выбран *OLS* или для стандартизации (рамка **Standardization)** используется «старый метод» (*Old*), установка этого флажка невозможна.

16.5. Запуск процедуры оценивания. Анализ результатов

После того как модель записана на языке *PATH1* и параметры анализа установлены, можно произвести вычисления. Для этого в стартовом окне модуля запустите программу, щелкнув кнопкой **OK**, **Run model** (оценить модель). Программа начнет итерационную процедуру оценки неизвестных параметров. На рис. 16.18 представлена таблица (*Iteration Results*) результатов успешно завершившегося итерационного процесса, в которой приняты следующие обозначения.

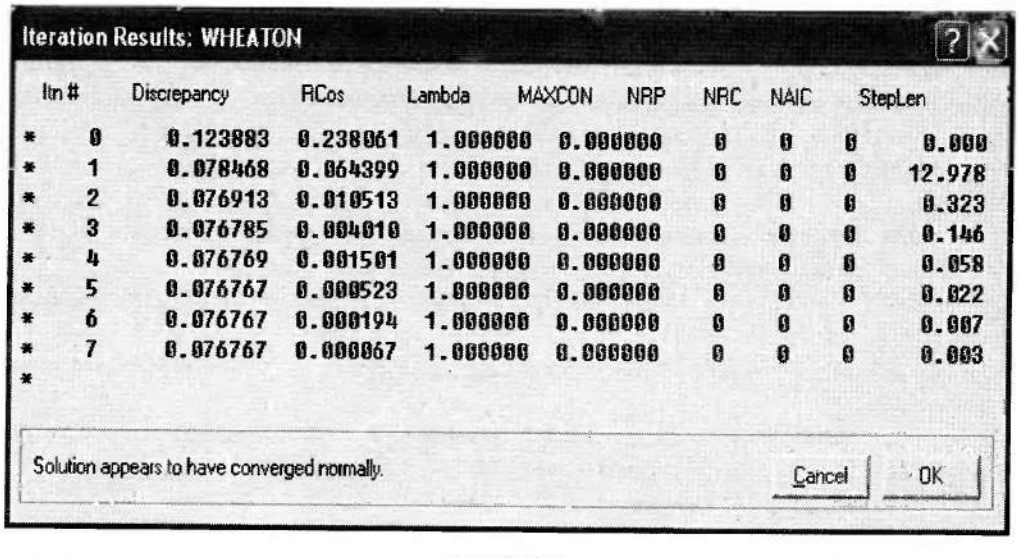


Рис. 16.18

Itn # (итерация). Номер произведенной итерации.

Discrepancy (несогласие). Текущее значение минимизируемой функции несогласия.

RCos. Текущее значение критерия максимума косинусов остатков.

Lambda (лямбда). Значение множителя приращения, использованного на текущем шаге итерации. Значение 1,0 означает, что первый «полный» шаг уменьшил значение функции несогласия «достаточно», чтобы перейти к следующей итерации. Значение меньше 1,0 означает, что программа использовала линейный поиск в выбранном направлении для выбора конкретной точки, уменьшающей значение функции несогласия. Очень маленькие значения обычно говорят о том, что проведение итерации было практически невозможно.

MAXCON. Максимальное значение функций ограничений. Это значение отлично от нуля только при проведении ограниченного оценивания. Ограниченное оценивание используется, если был выбран способ *New* в рамке **Standardization** или *Correlations* в рамке **Data to analyze**. Если процесс корреляции протекает нормально, то это значение уменьшается и приближается к нулю.

NRP. Значение *NRP* показывает число избыточных параметров. Программа определяет их число в процессе итераций, и если такие параметры имеются, *NRP* будет отлично от нуля.

NRC. Число избыточных ограничений. Программа определяет их число в процессе итераций, и если такие ограничения имеются, *NRC* будет отлично от нуля.

NAIC. Число активных ограничивающих неравенств или условий — это то, что использует программа в процессе итерации. Программа использует некоторые ограничивающие неравенства, чтобы предотвратить появление «невозможных» значений параметров. Например, недопустимы отрицательные значения дисперсий. Если программа обнаруживает, что на следующем шаге итерации дисперсия может принять отрицательное значение, она устанавливает ограничивающее неравенство (неотрицательной дисперсии) и после этого минимизирует функцию несогласия, используя остальные параметры. В обычном факторном анализе случаем Хейвуда называется ситуация, когда минимум функции несогласия достигается при отрицательных значениях одного или нескольких параметров, соответствующих дисперсиям некоторых параметров. Такие значения недопустимы. При возникновении случая Хейвуда во время проведения подтверждающего факторного анализа параметр *NAIC* будет отличен от нуля.

StepLen (длина шага). Длина полного шага при проведении текущей итерации. Если рядом стоит звездочка, это означает, что был достигнут максимальный размер шага (допустимый размер шага можно изменить в рамке **Global iteration parameters**.

Появившаяся внизу окна строка *Solution appears to have converged normally* (решение сошлось нормально) показывает, что итерационный процесс сошелся.

Щелкните кнопкой **OK**. Появится диалог анализа результатов структурного моделирования (рис. 16.19), который состоит из двух частей: верхней — информационной, где содержится основная информация о результатах оценивания, и нижней — функциональной, где выделены группы кнопок, позволяющие всесторонне просмотреть результаты, сохранить их в удобном виде, представить графически и т.д.

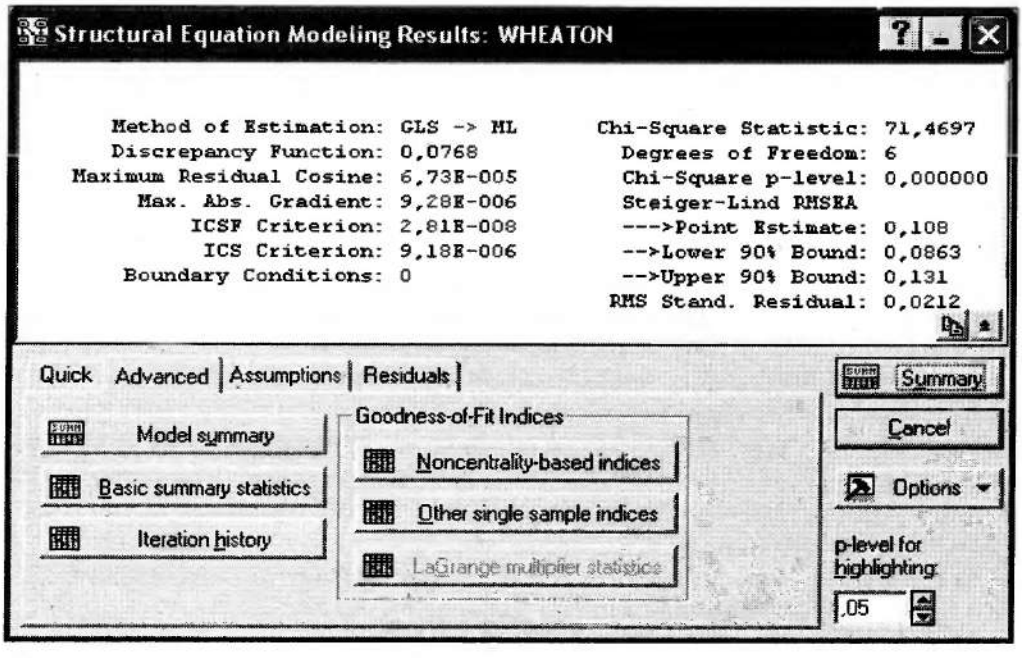


Рис. 16.19

В верхней информационной части выводятся следующие показатели.

Method of estimation (метод оценивания). Здесь выводится вид использованной функции несогласия.

Discrepancy Function (функция несогласия). Выводится окончательное значение, принимаемое функцией несогласия.

Maximum Resedual Cosine (максимум косинуса остатков). Если процесс итерации сошелся успешно, то значение близко к 0.

Maximum Absolute Gradient (максимальная по модулю компонента градиента). Это значение равно максимальной по модулю компоненте градиента.

Структурная модель называется устойчивой к умножению на постоянный множитель масштаба (УУПММ), если степень согласия модели и данных не меняется при одновременном умножении всех переменных на одну и ту же константу. Большинство, но не все встречающиеся на практике модели устойчивы к умножению на постоянный множитель масштаба. Рассматриваемый модуль выводит значения двух индексов, позволяющих судить о степени устойчивости подгоняемой модели.

ICSF Criterion. Значение этого критерия должно быть близким к нулю, если структурная модель устойчива к умножению на постоянный множитель масштаба.

ICS Criterion. Значение данного критерия должно быть близким к нулю, если рассматриваемая структурная модель устойчива к изменениям масштаба. При проведении анализа корреляции этот индекс должен быть близким к нулю.

Boundary Conditions (граничные условия). Выводится число ограничивающих неравенств, используемых при управлении сходимостью. Оно должно равняться

нулю, если только в анализируемой модели не возникла ситуация, подобная случаю Хейвуда. Если число не равно нулю, статистика χ^2 будет иметь неверное распределение, и ее использование дает ненадежные результаты.

Chi-square Statistic (статистика χ^2). Для всех функций несогласия, кроме *LS* (МНК), эта статистика, имеющая асимптотическое распределение χ^2, содержит информацию об истинности нулевой гипотезы, т.е. о точном соответствии построенной модели исходным данным. В моделях с несколькими группами эта статистика содержит итоговое значение для всех групп.

Degrees of Freedom (число степеней свободы) — число степеней свободы рассчитываемой статистики χ^2.

Chi-square p-level (*p-уровень* статистики χ^2) — вероятностный уровень статистики χ^2. Напомним, что p — это вероятность ошибочного отклонения нулевой гипотезы. Поэтому если p принимает значение, близкое к 0, то вероятность ошибочного отклонения нулевой гипотезы мала и можно ее отвергнуть. Это означает, что чем больше значение p и меньше значение критерия χ^2, тем достовернее нулевая гипотеза, а значит и адекватнее модель. Если $p < 0{,}05$, то отклоняем нулевую гипотезу при уровне значимости $p = 0{,}05$.

Steiger-Lind RMSEA. Здесь выводится *Point Estimate* (точечная оценка) и 90% *Bound* (90% доверительный интервал) для *Steiger-Lind RMSEA*.

RMS Stand. Residual (Root Mean Square Standardized Residual). Этот индекс также показывает качество подгонки модели. Если значение индекса меньше 0,05, подгонка очень хорошая, больше 0,1 — модель неадекватно описывает данные.

Рассмотрим функциональное назначение кнопок нижней части окна на вкладке **Advanced.**

Model Summary (итоговая) модель. Нажмите эту кнопку, откроется таблица с результатами оценивания. На рис. 16.20 приведен фрагмент таблицы. Строки таблицы соответствуют записи на языке *PATH1* очередного пути, в столбцах приведены оценки свободного параметра, стандартные ошибки, значения *t-статистик*, *p-уровни* значимости статистик. Значимые *t-статистики* ($p < 0{,}05$) выделены красным цветом. Если *t-статистика* значима, то верна гипотеза о неравенстве нулю оценки соответствующего свободного параметра.

	Model Estimates (Wheaton)			
	Parameter Estimate	Standard Error	T Statistic	Prob. Level
(EPSILON3)-9-(EPSILON3)	3,701	0,373	9,910	0,000
(EPSILON4)-10-(EPSILON4)	3,625	0,292	12,412	0,000
(ZETA1)-->(AL67)				
(ZETA2)-->(AL71)				
(ZETA1)-11-(ZETA1)	5,307	0,473	11,230	0,000
(ZETA2)-12-(ZETA2)	3,741	0,388	9,653	0,000
(SES)-13->(AL67)	-1,585	0,120	-13,220	0,000
(SES)-14->(AL71)	-0,450	0,136	-3,303	0,001
(AL67)-15->(AL71)	0,705	0,054	13,164	0,000

Рис. 16.20

В этой таблице даны оценки параметров регрессионной модели, связывающие факторы *SES*, *AL67*, *AL71*, т.е. имея только явные переменные, программа построила регрессионную модель, связывающую скрытые общие факторы. Ни в каком другом модуле *STATISTICA* этого сделать непосредственно нельзя.

Basic Summary Statistics (основные итоговые статистики). При нажатии на эту кнопку появятся основные статистики, которые даны в информационной части окна. Однако представление их в таблице удобно, так как далее они могут быть сохранены, распечатаны, представлены графически и т.д.

Iteration history (отчет об итерациях). В процессе проведения итераций программа выводит информацию о них в окне **Iteration Results**. Открыв эту таблицу, можно снова просмотреть полученные результаты, а также сохранить их в файле данных или провести их графический анализ.

Noncentraility-based indices (индексы нецентральности). Эти индексы показывают степень адекватности модели на основе оценки параметра нецентральности статистики χ^2. Щелкните этой кнопкой. Откроется таблица (рис. 16.21), в которой последовательно даны: нижняя граница 90% доверительного интервала, точечная оценка индекса и верхняя граница 90% доверительного интервала. В таблице представлены следующие индексы: параметр нецентральности распределения, Стингера-Линда — значения этих индексов меньшие чем 0,05 говорят о хорошей подгонке модели, значения этих индексов меньшие чем 0,01 говорят об отличной подгонке модели; нецентральности МакДональда, гамма, скорректированный гамма-индекс — хорошей подгонке соответствуют значения этих индексов большие чем 0,95.

Кнопка **Other Single Sample Indexes** (другие одновыборочные индексы). Если нажать на эту кнопку, появятся некоторые наиболее известные одновыборочные индексы подгонки, а также некоторые связанные с ними величины: Джорескога, скорректированный индекс Джорескога, информационный критерий Акаике, байесовский критерий Шварца.

	Noncentrality Fit Indices (Wheaton)		
	Lower 90% Conf. Bound	Point Estimate	Upper 90% Conf. Bound
Population Noncentrality Parameter	0,045	0,070	0,103
Steiger-Lind RMSEA Index	0,086	0,108	0,131
McDonald Noncentrality Index	0,950	0,966	0,978
Population Gamma Index	0,967	0,977	0,985
Adjusted Population Gamma Index	0,884	0,920	0,949

Рис. 16.21

Для того чтобы проверить выполнимость условий применения модуля **SEPATH**, надо воспользоваться вкладкой **Assumptions** (проверка предположений, рис. 16.22). Большинство процедур *SEPATH* работает в предположении нормальности выборки.

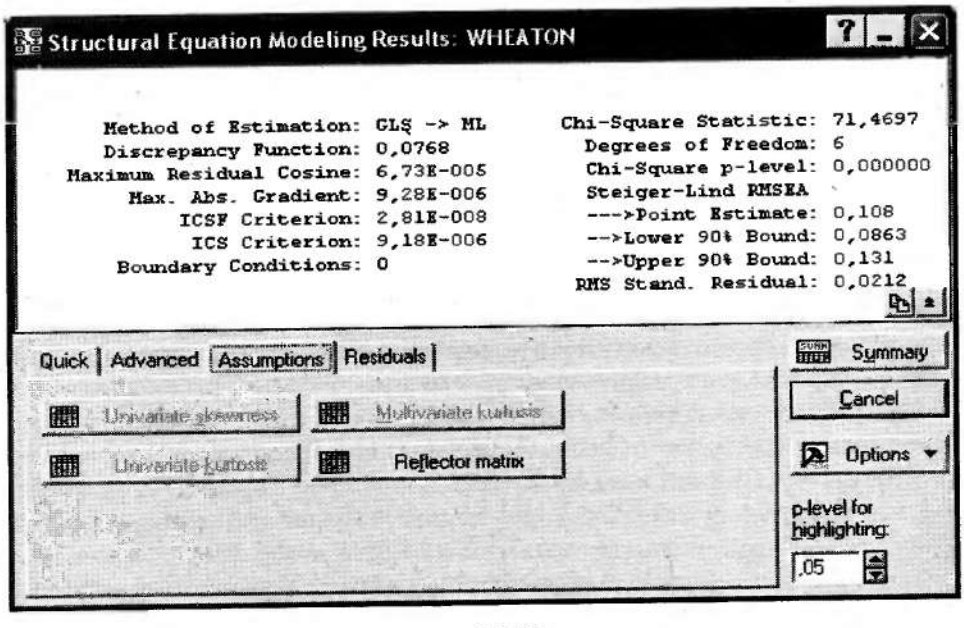


Рис. 16.22

Univariate skwness (одномерная асимметрия). Нажатием на эту кнопку получим таблицу результатов с различными величинами одномерной асимметрии.

Univariate kurtosis (одномерный эксцесс). После нажатия этой кнопки на экране отображается таблица результатов с тремя величинами одномерного эксцесса: эксцесс, уточненный эксцесс и нормированный эксцесс.

Multivariate kurtosis (многомерный эксцесс). Эта кнопка открывает таблицу результатов с различными величинами многомерного эксцесса, включая те, которые обеспечивают совместимость с другими программами структурного моделирования. Напомним, что чем ближе значение асимметрии и эксцесса к 0, тем более распределение соответствует нормальному закону.

Reflector matrix (матрица-рефлектор). Матрица-рефлектор (рис. 16.23) применяется для оценивания инвариантных свойств модели, т.е. устойчивости модели к изменению масштаба измерения исходных данных. Чем ближе значения элементов матрицы, тем более устойчива модель к изменению масштаба.

	Reflector Matrix (Wheaton)					
	ANOMIA67	POWLES67	ANOMIA71	POWLES71	EDUCATN	SEINDEX
ANOMIA67	-0,000	0,000	-0,143	0,117	-0,065	0,051
POWLES67	-0,000	0,000	0,144	-0,113	0,078	-0,139
ANOMIA71	-0,147	0,106	-0,000	0,000	-0,052	-0,001
POWLES71	0,150	-0,104	-0,000	0,000	0,056	0,087
EDUCATN	-0,045	0,050	-0,037	0,029	0,000	-0,000
SEINDEX	0,003	-0,004	0,003	-0,000	0,000	-0,000

Рис. 16.23

Для того чтобы оценить адекватность модели, перейдите на вкладку **Residuals** (анализ остатков, рис. 16.24). Остатки — это разность между наблюдаемыми данными и значениями, прогнозируемыми с помощью модели. Просмотр остатков позволяет оценить качество подгонки.

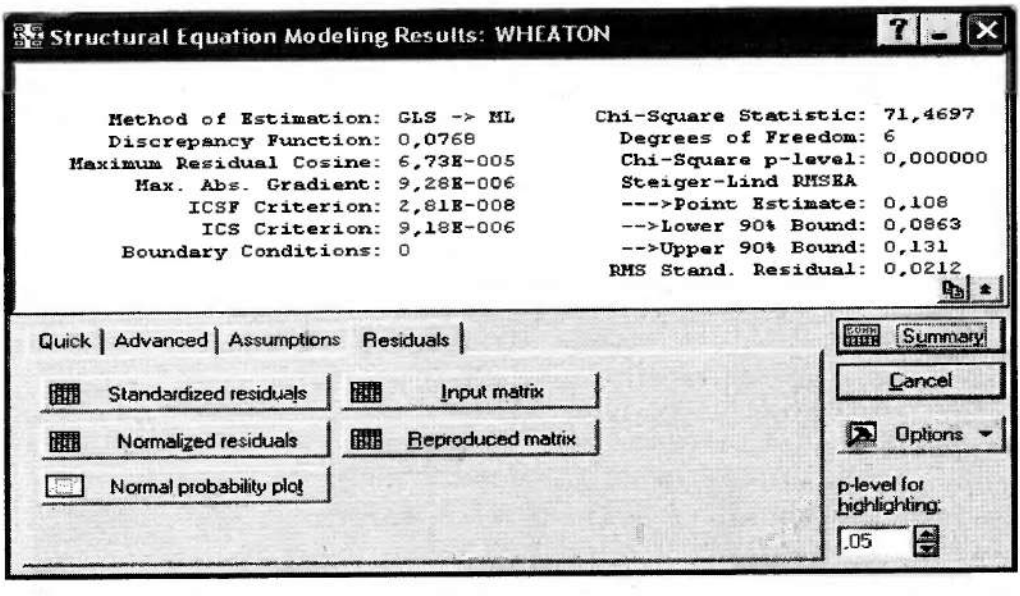


Рис. 16.24

В окне имеются следующие кнопки для исследования остатков:

1. **Standartized Residuals** (стандартизованные остатки) — просмотр стандартизованных остатков;
2. **Normalized Residuals** (нормализованные остатки) — просмотр нормализованных остатков;
3. **Normal Probability Plot** (нормальный вероятностный график, рис. 16.25); напомним, что одним из признаков адекватности модели является соответствие закона распределения остатков нормальному закону; чем плотнее располагаются точки на прямой, тем более закон распределения остатков соответствует нормальному закону;
4. **Input matrix** (исходная матрица);
5. **Reproduced Matrix** (воспроизведенная матрица, рис. 16.26).

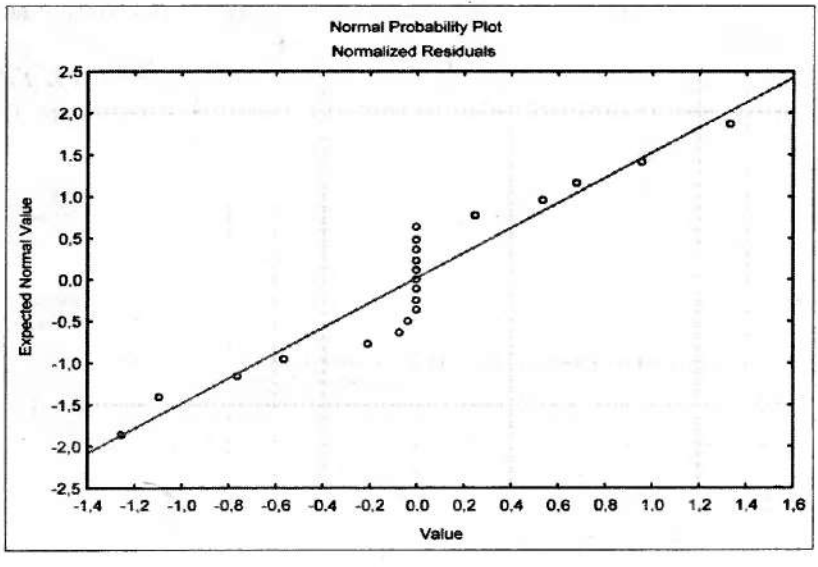


Рис. 16.25

	Reproduced Matrix (Wheaton)					
	ANOMIA67	POWLES67	ANOMIA71	POWLES71	EDUCATN	SEINDEX
ANOMIA67	11,834	6,947	6,223	5,281	-4,092	-21,805
POWLES67	6,947	9,364	5,529	4,692	-3,635	-19,373
ANOMIA71	6,223	5,529	12,532	7,495	-4,045	-21,554
POWLES71	5,281	4,692	7,495	9,986	-3,432	-18,292
EDUCATN	-4,092	-3,635	-4,045	-3,432	9,610	35,522
SEINDEX	-21,805	-19,373	-21,554	-18,292	35,522	450,288

Рис. 16.26

Последние две кнопки позволяют просмотреть и сравнить матрицы: исходную (рис. 16.4), которая является исходной для анализа, и воспроизведенную, которая подсчитана на модели с оцененными параметрами. Чем более отличны элементы этих матриц, тем менее адекватна модель.

Подведем итог исследования построенной модели. Итерационный процесс (рис. 16.18) сошелся успешно, значения *ICSF Criterion* и *ICS Criterion* близки к 0, значение функции несогласия *(Discrepancy Function)* мало, значения максимума косинуса остатков *(Maximum Resedual Cosine)*, максимальная по модулю компонента градиента *(Maximum Absolute Gradient)* близки к 0. Все перечисленные факторы свидетельствуют об адекватности построенной модели.

Тем не менее, значение статистики χ^2 и уровень значимости критерия p, нормальный вероятностный график, значение *RMS Stand. Residual* (больше чем 0,05) свидетельствуют о недостаточной адекватности модели.

Таким образом, по-видимому, будет верным общий итог анализа — качество подогнанной модели невысокое. Попытаемся улучшить результат структурного

моделирования, изменив структуру связей в модели, усложнив первоначальную структуру.

Установите связь между переменными *EPSILON1* и *EPSILON3, EPSILON2* и *EPSILON4*. Для этого при помощи конструктура путей (**Path tool**, рис.16.27) добавьте в модель новые связи.

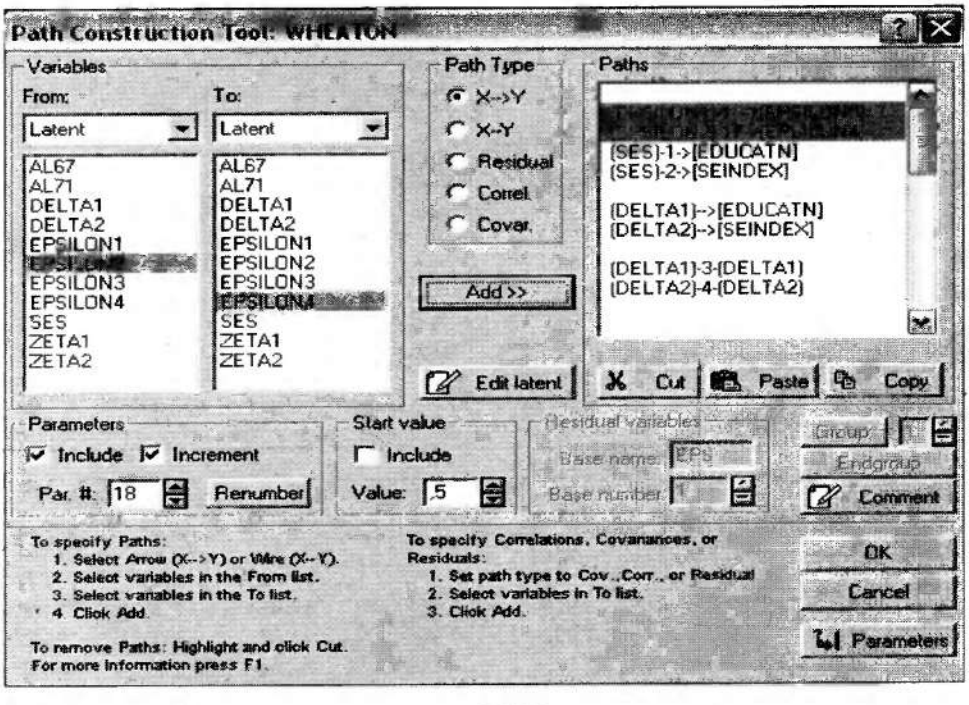


Рис. 16.27

На рис. 16.28 приведены информационные части окон результатов анализа исходной модели (первый вариант) и модели с измененной структурой связей (второй вариант).

Таким образом, за счет небольших структурных изменений нам удалось более успешно реализовать процесс моделирования структурными уравнениями и построить вполне адекватную модель.

Используя вычисленные параметры модели (рис.16.29), можно записать и проанализировать интересующие нас соотношения между социальными характеристиками человека.

Первый вариант

```
Method of Estimation: GLS -> ML
Discrepancy Function: 0,0768
Maximum Residual Cosine: 5,75E-005
Max. Abs. Gradient: 7,49E-006
ICSF Criterion: 7,84E-009
ICS Criterion: 7,52E-006
Boundary Conditions: 0

Chi-Square Statistic: 71,4697
Degrees of Freedom: 6
Chi-Square p-level: 0,000000
Steiger-Lind RMSEA
--->Point Estimate: 0,108
-->Lower 90% Bound: 0,0863
-->Upper 90% Bound: 0,131
RMS Stand. Residual: 0,0212
```

Второй вариант

```
Method of Estimation: GLS -> ML
Discrepancy Function: 0,00508
Maximum Residual Cosine: 4,94E-005
Max. Abs. Gradient: 1,36E-006
ICSF Criterion: 4,53E-008
ICS Criterion: 1,63E-006
Boundary Conditions: 0

Chi-Square Statistic: 4,73018
Degrees of Freedom: 4
Chi-Square p-level: 0,316119
Steiger-Lind RMSEA
--->Point Estimate: 0,0139
-->Lower 90% Bound: 0
-->Upper 90% Bound: 0,0531
RMS Stand. Residual: 0,00745
```

Рис. 16.28

EDUCATN = 2,609SES + 2,804;
SEINDEX = 13, 616SES + 264,885,
ANOMIA67 = AL67 +4,736;
POWLESS67 = 0,979AL67 +2,566;
ANOMIA71 = AL71 +4,404;
POWLESS71 = 0,922AL71 +3,073.

Из приведенных уравнений следует, что с увеличением социоэкономического статуса возрастает уровень образованности, коммуникабельности человека; с увеличением отчужденности от бщества возрастают аномальность в поведении человека, показатель отсутствия способностей.

Линейные регрессионные модели, выражающие зависимость между отчужденностью человека от общества и его социоэкономическим статусом соответственно в 1967 г. и 1971 г. имеют вид:

AL67 = – 1,5SES + 4,847,
AL71= – 0,592SES + 0,607AL67 + 4,088.

Из приведенных уравнений следует вывод: чем выше социоэкономический статус человека, тем меньше его отчужденность от общества и менее аномальным становится его поведение. Причем величина отчужденности человека от общества в 1971 г. зависит от величины отчужденности от общества в 1967 г.

	Model Estimates (Wheaton)			
	Parameter Estimate	Standard Error	T Statistic	Prob. Level
(SES)-1->[EDUCATN]	2,609	0,125	20,939	0,000
(SES)-2->[SEINDEX]	13,616	0,791	17,219	0,000
(DELTA1)-->[EDUCATN]				
(DELTA2)-->[SEINDEX]				
(DELTA1)-3-(DELTA1)	2,804	0,508	5,521	0,000
(DELTA2)-4-(DELTA2)	264,885	18,156	14,590	0,000
(AL67)-->[ANOMIA67]				
(AL67)-5->[POWLES67]	0,979	0,062	15,887	0,000
(AL71)-->[ANOMIA71]				
(AL71)-6->[POWLES71]	0,922	0,060	15,490	0,000
(EPSILON1)-->[ANOMIA67]				
(EPSILON2)-->[POWLES67]				
(EPSILON3)-->[ANOMIA71]				
(EPSILON4)-->[POWLES71]				
(EPSILON1)-7-(EPSILON1)	4,736	0,454	10,435	0,000
(EPSILON2)-8-(EPSILON2)	2,566	0,404	6,356	0,000
(EPSILON3)-9-(EPSILON3)	4,404	0,516	8,537	0,000
(EPSILON4)-10-(EPSILON4)	3,073	0,435	7,067	0,000
(ZETA1)-->(AL67)				
(ZETA2)-->(AL71)				
(ZETA1)-11-(ZETA1)	4,847	0,468	10,354	0,000
(ZETA2)-12-(ZETA2)	4,088	0,405	10,099	0,000
(SES)-13->(AL67)	-1,500	0,124	-12,086	0,000
(SES)-14->(AL71)	-0,592	0,131	-4,515	0,000
(AL67)-15->(AL71)	0,607	0,051	11,892	0,000
(EPSILON1)-16-(EPSILON3)	1,625	0,314	5,173	0,000
(EPSILON2)-17-(EPSILON4)	0,339	0,261	1,297	0,195

Рис. 16.29

Таким образом, путем моделирования структурными уравнениями показана причинная связь между такими факторами, как социоэкономический статус человека и отчужденность от общества. Еще раз подчеркнем, что важность и значимость результатов моделирования заключаются в том, что эти факторы являются скрытыми (неявными), так как в файле данных по этим факторам отсутствуют данные.

Естественно, невозможно объективно судить о качестве структурной модели на основе одной реализации модуля. Необходима серия экспериментов, чтобы накопить определенную статистику и сделать более достоверный вывод об адекватности построенной модели.

В модуле **SEPATH** имеются большие возможности для моделирования методом статистических испытаний Монте-Карло и оценивания адекватности модели [2].

Вернемся в диалог **Structural Equation Modeling** и нажмем кнопку **Monte-Carlo**. Откроется окно **Monte-Carlo Analysis** (рис. 16.30). Рассмотрим функциональное назначение кнопок и установочных опций диалога на вкладке **Advanced** [6].

Seed1 (начальное значение 1). Это главное начальное значение, которое используется при проведении всех экспериментов по методу Монте-Карло. Оно должно находиться в границах от 1 до 2147483647.

Seed2 (начальное значение 2). Это начальное значение используется только при проведении экспериментов, требующих генерации нормального распределения, содержащего выбросы. Оно должно находиться в границах от 1 до 2147483647.

Number of replications (число повторений). В этом поле задается число повторений анализа по методу Монте-Карло. Допустимым является любое целое значение из диапазона от 1 до 1000.

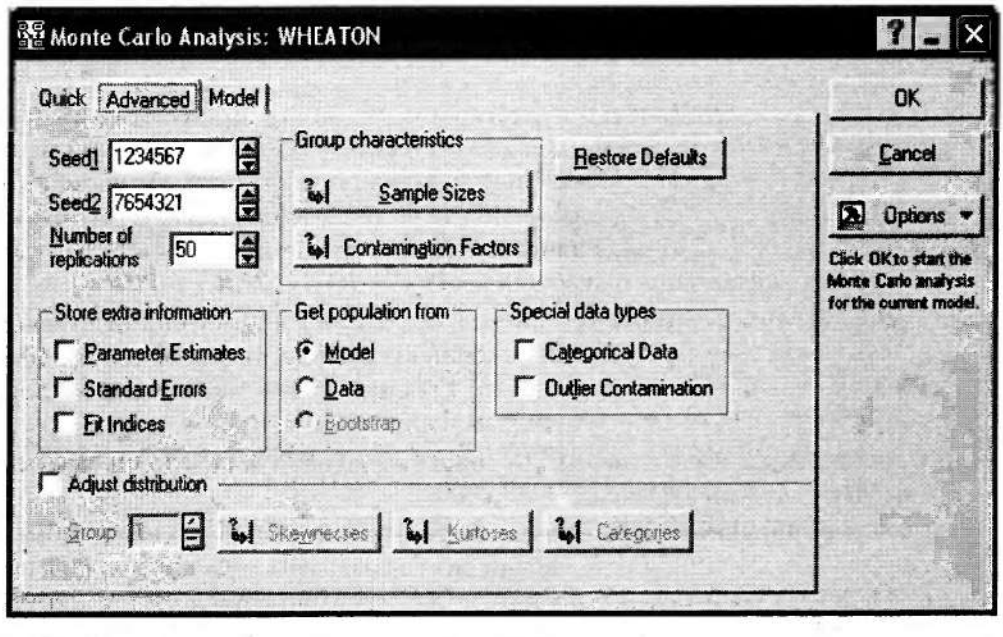


Рис. 16.30

Рамка **Group characteristics** (групповые характеристики). В этой части окна можно выбрать характеристики групп, используемых при анализе по методу Монте-Карло.

1. **Sample Sizes** (объемы выборок). Эта кнопка открывает окно **Set Monte-Carlo Sample Sizes** (объемы выборок для метода Монте-Карло), в котором можно задать размеры выборок по каждой анализируемой группе. Размер выборки должен быть больше числа анализируемых переменных, и может варьироваться для разных групп.
2. **Contamination Factors** (засоряющие факторы). В окне **Set Contamination parameters** (задание парметров выбросов) для каждой группы можно ввести процент выбросов *p* (*probability Contamination*) и множитель *k* (*multiplier*) для получения выбросов. Для имитации наличия выбросов модуль анализа по методу Монте-Карло использует стандартную технику создания выборок

из смешанных распределений. Далее, предположим, что новые данные были получены из распределения, имеющего среднее значение μ и ковариационную матрицу Σ. Засоренная выборка получается заменой указанного процента наблюдений p случайными величинами со средним μ и матрицей ковариаций $k\Sigma$, где k — сравнительно большой множитель (скажем, 10). Для этого выбросы вычисляются как сумма среднего значения и смещения от среднего, умноженного на квадратный корень из k.

Для каждой группы пользователь определяет параметры выбросов k и p. Таким образом, доля p наблюдений в полученной выборке будут иметь (при большом размере выборке) среднее значение μ и матрицу ковариаций $k\Sigma$.

В рамке **Store extra information** (дополнительная информация) можно задать в качестве опции сохранение дополнительной информации о результатах каждого анализа.

1. *Parameter Estimates* (оценки параметров). Эта опция позволяет сохранить значения свободных параметров. Для их обозначения используются имена, заданные в программе *PATH1*. Они имеют вид *PAR_#*, где # обозначает номер параметра. Таким образом, если в программе использованы свободные параметры 1, 2, 3, 6, в окно **Summary Residuals** (итоговые результаты) будет добавлена информация о переменных *PAR 1*, *PAR 2*, *PAR 3*, и *PAR 6*.
2. *Standard Errors* (стандартные ошибки). Использование этой опции позволяет сохранить оценки стандартных ошибок для каждого свободного параметра с использованием следующей схемы. Каждая стандартная ошибка обозначается *SE_#*, где # — номер свободного параметра.
3. *Fit indices (*индексы согласия). Эта опция позволяет сохранить индексы согласия для каждого проведенного анализа.

В рамке **Get population from** (использовать распределение из) надо выбрать одну из нескольких опций определяющих, из какого распределения должны выбираться данные для проведения анализа по методу Монте-Карло.

1. *Model* (модель). Если выбрана эта опция, программа генерирует данные в соответствии с текущей моделью. При этом модель, активная в текущем окне, будет использована для генерации ковариационной матрицы нового распределения. Конкретнее значения свободных параметров могут быть изменены непосредственным редактированием их начальных значений в фигурных скобках после номера параметра. Так, например, если нужно, чтобы распределение основывалось на факторной модели, где первый фактор нагружает 0,5 на первую переменную, следует добавить похожую строчку в файл **.cmd*:

$$(F1)-1\{0,5\}->[X1].$$

2. *Data* (данные). Если выбрана эта опция, программа использует для создания новых выборок параметры имеющихся входных данных. При этом она производит вычисления соответствующих матриц (ковариации, корреляции, смешанных моментов) и генерирует новые данные, основанные на характеристиках выбранного распределения.

3. *Bootstrap* (бутстреп). Эта опция может быть выбрана, только если фай входных данных содержит необработанные данные. При использовани этого метода новая выборка размера N получается с помощью рассмотре ного текущего файла данных как множества значений многомерного ди кретного распределения, где каждое значение имеет равную вероятнос появления в выборке. Предположим, что текущий файл данных содерж 100 наблюдений по 10 переменным. Если была выбрана опция бутстре с размером выборки, равным 50, программа произведет случайную выбор 50 целых чисел из диапазона от 1 до 100. Полученный список использует для определения того, какие из наблюдений, содержащихся в файле исхо ных данных, будут включены в новую бутстреп-выборку.

В рамке **Special data types** (специальные виды данных) предоставляется во можность генерации данных, имеющих распределение специального (отлично от нормального) вида.

1. *Categorical Data* (категориальные данные). Программа после генераци данных обычным способом преобразует данные из непрерывных в катег риальные, используя заданные правила категоризации. Можно задать от до 10 категорий, а также задать соответствующие им ограничения.
2. *Outlier Contamination* (содержащие выбросы). Эта опция позволяет создат выборку, содержащую выбросы, используя стандартную технику генер ции смешанных распределений, которая была описана в **Contaminatio Factors.**

В рамке **Adjust distribution** (дополнительные) можно задать параметры ра пределения искусственно создаваемых данных, отличных от многомерного но мального. Эти параметры могут быть заданы отдельно для каждой группы.

1. **Group** (группа). В данном поле вводится число групп, характеристики к торых будете менять впоследствии.
2. **Skewnesses** (асимметрия). Указанная кнопка активирует окно, в которо можно задать требуемые значения для асимметрии переменных в кажд группе.
3. **Kurtoses** (эксцесс). Эта кнопка активирует окно, в котором можно зада требуемые значения для эксцесса переменных в каждой группе.
4. **Categories** (категории). Кнопка открывает интерактивную таблицу резул татов, в которую можно ввести параметры преобразования переменны в категориальную форму. Для каждой переменной необходимо задать чи ло точек разбиения ее области значений на категории. Если предполагает определить k категорий, необходимо ввести $k-1$ точек разбиения. Все зн чения непрерывной переменной, меньше первого введенного значения, б дут отнесены к категории 1. Все значения, больше первого значения и мен ше второго, получат категориальное значение 2 и т.д.

Кнопка **Restore Defaults** (установки по умолчанию) восстанавливает значе ния по умолчанию.

Выберите установки опций диалога **Monte-Carlo Analysis** по умолчанию и щелкните кнопкой **OK**, откроется окно **Monte-Carlo Results** (рис. 16.31). Нажмите кнопку **Summary Display overall results**, появится таблица результатов испытаний Монте-Карло. На рис. 16.32 приведен фрагмент таблицы, в которой отображены результаты 50 повторений метода Монте-Карло. Номер строки соответствует номеру повторения.

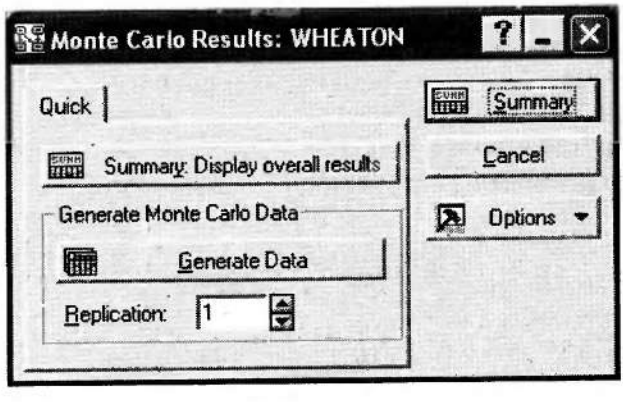


Рис. 16.31

	Monte Carlo Results (WHEATON)						
	SEED1	TERMCOD	DISCREF	RCOS	GRADIEN	NUM_ITEF	ICSF
1	1234567	0	0,0487	0,0000	0,0000	5	0,00000
2	741284129	0	0,0743	0,0001	0,0000	5	0,00000
3	249253794	0	0,0310	0,0001	0,0000	4	0,00000
4	193295108(	0	0,0453	0,0001	0,0000	4	0,00000
5	186489026(	0	0,0404	0,0000	0,0000	5	0,00000
6	110543690(	0	0,0545	0,0001	0,0000	5	0,00000
7	738223973	0	0,0973	0,0001	0,0023	7	0,00000
8	845987535	0	0,1278	0,0001	0,0003	12	0,00000
9	146087487(	0	0,0412	0,0000	0,0000	5	0,00000
10	520697505	0	0,0559	0,0001	0,0125	5	0,00000
11	196108812	0	0,0617	0,0001	0,0000	5	0,00000
12	1383124434	0	0,0872	0,0000	0,0000	8	0,00000
13	641502649	0	0,0931	0,0000	0,0000	10	0,00000

Рис. 16.32

В столбцах таблицы указаны результаты анализа для каждого повторения: *SEED1* (начальное значение), *TERMCODE* (код завершения), *DISCREP* (значение функции несогласия), *RCOS, GRADIENT* (максимальная по модулю компонента градиента), *NUM. ITER* (число итераций), *ICSF, ICS, RED PAR* (число избыточных параметров), *RED CON* (число избыточных ограничений), *BOUNDARY* (число активных ограничивающих неравенств), *CHI SQR* (критерий χ^2), *DF* (число степеней свободы), *P LEVEL* (*p-уровень*).

Анализ таблицы позволяет сделать вывод, что практически для всех реализаций метода Монте-Карло построена адекватная модель. Только для одного повторения (№ 8) неверна гипотеза о точном соответствии построенной модели исходным данным (*p level* меньше чем 0,05).

Таким образом, методом Монте-Карло получено дополнительное подтверждение адекватности построенной структурной модели.

Глава 17
Методы анализа выживаемости

17.1. Основные понятия

Методы анализа выживаемости *(Survival Analysis)* первоначально были развиты в медицинских, биологических исследованиях и страховании, но затем стали широко применяться в социальных и экономических науках, а также в промышленности в инженерных задачах (анализ надежности и времен отказов) [16, 10].

На Западе методы анализа выживаемости (длительностей до момента прекращения) находят все большее практическое применение.

Модели, построенные на основе длительности безработицы, используются для оценки эффективности проведения различных государственных программ по переквалификации и профессиональному обучению в Швеции.

Формирование государственной политики США в отношении табачной промышленности и здравоохранения частично опираются на исследования продолжительности употребления молодыми людьми табачных изделий.

Данные методы позволили провести статистический анализ, в котором были представлены результаты выявления наиболее значимых факторов, влияющих на продолжительность безработицы в Восточной Германии.

В России методы анализа длительности до момента прекращения широко пока еще не используются, а если и используются, то исключительно в области медицины.

Представьте, что изучается эффективность нового метода лечения или лекарственного препарата. Очевидно, наиболее важной и объективной характеристикой является средняя продолжительность жизни пациентов с момента поступления в клинику или средняя продолжительность ремиссии заболевания. Для описания средних времен жизни или ремиссии можно было бы использовать стандартные параметрические и непараметрические методы. Однако в анализируемых данных есть существенная особенность — могут найтись пациенты, которые в течение всего периода наблюдения выжили, а у некоторых из них заболевание все еще находится в стадии ремиссии. Также может образоваться группа больных, контакт с которыми был потерян до завершения эксперимента (например, их перевели в другие клиники). При использовании стандартных методов оценки среднего эту группу пациентов пришлось бы исключить, тем самым потеряв с трудом собранную важную информацию. К тому же большинство этих пациентов являются выжившими (выздоровевшими) в течение того времени, которое их наблюдали, что свидетельствует в пользу нового метода лечения (лекарственного препарата).

Такого рода информация, когда нет данных о наступлении интересующего нас события, называется неполной *(censored)*. Примеры неполной информации: «пациент *A* был жив, по крайней мере, 4 месяца до того, как был переведен в другую клинику и контакт с ним был потерян»; или, «пациент *A* жив до настоящего времени».

Если есть данные о наступлении интересующего нас события, то информация называется полной *(complete)*. Например: «пациент *A* прожил 5 лет после проведенного лечения»; «пациент *A* вновь заболел через 3 месяца после лечения».

Наблюдения, которые содержат неполную информацию, называются цензурированными наблюдениями. Цензурированные наблюдения типичны, когда наблюдаемая величина представляет время до наступления некоторого критического события, а продолжительность наблюдения ограничена по времени.

Цензурированные наблюдения встречаются во многих областях. Например, в социальных науках мы можем изучать «длительность» брака, интенсивность выбытия студентов из высшего учебного заведения (времени до выбытия), динамику численности работников в некоторых организациях и т.п. В рассмотренных примерах в конце периода наблюдения одни субъекты остаются состоящими в браке, другие продолжают учебу, а третьи продолжают работать в компании; таким образом, данные об этих субъектах являются цензурированными. Мы не можем дождаться того момента, когда все выбранные студенты покинут учебное заведение, а сотрудники компанию.

В экономике можно изучать «выживание» новых предприятий, или времена «жизни» продуктов, изделий производства. В задачах контроля качества типично изучение «выживания» элементов изделий «под нагрузкой» (анализ времен отказов).

В актуарной математике в качестве объекта исследований обычно используют таблицы смертности, содержащие данные о смертности лиц определенных категорий (например, мужчин старше 30 лет) за выбранные интервалы времени. В медицине, например, возникновение рецидива после проведенного лечения, наступление выздоровления или смерти.

Использование цензурированных наблюдений составляет специфику рассматриваемого метода — анализа выживаемости. В данном методе исследуются

вероятностные характеристики интервалов времени между последовательным возникновением критических событий. Такого рода исследования называются анализом длительностей до момента прекращения *(duration или time until failure)*, которые можно определить как интервалы времени между началом наблюдения за объектом и моментом прекращения *(failure)*, при котором объект перестает отвечать заданным для наблюдения свойствам. Цель исследований — определение условных вероятностей, связанных с длительностями до момента прекращения.

Таким образом, методы анализа выживаемости в основном применяются к тем же статистическим задачам, что и другие методы, однако их особенность в том, что они применяются к цензурированным или, как иногда говорят, неполным данным. Отметим также, что более часто, чем обычная функция распределения, в этих методах используется так называемая функция выживания, представляющая собой вероятность того, что объект проживет время больше t.

Построение таблиц времен жизни, подгонка распределения выживаемости, оценивание функции выживания с помощью процедуры Каплана-Мейера относятся к описательным методам исследования цензурированных данных. Некоторые из предложенных методов позволяют сравнивать выживаемость в двух и более группах.

Наконец, анализ выживаемости содержит регрессионные модели для оценивания зависимостей между многомерными непрерывными переменными со значениями, аналогичными временам жизни.

17.2. Описание модуля *Survival Analysis*. Таблицы выживаемости

Наиболее естественный способ описания функции выживания в выборке — построение таблиц времен жизни. Техника таблиц времен жизни - один из старейших методов анализа данных о выживаемости (времен отказов). Такую таблицу можно рассматривать как «расширенную» таблицу частот. Область возможных времен наступления критических событий (смертей, отказов и др.) разбивается на некоторое число интервалов. Для каждого интервала вычисляется число и доля объектов, которые в начале рассматриваемого интервала были «живы», число и доля объектов, которые «умерли» в данном интервале, а также число и доля объектов изъятых или цензурированых в каждом интервале. В отличие от таблицы частот, в таблице времен жизни учтены как полные, так и неполные наблюдения.

Из библиотеки **Example → Datasets** откройте файл данных **Heart**. В файле приведены данные о выживаемости 65 пациентов, которым была проведена трансплантация сердца. Первые шесть переменных представляют собой даты, а именно дату трансплантации (месяц/день/год) и дату, когда соответствующий пациент умер или был исключен из наблюдения (не было возможности связаться с ним для получения информации о его здоровье). Переменная *CENSORED* — это цензурированная индикаторная переменная, содержащая коды, определяющие либо конкретное наблюдение за пациентом, либо цензурированное наблюдение (0 — *COMPLETE;* 1 —

CENSORED). Переменная *AGE* означает возраст пациентов. Переменные *ANTIGEN, MISMATCH* содержат специальную медицинскую информацию об антигенной несовместимости и несовместимости тканей. Переменная *HOSPITAL* (условная переменная) определяет то, к какому из трех госпиталей относится определенный пациент. Фрагмент файла приведен на рис. 17.1.

Для запуска модуля **Survival Analysis** в меню **Statistics** щелкните по **Advanced Linear/Nonlinear Models** и выберите команду **Survival Analysis**. Стартовое окно модуля **Survival and Failure Time Analysis** будет иметь вид, изображенный на рис. 17.2.

В стартовом окне представлены основные процедуры модуля: **Life tables & Distributions** (таблицы времен жизни и распределения); **Kaplan & Meier product-limit method** (метод множительных оценок Каплана-Мейера); **Comparing two samples** (сравнение двух выборок); **Comparing multiple samples** (сравнение нескольких выборок); **Regression models** (регрессионные модели), **Time-dependent covariates** (зависящие от времени коварианты).

Heart transplant data from Crowley and Hu, stratified

	1 MONTH_1	2 DAY_1	3 YEAR_1	4 MONTH_2	5 DAY_2	6 YEAR_2	7 CENSORED	8 AGE	9 ANTIGEN	10 MISM
1	JANUARY	6	68	JANUARY	21	68	CENSORED	54	0	1,11
2	MAY	2	68	MAY	5	68	CENSORED	40	0	1,66
3	AUGUST	31	68	MAY	17	70	COMPLETE	51	0	1,32
4	AUGUST	22	68	OCTOBER	7	68	COMPLETE	42	0	0,61
5	SEPTEMB	9	68	JANUARY	14	69	CENSORED	48	0	0,36
6	OCTOBEF	5	68	DECEMBE	8	68	COMPLETE	54	0	1,89
7	OCTOBEF	26	68	JULY	7	72	COMPLETE	54	0	0,87
8	NOVEMBE	22	68	AUGUST	29	69	COMPLETE	49	0	1,12

Рис. 17.1

Выделите процедуру **Life tables & Distributions** и нажмите **OK**. Откроется окно диалога **Life Table & Distribution of Survival Times** (таблицы и распределение времен жизни). Диалог имеет две вкладки: **Raw data** (исходные данные) и **Table of survival times** (таблица времен жизни). Первая вкладка соответствует случаю, когда в качестве входных данных используются необработанные данные — обычная таблица программы *STATISTICA* (строки — наблюдения, столбцы — переменные), вторая вкладка — случаю, когда в качестве входных данных анализа выбрана ранее вычисленная таблица времен жизни.

Выберите вкладку **Raw data.** Рассмотрим функциональное назначение кнопок и основных установок этой вкладки (рис. 17.3).

Кнопка **Variables (survival times & censoring indicator)** (переменные, времена жизни и индикатор цензурирования) вызывает стандартное диалоговое окно **Life Table & Distribution of Survival Times** выбора переменных, содержащих времена жизни, и цензурирующей переменной — индикатора цензурирования. Имеется два способа задания времен жизни. Можно выбрать одну переменную, содержащую

времена жизни (например, число прожитых недель), либо выбрать шесть переменных с датами. В частности, эти переменные должны содержать месяц (от 1 до 12), день (от 1 до 31) и год начала наблюдения (например, поступления пациента в больницу) и месяц, день и год завершения наблюдения (в связи со смертью или цензурированием, например, переводом в другую больницу). При обработке данных модуль **Survival Analysis** вычисляет число дней между датами (поступления и завершения) и далее проводит анализ на основе этой величины (продолжительности). Заметим, что если значение года меньше 100, *STATISTICA* воспринимает это как год XX столетия; например, значение 88 понимается как 1988 [6].

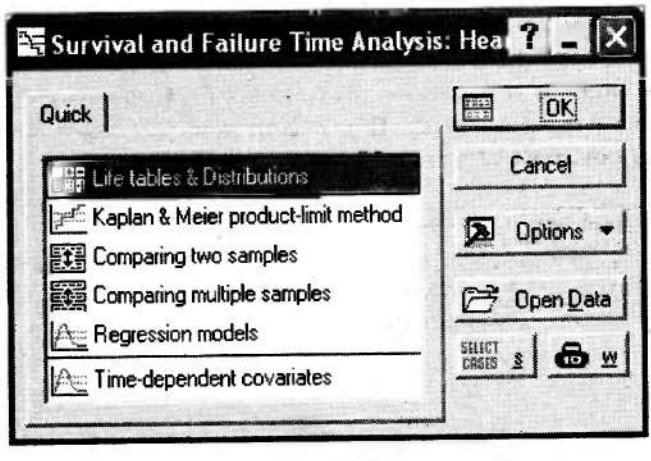


Рис. 17.2

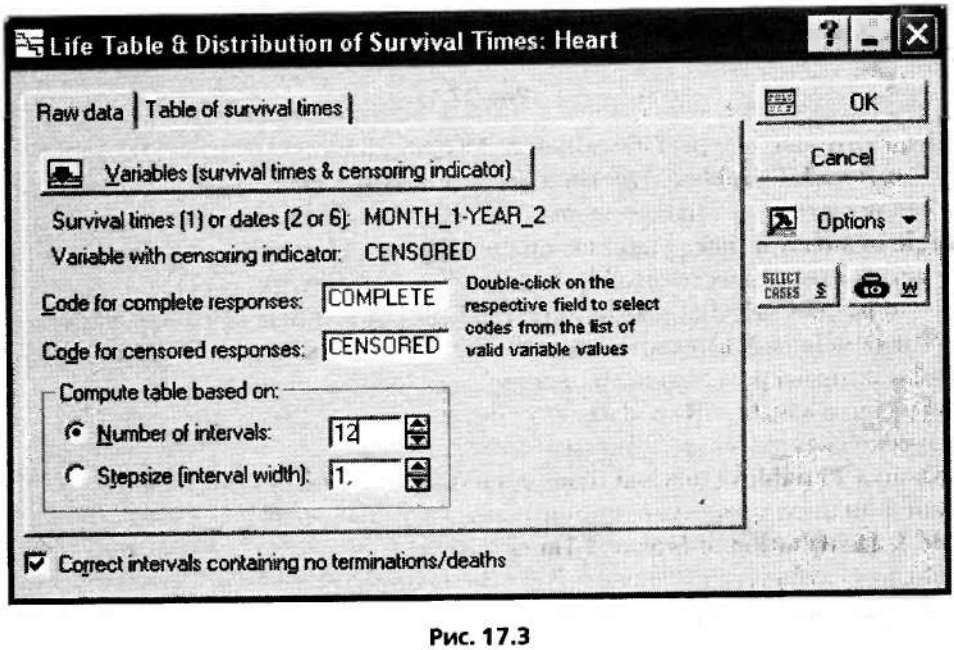


Рис. 17.3

Поле **Code for complete responses** (код для полных наблюдений) предназначено для введения кодов или текстовых значений для цензурирующей индикаторной переменной, однозначно идентифицирующих завершенные (нецензурированные) наблюдения.

В поле **Code for censored responses** (код для цензурированных наблюдений) вводятся коды или текстовые значения для цензурирующей индикаторной переменной, однозначно идентифицирующие незавершенные (цензурированные) наблюдения. Чтобы просмотреть все коды соответствующей переменной, надо дважды щелкнуть по этому полю ввода.

В рамке **Compute table based on** (построить таблицы исходя из) производятся установки для выбора числа интервалов *(Number of intervals)* или шага (длины интервала) *(Stepsize (interval width))*. Процедура подгонки теоретического распределения к данным невозможна при наличии интервалов, не содержащих ни смертей (отказов), ни изъятых наблюдений. Если необходимо сделать подгонку, то поставьте флажок возле *Correct intervals containing no terminations/deaths* (скорректировать интервалы, не содержащие отказов/смертей). Другой способ — уменьшить число интервалов таким образом, чтобы за счет увеличения длин интервалов не было интервалов, не содержащих отказов или изъятых наблюдений. Если таблица времен жизни используется только в описательных целях и не предполагается подгонка распределения, то корректировку интервалов делать не нужно.

Нажмите кнопку **Variables (survival times & censoring indicator)**. В открывшемся окне **Select survival times /dates and censoring indicator** (рис. 17.4) выберите первые 6 переменных в левом поле и *CENSORED* в правом поле. Нажмите **OK** для того, чтобы вернуться в диалоговое окно **Life Table & Distribution of Survival Times**.

Щелкните два раза в поле **Code for complete responses**, появится диалоговое окно **Variable 7**. Выберите код для полных наблюдений — *COMPLETE* и нажмите **OK**. Точно так же двойным щелчком на поле **Code for censored responses** выберите код для неполных наблюдений — *CENSORED*. Нажмите **OK**. Остальные установки сделайте по умолчанию и нажмите **OK**. Откроется окно с результатами **Life Table & Survival Time Distribution Results** (рис. 17.5).

В информационной части окна указаны:

1) в строке *Variable* — переменной величиной является количество дней, вычисленное по исходным данным;
2) в строке *Variable with censoring indicator* — индикатором цензурирования является переменная *censored;*
3) в строке *Total number of valid observations* — число допустимых наблюдений равно 65;
4) в строке *uncensored количество* — количество нецензурированных объектов равно 29;
5) в строке *censored* — количество цензурированных объектов равно 36. В скобках приведены доли от общего числа объектов.

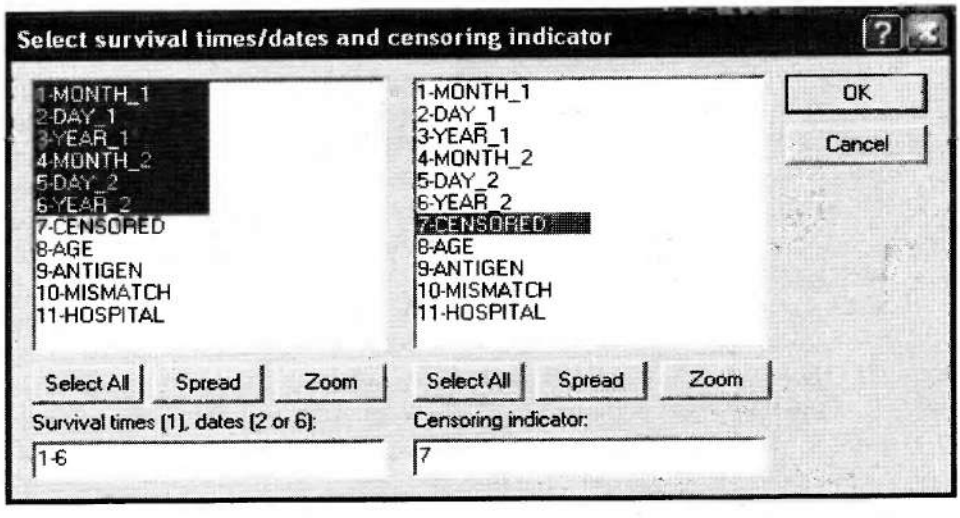


Рис. 17.4

Перейдите на вкладку **Advanced**. Нажмите кнопку **Summary: Life table** (итоговая таблица времен жизни), чтобы появилась полная развернутая таблица результатов (на рис. 17.6 приведен фрагмент таблицы).

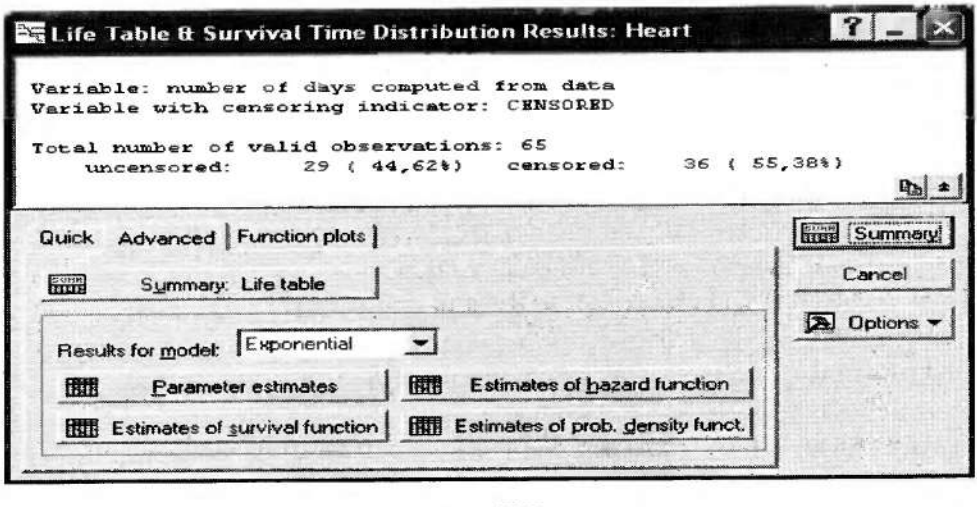


Рис. 17.5

В дополнение к стандартным описательным статистикам эта электронная таблица будет содержать оценки кумулятивной функции выживаемости, плотности распределения и медианы условного распределения продолжительности жизни в каждом интервале. Вычисляются также стандартные ошибки этих оценок. Опишем подробнее столбцы таблицы.

Interval Start — начало интервала.

Mid Point — средняя точка интервала.

Interval	Life Table (Heart) Log-Likelihood for data: -68,6809								
	Interval Start	Mid Point	Interval Width	Number Entering	Number Withdrwn	Number Exposed	Number Dying	Proportn Dead	Proportn Survivng
Intno.1	0,00	80,68	161,36	65	14	58	19	0,3276	0,6724
Intno.2	161,36	242,05	161,36	32	4	30	4	0,1333	0,8667
Intno.3	322,73	403,41	161,36	24	4	22	0	0,0227	0,9773
Intno.4	484,09	564,77	161,36	20	4	18	1	0,0556	0,9444
Intno.5	645,45	726,14	161,36	15	1	14,5	1	0,0690	0,9310
Intno.6	806,82	887,50	161,36	13	3	11,5	1	0,0870	0,9130
Intno.7	968,18	1048,86	161,36	9	1	8,5	2	0,2353	0,7647
Intno.8	1129,55	1210,23	161,36	6	1	5,5	0	0,0909	0,9091
Intno.9	1290,91	1371,59	161,36	5	1	4,5	1	0,2222	0,7778
Intno.10	1452,27	1532,95	161,36	3	2	2	0	0,2500	0,7500
Intno.11	1613,64	1694,32	161,36	1	0	1	0	0,5000	0,5000
Intno.12	1775,00			1	1	0,5	0	1,0000	0,0000

Рис. 17.6

Interval Width — длина интервала.

Number Entering — число вначале. Это число объектов, которые были «живы» в начале рассматриваемого временного интервала.

Number Exposed — число изучаемых объектов, т.е. число объектов, которые были «живы» в начале рассматриваемого временного интервала, минус половина числа изъятых или цензурированных объектов.

Number Dying — число умерших, это число объектов, умерших (отказавших) на данном временном интервале. В файле исходных данных этим объектам в столбце *CENSORED* соответствует текстовое значение *COMPLETE*.

Proportion Dead — доля умерших, это отношение числа объектов, умерших (отказавших) в соответствующем интервале, к числу объектов, изучаемых на этом интервале.

Proportion Surviving — доля выживших. Она равна единице минус доля умерших.

Cumulative proportion surviving — кумулятивная доля выживших. Это кумулятивная доля выживших к началу соответствующего временного интервала. Поскольку вероятности выживания считаются независимыми на разных интервалах, эта доля равна произведению долей выживших объектов по всем предыдущим интервалам. Полученная доля как функция от времени называется также выживаемостью, или функцией выживания (точнее, это оценка функции выживания). Функция выживания является оценкой вероятности того, что объект «переживет» данный интервал.

Probability density — плотность вероятностей. Это оценка вероятности смерти (отказа) в соответствующем интервале, определяемая таким образом:

$$f_i = (P_i - P_{i+1}) / h_i,$$

где f_i — оценка плотности вероятности смерти (отказа) в i-м интервале; P_i, P_{i+1} — кумулятивные доли выживших объектов (функция выживания) соответственно к началу i-го и $i+1$-го интервалов; h_i — ширина i-го интервала, $P_i - P_{i+1}$ — доля умерших.

Hazard rate (функция интенсивности отказов или функция мгновенного риска). Функция интенсивности определяется как оценка вероятности того, что объект, выживший к началу соответствующего интервала, умрет (откажет) в течение этого интервала. Оценка функции интенсивности вычисляется как число смертей (отказов), приходящихся на единицу времени соответствующего интервала, деленное на среднее число объектов, доживших до момента времени, находящегося в середине интервала. Заметим, что именно функция риска используется для прогностических целей [10].

Std. Err. Cum. Surv — стандартная ошибка кумулятивной доли выживших.

Std. Err. Prob. Den — стандартная ошибка плотности вероятностей.

Std. Err. Haz. Rate — стандартная ошибка функции интенсивности.

Исследователя интересует функции выживания и функции риска. Однако реально программа вычисляет лишь оценки этих функций. Естественно, доверять надо тем оценкам, у которых малы стандартные ошибки. Например, не следует доверять тем оценкам, погрешность которых имеет тот же порядок, что и сама оценка.

Median life expected — медиана ожидаемого времени жизни. Это точка на временной оси, в которой кумулятивная функция выживания равна 0,5. Другие процентили (например, 25- и 75-процентиль или квартили) кумулятивной функции выживания вычисляются по такому же принципу. Отметим, что 50-процентиль (медиана) кумулятивной функции выживания обычно не совпадает с точкой выживания 50% выборочных наблюдений. Совпадение происходит только тогда, когда за прошедшее к этому моменту время не было цензурированных наблюдений.

Std. Err. Life exp — стандартная ошибка медианы ожидаемого времени жизни.

Заметим, чтобы получить надежные оценки трех основных функций (функции выживания, плотности вероятности и функции интенсивности) и их стандартных ошибок на каждом временном интервале, рекомендуется использовать не менее 30 наблюдений.

Results for model. Это поле списка используется для подгонки к данным одного из четырех общих семейств теоретических распределений времени жизни: *Exponential* (экспоненциального), *Linear Hazard* (семейства с линейной интенсивностью), *Gompertz* (Гомпертца) и *Weibull* (Вейбулла). При помощи несложных преобразований все 4 семейства сводятся к линейному уравнению вида $y = ax + b$.

Кнопка **Parameter Estimates** — оценки параметров выдает электронную таблицу результатов с оценками параметров для соответствующего теоретического семейства распределений продолжительности жизни. С помощью поля списка **Results for model** может быть выбрано теоретическое распределение из каждого семейства, наиболее подходящее к данным. Модуль **Survival Analysis** подгоняет теоретическое распределение с помощью обычного метода наименьших квадратов, т.е. с весами, равными 1 — *Weight1*, и двух методов взвешенных квадратов *Weight2* и *Weight3* (поэтому в таблице выводится значение трех весов):

$$WSS = \sum w_i (y_i - a - bx_i)^2,$$

где $w_i = 1$ (невзвешенные наименьшие квадраты);
$w_i = 1/v_i$ (взвешенные наименьшие квадраты);
$w_i = n_i * h_i$ (взвешенные наименьшие квадраты);
v_i - (оцененная) дисперсия интенсивности,
h_i *и* n_i — ширина i-го интервала и число объектов в начале i-го интервала соответственно.

Данная таблица позволяет при заданном в поле **Results for model** теоретическом законе распределения определить, какой метод наименьших квадратов дает наилучшее соответствие теоретической функции распределения эмпирической. В поле **Results for model** выберите распределение Вейбулла и нажмите кнопку **Parameter Estimates**. Из таблицы на рис.17.7 следует, что наилучшая подгонка соответствует весу 3 (*Weight3*), так как этому весу соответствует наименьшее значение критерия (7,757) и уровень значимости p критерия (0,55881) значительно больше 0,05.

	Parameter Estimates, Model: Weibull (Heart) Note: Weights: 1=1., 2=1./V, 3=N(I)*H(I)							
Estimatn Method	Lambda	Variance Lambda	Std.Err. Lambda	Covarnce Gam-Lamd	Log-Likelihd.	Chi-Sqr.	df	p
Weight 1	0,00031	0,00000	0,00057	-0,00015	-84,342	31,3240	9	0,00026
Weight 2	0,01600	0,00032	0,01795	-0,00292	-75,434	13,5075	9	0,14100
Weight 3	0,05110	0,00521	0,07223	-0,01370	-72,559	7,7570	9	0,55881

Рис. 17.7

Нажатие любой из кнопок **Estimates of survival function, Estimates of hazard function, Estimates prob. density funct** приводит к построению отдельных таблиц результатов с соответствующими оценками функций выживания, риска, плотности вероятности. Исходные значения этих функций выводятся в общей таблице результатов *Life table*. В информационной части таблиц выводятся формулы вычисления весов. Конкретная модель теоретического распределения выбирается в поле **Results for model**.

На вкладке **Function plots** (рис. 17.8) при помощи кнопок **Plots of hazard function, Plots of probability density function, Plots of survival function** можно посмотреть графики соответственно функций риска, плотности вероятностей и выживаемости.

На графиках (рис. 17.9–17.11) сплошной линией изображено исходное (эмпирическое) распределение, пунктирными цветными линиями — теоретические распределения, соответствующие различным весам. Из графика функции риска следует, что вероятность смерти резко уменьшается с 1-го дня после операции до 322 дня, затем эта вероятность незначительно возрастает до 806 дня и резко возрастает до 968 дня. Затем наблюдается спад вероятности смерти до 1129 дня, после чего вероятность смерти вновь резко возрастает. Наилучшее приближение теоретического распределения эмпирическому распределению дает метод наименьших квадратов с взвешенным весом 3. По-видимому, для большего соответствия распределений необходимо использовать кусочную аппроксимацию — отдельно для левой, средней и правой частей графика.

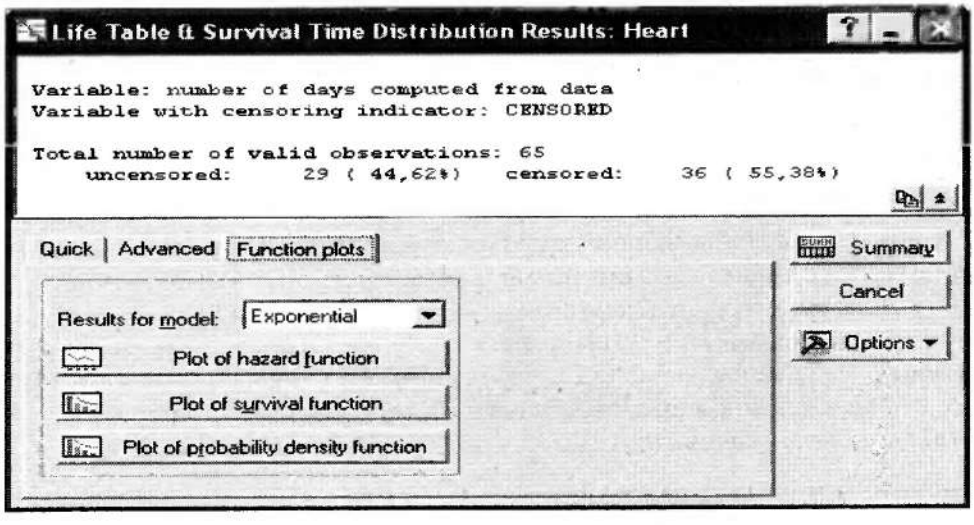


Рис. 17.8

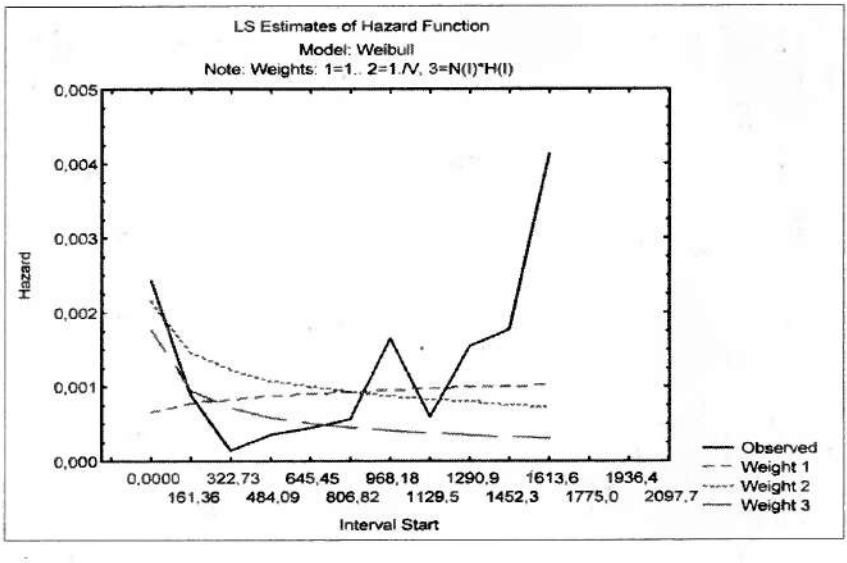


Рис. 17.9

Из гистограммы, изображенной на рис. 17.11 и таблицы *Life table* (рис. 17.6) следует, что через 161 день после операции кумулятивная доля выживших составила 67,2%, а через 322 дня – уже 58,3%, далее уменьшение доли выживших продолжается, но замедляется темп. Резкий спад доли выживших наблюдается через 1129 дней и составляет 34,96%. К концу рассматриваемого периода (через 1775 дней) доля выживших составляет всего 9,3%. Из рис. 17.10 следует, что наибольшая вероятность смерти больных в первые 161 день после операции, наименьшая – с 322 по 484 день.

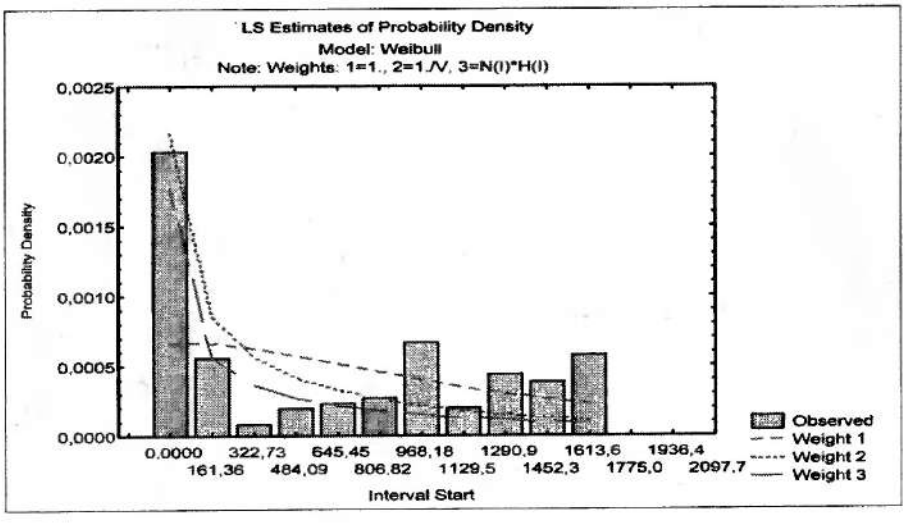


Рис. 17.10

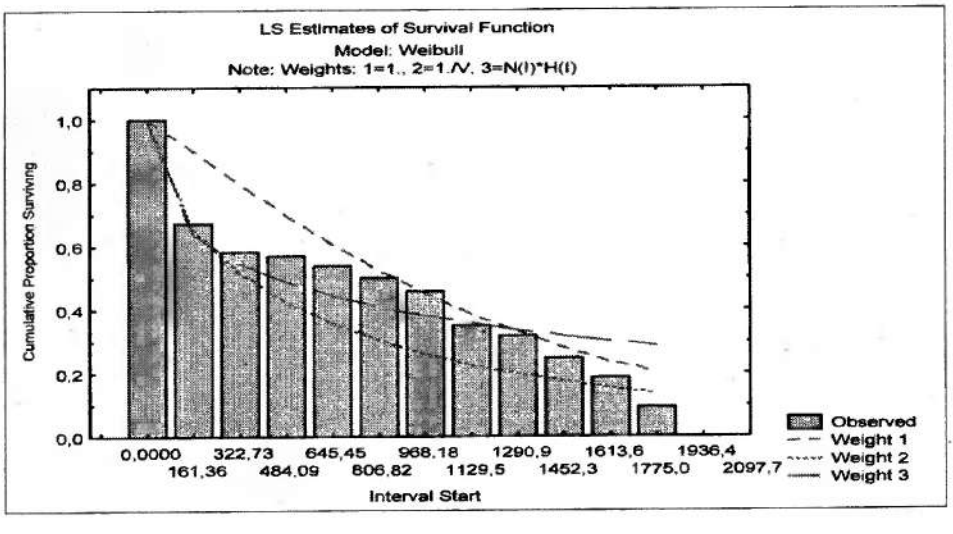


Рис. 17.11

К сожалению, в модуле не предусмотрена возможность прогноза функций риска, выживаемости и плотности вероятностей при помощи строящихся теоретических распределений, так как программа не выводит значения параметров этих распределений.

17.3. Метод множительных оценок Каплана–Мейера

Метод таблиц распределения времен жизни — один из самых старых и наиболее используемых способов оценки функции выживания (а также функции риска

и плотности вероятности функции). Однако точные оценки методом таблиц будут зависеть от выбора числа и ширины интервалов времени выживания. Метод множительных оценок Каплана-Мейера производит оценку функции выживания для цензурированных данных, используя непосредственно время выживания, без обработки (группировки данных).

Предположим, что задан файл, в котором записаны в хронологическом порядке отдельные события. Создадим новый файл данных, упорядочив наблюдения по количеству дней, проведенных объектом под наблюдением (для полных данных — это число дней до отказа (смерти), для цензурированных — это число дней до потери контакта с объектом). Каплан и Мейер предложили следующую оценку функции выживания [6]:

$$S(t) = \Pi_{j-1}^{t}[(n-j)/(n-j+1)]^{\delta_j}.$$

В этом выражении $S(t)$ — оценка функции выживаемости, n — общее число событий, Π — произведение (геометрическая сумма) по всем наблюдениям, завершившимся к моменту t; δ_j равно 1, если j-е наблюдение нецензурированное (законченное), и равно 0, если это наблюдение потеряно (цензурированное). Данная оценка функции выживаемости называется еще множительной оценкой. Заметим, что j — это не номер наблюдения в исходном файле данных, а номер наблюдения в новом файле, где произведено упорядочивание по количеству проведенных под наблюдением дней. Преимущество метода Каплана-Мейера (по сравнению с методом таблиц жизни) состоит в том, что оценки не зависят от разбиения времени наблюдения на интервалы, т.е. от группировки.

Метод множительных оценок и метод таблиц времен жизни приводят, по существу, к одинаковым результатам, если временные интервалы содержат минимум по одному наблюдению.

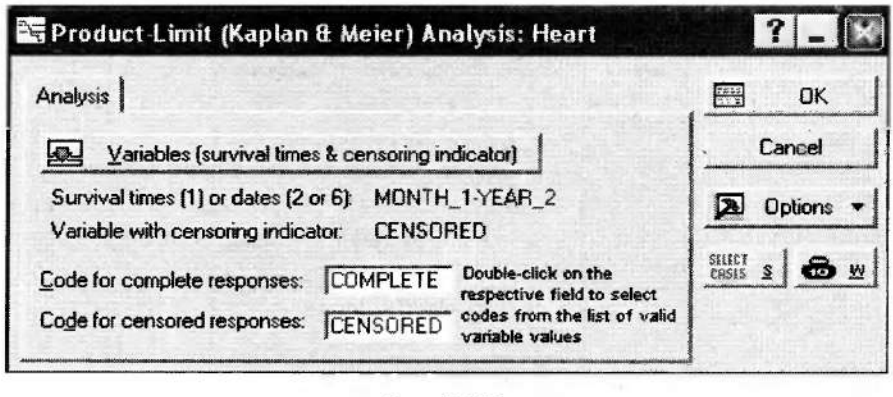


Рис. 17.12

В диалоге **Survival and Failure Time Analysis** модуля **Survival Analysis Survival and Failure Time Analysis** выделите процедуру **Kaplan & Meier product-limit method** и нажмите **OK**. В появившемся окне диалога (рис. 17.12)

укажите имена и коды переменных так, как это было сделано в диалоге **Life Table & Distribution of Survival Times**.. Нажмите **OK**, в открывшемся окне результатов анализа **Product-limit (Kaplan-Meier) Analysis Results** (рис. 17.13) нажмите кнопку **Summary: Product-limit survival analysis** для того, чтобы посмотреть оценки функции выживания. На рис. 17.14 приведен фрагмент таблицы.

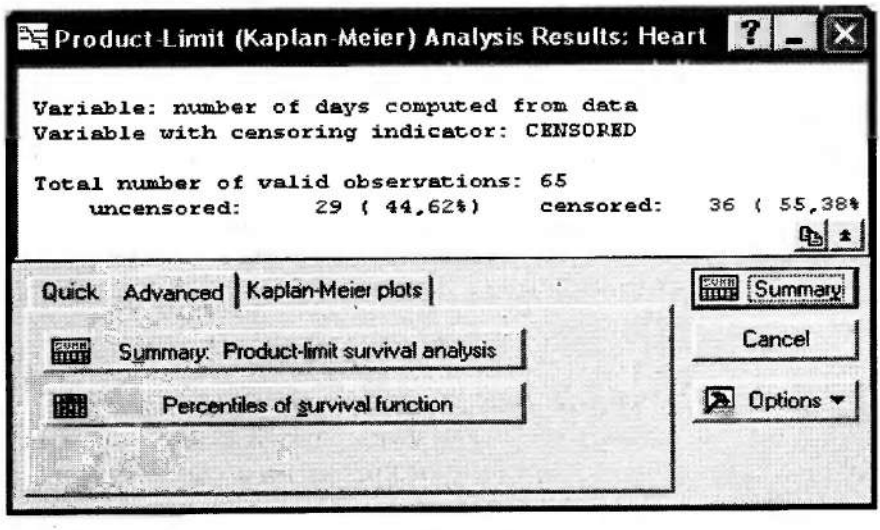


Рис. 17.13

Обратите внимание, что цензурированные наблюдения в таблице помечены знаком «+». Данные упорядочены по количеству дней, прожитых в больнице. В первом столбце указаны номера больных в исходном файле данных, во втором — время, проведенное пациентами в больнице.

	Kaplan-Meier (Product-limit) analysi Note: Censored cases are marked v		
Case Number	Time	Cumulatv Survival	Standard Error
23+	0,000		
16+	1,000		
65+	1,000		
2+	3,000		
10	10,000	0,983607	0,016259
46+	12,000		
64+	13,000		
1+	15,000		
9+	23,000		
42	25,000	0,966042	0,023622
58+	26,000		
49	29,000	0,948152	0,029183
59+	30,000		

Рис. 17.14

Прокрутив таблицу, можно увидеть, что эта величина изменяется от 0 до 1775 дней. Меньше всего провел в больнице больной с порядковым номером 23, так как в первый же день после операции покинул ее. Больше всего дней провел в больнице пациент с номером 15, через 1775 дней после операции он также покинул больницу. Если в столбце *Cumulative Survival* отсутствует значение — значит больной выбыл из больницы (цензурированное наблюдение), если есть значение — значит больной умер, прожив количество дней, указанное в столбце *Time*. Так, пациенты под номерами 23 и 15 являются цензурированными наблюдениями. Первый умерший после операции больной под номером 10, он прожил после операции 10 дней. При этом значение в столбце *Cumulative Survival* означает вероятность того, что произвольный больной проживет больше дней, чем указано в соответствующей строке столбца *Time*. Эта вероятность просчитана по формуле Каплана-Мейера:

$$S(10) = (65 - 5)/(65 - 5 + 1) = 0{,}98360.$$

Стандартные ошибки оценок функции выживания малы, что свидетельствует о достоверности оценок.

Чтобы увидеть графическое изображение оцененной функции выживания, нажмите кнопку **Survival times vs. cum. proportion surviving**. Из графика, изображенного на рис. 17.15, видно, что значение функции выживания резко падает в течение первых 100 дней, после того как была произведена трансплантация сердца. Начиная с этого времени, функция убывает менее резко. Поэтому можно сделать вывод, что первые 100 дней после трансплантации сердца — наиболее критические (опасные для жизни). Для удобства интерпретации результатов полные наблюдения помечены точками, неполные — крестиками.

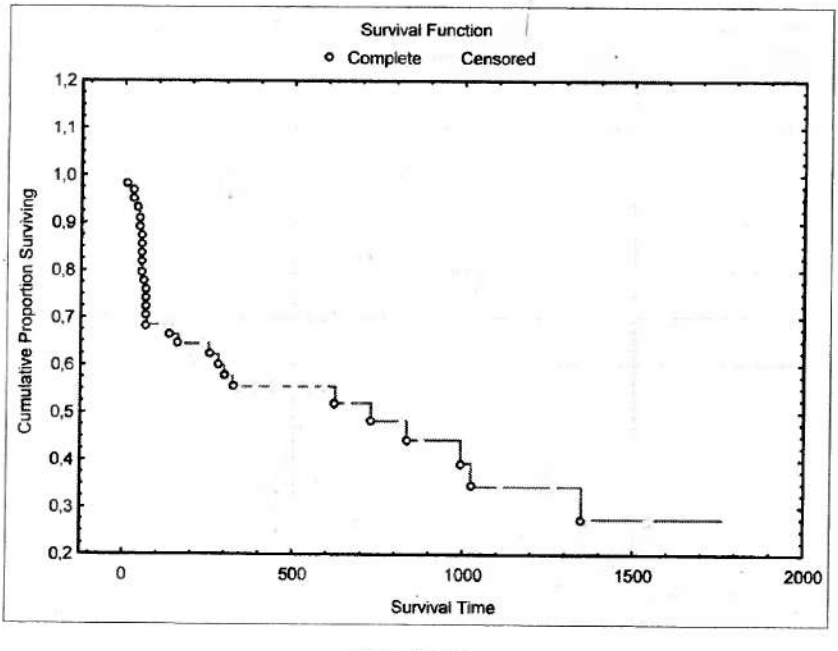


Рис. 17.15

Нажмите кнопку **Percentiles of survival function** с целью просчитать и вывести процентили в таблице (рис. 17.16). Указанные процентные отношения снова отражают природу распределения. Так, 25% (нижняя квартиль) — пациенты умирающие в течение первых 64 дней после трансплантации сердца, 50% (медиана) — пациенты живущие дольше (менее) 679 дней.

Percentiles	Percentile the Surviv Survival Time
25'th percentile (lower quartile)	63,5140
50'th percentile (median)	679,1255
75'th percentile (upper quartile)	

Рис. 17.16

17.4. Сравнение выживаемости в двух группах

Определенный интерес представляет сравнение выживаемости в нескольких группах наблюдений. Для этой цели в модуле **Survival Analysis** предусмотрены две процедуры: **Comparing two samples** — для сравнения выживаемости в двух группах и **Comparing multiple samples** — для сравнения выживаемости более чем в двух группах (рис. 17.2).

Для сравнения выживаемости имеется пять различных (в основном непараметрических) критериев: обобщенный Геханом критерий Вилкоксона, *F-критерий* Кокса, критерий Кокса-Ментеля, логарифмический ранговый критерий, критерий Вилкоксона-Пето. Большинство этих критериев вычисляют соответствующие *z-значения* стандартного нормального распределения; эти *z-значения* могут быть использованы для статистической проверки любых различий между группами. Однако критерии дают надежные результаты лишь при достаточно больших объемах выборок. При малых объемах выборок эти критерии не столь надежны [10]. Поэтому желательно применение числовых критериев сравнения сопровождать визуализацией функций времени жизни.

Не существует твердо установленных рекомендаций по применению определенных критериев.

Однако известно, что *F-критерий* Кокса обычно более мощный, чем критерий Вилкоксона-Гехана, если:

- выборочные объемы малы (т.е. объем группы n меньше 50);
- выборки извлекаются из экспоненциального распределения или распределения Вейбулла;
- нет цензурированных наблюдений.

Показано [10], что критерий Кокса-Ментела и логарифмически ранговый критерий являются более мощными, если выборки извлечены из экспоненциального распределения или распределения Вейбулла; при этих условиях между этими критериями почти нет различия.

Если сравниваются группы, то важно проверить доли цензурированных наблюдений в каждой. В частности, в медицинских исследованиях степень цензурирования может зависеть, например, от различий в методе лечения: пациенты, которым стало много лучше или хуже, с большой вероятностью теряются из наблюдения. Различие в степени цензурирования может привести к смещению в статистических выводах.

Операции по пересадке сердца были осуществлены в трех клиниках: *Hillview, St_Andrea s., Biner.* Сравним функции выживаемости в двух клиниках — *Hillview, St_Andrea s.* В стартовом окне **Survival and Failure Time Analysis** два раза щелкните по строке **Comparing two samples** (сравнение двух групп), появится окно **Comparing Survival in Two Groups**. Нажмите кнопку **Variable** и произведите выбор всех переменных, как было сделано ранее, выбрав в качестве группирующей переменную *Hospital.* Щелкните два раза в поле **Code for first group**, в появившемся окне выберите первую группу — *Hillview*, нажмите **OK**. Так же в поле **Code for second group** выберите вторую группу — *St_Andrea s.* В диалоговом окне **Comparing Survival in Two Groups** (рис. 17.17) нажмите **OK**, откроется окно результатов **Two-sample Tests Results** (результаты двух выборочных критериев, рис. 17.18). На вкладке **Two-sample Tests** названия первых 5 кнопок соответствуют названиям используемых критериев [6].

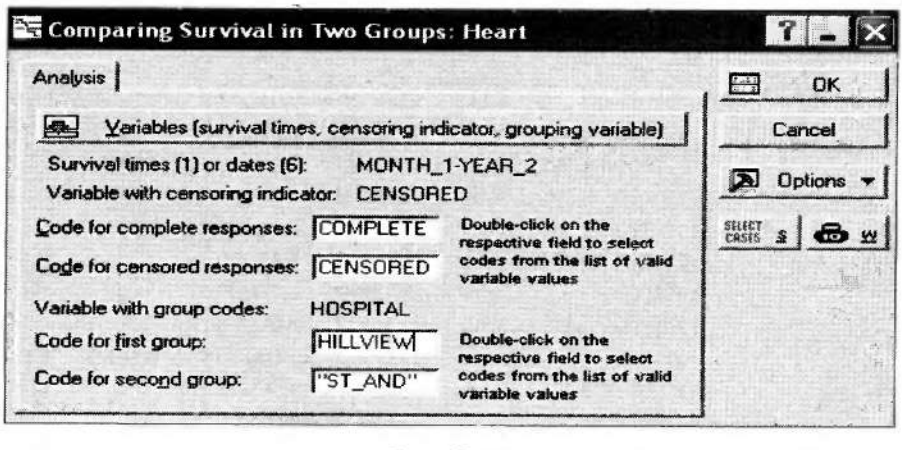


Рис. 17.17

Gehan's Wilcoxon test. Процедура строит таблицу результатов (рис. 17.19) со значениями обобщенного Геханом критерия Вилкоксона. Кроме статистики критерия Гехана-Вилкоксона таблица содержит также список всех упорядоченных по возрастанию времен жизни наблюдений (столбец 1); группу, которой принадлежит наблюдение; число полных наблюдений плюс 1, меньших, чем данное (столбец *R1*); для полных наблюдений — число наблюдений, больших, чем данное (столбец *R2*). Для цензурированных наблюдений в столбце 2 стоят единицы. Отметим, что цензурированные наблюдения помечены знаком плюс (+).

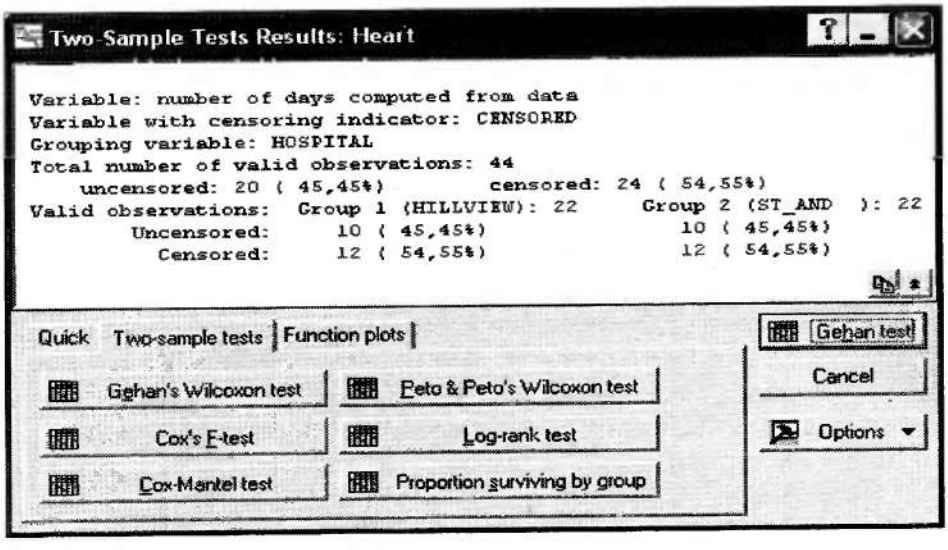


Рис. 17.18

Survival Time	Gehan's Wilcoxon Test (Heart) WW = -2,000 Sum = 12456, Var = Test statistic = -,026573 p = ,9788 Group No	R1	R2
0,0000+	ST_AND	1,00000	1,00000
1,0000+	ST_AND	1,00000	1,00000
3,0000+	HILLVIEW	1,00000	1,00000
10,000	HILLVIEW	1,00000	41,00000
12,000+	HILLVIEW	2,00000	1,00000
13,000+	HILLVIEW	2,00000	1,00000
15,000+	HILLVIEW	2,00000	1,00000
23,000+	HILLVIEW	2,00000	1,00000
25,000	ST_AND	2,00000	36,00000
26,000+	ST_AND	3,00000	1,00000
29,000	ST_AND	3,00000	34,00000
30,000+	ST_AND	4,00000	1,00000

Рис. 17.19

Cox's F-testF. Процедура строит таблицу результатов (рис. 17.20) со значениями *F-критерия* Кокса. Кроме значений *F-статистики* таблица содержит *Distinct Failures* – список всех различных времен отказов (смертей, столбец 1), число объектов, подвергнутых риску в соответствующие моменты отказа (столбец *R(I)*), кратности отказов в каждый момент наступления отказов (столбец *M(I)*), отношение этих величин (*M/R*) и оценки Каплана-Мейера функции выживаемости (столбец 5).

Distinct Failures	Cox's F-Test (Heart) T1 = 10,52661 T2 = 10,47339 F(20, 20) = 1,005081 p = ,49553 R(I)	M(I)	M/R	Kap/Meir Estimate
10,000	41,00000	1,000000	0,024390	1,000000
25,000	36,00000	1,000000	0,027778	0,975610
29,000	34,00000	1,000000	0,029412	0,948510
39,000	32,00000	1,000000	0,031250	0,920612
46,000	30,00000	1,000000	0,033333	0,891843
47,000	29,00000	1,000000	0,034483	0,862115
50,000	28,00000	1,000000	0,035714	0,832387
51,000	27,00000	2,000000	0,074074	0,802659
54,000	25,00000	1,000000	0,040000	0,743202
60,000	24,00000	1,000000	0,041667	0,713474
64,000	23,00000	1,000000	0,043478	0,683746
65,000	22,00000	1,000000	0,045455	0,654018

Рис. 17.20

Cox-Mantel test. Процедура открывает электронную таблицу (рис. 17.21) с результатами применения критерия Кокса-Ментела. Кроме статистики Кокса-Ментела таблица выдает список множеств риска в каждый момент отказа (число объектов, которые были живыми перед моментом соответствующего отказа, столбец 1) и долю объектов в группе 2, принадлежащих соответствующему множеству риска (столбец 2).

Peto & Peto's Wilcoxon test. Процедура строит таблицу результатов (рис. 17.22) со значениями обобщенного Пето критерия Вилкоксона. В дополнение к статистике критерия таблица выдает список всех наблюдений (времен жизни; при этом цензурированные наблюдения отмечаются знаком +, столбец 1) и вклады, которые используются при вычислении статистики этого критерия.

Log-rank test. Процедура строит таблицу результатов (рис. 17.23) для лог-рангового критерия. Вместе со статистикой критерия таблица содержит список всех наблюдений (времена жизни; при этом цензурированные наблюдения помечены знаком +, столбец 1); группу, которой принадлежит каждое наблюдение (столбец 2), значения вкладов, используемых при вычислении критерия (столбец 3).

Proportion surviving by. Процедура открывает таблицу результатов (рис. 17.24) со сравнительными данными. Для групп 1 и 2 приведены количество наблюдений в начале интервала (*№ Enter*), число цензурированных (*№ Cnsrd*), число умерших (*№ Dying*), процент выживших (*%Srwng*), кумулятивный процент выживших (*%ComSrwng*).

Survival Time	Peto & P WW = ,0 Test stati Score
0,0000+	0,00000
1,0000+	0,00000
3,0000+	0,00000
10,000	0,97561
12,000+	-0,02439
13,000+	-0,02439
15,000+	-0,02439
23,000+	-0,02439
25,000	0,92411
26,000+	-0,05149
29,000	0,86912
30,000+	-0,07938

Рис. 17.21

Survival Time	Peto & F WW = ,0 Test stat Score
0,0000+	0,0000
1,0000+	0,0000
3,0000+	0,0000
10,000	0,9756
12,000+	-0,0243
13,000+	-0,0243
15,000+	-0,0243
23,000+	-0,0243
25,000	0,9241
26,000+	-0,0514
29,000	0,8691
30,000+	-0,0793

Рис. 17.22

Survival Time	Log-Rank Test (Heart WW = ,13727 Sum = Test statistic = ,0625 Group No	Score
0,0000+	ST_AND	0,00000
1,0000+	ST_AND	0,00000
3,0000+	HILLVIEW	0,00000
10,000	HILLVIEW	0,97561
12,000+	HILLVIEW	-0,02439
13,000+	HILLVIEW	-0,02439
15,000+	HILLVIEW	-0,02439
23,000+	HILLVIEW	-0,02439
25,000	ST_AND	0,94783
26,000+	ST_AND	-0,05217
29,000	ST_AND	0,91842
30,000+	ST_AND	-0,08158

Рис. 17.23

	Life Table for Group 1 and Group 2 (Heart) Group 1: HILLVIEW Group 2: ST_AND				
Lower Limit	Group 1: No.Enter	Group 2: No.Enter	Group 1: No.Cnsrd	Group 2: No.Cnsrd	Group 1: No.Dying
0,000000	22	22	5	6	8
197,2222	9	8	3	1	1
394,4445	5	6	2	2	0
591,6667	3	4	0	1	1
788,8889	2	2	0	1	0
986,1111	2	1	1	0	0
1183,333	1	1	0	0	0
1380,556	1	1	1	0	0
1577,778	0	1	0	0	0
1775,000	0	1	0	1	0

Рис. 17.24

Из величин уровня значимости p всех пяти критериев, приведенных в информационной части таблицы следует, что верна гипотеза о равенстве средних продолжительности жизни больных в обеих клиниках (так как все величины значительно больше 0,05). То есть по всем критериям следует, что нет существенной разницы между выживаемостью больных в клиниках *Hillview, St_Andrea s.*

Чтобы дополнительно в этом убедиться, перейдите на вкладку **Function plots** и нажмите кнопку **Cum.prop. Surviving by group (Koplan Meier)** (кумулятивная доля выживших). Программа построит графики кумулятивной функции выживаемости для обеих групп. На этом графике полные наблюдения отмечены кружками,

а цензурированные наблюдения отмечены крестиками. Из графика, изображенного на рис. 17.25, видно, что кумулятивные функции выживаемости отличаются незначительно.

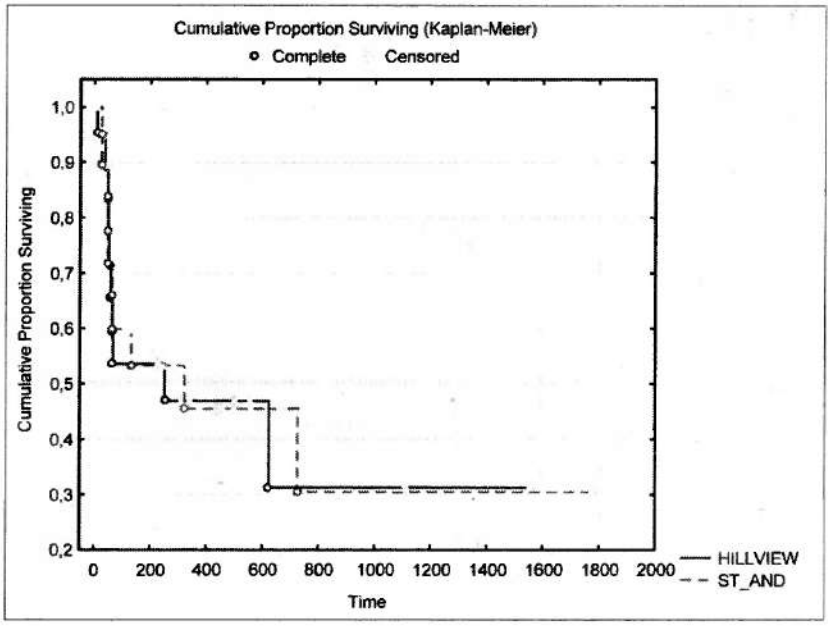


Рис. 17.25

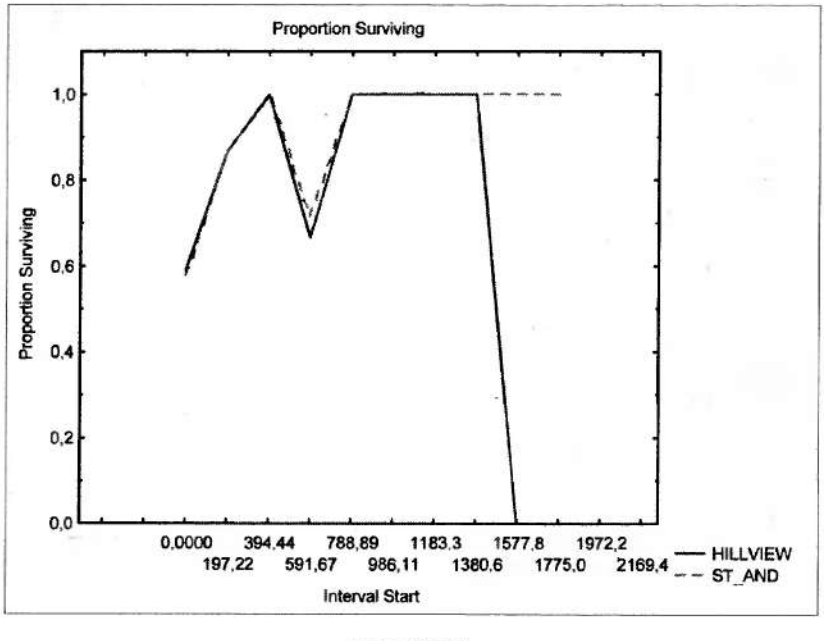


Рис. 17.26

Чтобы построить график доли выживших, нажмите кнопку **Plot of proportion surviving for each group** (график доли выживших в каждой группе). Как видно из рис. 17.26, здесь также не просматривается существенное различие со дня операции по 1380 день, а затем доля выживших в больнице *Hillview* резко уменьшается с 1 до 0, а доля выживших в больнице *St_Andrea s* остается неизменной и равной 1.

Несколько иной получится результат, если сравнить клиники *Hillview* и *Biner* (проведите анализ самостоятельно).

17.5. Сравнение выживаемости в более чем двух группах

Проведем анализ выживаемости в трех клиниках в совокупности. В стартовом окне **Survival and Failure Time Analysis** два раза щелкните по строке **Comparing multiple samples**, появится окно **Comparing Survival in Multiple Groups**. Нажмите кнопку **Variable** и произведите выбор всех переменных, как было сделано ранее, выбрав в качестве группирующей переменную **Hospital**. Щелкните два раза в поле **Codes [for groups]**, в появившемся окне нажмите кнопку **ALL**, далее нажмите **OK**. В диалоговом окне **Comparing Survival in Multiple Groups** нажмите **OK**, откроется окно результатов. Перейдите на вкладку **Advanced** (рис. 17.27).

Summary: Survival times & scores (итоговые времена жизни и их вклады). Процедура открывает электронную таблицу результатов со всеми временами жизни (цензурированные наблюдения отмечены знаком +) и соответствующими им вкладами, которые вычисляются с помощью процедуры Ментела и используются для вычисления статистики его критерия (рис. 17.28).

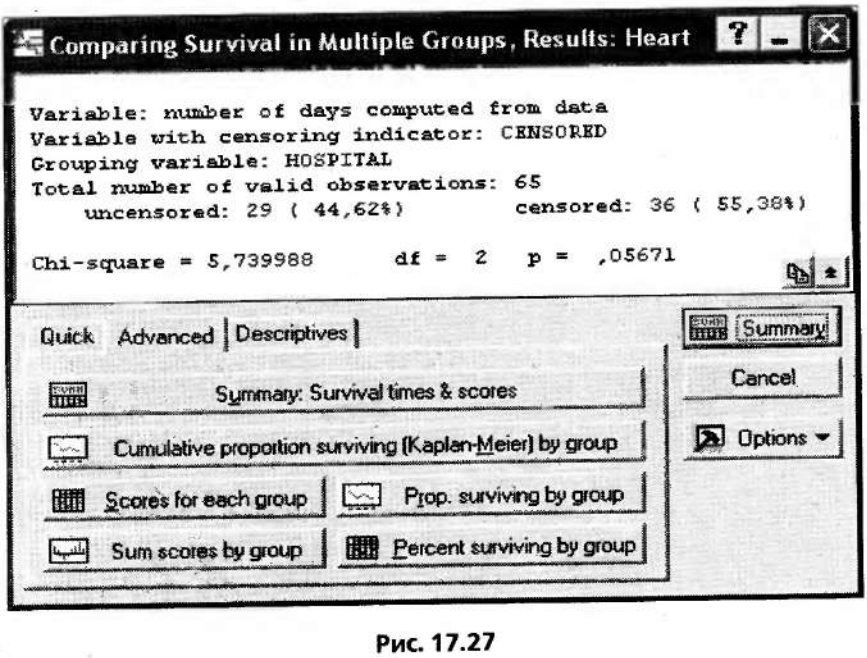


Рис. 17.27

Survival Time	Variables: Survival tim Censoring var.: CENS Chi2 = 5,73999 df = 3 Group	Score
0,0000+	ST_AND	0,0000
1,0000+	ST_AND	0,0000
1,0000+	BINER	0,0000
3,0000+	HILLVIEW	0,0000
10,000	HILLVIEW	-60,0000
12,000+	HILLVIEW	1,0000
13,000+	HILLVIEW	1,0000
15,000+	HILLVIEW	1,0000
23,000+	HILLVIEW	1,0000
25,000	ST_AND	-54,0000
26,000+	ST_AND	2,0000
29,000	ST_AND	-51,0000

Рис. 17.28

Cumulative proportion surviving (Kaplan-Meier) by group (кумулятивная доля выживших по группам). Процедура выводит график кумулятивной функции выживаемости для каждой группы. На этом графике завершенные наблюдения отмечены кружками, а цензурированные наблюдения отмечены крестиками (рис. 17.29).

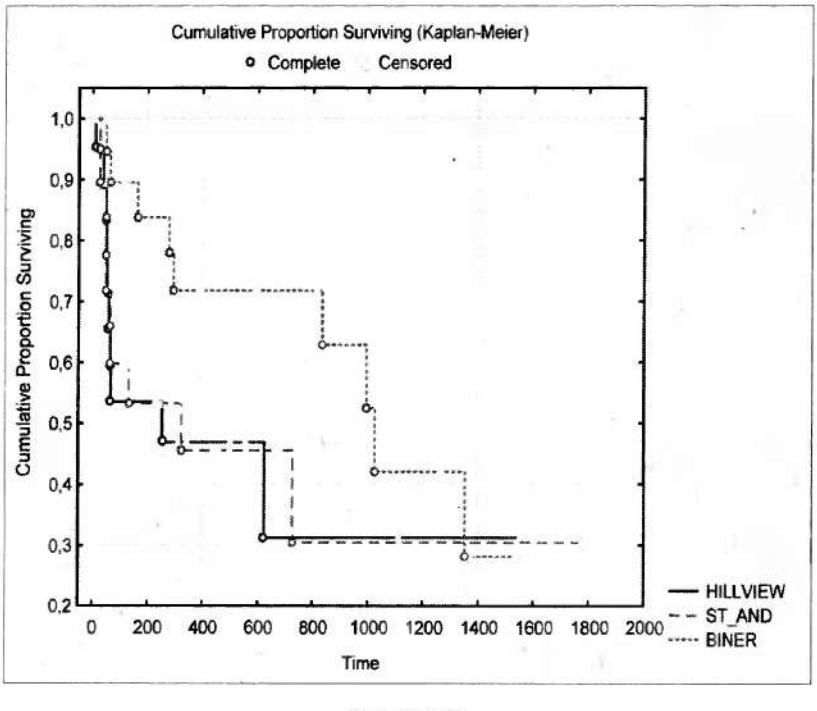


Рис. 17.29

Scores for each group (вклады по группам). Процедура открывает таблицу результатов с суммами вкладов для каждой группы (полученных с помощью процедуры Ментела). Эти вклады используются для вычисления статистики критерия Ментела. Кроме того, на экран выводятся число и процент завершенных и цензурированных наблюдений в каждой группе.

Prop. surviving by group (график доли выживших по группам). Процедура строит график доли выживших для каждой группы (рис. 17.30).

Sum scores by group (гистограмма суммы вкладов по группам). Процедура строит гистограмму суммы вкладов по группам (вычисляемых с помощью процедуры Ментела).

Percent surviving by group (доля выживших по группам). Процедура открывает таблицу времен жизни для каждой группы в процентах.

В информационной части окна **Summary: Survival times & scores** (рис. 17.28) приведены значения критерия (5,73999) и величина уровня значимости критерия p (0,05671). Так как величина p незначительно больше 0,05, то можно сделать вывод, что между тремя клиниками, делающими операции по пересадке сердца, существуют отличия в выживаемости пациентов.

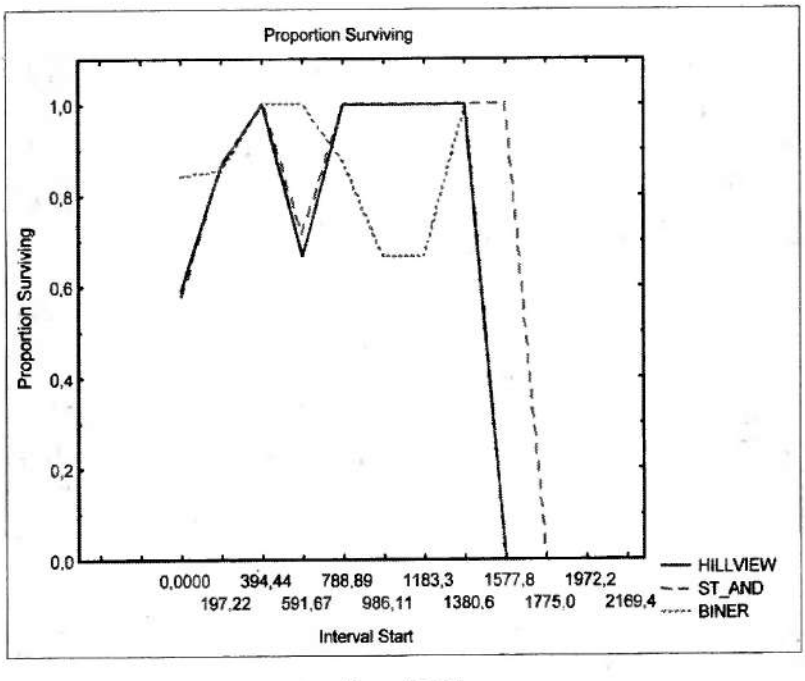


Рис. 17.30

Приведенные таблицы и графики показывают, что выживаемость пациентов в клиниках *Hillview, St_Andrea s* примерно одинакова и отличается от выживаемости в клинике *Biner*. При этом кумулятивная функция выживания в больнице *Biner* не резко убывает в первое время по сравнению с двумя другими больницами, и на протяжения длительного периода значения функции превосходят значения функции, соответствующие клиникам *Hillview u St_Andrea s*. В первые

787 дней график доли выживших в клинике *Biner* расположен значительно выше графиков доли выживших для клиник *Hillview, St_Andrea s.* Поэтому делаем вывод, что по каким-то причинам пациенты в больнице *Biner* имеют бо́льшие шансы выжить не только в первые критические дни после трансплантации сердца, но и в последующие.

17.6. Регрессионные модели

При анализе времен жизни особую актуальность приобретает выяснение того, являются ли некоторые переменные связанными с наблюдаемыми временами жизни. При наличии такой зависимости необходимо подобрать подходящую математическую модель и оценить значения параметров модели. Такую модель проблематично строить при помощи классической множественной регрессии по двум причинам. Во-первых, времена жизни обычно не являются нормально распределенными, а это серьезное нарушение предположений для оценивания множественной регрессии по методу наименьших квадратов. Времена жизни обычно имеют экспоненциальное распределение или распределение Вейбулла. Во-вторых, имеется проблема с цензурированными, т.е. незавершенными наблюдениями [6].

В модуле **Survival Analysis** реализованы 5 регрессионных моделей зависимости времен жизни от независимых переменных: регрессионная модель пропорциональных интенсивностей Кокса, экспоненциальная регрессия, логнормальная линейная регрессия, нормальная линейная регрессия и модель Кокса с зависящими от времени ковариатами. Для каждой из этих моделей программа вычисляет оценки максимального правдоподобия, по которым можно судить об адекватности модели, а также параметры модели.

Модель пропорциональных интенсивностей Кокса — наиболее общая регрессионная модель, поскольку она не связана с какими-либо предположениями относительно распределения времени выживания. Эта модель предполагает, что функция интенсивности имеет некоторый уровень y, являющийся функцией независимых переменных $z_1, z_2, ..., z_m$, называемых ковариатами. Модель представляется в виде следующего соотношения:

$$h(t) = h_0(t)\, y\, (z_1, z_2, ...z_m),$$

где множитель $h_0(t)$ называется базовой функцией интенсивности.

Никаких предположений о виде функции интенсивности не делается. Поэтому модель Кокса может рассматриваться в некотором смысле как непараметрическая. Модель может быть параметризована и записана, например, в следующем виде:

$$h[(t), (z_1, z_2, ..., z_m)] = h_0(t)exp(b_1z_1 + ... + b_mz_m),$$

где $h[t, ...]$ обозначает результирующую интенсивность, при заданных для соответствующего наблюдения значениях m ковариат и соответствующем времени жизни t. Множитель $h_0(t)$ называется базовой функцией интенсивности, она равна интенсивности в случае, когда все независимые переменные равны нулю. Обратите внимание,

что в правой части уравнения стоит произведение двух функций, каждая из которых зависит от своего множества переменных.

Можно линеаризовать модель пропорциональных интенсивностей Кокса, поделив обе части соотношения на $h_0(t)$ и взяв натуральный логарифм от обеих частей:

$$ln[\, h\{(t), (z...)\}/h_0(t)] = b_1z_1 + ... + b_mz_m.$$

Получена достаточно простая линейная модель, которая легко поддается изучению.

В то время как никаких прямых предположений о виде функции интенсивности ранее не делалось, приведенная параметризованная модель Кокса все же подразумевает два предположения. Во-первых, зависимость между функцией интенсивности и логлинейной функцией ковариат является мультипликативной. Это соотношение называется также предположением (гипотезой) пропорциональности. Реально оно означает, что для двух заданных наблюдений с различными значениями независимых переменных отношения их функций интенсивности не зависят от времени. Второе предположение состоит именно в логарифмической линейности соотношения между функцией интенсивности и независимыми переменными.

В своей основе модель экспоненциальной регрессии предполагает, что распределение продолжительности жизни экспоненциально и связано со значениями некоторого множества независимых переменных (ковариат) z_i. Модель имеет вид

$$S(z) = exp(a + b_1z_1 + b_2z_2 + ... + b_mz_m),$$

где $S(z)$ — время жизни; a — константа; а b_i — параметры регрессии.

Для оценки адекватности модели вычисляется значение критерия как функции логарифма правдоподобия для модели со всеми оцененными параметрами (L_1) и логарифма правдоподобия модели, в которой все ковариаты обращаются в 0 (L_0). Если величина χ^2 статистически значима (уровень значимости p меньше чем 0,05), отвергаем нулевую гипотезу и принимаем, что независимые переменные значимо влияют на время жизни.

Другой способ проверки предположения экспоненциальности — построение остатков времен жизни и сравнение их со значениями стандартных экспоненциальных порядковых статистик α (альфа), θ (тэта).

В моделях нормальной и логнормальной регрессии предполагается, что времена жизни (или их логарифмы) имеют нормальное распределение. Модель в основном идентична обычной модели множественной регрессии и может быть описана следующим образом:

$$S(z) = a + b_1z_1 + b_2z_2 + ... + b_mz_m.$$

Если выбираем модель логнормальной регрессии, то $S(z)$ заменяется $lnS(z)$. Модель нормальной регрессии особенно полезна, поскольку часто данные могут быть преобразованы в нормальные путем применения нормализующих аппроксимаций. Таким образом, в некотором смысле это наиболее общая параметрическая модель (в противоположность модели пропорциональных интенсивностей Кокса, которая является непараметрической), ее оценки могут быть получены для большого разнообразия исходных распределений времен жизни.

В качестве примера использования процедуры **Regression model** установите связь между продолжительностью жизни пациента и возрастом пациента на момент трансплантации (*AGE*), показателем антигенной несовместимости (*ANTIGEN*), показателем несовместимости тканей (*MISMATCH*). Воспользуйтесь наиболее общей регрессионной моделью (которая не предполагает никаких данных о природе или форме функции выживаемости) — пропорциональной моделью интенсивностей Кокса.

В диалоговом окне **Survival and Failure Time** (рис. 17.2) двойным щелчком по **Regression models** откройте окно **Regression Models for Censored Data** (рис. 17.31). Для выбора переменных нажмите кнопку **Variables**, чтобы появилось окно **Variable selection**. Выберите в первом поле слева первые 6 переменных, во втором — переменные *AGE, ANTIGEN, MISMATCH* и в третьем — *CENSORED* (рис. 17.32).

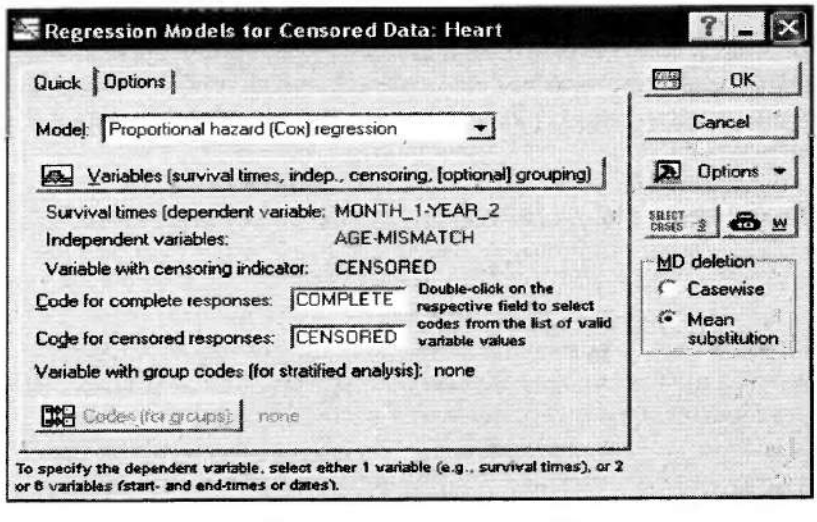


Рис. 17.31

Рис. 17.32

Дважды щелкнув в полях **Code for complete responses**, **Code for censored responses**, выберите соответственно *COMPLETE* и *CENSORED*. Так как в поле **Model** установлено по умолчанию *Proportional hazard (Cox) regression*, то можно приступать к анализу.

Нажмите **OK**, программа запустит итерационную процедуру оценивания параметров, которая максимизирует логарифмическую функцию правдоподобия регрессионной модели посредством метода Ньютона-Рафсона. После того как программа найдет наиболее подходящие оценки параметров и появится окно **Regression Results**, перейдите на вкладку **Advanced** (рис. 17.33).

Из информационной части диалога видно, что значение (*Chi-Square*) статистически значимо (уровень значимости $p = 0{,}00006$, что существенно меньше 0,05), поэтому можно сделать вывод, что некоторые переменные связаны с выживаемостью.

Нажмите кнопку **Summary: Parameter estimates**, чтобы посмотреть оценки

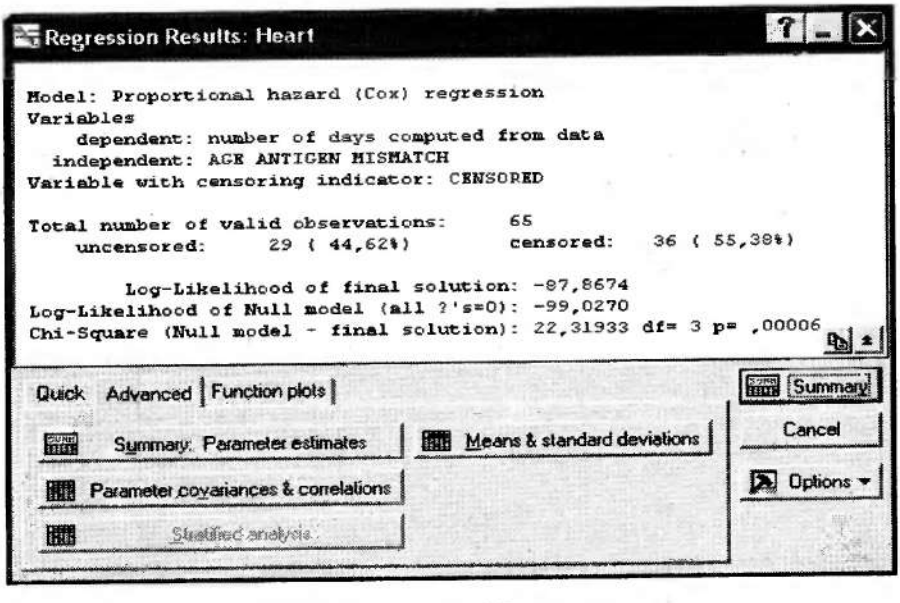


Рис. 17.33

параметров и стандартные ошибки оценок параметров (рис. 17.34). В первом столбце (*Beta*) приведены оценки параметров (коэффициенты при соответствующих переменных в регрессионном уравнении), во втором столбце (*Standard Error*) — стандартные ошибки, в третьем столбце — значения *t-критерия*, которые равны отношению соответствующих элементов первого и второго столбцов. В пятом и шестом столбцах приведены значения статистики Вальда (*Wald Statist.*) и уровень значимости (p). Обычно статистически значимыми (при $p<0{,}05$) считаются такие оценки параметра, которые, по крайней мере, в два раза превышают ошибку этого параметра ($t>2{,}0$).

N=65	Dependent Variable: Survival times in days (Heart) Censoring var.: CENSORED Chi2 = 22,3193 df = 3 p = ,00006					
	Beta	Standard Error	t-value	exponent beta	Wald Statist.	p
AGE	0,109096	0,033293	3,276836	1,115269	10,73766	0,001051
ANTIGEN	-0,048782	0,471644	-0,103431	0,952388	0,01070	0,917622
MISMATCH	1,063761	0,394599	2,695804	2,897246	7,26736	0,007026

Рис. 17.34

Следовательно, из данных таблицы делаем вывод, что возраст (*AGE*) и несовместимость тканей (*MISMATCH*) являются наиболее важными (значимыми) предикторами функции интенсивности (мгновенного риска). Причем несовместимость тканей — более важный предиктор для риска, чем возраст (коэффициент при *MISMATCH* в 10 раз больше коэффициента при *AGE*), а коэффициент при переменной *ANTIGEN* в пропорциональной модели интенсивностей Кокса можно считать равным 0, так как соответствующая ему величина *t-критерия* значительно меньше 2 и значение *p* существенно больше 0,05.

Кнопка **Parameter covariances & correlations** открывает таблицу результатов с ковариационной и корреляционной матрицами оценок параметров. Кнопка **Means & standard deviations** открывает таблицу результатов со средними и стандартными отклонениями независимых переменных (ковариат) и времен жизни (рис. 17.35).

variable	Means and Standard Deviations (Heart)			
	mean	st. dev.	minimum	maximum
AGE	45,6769	9,1858	19,00000	64,000
ANTIGEN	0,2615	0,4429	0,00000	1,000
MISMATCH	1,1646	0,6233	0,00000	3,050
No.days	382,6769	463,2327	0,00000	1775,000

Рис. 17.35

Кнопка **Stratified analysis** предназначена для просмотра результатов стратифицированного анализа (анализа по группам). Она активна, если в окне **Variable selection** в последнем столбце была выбрана группирующая переменная, например, *HOSPITAL* и если не выбрана опция *Equal coefficients, different baseline* h_0 (одинаковые коэффициенты, различные базовые функции h_0) на вкладке **Options** (опции) диалогового окна **Regression Models for Censored Data**.

К полученным значениям оценок параметров можно построить графики выживаемости как функции независимых переменных. Перейдите на вкладку **Function plots** и нажмите кнопку **Graph survival function for means**, программа построит график выживаемости (рис. 17.36) при средних значениях переменных *AGE, ANTIGEN, MISMATCH* (первый столбец, табл. на рис. 17.35).

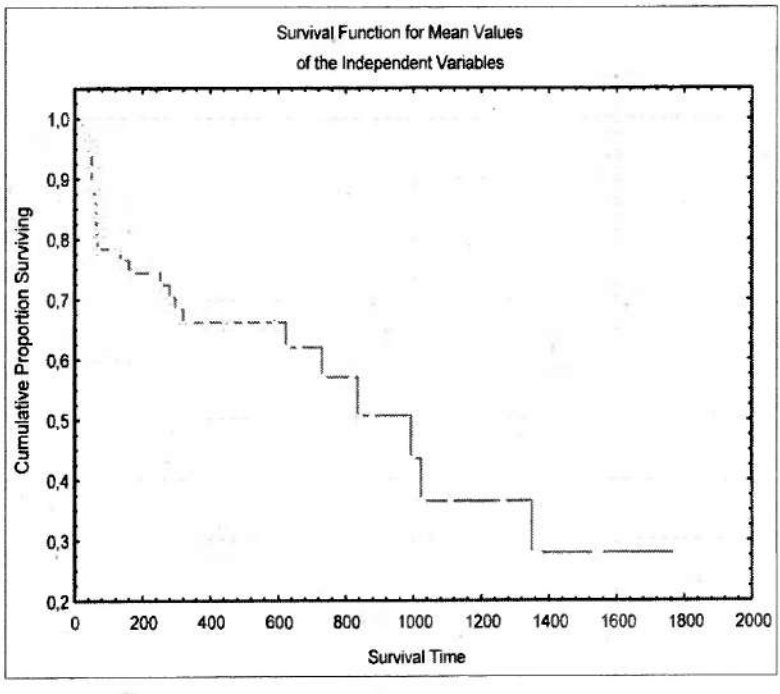


Рис. 17.36

В программе предусмотрена возможность построения графика функции выживания, когда значения ковариат задаются пользователем. Нажмите кнопку **Graph survival function for spec. vals.**, откроется окно **Independent Variable Values** с заданными по умолчанию средними значениями переменных. Проверьте, как влияет возраст пациента на функцию выживания. Увеличьте возраст до 65 лет (рис. 17.37), не изменяя значения других переменных, и нажмите **OK.** Из графика выживаемости, изображенного на рис. 17.38, следует, что существенно уменьшились значения функции выживания, вероятность пациента прожить после операции более 50 дней примерно равна 0,13, а вероятность прожить более 800 дней практически равна 0, в то время как для пациентов с возрастом 45 лет эта вероятность равно 0,5.

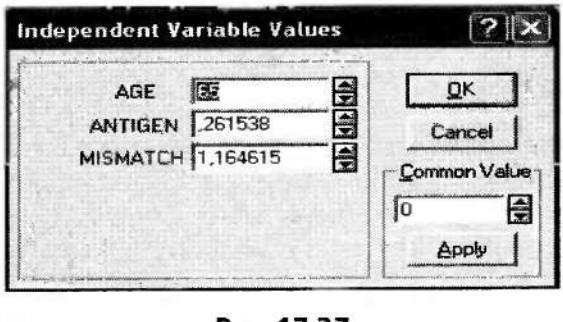


Рис. 17.37

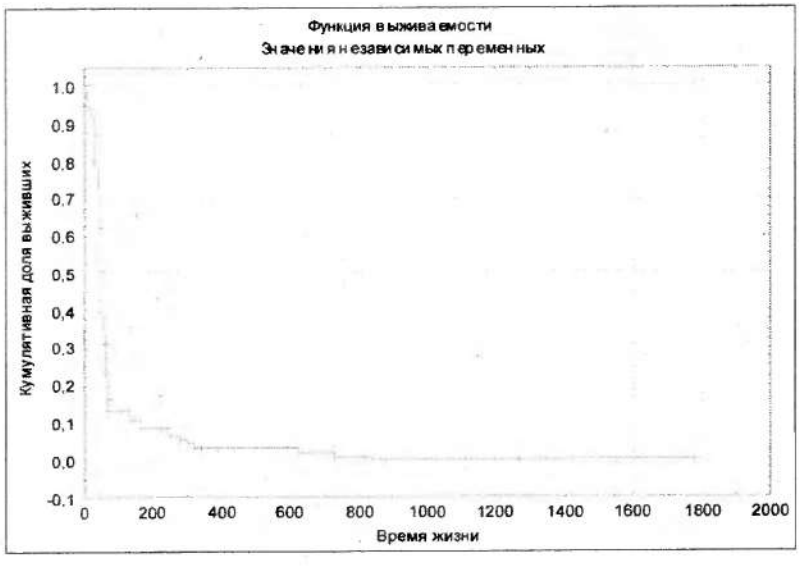


Рис. 17.38

Если нужны числовые значения функции выживания, щелкните правой кнопкой мыши на изображении графика и в появившемся контекстном меню выберите команду *Graph Data Editor*.

Постройте при помощи контекстного меню график линии (*Line Plot*) столбца *Number Exposed* (рис. 17.6) — содержащего число объектов, которые были «живы» в начале рассматриваемого временного интервала, минус половина числа изъятых или цензурированных объектов.

Из данного графика (рис. 17.39) можно сделать вывод, что гипотеза о том, что распределение продолжительности жизни является экспоненциальным, небезосновательна. Поэтому имеет смысл построить экспоненциальную регрессионную модель зависимости продолжительности времени жизни от переменных *AGE, ANTIGEN, MISMATCH*.

В диалоговом окне **Regression Models for Censored Data** (рис. 17.31) в поле **Model** установите *Exponential regression*, остальные установки оставьте без изменения и щелкните **OK**.

Из информационной части окна **Regression Results** (рис. 17.40) следует, что модель **Exponential regression** более адекватна, чем модель *Proportional hazard (Cox) regression*, так как оценки максимального правдоподобия и существенно больше, а уровень значимости p меньше. Таблица с оценками параметров приведена на рис. 17.41.

В данной таблице, как и в случае пропорциональной модели интенсивностей Кокса, несовместимость тканей — более важный предиктор времени жизни, чем возраст, а коэффициент при переменной *ANTIGEN* также можно считать равным 0.

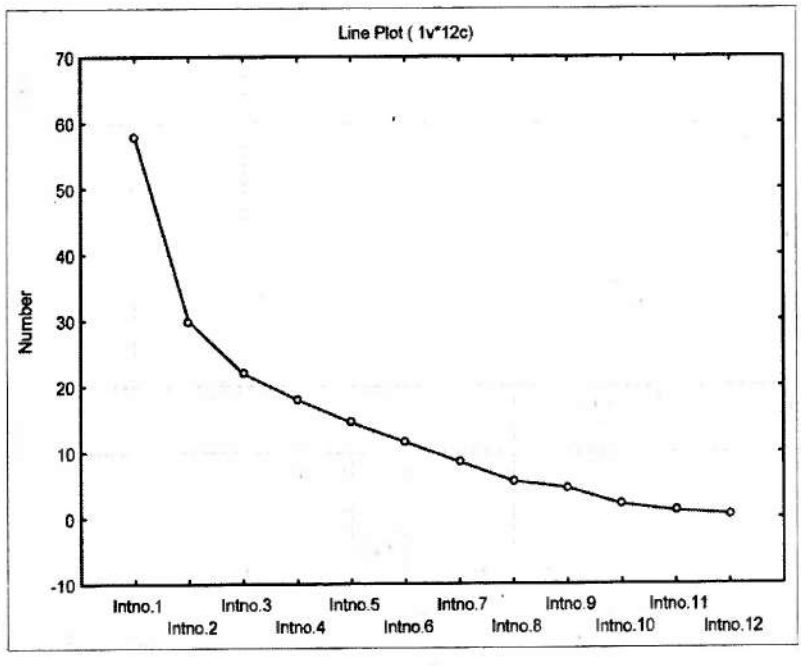


Рис. 17.39

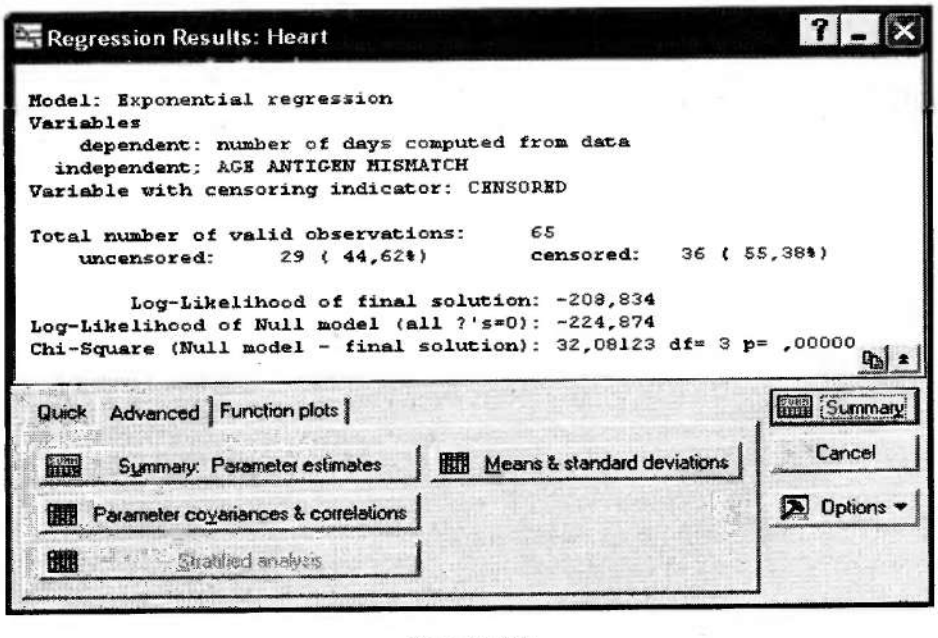


Рис. 17.40

N=65	Dependent Variable: Survival tin Censoring var.: CENSORED Chi2 = 32,0812 df = 3 p = ,0000 Beta	Standard Error	t-value
AGE	-0,12418	0,032857	-3,77938
ANTIGEN	-0,06363	0,459082	-0,13861
MISMATCH	-1,22008	0,365911	-3,33436
Constant	14,08542	1,686244	8,35313

Рис. 17.41

17.7. Модель пропорциональных интенсивностей Кокса с зависящими от времени ковариатами

Модель Кокса с зависящими от времени ковариатами целесообразна, если обоснованность гипотезы пропорциональности рисков подвергается сомнению. Например, рассмотрим исследование, в котором ковариатой является категориальная (групповая) переменная, а именно индикатор того, подвергнут или нет некоторый пациент хирургической операции. Пусть пациент 1 подвергнут операции, в то время как пациент 2 — нет. Согласно предположению пропорциональности отношение функций интенсивностей для обоих пациентов не зависит от времени и означает, что риск для пациента, подвергнутого операции, постоянно более высокий (или более низкий), чем риск пациента, не подвергнутого операции (при условии, что оба дожили до рассматриваемого момента). Однако обычно более реалистична другая модель, а именно: сразу после операции риск прооперированного пациента выше, однако при благоприятном исходе операции с течением времени убывает и становится меньше риска неоперированного пациента. Можно привести много других примеров, где гипотеза о пропорциональности неприемлема. Так, при изучении физического здоровья возраст служит одним из факторов выживаемости после хирургической операции. Ясно, что возраст — более важный предиктор для риска сразу после операции, чем по прошествии некоторого времени после операции (например, после первых признаков выздоровления). В ускоренных испытаниях на надежность иногда используют нагрузочную ковариату (например, уровень напряжения), которую медленно наращивают со временем вплоть до отказа прибора, например, до пробоя изоляции. В этом случае влияние ковариаты опять зависит от времени.

Если гипотеза о пропорциональности несправедлива, то в этом случае предпочтительнее ковариаты, зависящие от времени, т.е. можно явно определить ковариаты как функции времени. Рассмотрим пример такой модели из медицины [10]. Пусть изучается воздействие некоторого препарата на состояние больного, z — категориальная переменная со значениями 1 (для больных, принимавших препарат в процессе лечения) и 0 (для больных, не принимавших препарат). Тогда функцию риска можно записать в виде

$$h(t, z) = h_0(t)exp[\beta_1 z + \beta_2 z(lnt - 100)],$$

где t — время, прошедшее после начала приема лекарства. Константа 100 использована как нормировка, поскольку среднее логарифма жизни для этого множества данных равно 100. Зная оценки параметров β_1, β_2 и функцию интенсивности $h_0(t)$, можно оценить значение функции мгновенного риска через время t после начала приема лекарственного препарата. В этом примере функция интенсивности в момент t есть функция базовой функции интенсивности h_0, ковариаты z *и z-кратного* логарифма времени. Другими словами, функция риска в каждый момент есть функция ковариаты и времени; таким образом, влияние ковариаты на выживаемость зависит от времени; отсюда название — ковариата, т.е. зависящая от времени. Эта модель позволяет использовать специфический критерий проверки гипотезы пропорциональности. Если параметр β_2 статистически значим (например, если он, по крайней мере, в два раза больше своей стандартной ошибки), то можно сделать вывод, что ковариаты z действительно зависят от времени и поэтому гипотеза пропорциональности неверна.

Если предположение о пропорциональности не выполняется, то для того, чтобы воспользоваться моделью пропорциональных интенсивностей Кокса, можно в случае категориальных ковариат (например, учитывающих, был или не был больной прооперирован, принимал или не принимал лекарственный препарат) обратиться к стратифицированному анализу выживаемости, в котором исследователь разбивает наблюдения на однородные по фактору риска группы. И тогда можно провести подгонку модели пропорциональных интенсивностей отдельно для каждой группы наблюдений.

	Pike, 1966; Survival in two groups of		
	1 SURVIVAL	2 CENSORED	3 GROUP
1	143	complete	Group_1
2	164	complete	Group_1
3	188	complete	Group_1
4	188	complete	Group_1
5	190	complete	Group_1
6	192	complete	Group_1
7	206	complete	Group_1
8	209	complete	Group_1
9	213	complete	Group_1
10	216	complete	Group_1
11	220	complete	Group_1
12	227	complete	Group_1
13	230	complete	Group_1
14	234	complete	Group_1
15	246	complete	Group_1
16	265	complete	Group_1
17	304	complete	Group_1
18	216	censored	Group_1

Рис. 17.42

Проверим верность гипотезы о пропорциональности для файла данных **Pike** из библиотеки **Example → Datasets**, в котором описана продолжительность жизни в двух группах крыс (рис. 17.42). Одна группа была контрольной, другая — подвержена воздействию канцерогена.

Предположим, что эффективность лечения (подверженность раку закодирована в переменной *GROUP*) основной болезни не равна константе, а это значит, что предположение о прямой пропорциональной зависимости может быть неверным. Для того чтобы проверить, является ли предположение логичным (имеет ли основание), используемую модель сведем к данным, состоящим из фиксированной ковариаты *GROUP* и зависящей от времени переменной:

$$TIME\ DEPENDENT\ GROUP = GROUP(Ln(TIME) - 5{,}4).$$

Значение 5,4 используется только в целях нормировки, так как среднее логарифмической функции выживания примерно равняется 5,4.

В стартовом окне **Survival and Failure Time Analysis** (рис. 17.2) выделите процедуру **Time-dependent covariates** (зависящие от времени ковариаты) и нажмите **OK**. Появится окно **Proportional Hazard Model with Time-Dependent Covariates** (модель пропорциональных интенсивностей с зависящими от времени параметрами). Нажмите кнопку **Variables (survival times, censoring [optional] grouping)** — появится стандартное окно **Variable selection**. Выберите здесь переменную *SURVIVAL* в левом окошке, переменную *CENSORED* в среднем окошке и нажмите **OK**. Щелкните дважды в поле **Code for complete responses** (код для завершенных наблюдений) — появится окно **Variable 2**. Выберите здесь *COMPLETE* и нажмите **OK**. Также щелкните по полю **Censored** и выберите *CENSORED*.

Для определения зависящих от времени ковариат в приведенном окне есть два изменяемых поля. Левое поле **Covariate** — для ввода названий ковариат, используемых выходными данными. Правое (широкое) поле **Expression** — для определения соответствующей ковариаты через математические (арифметические) выражения. Нажмите кнопку **Select variables** — появится стандартное окно выбора переменных — **Select variables to be transferred to expressions**. Переменные, которые вы выберите, будут автоматически вставлены в оба поля: и в **Covariate**, и в **Expression**. Выберите переменную *GROUP* и нажмите **OK**. Переменная *GROUP* будет помещена в первую строку обоих полей. Повторите эту операцию для того, чтобы во второй строке каждого из полей появилось такая же переменная. Далее измените вторую строку, как показано на рис. 17.43. Обратите внимание, что текст во второй строчке после знака «точка с запятой» воспринимается программой как комментарий. Задав какие-то переменные и их значения, можно сохранить, а позже вернуться к сохраненным установкам с помощью кнопок **Save expressions** и **Open expressions** соответственно. Нажмите **OK**. На несколько секунд появится окно **Model Parameter Estimation**, и после того как программой будут найдены наиболее подходящие параметры, итерационная процедура завершится и появится окно **Regression Results**.

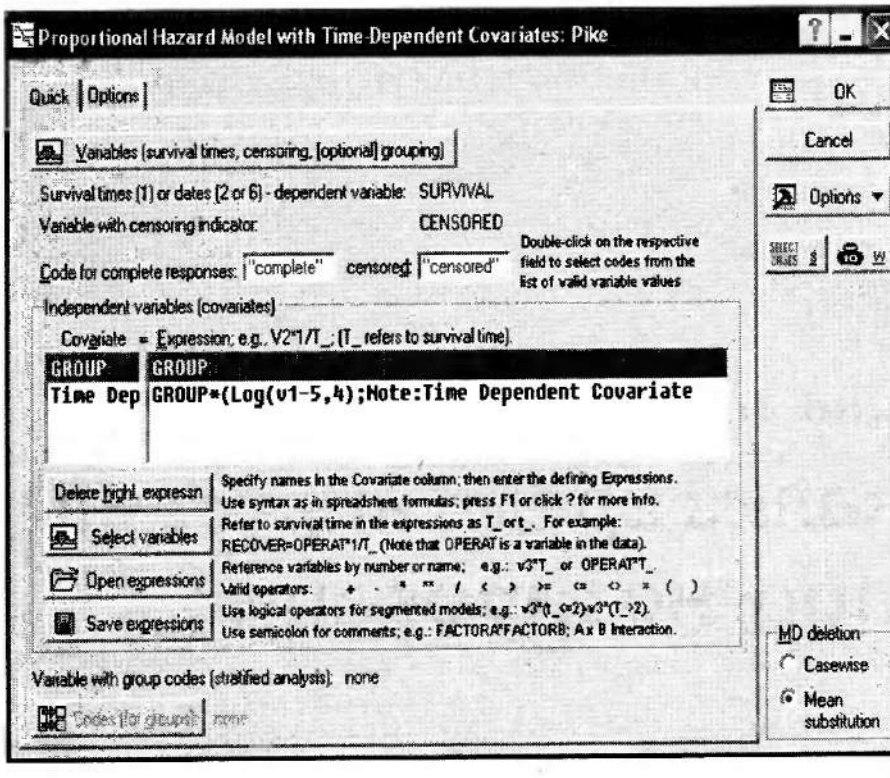


Рис. 17.43

Нажмите кнопку **Summary: Parameter estimates** на вкладке **Quick**. Как видно из таблицы (рис. 17.44), по критерию Вальда параметр β_2 статистически значим, при этом его значение (–9,4781) более чем в 2 раза больше стандартной ошибки (1,8787). Поэтому можно сделать вывод, что гипотеза о пропорциональности несостоятельна.

	Dependent Variable: SURVIVAL (Pike) Censoring var.: CENSORED Chi2 = 34,5726 df = 2 p = ,00000					
N=40 Exp.Name	Beta	Standard Error	t-value	exponent beta	Wald Statist.	p
GROUP	0,59127	0,395066	1,49664	1,806285	2,23994	0,134496
DEPENDENT	-9,47818	1,878730	-5,04500	0,000077	25,45198	0,000000

Рис. 17.44

Глава 18

Анализ временных рядов и прогнозирование

Временные ряды — это наиболее интенсивно развивающееся, перспективное направление математической статистики.

Под временным (динамическим) рядом подразумевается последовательность наблюдений некоторого признака X (случайной величины) в последовательные равноотстоящие моменты t [11]. Отдельные наблюдения называются уровнями ряда и обозначаются x_t, $t = 1, ..., n$.

При исследовании временного ряда выделяются несколько составляющих:

$$x_t = u_t + \gamma_t + c_t + \varepsilon_t, t = 1, ..., n, \qquad (18.1)$$

где u_t — тренд, плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов (убыль населения, уменьшение доходов и т.д.); γ_t — сезонная компонента, отражающая повторяемость процессов в течение не очень длительного периода (дня, недели, месяца и т.д.); c_t — циклическая компонента, отражающая повторяемость процессов в течение длительных периодов времени свыше одного года; ε_t — случайная компонента, отражающая влияние не поддающихся учету и регистрации случайных факторов.

Первые три компоненты представляют собой детерминированные составляющие. Случайная составляющая образована в результате суперпозиции большого

числа внешних факторов, оказывающих каждый в отдельности незначительное влияние на изменение значений признака X.

Как и в математической статистике, вариационный ряд рассматривается как одна из реализаций случайной величины X, а временной ряд — как одна из реализаций случайного процесса $X(t)$ (t — время). Но в отличие от вариационного ряда, члены временного ряда, как правило, не являются статистически независимыми и одинаково распределенными.

Временной ряд x_t, $t = 1, ..., n$ называется строго стационарным, если совместное распределение вероятностей n наблюдений $x_1, x_2, ..., x_n$ — такое же, как и наблюдений $x_{1+\tau}, x_{2+\tau}, ..., x_{n+\tau}$ при любых n, t, τ. То есть у стационарных временных рядов вероятностные характеристики не зависят от момента t. Поэтому математическое ожидание $M(x_t) = a$ и среднеквадратическое отклонение могут быть оценены по значениям x_t по формулам

$$a = \bar{x}_t = \sum_{t=1}^{n} x_t / n, \qquad \sigma = \sqrt{\frac{\sum_{t=1}^{n} (x_t - \bar{x}_t)^2}{n}}. \qquad (18.2)$$

Степень тесноты связи между последовательностями наблюдений временного ряда $x_1, x_2, ..., x_n$ и $x_{1+\tau}, ..., x_{n+\tau}$, сдвинутых относительно друг друга на единиц, может быть определена при помощи коэффициента корреляции

$$\rho(\tau) = \frac{M[(x_t - M(x_t))(x_{t+\tau} - M(x_{t+\tau}))]}{\sigma_{x_t} \sigma_{x_{t+\tau}}} = \frac{M[(x_t - a)(x_{t+\tau} - a)]}{\sigma^2}. \qquad (18.3)$$

Параметр τ называют лагом ряда. Так как $\rho(\tau)$ измеряет корреляцию между членами одного и того же ряда, его называют коэффициентом автокорреляции, а зависимость $\rho(\tau)$ — автокорреляционной функцией. В силу стационарности временного ряда $\rho(\tau)$ зависит только от τ и $\rho(\tau) = \rho(-\tau)$.

Можно выделить следующие этапы анализа временных рядов:

1. Графическое представление и анализ поведения временного ряда.
2. Выделение и анализ детерминированных составляющих ряда.
3. Сглаживание и фильтрация (удаление низко- или высокочастотных составляющих) временного ряда.
4. Исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели ее описания.
5. Прогнозирование поведения временного ряда на основе проведенных исследований.

Задача прогнозирования состоит в том, чтобы по значениям наблюдений, собранных к данному моменту, определить значения в следующие моменты [21].

В модуле **Time Series/Forecasting** (Временные ряды и прогнозирование) представлен широкий набор процедур, реализующих все перечисленные этапы

анализа. Наиболее полно эти процедуры применительно к версии 5.1 описаны в [21]. Все процедуры полностью интегрированы. Результаты анализа одной модели можно использовать для дальнейшего анализа. Программа автоматически отмечает все этапы анализа временного ряда и сохраняет полную историю преобразований и полученные результаты. Поэтому пользователь всегда имеет возможность вернуться к более раннему этапу анализа или отобразить на графике исходный ряд и его преобразования. Информация о последовательных преобразованиях хранится в виде длинных меток переменных, поэтому при сохранении вновь полученных рядов в файле данных автоматически сохраняется вся «история» каждого из рядов.

С помощью различных преобразований исходного временного ряда можно понять его структуру и имеющиеся в нем закономерности, привести его к виду, пригодному для моделирования (например, добиться стационарности). В модуле реализованы такие часто используемые преобразования, как: удаление тренда, удаление автокорреляций, взятие разностей, суммирование, вычисление остатков, сдвиг, преобразование Фурье, обратное преобразование Фурье и др.

Реализованы различные методы сглаживания: сглаживание скользящими средними (не взвешенными или взвешенными — с весами, заданными пользователем или вычисленными по методам Даниеля, Тьюки, Хэмминга, Парзена и Бартлета), медианное сглаживание, простое экспоненциальное сглаживание, 4253H-сглаживание, косинус-сглаживание. Можно выполнить анализ автокорреляций, частных автокорреляций.

Для активизации модуля в меню **Statistica** надо выбрать последовательность команд **Advanced Linear/Nonlinear Models → Time Series/Forecasting.** Откроется стартовое окно модуля (рис.18.1). Описание модуля и основные принципы работы с ним проиллюстрируем на примере файла из **Examples → Datasets — Series G**, в котором приведен ряд месячных авиаперевозок с января 1949 г. по декабрь 1960 г., всего 144 наблюдения. При помощи кнопки **Variables** введите имя переменной *SERIES G*. После открытия файла и выбора переменной в информационной части окна в поле **Variable** появятся имена переменных, а в поле **Long variable name** — расширенное имя *Monthly passenger totals* (общие месячные пассажирские перевозки).

Значок слева от имени переменной означает, что переменные закрыты на ключ и не могут быть удалены без прерывания анализа. Весь дальнейший диалог происходит именно с этими переменными, которые можно преобразовывать, анализировать, но нельзя удалять из текущего анализа. В процессе работы для выбора наиболее подходящего преобразования ряды многократно преобразовываются, и чтобы не хранить лишнюю информацию (неудачные преобразования), их следует удалить. Для этого служит кнопка **Delete highlighted variable** (удалить высвеченные переменные). Если нам для проведения дальнейших исследований (возможно и в других модулях *STATISTICA*) необходимо сохранить некоторые преобразования, надо воспользоваться кнопкой **Save variables** (сохранить переменные).

В поле **Number of backups per variable** (число резервов для переменных) определяют число преобразований текущего диалога в информационной части окна. Если число преобразований превысит указанное число, то система сделает запрос — сохранить ли очередное преобразование.

Кнопка **Select cases** (выбрать наблюдения) предназначена для выбора подмножества случаев для анализа.

Кнопка **OK (transformations, autocorrelations, plots)** (преобразования, автокорреляции, графики) открывает специальное окно для преобразования ряда.

Остальные кнопки, расположенные в центре стартового окна, — функциональные, определяющие различные методы (процедуры) анализа временных рядов.

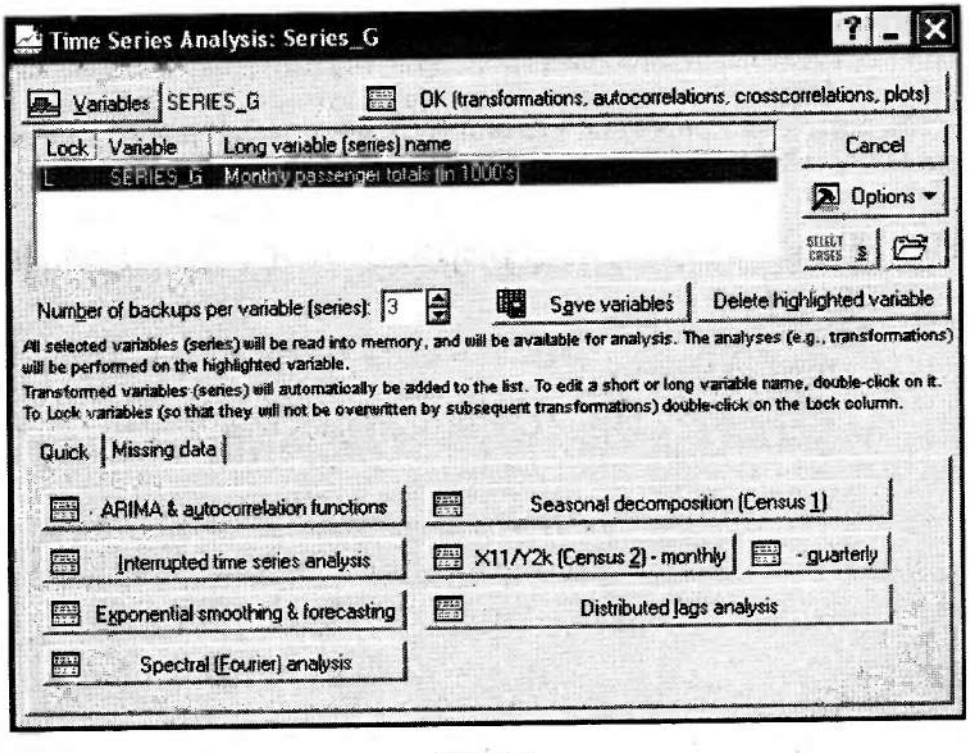


Рис. 18.1

ARIMA & autocorrelation functions (модель авторегрессии и проинтегрированного скользящего среднего — сокращенно АРПСС).

Interrupted time series analysis (анализ прерванных временных рядов или модели АРПСС с интервенцией).

Exponential smoothing & forecasting (экспоненциальное сглаживание и прогнозирование).

Seasonal decomposition (Census 1) (сезонная декомпозиция).

Spectral (Fourier) analysis (спектральный (Фурье) анализ).

XII/Y2k (Census 1) — montly (12-месячная сезонная корректировка).

Quartenly (квартальная сезонная корректировка).

Distributed Lags Analysis (анализ распределенных лагов).

На вкладке **Missing Data** система предлагает различные возможности для заполнения пропущенных значений:

- *Overall mean* — общее среднее.
- *Interpolation from adjacent points* — интерполяция по соседним точкам.
- *Mean of N adjacent points* — среднее по соседним точкам.
- *Median of N adjacent points* — медиана соседних точек.
- *Predicted values from linear trend regression* — предсказанные значения с учетом линейной регрессии.

Прогнозирование временных рядов — и наука, и искусство. для того чтобы делать правильные (адекватные) прогнозы, необходимы знания и опыт. Надо строить прогнозы различными способами, сопоставлять результаты и только после этого выбирать ту наилучшую модель, которая наиболее правдоподобно прогнозирует ряд.

Рассмотрим прогнозирование при помощи модели авторегрессии и проинтегрированного скользящего среднего.

18.1. Модель проинтегрированного скользящего среднего

Этот важный класс параметрических моделей, описывающих и нестационарные ряды, имеет большое практическое значение. В программе *STATISTICA ARIMA* реализована в методологии Бокса и Дженкинса. Большинство временных рядов, например в экономике, описываются моделью *ARIMA — АРПСС*. Модель может включать константу. Перед построением модели ряд можно подвергнуть преобразованию, которое автоматически будет отменено после построения прогноза по АРПСС, при этом предсказанные значения и их стандартные ошибки будут выражены через значения исходного (а не преобразованного) ряда. Уникальной особенностью модели АРПСС является способность анализировать модели с длинными периодами сезонности (с лагом до 30). Стандартный набор результатов содержит оценки параметров, стандартные ошибки и корреляции. Предсказанные значения могут быть представлены в числовой и графической форме и добавлены к исходному ряду. Имеются многочисленные дополнительные функции для исследования остатков модели АРПСС, в том числе большой набор графических средств.

Нажмите кнопку **ARIMA & autocorrelation functions.** Откроется диалоговое окно процедуры (рис.18.2).

Анализ временного ряда начнем с графической иллюстрации ряда. Выберите вкладку **Review Series** (просмотреть серии). Верхняя часть окна — информационная, здесь записывается имя ряда и его преобразования.

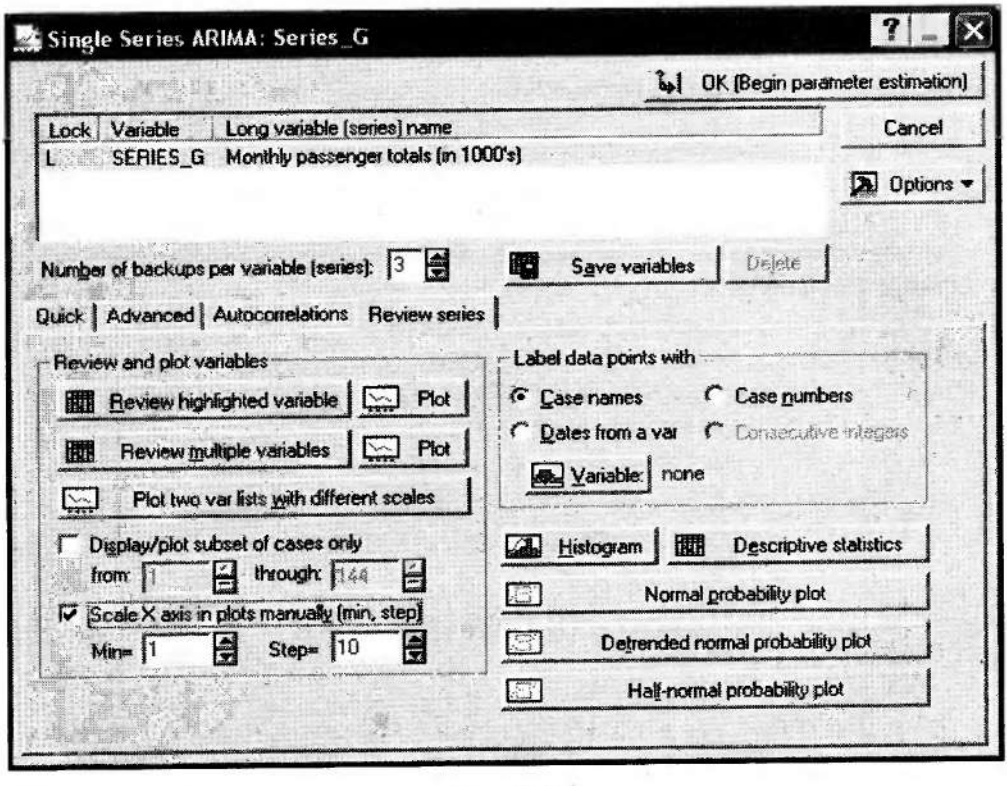


Рис. 18.2

В левой части окна в рамке **Review and plot variables** (просмотреть переменные и построить график) расположены следующие кнопки и опции:

- **Display/plot subset only** — показать на экране / построить график только подмножества;
- **Review highlighted variables** — просмотреть высвеченные переменные;
- **Review multiple variables** — просмотреть несколько переменных.
- **Plot** — график;
- **Plot two var list with different scales** — графики переменных из двух списков в различных шкалах.

В правой части окна в рамке **Label data points with** (метки точек данных) расположены опции обозначений наблюдений на графике (*Case names* — имена наблюдений, *Case numbers* — номера наблюдений, *Dates from a var* — значения переменной, имя которой указывается при помощи кнопки **Variable**) и кнопки для построения различных типов вероятностных графиков.

Перед построением графиков необходимо выставить нужные обозначения шкал *X* и *Y*. Для этого надо отметить нужные значения минимума (номер наблюдения, с которого строится график) и шага по шкале *X* в опции *Scale X axis in plots manually (min, step)*, а также выбрать метку *Case names* (пометить именами наблюдений) для шкалы *Y* в опциях *Label data points with*, как показано на рис. 18.2.

Далее с помощью кнопки **Plot** рядом с кнопкой **Review highlighted variables** (просмотр выделенной переменной) постройте график ряда (рис.18.3), который позволит произвести предварительную, визуальную оценку данных ряда.

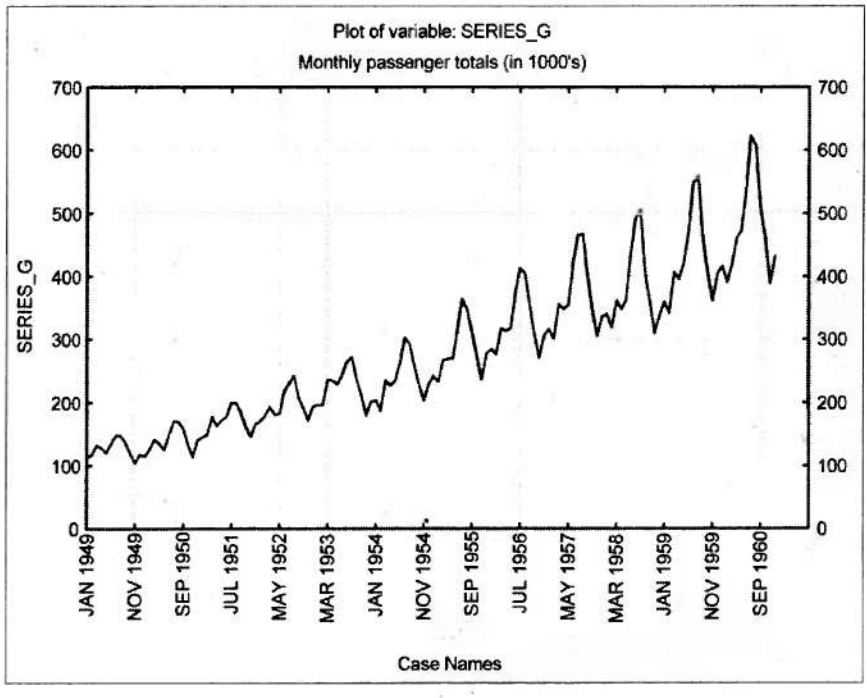


Рис. 18.3

Для построения «качественного» прогноза уровни исследуемого показателя (в нашем случае объемы перевозок) обязательно должны быть *сопоставимы* и *однородны*, а для выявления тенденций, кроме этого, *устойчивы* и *полны*, т.е. количество наблюдений должно быть достаточно велико.

Сопоставимость предполагает формирование всех уровней по одной и той же методике, использование одинаковой единицы измерения и шага наблюдений.

Требование *однородности* данных предполагает отсутствие сильных изломов тенденций, а также нетипичных, аномальных наблюдений. При поиске тенденций бывает целесообразно отбросить часть прошлых данных, если они отражают уже утратившую силу закономерность прошлого развития. Наличие аномальных (резко выделяющихся) наблюдений приводит к искажению результатов. Формально она проявляется как сильный скачок (спад) с последующим приблизительным восстановлением предыдущего уровня.

Устойчивость характеризует преобладание закономерности над случайностью в изменении уровней ряда. На графиках устойчивых временных рядов даже визуально прослеживается закономерность, на графиках неустойчивых рядов изменения последовательных уровней представляются хаотичными, и поэтому поиск закономерностей в формировании значений уровней таких рядов лишен смысла.

Требование *полноты* данных обусловливается тем, что закономерность может быть обнаружена лишь при наличии минимально допустимого объема наблюдений.

Из построенного графика видно, что ряд удовлетворяет всем перечисленным требованиям, при этом нет резких скачков, просматривается тренд ряда, который выражается в плавном увеличении объемов перевозок, и некоторая сезонность, проявляемая в периодичности увеличения и уменьшения объемов перевозок. Таким образом, есть достаточно веские аргументы, чтобы рассматривать динамику роста объема перевозок как процесс, имеющий регулярную сезонную составляющую.

Для нахождения такого рода периодичностей используйте спектральный анализ. Для этого воспользуйтесь диалогом **Spectral (Fourier) analysis** (Фурье, спектральный анализ). Появится одноименное диалоговое окно. Нажмите кнопку **OK** (одномерный анализ Фурье). Откроется диалоговое окно с результатами анализа (рис. 18.4). Установите опцию *Plot by* (график) в положение *Period* (период) и нажмите кнопку **Periodogram** (периодограмма). Программа построит график, изображенный на рис. 18.5. Узкий высокий пик на этом графике свидетельствует о наличии регулярных циклов, а широкие пики соответствуют нерегулярным, неустойчивым циклам. График служит подтверждением того, что в ряду наблюдается тренд с переменным периодом. На графике имеется два ярко выраженных пика, причем второй значительно выше первого, что дает основание предположить возможность существования тенденции к формированию устойчивого сезонного цикла с периодом в 18 месяцев.

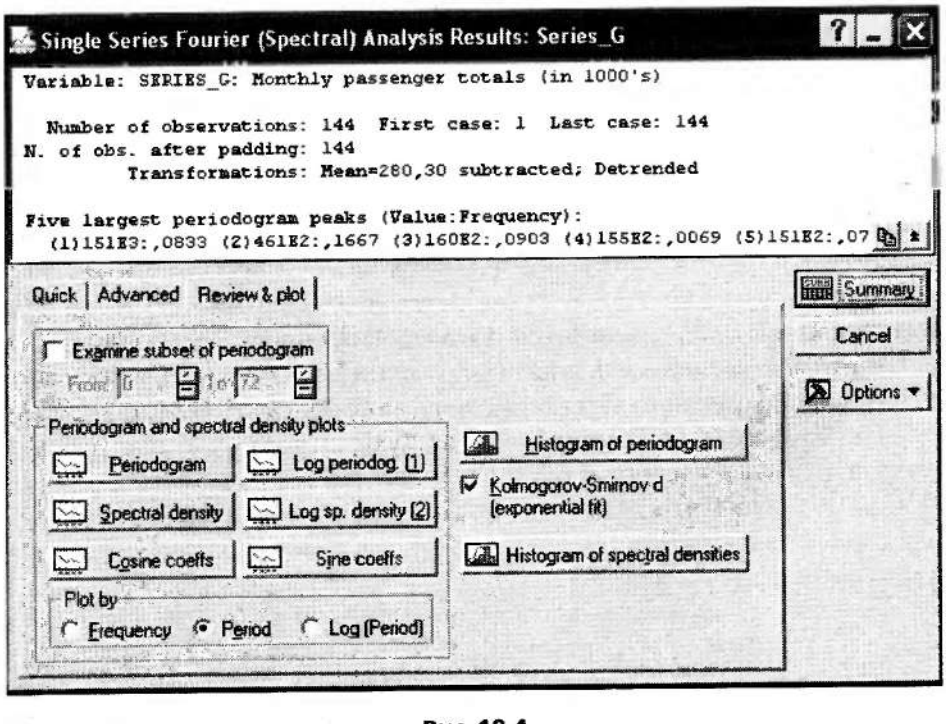


Рис. 18.4

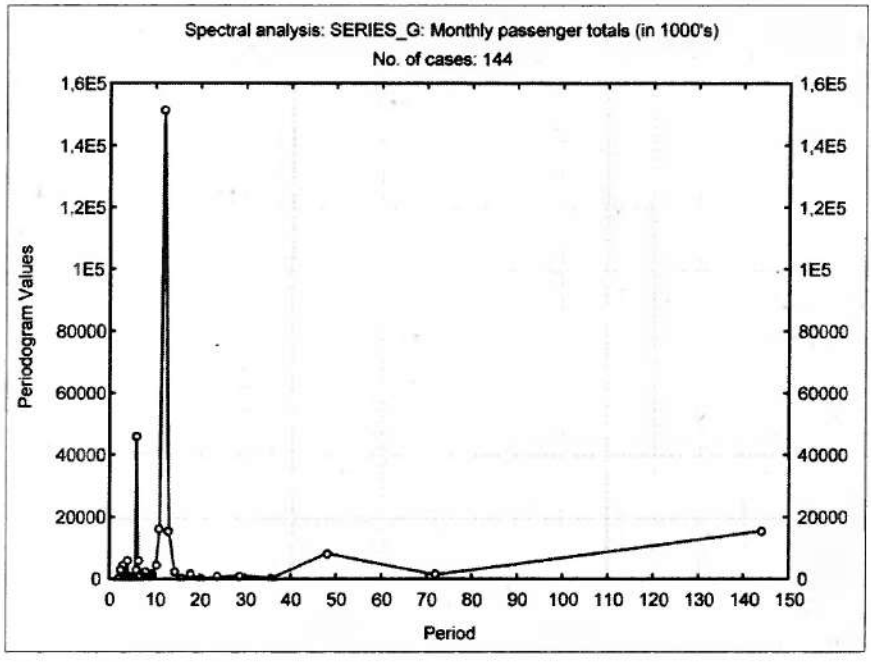


Рис. 18.5

Вернемся к диалогу на вкладке **Advanced** (рис. 18.6) и опишем назначение основных кнопок и опций.

В рамке **Arima model parameters** (параметры АРПСС) расположены опции, задающие число различных параметров модели:

- *p-Autoregressive* — параметр авторегрессии (регулярный);
- *P-Seasonal* — сезонный параметр авторегрессии;
- *q-Moving average* — параметр скользящего среднего (регулярный);
- *Q-Seasonal* — сезонный параметр скользящего среднего.

В поле возле каждого параметра задается число параметров данного типа. По крайней мере, один из параметров должен быть определен. Задание параметров в опциях означает идентификацию, т.е. определение модели. После того как определено количество параметров в модели, можно перейти к их оцениванию.

В рамке **Transform variable prior to analysis** (преобразование переменной до анализа) задаются различные опции преобразования ряда:

- *Natural Log* — логарифмирование по натуральному основанию;
- *Difference* — вычитание;
- *Power Transform* — возведение в степень.

Указанные преобразования не сохраняются после выхода из этого окна, они устанавливаются после проведенных исследований и осуществляются при оценивании параметров модели АРПСС.

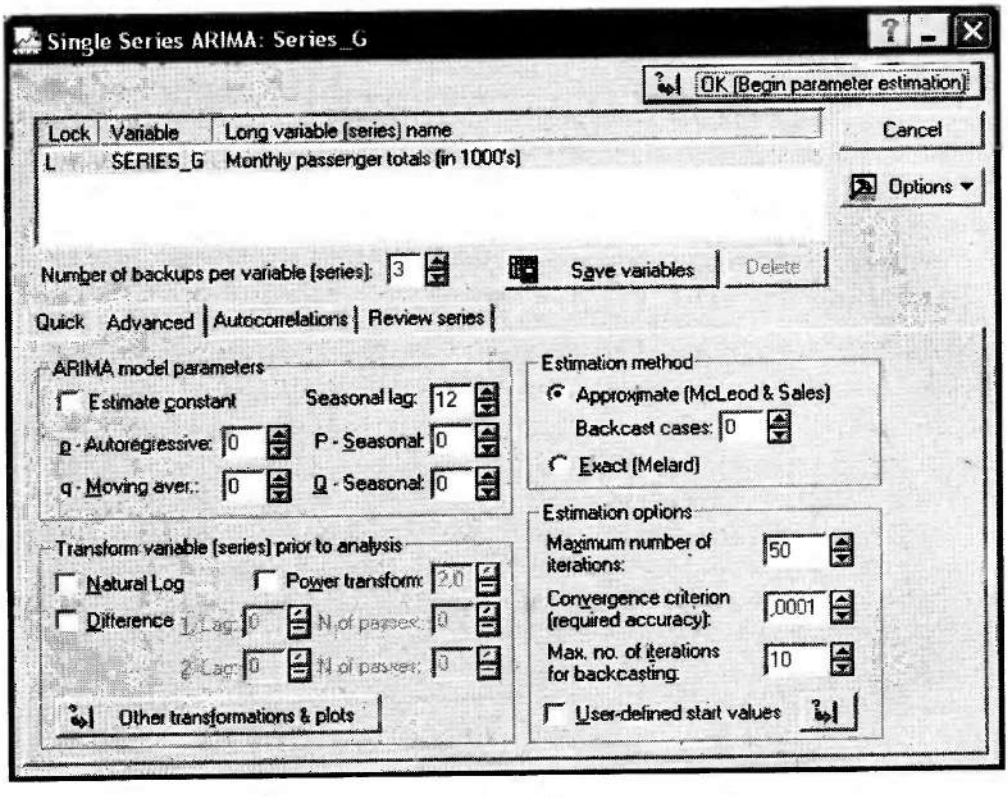


Рис. 18.6

В рамке **Estimation method** (метод оценивания) программа предлагает две вычислительные процедуры оценивания: *Approximate* (приближенная) и *Exact* (точная).

Ниже в рамке **Estimation options** (параметры оценивания) задаются начальные приближения для неизвестных параметров, указывается максимальное число итераций и устанавливается параметр для критерия сходимости процедуры оценивания. Все эти значения могут быть предложены системой.

Кнопка в правом верхнем углу **OK Begin parameter estimation** (начать оценивание параметров) запускает процедуру оценивания параметров.

Кнопка **Other transformation & plots** (другие преобразования и графики) открывает окно процедуры преобразования переменных, которые сохраняются после выхода из окна и высвечиваются в информационной части диалоговых окон. Нажмите эту кнопку, откроется окно (рис. 18.7) **Transformation of variables** (преобразование переменных).

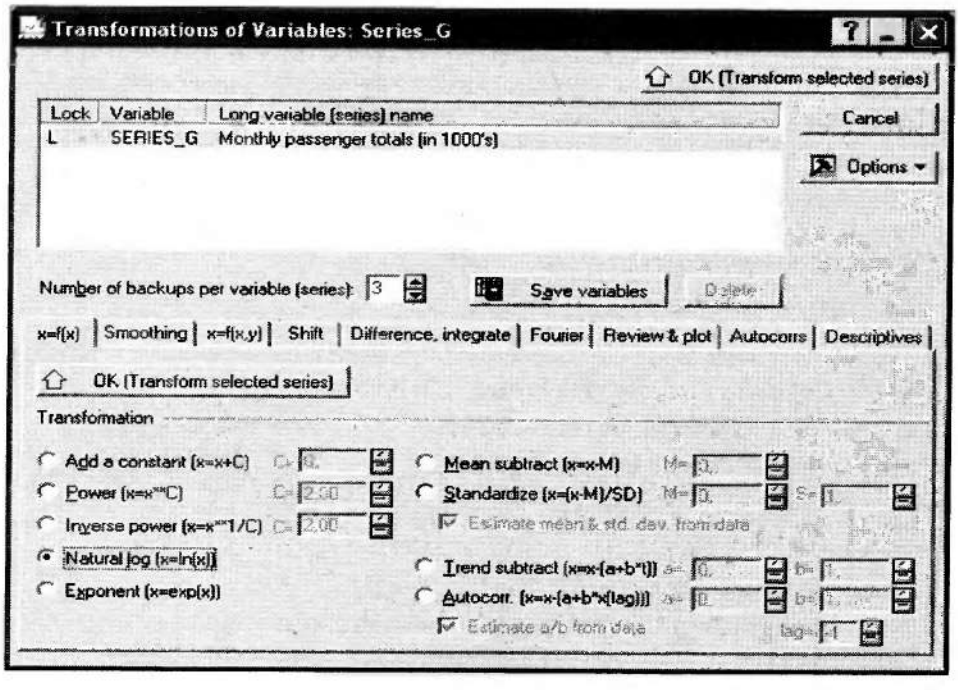


Рис. 18.7

На вкладке **x = f(x)** возможны следующие преобразования:

- *Add a constant* (прибавить константу к значениям ряда);
- *Power* (возвести в степень);
- *Inverse power* (возвести в обратную степень);
- *Natural log* (взять натуральный логарифм);
- *Exponent* (выполнить экспоненциальное преобразование).

В этом окне имеются также следующие опции.

Mean subtract (вычитание среднего). Из значений ряда вычитается среднее значение, подсчитанное по всем наблюдениям, либо численное значение, указанное в поле *M*.

Standardize (стандартизовать). Из значений ряда вычитается величина M и результат делится на *S*. Типичный случай *M* — среднее ряда, *S* = *SD* — стандартное отклонение. Если выбрана опция *Estimate mean & std. dev. from data* (оценить среднее и стандартное отклонение по данным), то значения *M* и *SD* оцениваются по траектории ряда, т.е. являются выборочными значениями.

Trend subtract (вычитание тренда). Из ряда вычитается линейный тренд, параметры которого либо оцениваются, либо задаются в поле *a, b*.

Autocorr ($x = x - (a + b \times x(lag))$) (автокорреляции). Это линейное преобразование, позволяющее занулить автокорреляции на определенном лаге, задаваемом в поле *lag*.

На вкладке **Smoothing** (сглаживание) возможны следующие преобразования ряда:

- *N-pts mov. aver.* (*N*-точечное скользящее среднее);
- *N-pts mov. median.* (*N*-точечное скользящая медиана);
- *Weighted* (усреднение с неравными весами);
- *Prior* (вычисления проводятся по предыдущим значениям ряда);
- *Simple exponential* (простое экспоненциальное сглаживание);
- *p, q 4253H Filter* (4253H Фильтр).

По шагам выполняются следующие преобразования ряда:

1) скользящее медианное сглаживание ряда по четырем точкам, слева и справа от текущей точки берутся по две точки ряда;
2) пятиточечное медианное сглаживание;
3) трехточечное медианное сглаживание;
4) трехточечное сглаживание скользящим средним с весами 0,25; 0,5; 0,25;
5) вычисляются остатки;
6) к остаткам применяются шаги 1–4;
7) преобразованные остатки добавляются к преобразованному ряду.

На вкладке $x = f(x,y)$ возможны следующие преобразования ряда:

- *Difference* (разность), вычисление нового значения ряда x по формуле: $x = x - y(lag)$, где значение *lag* (запаздывание) задается в поле *lag;*
- *Residualizing,* вычисляются новые значения ряда по формуле: $x = x - (a+by(lag))$, где параметры a и b либо задаются, либо оцениваются методом наименьших квадратов. В последнем случае следует выбрать опцию *Estimate a and b from data* (оценить параметры a и b из данных).

Опции этой вкладки доступны, когда анализируются, по крайней мере, два временных ряда (выбраны по меньшей мере две переменные из файла данных).

На вкладке **SHIFT** (сдвиг) доступно преобразование ряда.

Shift relative starting point of series (сдвинуть начальную точку ряда). Сдвигается ряд вперед (*forward*) или назад (*back*).

На вкладке **Differencing, Integrate** (вычитание, суммирование) вычисляются значения нового ряда x по формуле: $x = x - x(lag)$ или $x = x + x(lag)$.

На остальных вкладках доступны процедуры построения различных графиков, вычисление автокорреляций и описательных статистик преобразованных рядов.

Для уменьшения дисперсии анализируемого ряда *Series G* воспользуйтесь преобразованием *Natural log* на вкладке **x = f(x).** Щелкните кнопкой **OK (Transform selected series)**. Программа построит график (рис. 18.8) прологарифмированного по натуральному основанию ряда, дисперсия которого значительно меньше дисперсии исходного (рис. 18.9).

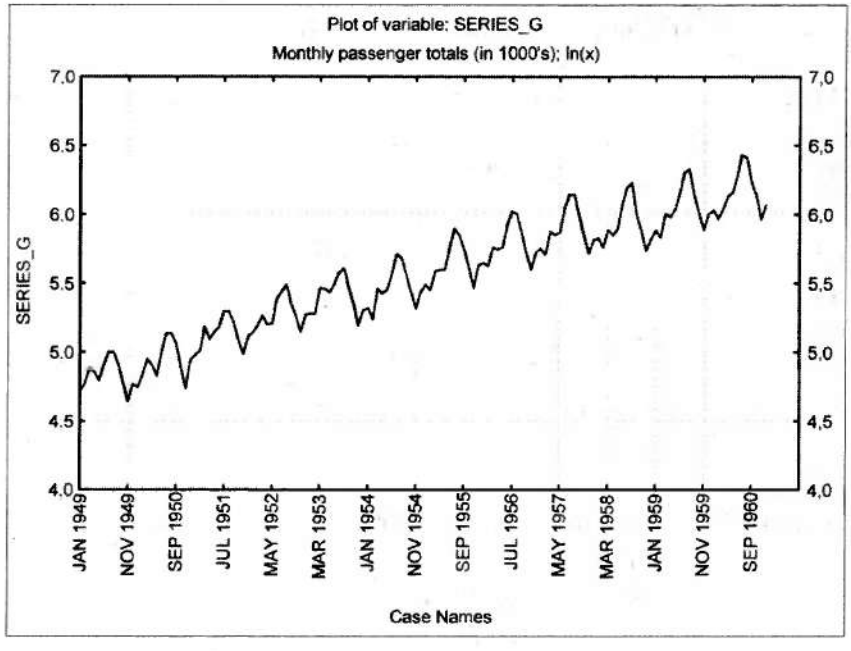


Рис. 18.8

Variable	Descriptive Statistics (Series_G		
	Mean	Std.Dv.	Minimum
SERIES_G: Monthly passenger totals (in 1000's)	280,2986	119,9663	104,0000
SERIES_G: Monthly passenger totals (in 1000's); ln	5,5422	0,4415	4,6444

Рис. 18.9

После уменьшения разброса надо идентифицировать параметры модели для возможности дальнейшего оценивания параметров модели АРПСС. В модели АРПСС имеются следующие типы параметров: p — порядок авторегрессии, d — порядок разности, q — порядок скользящего среднего, сокращенно — модель АРПСС (p, d, q). Идентифицировать модель АРПСС — значит определить эти параметры.

Различают идентификацию порядка разности модели АРПСС — d и идентификацию стационарного процесса или порядка смешанной модели — параметров p, q. Идентификация — довольно грубая процедура, с помощью которой получают прикидочные значения порядка модели. Довольно типично получение на этапе идентификации нескольких приемлемых моделей, которые с достаточной степенью точности подходят к наблюдаемым данным и в дальнейшем подвергаются детальному рассмотрению. Основной критерий идентификации — поведение автокорреляционной и частной автокорреляционной функций ряда. Но в действительности эти функции не известны, и мы имеем дело с их более

или менее точными оценками, которые называются выборочными автокорреляционными и частными автокорреляционными функциями.

Пусть d — неизвестный порядок модели, который нужно оценить. Прежде всего, визуализируем ряд и определим, является ряд стационарным или нет. Нестационарность ряда часто видна на глаз, например, если в ряду имеется ярко выраженный тренд. Особенно легко определить визуально наличие монотонного тренда: логарифмического, экспоненциального, линейного, параболического и др. Из графиков на рис. 18.3 и 18.8 видно, что ряд *Series G* имеет явную тенденцию к возрастанию значений при увеличении номера наблюдения, т.е. наблюдается монотонный тренд.

Наличие тренда, который хорошо виден, — первое свидетельство о нестационарности анализируемого ряда.

Если тренд не выражен ярко и нет других особенностей ряда, указывающих на нестационарность, то следует рассмотреть автокорреляционную функцию, точнее, выборочную автокорреляционную функцию.

Если автокорреляционная функция не имеет тенденции к затуханию, можно говорить о нестационарности ряда.

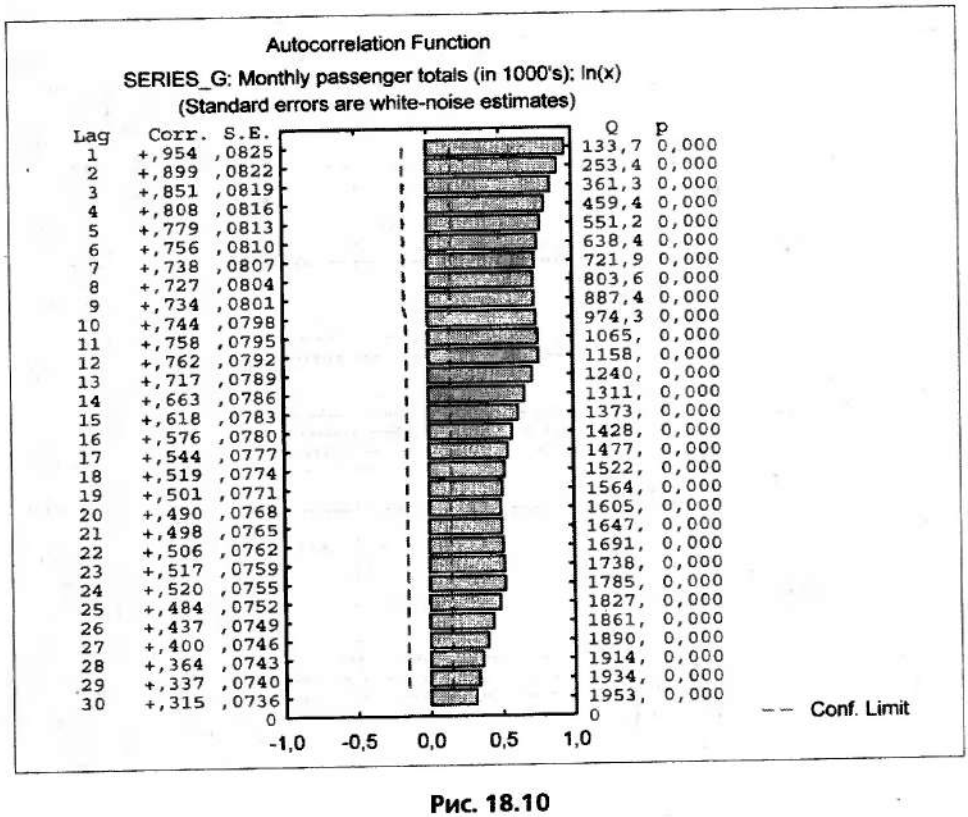


Рис. 18.10

Таким образом, критерий нестационарности выражается в отсутствии тенденции к затуханию у выборочной автокорреляционной функции ряда, т.е. тренд может быть не виден отчетливо на графике, однако нестационарность проявится с помощью данного критерия. На вкладке **Autocorrs** при помощи опции *Number of lag* увеличьте число лагов до 30 и нажмите кнопку **Autocorelations**. Появится график автокорреляционной функции (рис. 18.10). Из построенного графика видно, что автокорреляционная функция имеет слабую немонотонную тенденцию затухания (тот же вывод справедлив и для исходного, не прологарифмированного ряда).

Итак, рассмотрев график автокорреляционной функции, приходим к одному из следующих выводов: 1) ряд стационарен; 2) ряд нестационарен.

В применении к модели АРПСС первый случай означает, что $d = 0$ и следует перейти к определению остальных параметров модели, т.е. p, q. Второй случай означает нестационарность ряда. Тогда нужно рассмотреть разность первого порядка наблюдаемого ряда, предполагая, что ряд первых разностей будет стационарным.

Здесь опять возможны два вывода. Если приходят к заключению, что ряд первых разностей нестационарен, то вновь берут его разности первого порядка и используют критерий стационарности. Так как разности первого порядка применялись последовательно дважды, то это означает, что к исходному ряду применен разностный оператор второго порядка. На практике процедуру последовательного взятия разностей редко применяют больше двух раз, так как редко встречаются модели с порядком разности, большим 2. То есть процедура заканчивается на шаге k, если преобразованный ряд стал стационарным. Если применение процедуры закончено на шаге k, то полагают, что $d = k$. Порядок разности АРПСС определяется как k.

Следует обратить внимание на то, что критерий стационарности носит нестрогий характер, потому что в нем используются не точные автокорреляционные функции, а их оценки. Кроме того, используются не сами оценки, а графики функций, отсюда следует: критерий допускает довольно широкое толкование и, возможно, найдется несколько приемлемых значений для порядка разности d, что необходимо учитывать на практике.

Однако не нужно брать слишком большие значения d, по крайней мере, при начальном исследовании данных. Поэтому определим для ряда *Series G ln(x)* $d = k = 1$. На вкладке **Differencing, Integrate** (рис. 18.11) возьмите разность первого порядка, выделив опцию *Differencing* ($x = x - x(lag)$) и указав значение $lag = 1$.

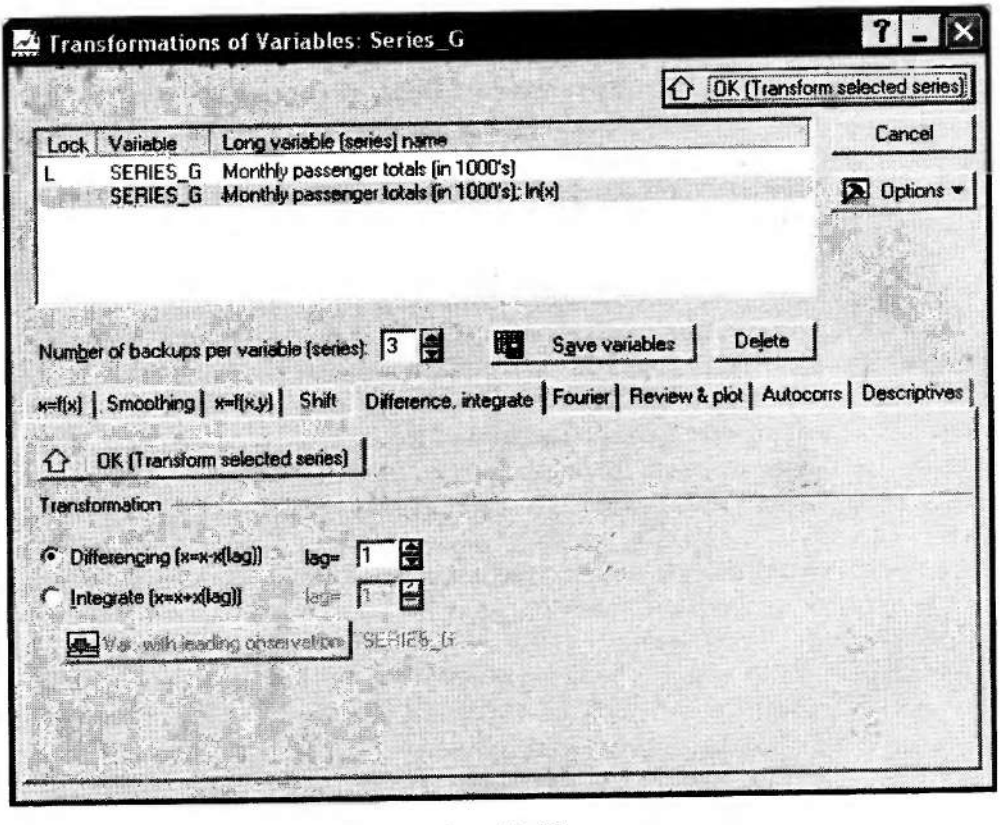


Рис. 18.11

Щелкните кнопкой **OK (Transform selected series)**. Из графика (рис. 18.12) видно, что ряд стал стационарным, так как нет тренда.

Автокорреляционная функция (рис. 18.13) имеет незначительный выброс на лаге 1 и слабую тенденцию к затуханию, если не считать пиков устойчивого сезонного цикла с периодом в 12 месяцев.

Частная автокорреляционная функция (рис. 18.14), осциллируя, экспоненциально приближается к нулю. Таким образом, после двух преобразований построена стационарная модель.

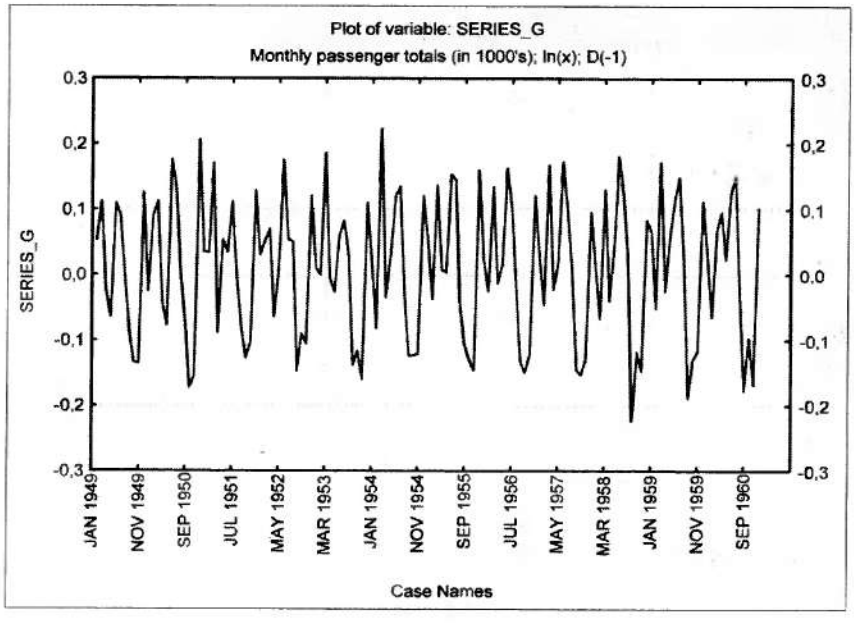


Рис. 18.12

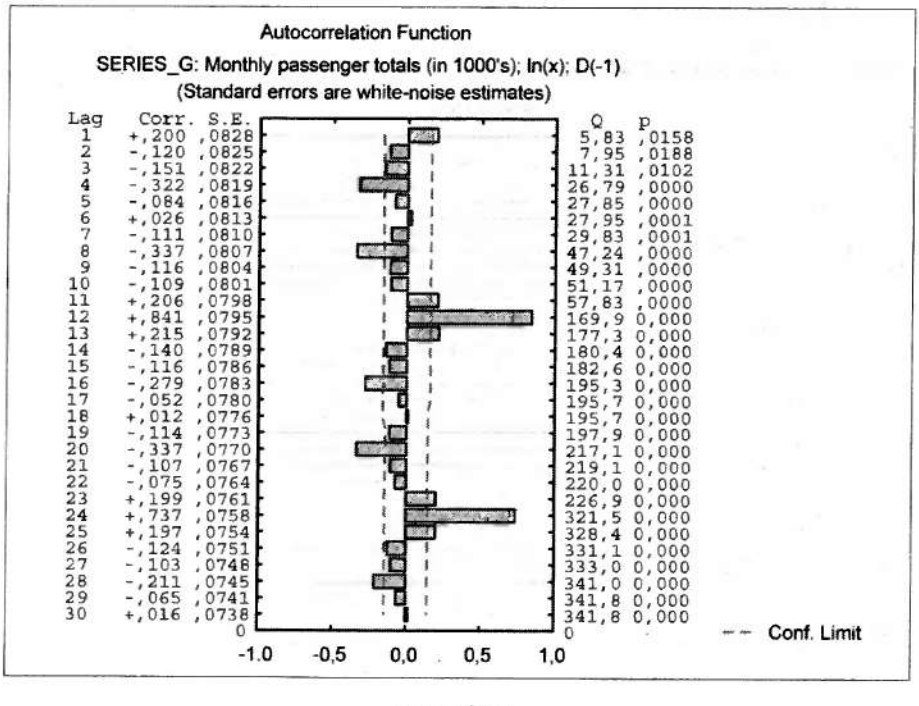


Рис. 18.13

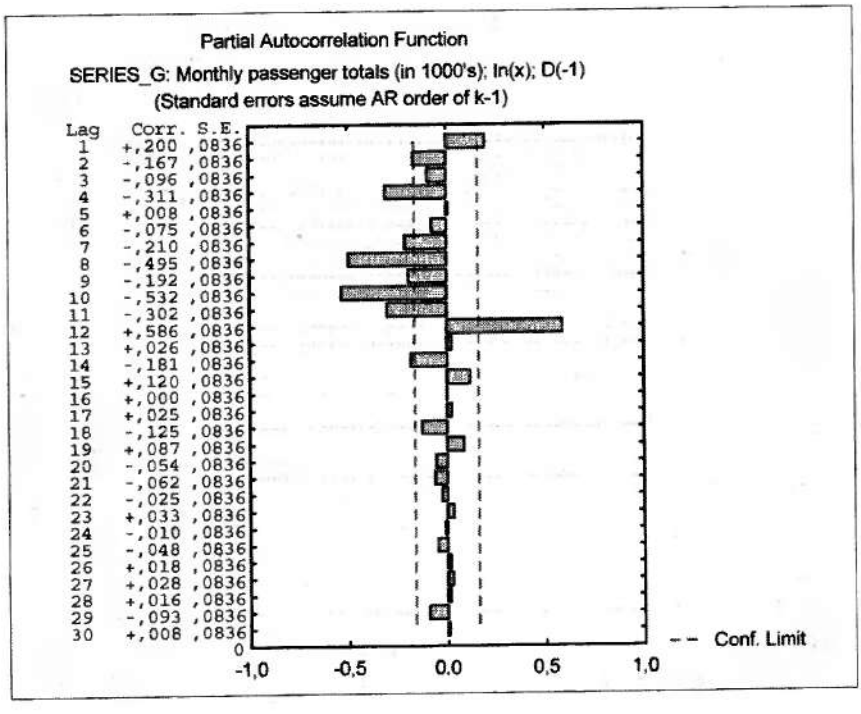


Рис. 18.14

Осуществим идентификацию построенной в результате трех преобразований стационарной модели в классе моделей смешанная авторегрессия — скользящее среднее, которые при определенных ограничениях на параметры более точно описывают стационарные временные ряды. Как уже было отмечено, идентификация модели заключается в определении параметров модели p и q. Для определения параметров p, q рассматривают поведение выборочных автокорреляционной и частной автокорреляционной функций ряда.

Пусть наблюдается процесс авторегрессии порядка p. Тогда его частная автокорреляционная функция обрывается на лаге p. Автокорреляционная функция плавно спадает. Пусть наблюдается процесс скользящего среднего порядка q. Тогда его автокорреляционная функция обрывается на лаге q. Частная автокорреляционная функция плавно спадает. Автокорреляционная функция в модели, у которой оба параметра не равны нулю, представляется в виде суммы экспонент и затухающих синусоид.

Практика показывает, что большинство наблюдаемых рядов, описываемых смешанной моделью авторегрессии и скользящего среднего, могут быть отнесены с достаточной степенью точности к одному из следующих пяти классов:

- модели авторегрессии с одним параметром: $p = 1, q = 0$;
- модели авторегрессии с двумя параметрами: $p = 2, q = 0$;
- модели скользящего среднего с одним параметром: $p = 0, q = 1$;

- модели скользящего среднего с двумя параметрами: $p = 0, q = 2$;
- модели авторегрессии с одним параметром и скользящего среднего с одним параметром: $p = q = 1$.

Прежде всего нужно попытаться отнести модель к одному из этих классов.

Имеются следующие практические критерии по определению этих моделей с помощью автокорреляционных и частных автокорреляционных функций ряда:

- один параметр авторегрессии: автокорреляционная функция экспоненциально затухает; частная автокорреляционная функция имеет выброс на лаге 1 (нет корреляций для других задержек);
- два параметра авторегрессии: автокорреляционная функция имеет форму затухающей синусоидальной волны или экспоненциально затухает; частная автокорреляционная функция имеет выброс только для сдвигов 1 и 2 (нет корреляций для других задержек);
- один параметр скользящего среднего: автокорреляционная функция имеет выброс на лаге 1 (нет корреляций для других задержек); частная автокорреляционная функция экспоненциально затухает — либо монотонно, либо осциллируя, т.е. меняя знак;
- два параметра скользящего среднего: автокорреляционная функция имеет выбросы на сдвигах 1 и 2 (нет корреляций для других задержек); частная автокорреляционная функция имеет форму синусоидальной волны или экспоненциально затухает;
- один параметр авторегрессии и один параметр скользящего среднего: автокорреляционная функция экспоненциально затухает, начиная с первой задержки (первое значение не нулевое), затухание может быть монотонное и колебательное; в частной автокорреляционной функции преобладает затухающий экспоненциальный член — либо монотонный, либо осциллирующий (первое значение не нулевое).

В начале анализа лучше использовать более простые критерии, однако для окончательного решения следует применять совокупность критериев. Критерии носят достаточно расплывчатый характер, возможно, с их помощью будет идентифицирована и не одна модель. Наличие нескольких подходящих моделей следует рассматривать не как фатальную ошибку, а как нормальный поисковый результат.

Критерии для чистых моделей авторегрессии и скользящего среднего двойственны в том смысле, что одни получаются из других заменой слов «автокорреляционная функция» на «частная автокорреляционная функция».

Заметим, что выборочные автокорреляционные и частные автокорреляционные функции являются состоятельными оценками теоретических автокорреляционных и частных автокорреляционных функций ряда. Точность этих оценок зависит как от длины ряда (при увеличении длины ряда выборочные автокорреляционная и частная автокорреляционная функции сходятся по вероятности к теоретической автокорреляционной и частной автокорреляционной функциям ряда), так и от его природы.

Как показывает практика, на этапе идентификации целесообразно определить несколько подходящих моделей и затем, оценив их параметры, и исследовав остатки, оценить адекватность моделей, после чего выбрать наилучшую модель из нескольких возможных.

Программа *STATISTICA* позволяет легко анализировать модели АРПСС. И с точки зрения временных затрат практически нет разницы: иметь дело лишь с одной моделью, оценивать далее ее параметры и строить прогноз или искать наилучшую среди нескольких подходящих.

Анализируя поведение автокорреляционной и частной автокорреляционной функций и учитывая приведенные критерии, можно сделать вывод, что наиболее подходящей моделью для ряда *Series G ln(x) D(–1)* будет модель — один параметр скользящего среднего, $p = 0$, $q = 1$. Учитывая, что $d = 1$, имеем несезонную модель АРПСС (0, 1, 1).

Так как ряд имеет ярко выраженную сезонную составляющую с периодом в 12 месяцев, надо внести в нашу модель сезонную корректировку. Сезонные модели АРПСС, реализованные в программе *STATISTICA*, являются обобщением обычных моделей АРПСС. Полная мультипликативная сезонная модель может быть представлена в виде АРПСС (*p*, *d*, *q*) (*Ps*, *Ds*, *Qs*), где к параметрам модели АРПСС *p*, *d*, *q* добавлены сезонные параметры: сезонный параметр авторегрессии — *Ps*, сезонная разность — *Ds*, сезонный параметр скользящего среднего — *Qs*. В целом идентификация полной модели АРПСС производится тем же способом, что и идентификация несезонной модели АРПСС. Поведение автокорреляционных и частных автокорреляционных функций на начальных лагах позволяет идентифицировать стандартным образом несезонную компоненту. Поведение автокорреляционных и частных автокорреляционных функций на лагах, кратных сезонному лагу, также стандартным образом позволяет идентифицировать сезонную составляющую. Приведенные практические критерии остаются в силе.

Для того чтобы учесть сезонные колебания с периодом в 12 месяцев, необходимо взять сезонную разность с лагом 12 ряда *Series G ln(x) D(–1)*. Вернитесь в окно **Transformations of variables.** На вкладке **Differencing, Integrate** (рис. 18.11) возьмите разность двенадцатого порядка, выделив опцию *Differencing* (*x = x – x(lag)*) и указав значение *lag* = 12. Щелкните кнопкой **OK (Transform selected series)**. На рис. 18.15–18.17 приведены графики построенного временного ряда, его автокорреляционной и частной автокорреляционной функций.

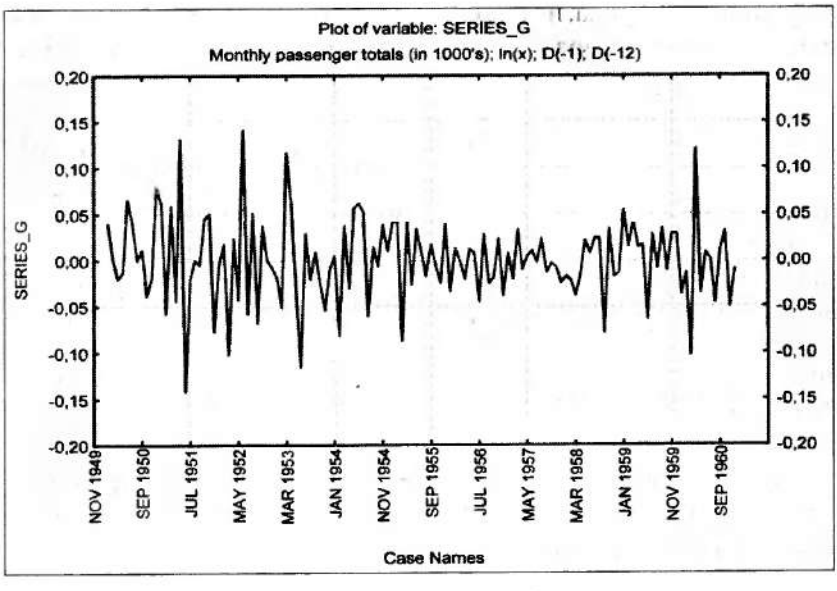


Рис. 18.15

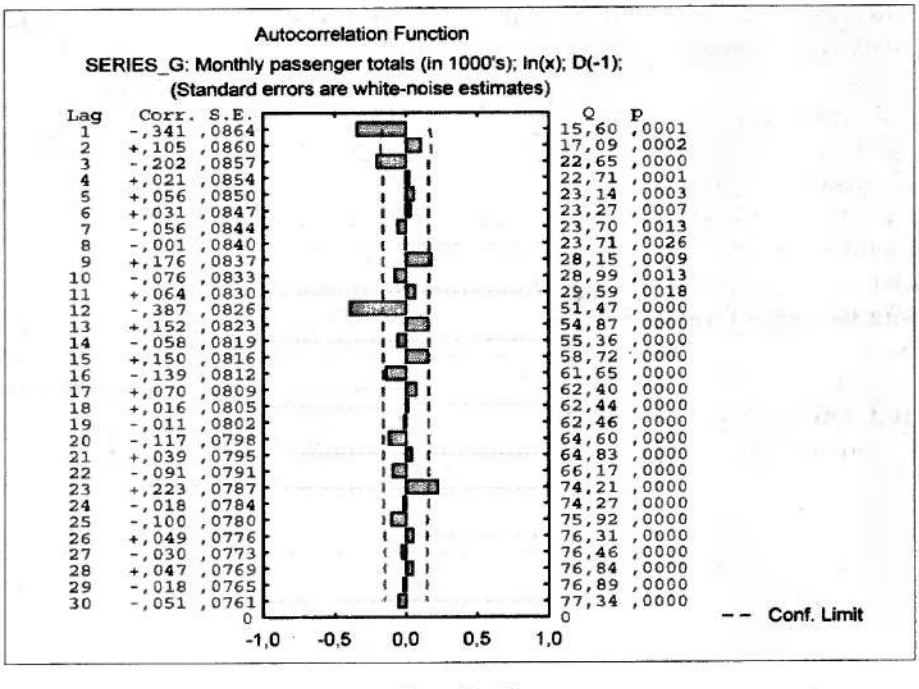


Рис. 18.16

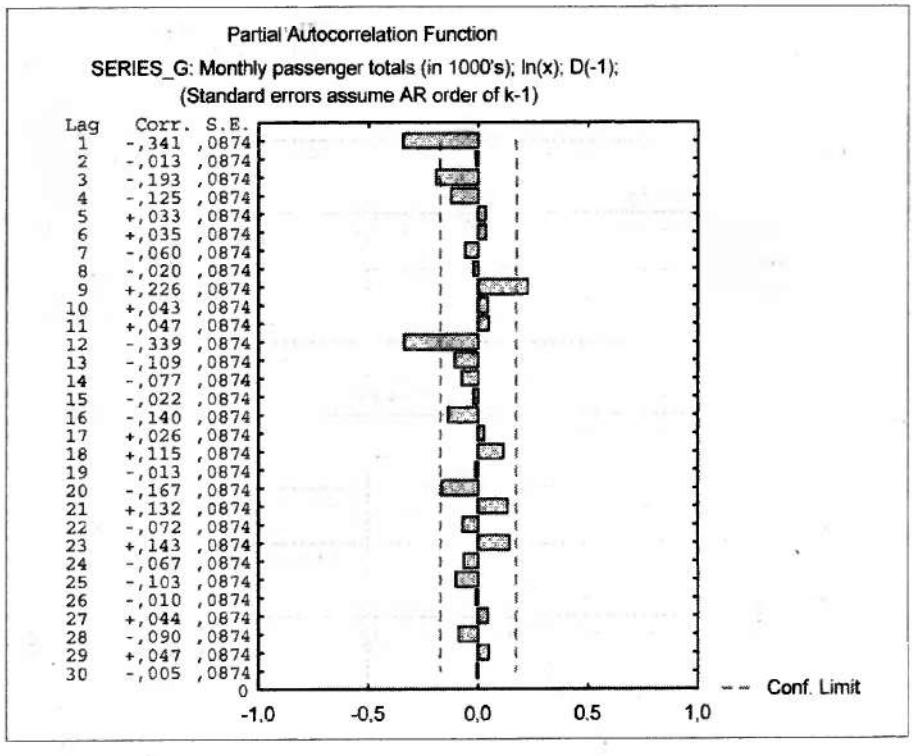


Рис. 18.17

Из графиков видно, что ряд является стационарным, автокорреляционная функция экспоненциально затухает, а частная автокорреляционная функция также экспоненциально затухает и имеет выброс на лаге 1.

Значит, можно определить сезонный параметр авторегрессии $Ps = 0$, сезонную разность $Ds = 1$, сезонный параметр скользящего среднего $Qs = 1$. Мы имеем дело с сезонной моделью АРПСС (0, 1, 1). Таким образом, полную модель определим как АРПСС (0, 1, 1), (0, 1, 1).

Ранее было выполнено логарифмическое преобразование данных и взяты две разности. Все эти преобразования уже выполнены и результаты просмотрены. Преобразованный ряд можно теперь непосредственно использовать в АРПСС. Однако в ситуациях, похожих на данную, рекомендуется анализировать исходный ряд и задать необходимые преобразования внутри АРПСС (эти преобразования будут частью спецификации АРПСС). Если вы захотите построить прогноз после оценки параметров АРПСС, то он будет вычислен из проинтегрированных рядов (интегрирование, более точно суммирование, в данном случае означает просто операцию, обратную взятию разностей с соответствующими лагами). Таким образом, проводя обратные преобразования, вы возвращаетесь к исходному ряду, и прогноз соответствует исходным данным, что обеспечивает более легкую интерпретацию результатов.

Заметим, внутри АРПСС доступны только преобразования: логарифм, возведение в степень и взятие сезонных/несезонных разностей. В некоторых случаях

определенные преобразования рекомендуется выполнять до работы в АРПСС. Речь идет о преобразованиях (например, сглаживание), не изменяющих диапазон данных, к которым не нужно применять обратные преобразования.

Теперь снова вернитесь в диалоговое окно **Single Series ARIMA**, нажав **Cancel** в окне **Transformation of variables**. В диалоговом окне высвечена исходная переменная *Series G*. В рамке **Arima model parameters** установите значения соответствующих параметров (рис. 18.18). Параметры АРПСС оцениваются максимизацией функции правдоподобия. Доступны два метода максимизации функции правдоподобия: *Approximate (McLeod & Sales)* — приближенный (МакЛеода и Сейлза) и *Exact (Melard)* — точный (Меларда). Далее нажмите **OK** (начать оценивание параметров) и запустите итеративную процедуру оценивания.

После того как процедура оценивания сойдется, откроется диалоговое окно (рис. 18.19) **Single Series ARIMA Results** (результаты одномерной АРПСС). Нажмите кнопку **Summary: Parameter estimates** (оценки параметров), чтобы увидеть таблицу (рис.18.20) результатов с оценками, стандартными ошибками, асимптотическими значениями *t-статистик* и т.д.

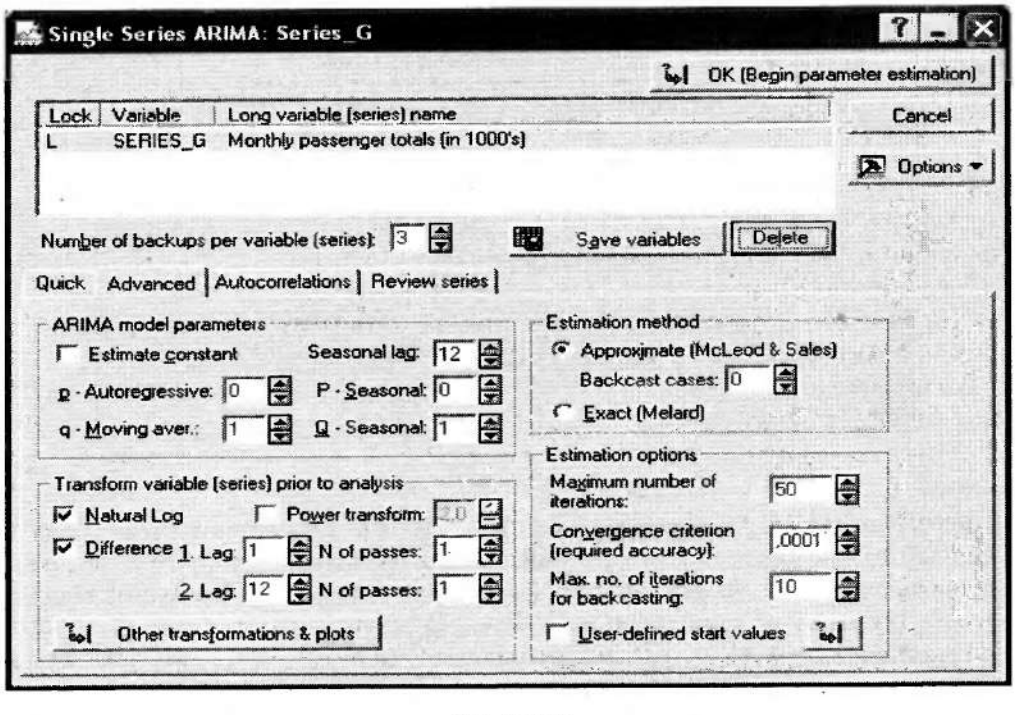


Рис. 18.18

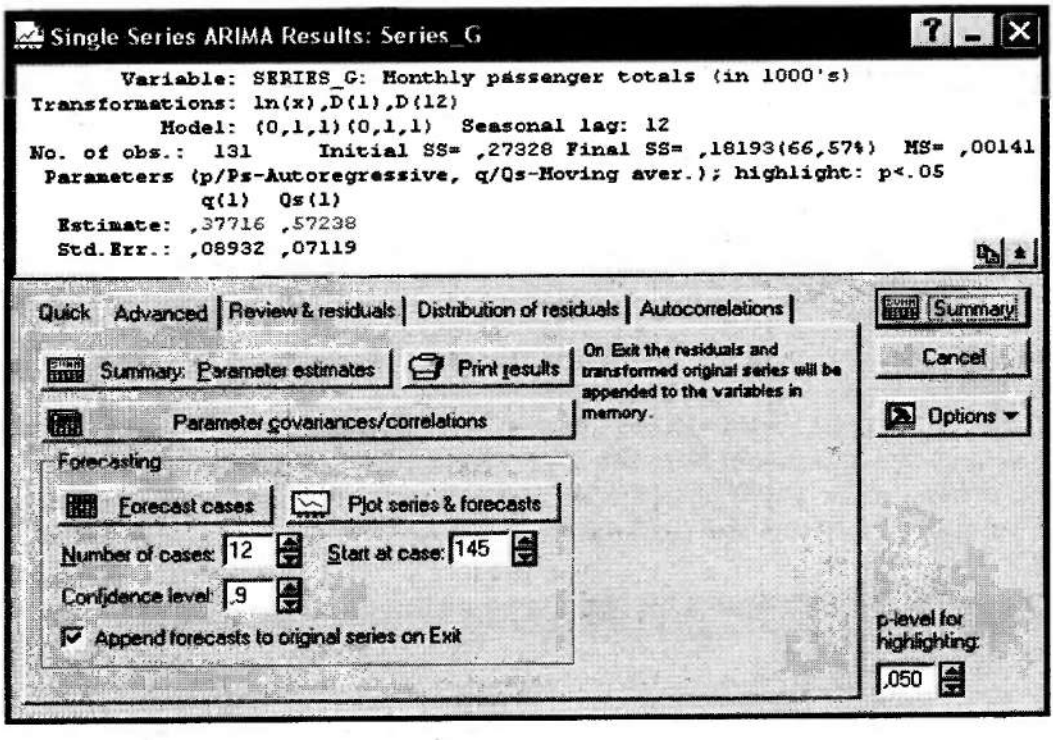


Рис. 18.19

Paramet.	Input: SERIES_G: Monthly passenger totals (in 1000's) (Series_ Transformations: ln(x),D(1),D(12) Model:(0,1,1)(0,1,1) Seasonal lag: 12 MS Residual=,00141					
	Param.	Asympt. Std.Err.	Asympt. t(129)	p	Lower 95% Conf	Upper 95% Conf
q(1)	0,377162	0,089318	4,222697	0,000045	0,200445	0,553880
Qs(1)	0,572379	0,071189	8,040233	0,000000	0,431529	0,713229

Рис. 18.20

Из таблицы (рис. 18.20) видно, что оценки обоих параметров высоко значимы (p значительно меньше 0,05). По умолчанию программа вычисляет прогнозы для одного полного сезонного цикла, начиная с последнего наблюдения. Прежде всего посмотрите прогнозы в таблице результатов.

Нажмите кнопку **Forecast cases** (прогноз). Таблица результатов (рис. 18.21) содержит прогнозы и их доверительные интервалы для наблюдений, начиная с 145-го (т.е. строит прогноз на 12 наблюдений). Нажмите кнопку **Plot series & forecasts** (график ряда и прогнозов). Программа построит прогноз на 12 месяцев (рис. 18.22).

Заметим, что если запросить построить прогнозы для имеющихся наблюдений (что также возможно), таблица результатов будет содержать наблюдаемые значения и остатки.

CaseNo.	Forecasts; Model:(0,1,1)(0,1,1) Se Input: SERIES_G: Monthly passer Start of origin: 1 End of origin: 144 Forecast	Lower 90,0000%	Upper 90,0000%
145	450,1171	422,9655	479,0117
146	425,6620	395,5777	458,0341
147	479,5240	441,3696	520,9766
148	492,0412	449,0088	539,1979
149	508,5479	460,4357	561,6874
150	583,0166	524,0264	648,6473
151	669,1520	597,3584	749,5742
152	666,4152	591,1003	751,3264
153	557,9980	491,9233	632,9478
154	496,7552	435,3899	566,7696
155	429,6965	374,5207	493,0009
156	477,1535	413,6613	550,3910

Рис. 18.21

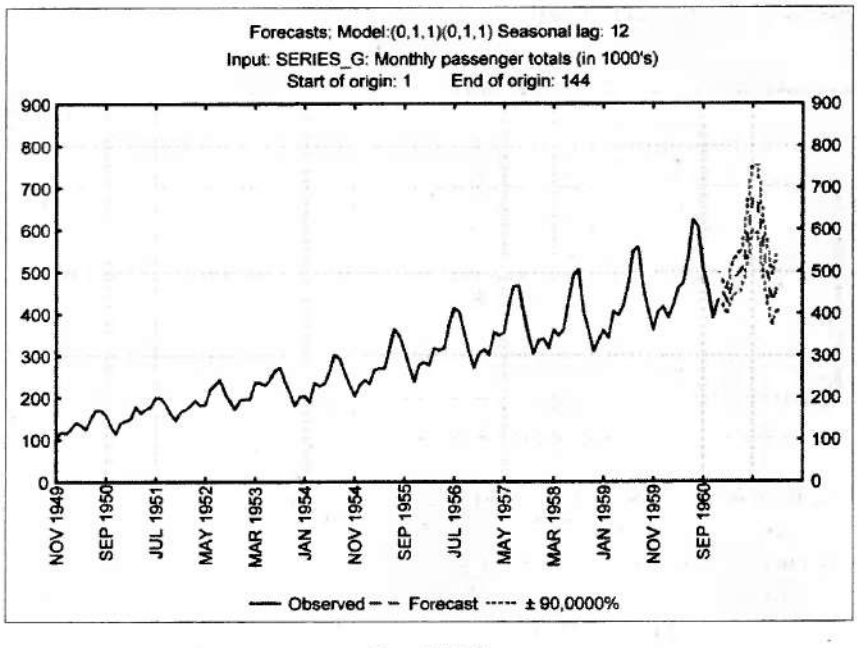


Рис. 18.22

Чтобы проверить, насколько хорошо построенная модель АРПСС прогнозирует последние 12 наблюдений установите в поле **Start at case** (начать с наблюдения) значение 133 и снова нажмите кнопку **Plot series & forecasts.** Из графика (рис. 18.23) видно, что прогнозная кривая практически повторяет фрагмент кривой исходного ряда, причем все наблюдаемые значения ряда попадают в доверительный интервал.

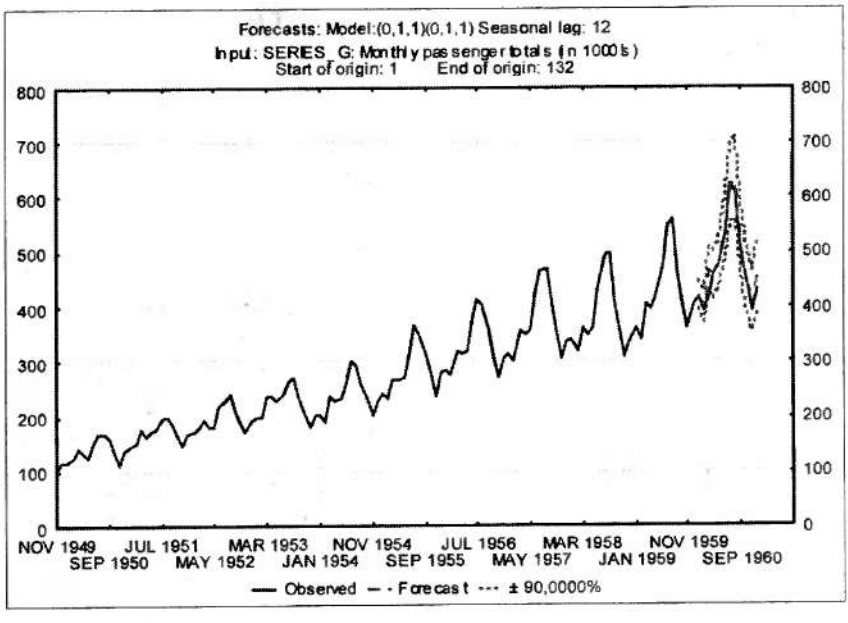


Рис. 18.23

Для анализа адекватности модели исследуют остатки, представляющие собой разности наблюдаемых значений и значений, предсказанных с помощью модели. В программе *STATISTICA* визуализацией гистограммы остатков, графиков автокорреляционных функций, графиков остатков оценивают адекватность модели. На вкладке **Distribution of residuals** нажмите кнопку **Histogram.** Из графика (рис. 18.24) видно, что выборочная плотность распределения остатков успешно аппроксимируется нормальным законом распределения, что является признаком адекватности построенной модели прогноза.

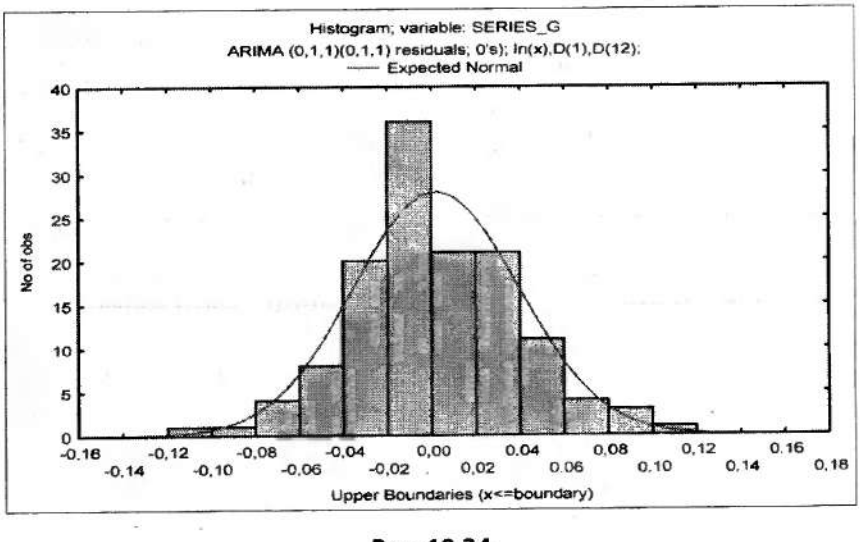


Рис. 18.24

Предположение о нормальности остатков может быть проверено с помощью **Normal probability plot** – нормальных вероятностных графиков. Стандартный нормальный вероятностный график строится следующим образом. Вначале происходит упорядочение отклонений от соответствующих средних (остатков). По этим рангам вычисляются стандартизованные значения нормального распределения и откладываются на оси *Y*. Если наблюдаемые значения (отложенные по оси *X*) нормально распределены, то значения попадут на прямую линию. Если распределение отлично от нормального, то на графике будет наблюдаться сильное отклонение от прямой. На этом графике можно отчетливо увидеть выбросы. Отличие нормальных вероятностных графиков без тренда *(Detrended normal probability plot)* от простых нормальных вероятностных графиков в том, что из данных исключается линейный тренд. На рис. 18.25 приведен *Normal probability plot*, из которого также следует возможность приближения плотности распределения остатков нормальным законом.

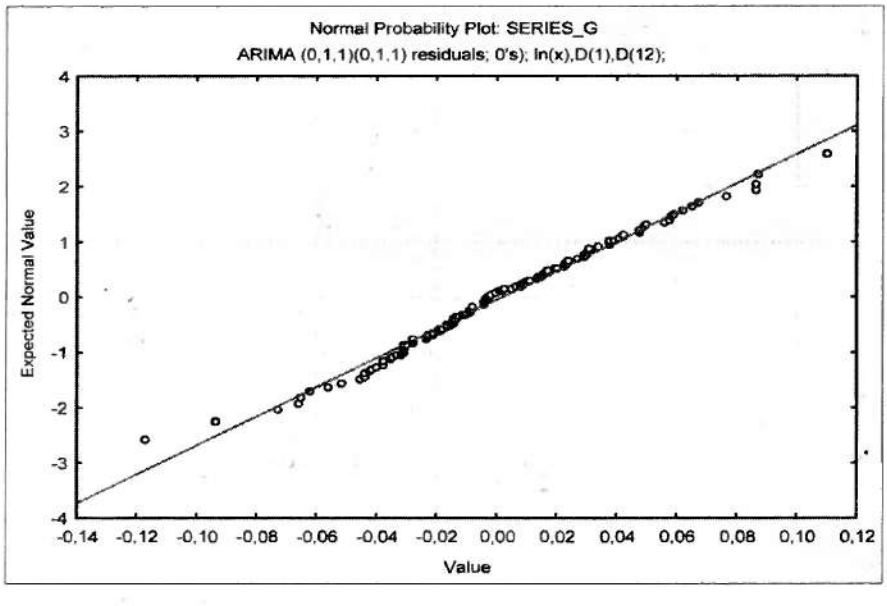


Рис. 18.25

Посмотрите график остатков и убедитесь, что остатки имеют примерно равную вариацию на всем протяжении ряда и нет очевидного тренда или сдвига в них. Для этого откройте вкладку **Review & residuals**. В опции *Review & plot variables* (просмотреть переменные и построить график) нажмите кнопку **Plot(3)** (график остатков) напротив кнопки **Review residuals.** Программа построит график остатков (рис. 18.26), из которого следует, что остатки имеют примерно равную вариацию на всем протяжении ряда и нет очевидного тренда или сдвига в них.

Таким образом, можно сделать вывод: остатки нормально распределены, практически не коррелированны, имеют примерно равную вариацию на всем протяжении ряда и нет очевидного тренда или сдвига в них.

В правильно подобранной модели остатки будут очень похожи на белый шум: в них не будет периодических колебаний, систематических смещений, между ними не будет сильных корреляций. На вкладке **Autocorrelations** нажмите кнопки **Autocorrelations и Partial Autocorrelations.** Из графиков (рис. 18.27–18.28) видно, что остатки практически являются белым шумом.

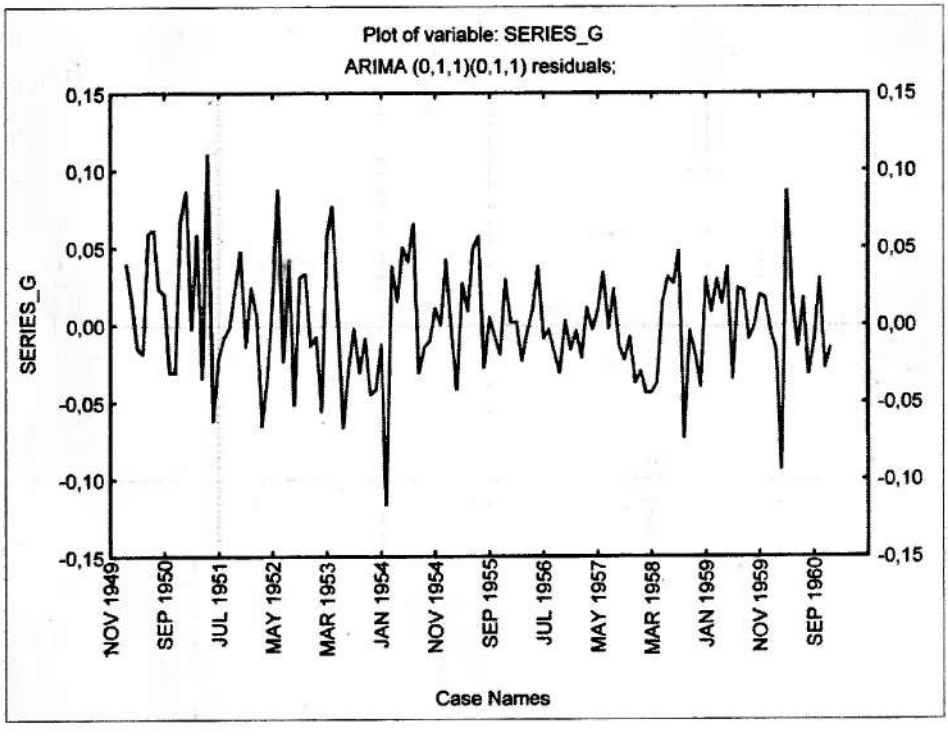


Рис. 18.26

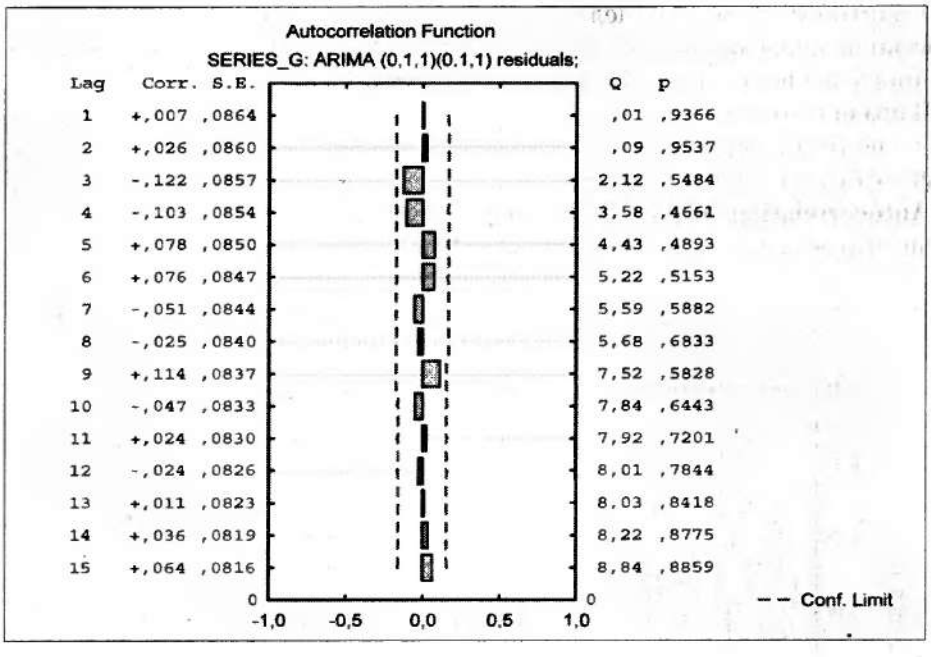


Рис. 18.27

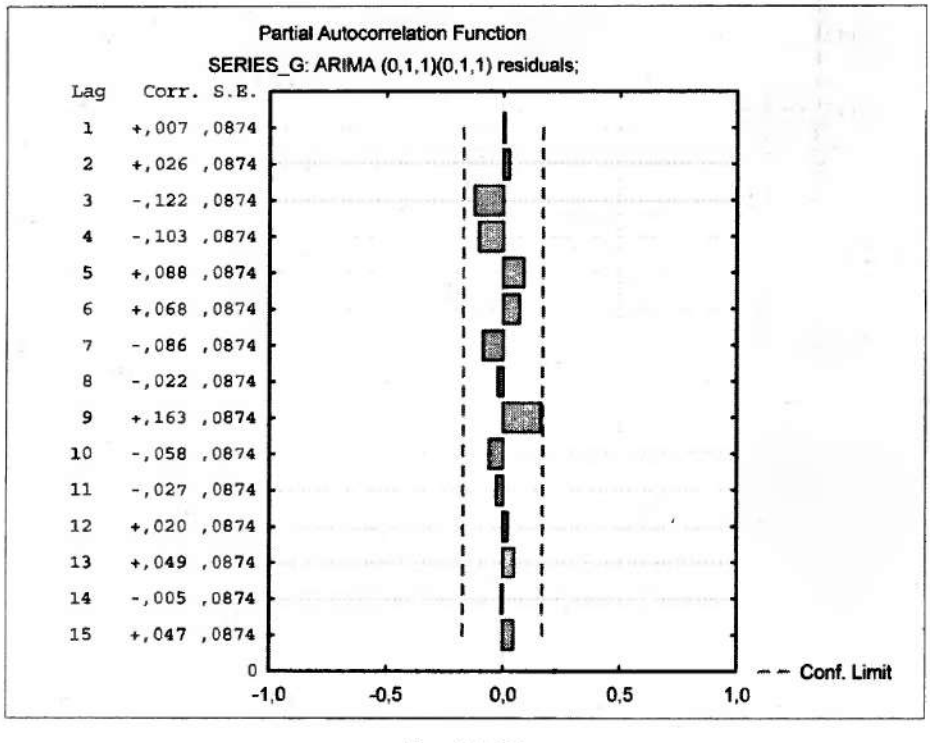


Рис. 18.28

Таким образом, всесторонний анализ остатков показал, что при помощи процедуры **ARIMA & autocorrelation functions** построена вполне адекватная модель прогноза объема месячных авиаперевозок по файлу данных **Series G** из библиотеки **Examples**.

18.2. Модель интервенции для АРПСС

Необходимость в анализе моделей с интервенцией возникает, когда с некоторого момента резко изменяется поведение ряда в силу внешних причин. Внешнее воздействие на ряд может быть кратковременным (импульсивным) и длительным (устойчивым). В момент воздействия траектория резко меняется, но далее вновь описывается моделью АРПСС. Имеется возможность использовать одновременно несколько различных интервенций (до 6). Доступны следующие виды интервенций: однопараметрические скачкообразные, двухпараметрические постепенные, временные (характер воздействия можно просмотреть на графике). Для всех прерванных рядов могут быть построены прогнозы, которые можно вывести на график (вместе с исходным рядом) и, если требуется, добавить прогнозы к исходному ряду. Краткое описание процедуры проведем на примере файла из **ExamplesDatasets — director**, в котором приведен ряд количества ежемесячных звонков помощнику директора.

В стартовой панели модуля **Time Series Analysis/Forecasting** (рис. 18.1) нажмите кнопку **Interrupted Time Series** (анализ прерванных временных рядов). Появится стартовое окно **Interrupted Time Series ARIMA (Intervention Analysis)** (прерванная АРИМА (анализ интервенций)).

Выберите переменную *CALLS* и на вкладке **Review series** (рис. 18.29). нажмите первую кнопку **Plot**, программа построит график ряда, изображенный на рис. 18.30. На этом графике хорошо видны резкие скачки значений ряда, характерные для прерванных рядов. При выборе вкладки **Advanced** откроется стартовое окно процедуры (рис. 18.31), где можно задать установки для подгонки к ряду модели АРПСС с интервенцией. Данное окно отличается от стартового окна АРПСС только в правой части, в группе опций **Specify times and type of interventions** (задать время и тип интервенции). С помощью этих опций задаются:

- *Intervention* — номер интервенции (можно задать до 6 различных интервенций).
- *At case number* — номер случая, в котором интервенция началась;
- *Type of intervention* — тип интервенции.

В процедуре предусмотрены три типа интервенции. Чтобы просмотреть их графики, выберите вкладку **Review impact patterns** (рис. 18.32).

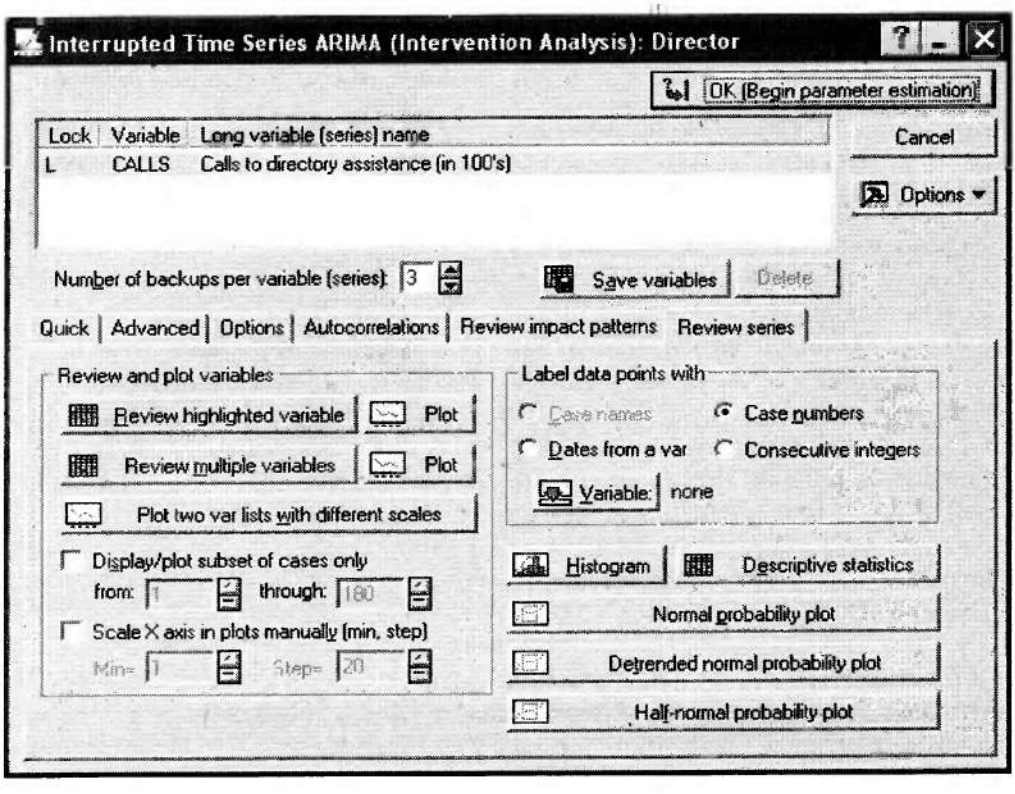


Рис. 18.29

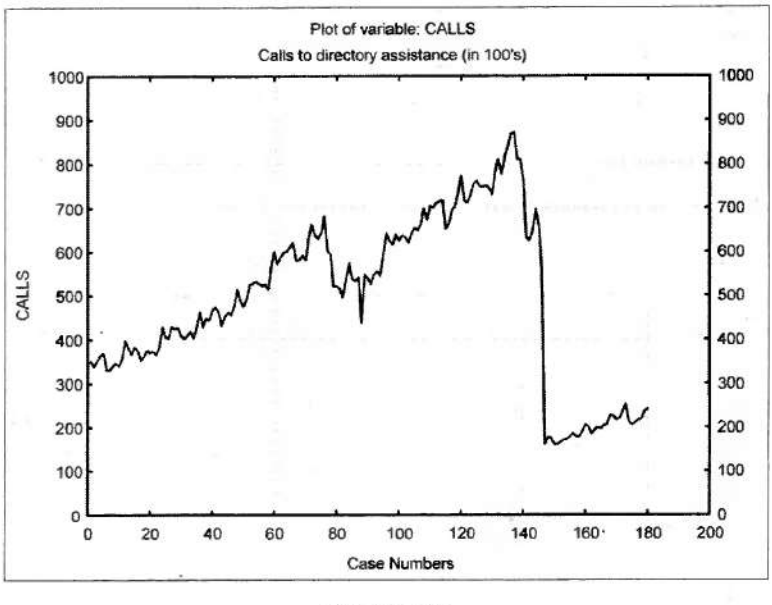


Рис. 18. 30

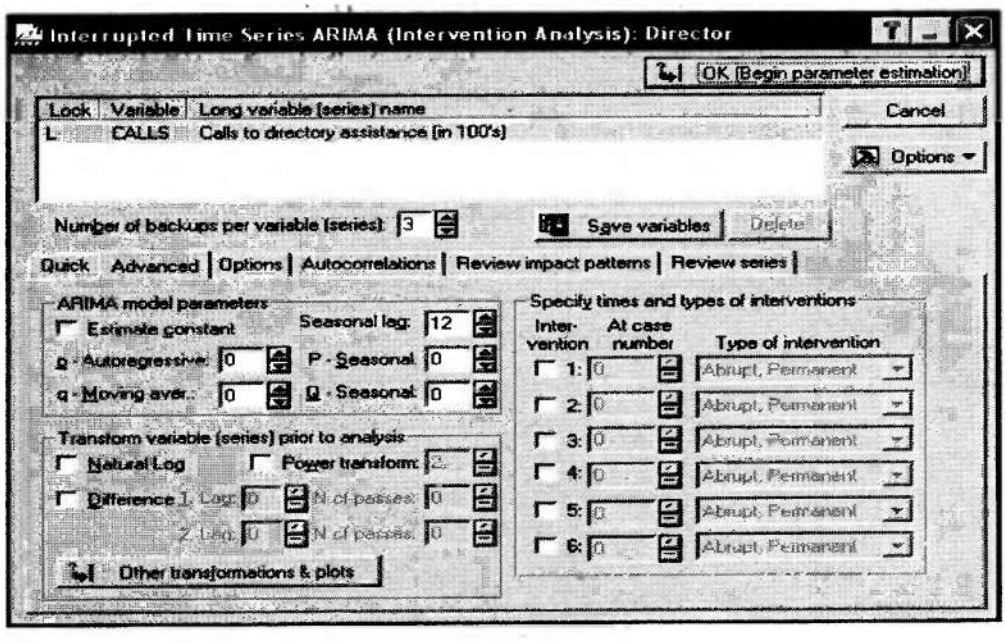


Рис. 18.31

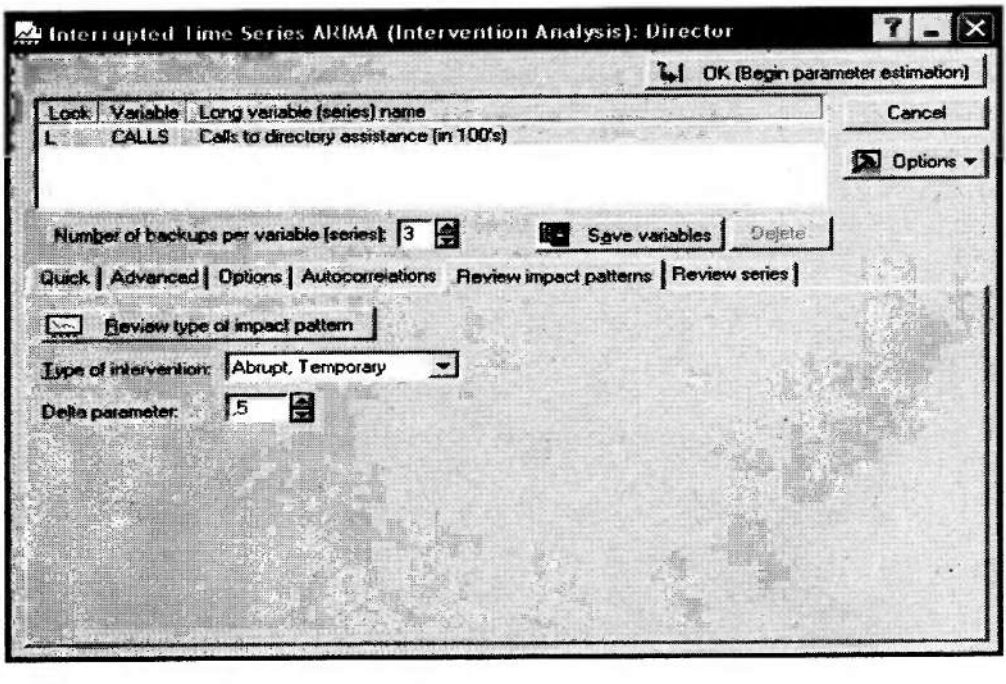


Рис. 18.32

В появившемся диалоге при помощи опции *Review type of impact pattern* можно просмотреть типы графиков интервенции. На рис. 18.33–18.35 приведены следующие типы графиков: *Abrupt, Permanent* (скачкообразный, устойчивый);

Gradual, Permanent (постепенный, устойчивый); *Abrupt, Temporary* (скачкообразный, временный). Выбрав наиболее подходящий тип интервенции, можно перейти к преобразованиям временного ряда и оцениванию параметров модели АРПСС, аналогично тому, как это было сделано в предыдущем разделе.

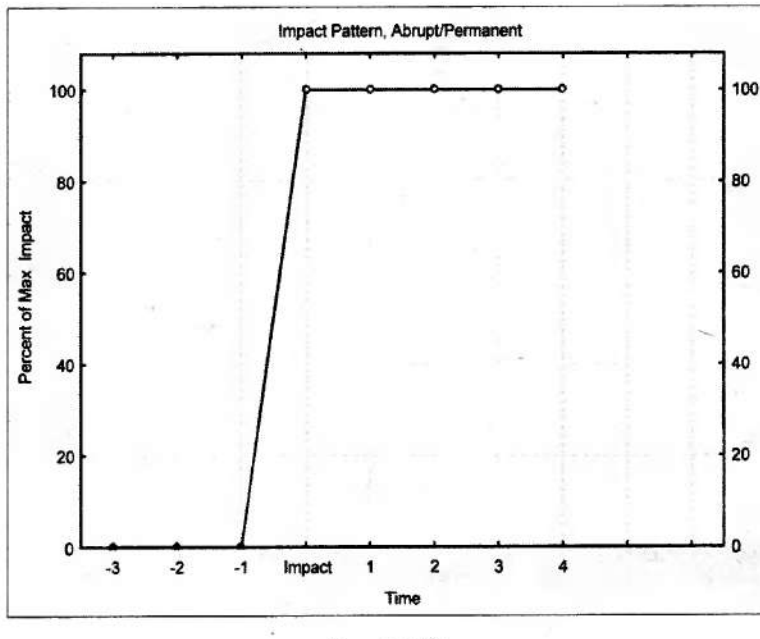


Рис. 18.33

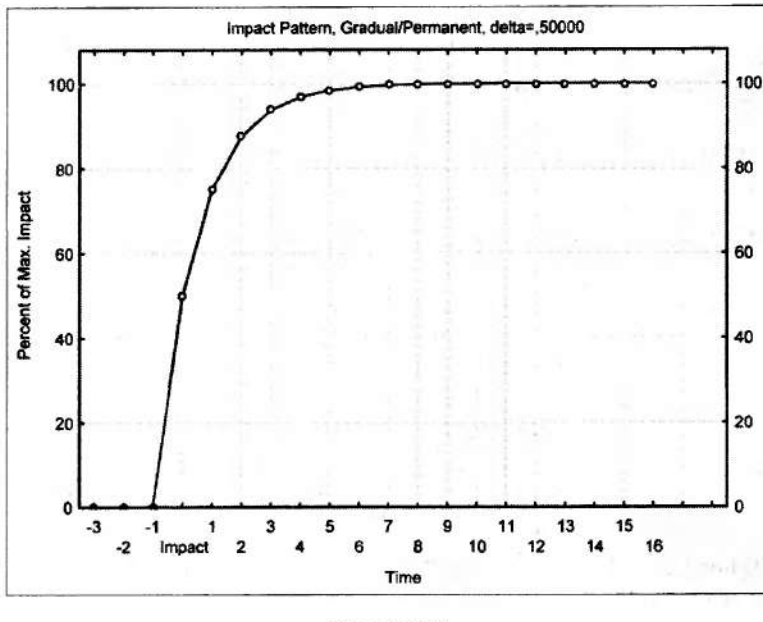


Рис. 18.34

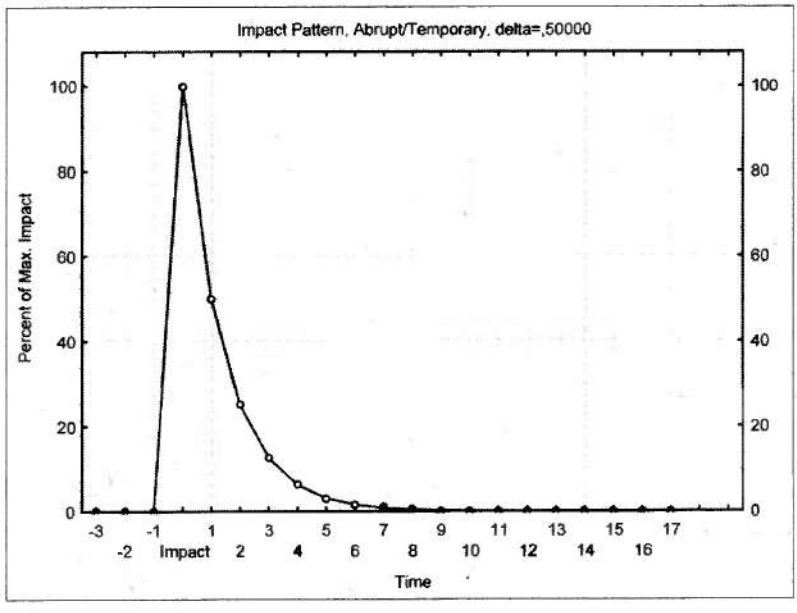


Рис. 18.35

18.3. Экспоненциальное сглаживание и прогнозирование

Простая и прагматически ясная модель временного ряда имеет следующий вид [6]:

$$X_t = \beta + E_t,$$

где β — константа; E_t — случайная ошибка. Константа относительно стабильна на каждом временном интервале, но может также медленно изменяться со временем. Один из интуитивно ясных способов выделения состоит в том, чтобы использовать сглаживание скользящим средним, в котором последним наблюдениям приписываются больший вес, чем предпоследним, предпоследним — еще больший вес, чем предпредпоследним и т.д. Простое экспоненциальное сглаживание именно так и устроено. Здесь более старым наблюдениям приписываются экспоненциально убывающие веса, при этом, в отличие от скользящего среднего, учитываются все предшествующие наблюдения ряда, а не те, что попали в определенное окно. Точная формула простого экспоненциального сглаживания имеет вид

$$S_t = \alpha X_t + (1-\alpha)\, S_{t-1}.$$

Когда эта формула применяется рекурсивно, каждое новое сглаженное значение (которое является также прогнозом) вычисляется как взвешенное среднее текущего наблюдения и сглаженного ряда. Очевидно, результат сглаживания зависит от параметра α (альфа). Если α равно 1, то предыдущие наблюдения полностью игнорируются. Если α равно 0, то игнорируются текущие наблюдения.

Значения α между 0 и 1 дают промежуточные результаты. Эмпирические исследования показали, что весьма часто простое экспоненциальное сглаживание дает достаточно точный прогноз.

Пользователь может задавать начальное значение параметров сглаживания, начальное значение тренда и (если требуется) сезонные факторы. Для тренда и сезонной составляющей могут быть заданы независимые параметры сглаживания. для оценки адекватности модели используются графики, на которых вместе с исходным рядом в подходящем масштабе по оси Y изображаются его сглаженный вариант, прогноз и ряд остатков.

Описание модуля и основные принципы работы с ним проиллюстрируем на примере файла **Series G**. В стартовой панели модуля **Time Series Analysis/Forecasting** (рис. 18.1) нажмите кнопку **Exponential Smoothing & Forecasting** (экспоненциальное сглаживание и прогнозирование). Появится окно диалога **Seasonal and Non-Seasonal exponential smoothing** (сезонное и несезонное экспоненциальное сглаживание) (рис. 18.36).

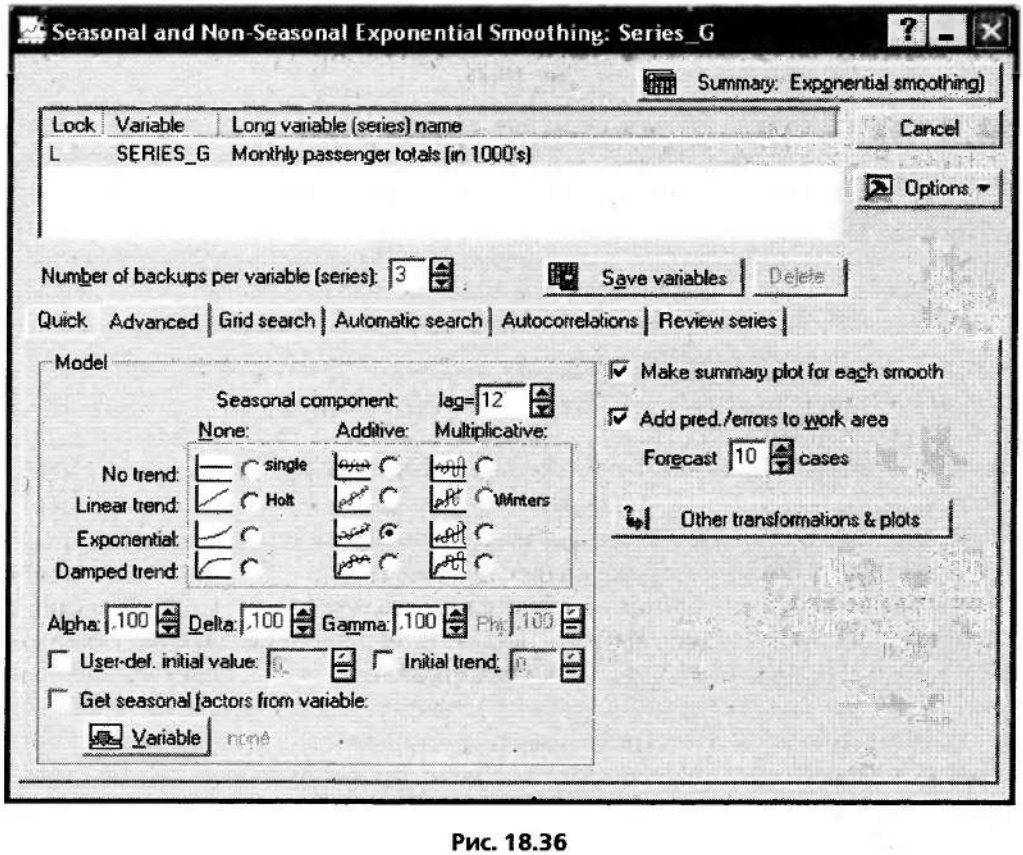


Рис. 18.36

Для определения модели экспоненциального сглаживания нужно задать сезонную компоненту, тренд и параметры сглаживания. Это можно сделать в следующих опциях:

- *Seasonal component* (сезонная компонента);
- *None* (нет);
- *Additive* – (аддитивная);
- *Multiplicative* (мультипликативная);
- *No trend* (нет тренда);
- *Linear trend* (линейный тренд);
- *Exponential* (экспоненциальный);
- *Damped trend* (демпфированный (затухающий) тренд).

Опция *User-def. Initial value* используется для задания начального значения *S(0)*.

Опция *Initial trend* (начальный тренд) задает начальное значение тренда. Если эта опция не используется, то начальное значений оценивается программой.

Опция *Get seasonal factors from variables* (оценить сезонные факторы на данных) запускает процедуру оценивания сезонных факторов.

В полях **Alpha, Delta, Gamma, Phi** задаются параметры экспоненциального сглаживания. Параметр *Alpha* необходим для всех моделей экспоненциального сглаживания. Остальные параметры нужны для специальных моделей. Параметр *Delta* – сезонный сглаживающий параметр, необходим лишь в сезонных моделях. Параметры *Gamma* и *Phi* являются параметрами сглаживания тренда. Параметр *Gamma* используется в моделях с линейным и экспоненциальным трендами и в моделях с демпфированным трендом в рядах без сезонной составляющей.

Параметр *Phi* используется в моделях с демпфированным трендом.

Следующие две опции относятся к представлению результатов на графиках:

- *Make summary plot for each smooth* (сделать итоговый график для каждого сглаживания);
- *Add pred/errors to work area* (добавить сглаженный ряд/остатки в рабочую область).

В опции *Forecast_cases* (прогноз случаев) указывается, на сколько случаев вперед будет прогнозироваться исходный ряд.

При помощи опции *Additive* задайте наиболее подходящий для этого ряда тренд (см. рис.18.3) – экспоненциальный тренд. В поле **Seasonal component** задайте *lag* = 12. Выберите вкладку **Grid Search**, появится поле параметров на сетке (рис. 18.37), где указаны начальные, предельные значения параметров и шаг.

Нажмите кнопку **Perform grid Search.** По этой команде программа переберет всевозможные значения параметров на сетке и укажет лучшие в первой строке появившейся таблицы (рис. 18.38).

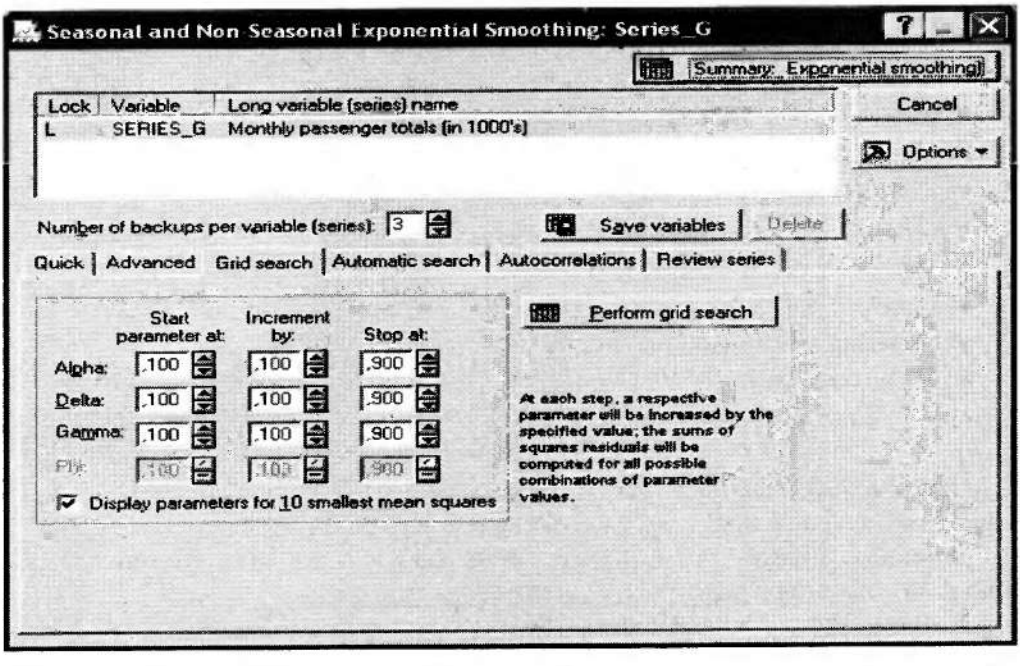


Рис. 18.37

Model Number	Parameter grid search (Smallest abs. errors are highlighted) (Series_ Model: Expon. trend, add.season (12); S0=120,6 T0=1,008 SERIES_G: Monthly passenger totals (in 1000's)					
	Alpha	Delta	Gamma	Mean Error	Mean Abs Error	Sums of Squares
154	0,200000	0,900000	0,100000	-0,539034	12,02541	34948,92
145	0,200000	0,800000	0,100000	-0,525827	12,25777	36472,68
73	0,100000	0,900000	0,100000	-0,625175	12,12782	36852,55
64	0,100000	0,800000	0,100000	-0,595069	12,17511	37643,18
155	0,200000	0,900000	0,200000	-0,621882	12,59774	38492,36
136	0,200000	0,700000	0,100000	-0,513936	12,60913	39135,24
55	0,100000	0,700000	0,100000	-0,563808	12,51606	39416,10
235	0,300000	0,900000	0,100000	-0,589378	13,22174	39448,90
74	0,100000	0,900000	0,200000	-0,588828	12,70447	39531,69
146	0,200000	0,800000	0,200000	-0,610063	12,80810	39545,50

Рис. 18.38

Щелкните по кнопке **Summary: Exponential smoothing.** Появятся две таблицы — таблица с исходными, прогнозными значениями ряда, с остатками ряда и прогнозными значениями на 12 наблюдений вперед (рис. 18.39) и таблица с различными оценками ошибки сглаживания, которая может быть использована для наилучшего подбора установок сглаживания.

Case	Exp. smoothing: Additive season (12) S0=120,6 T Expon.trend, add.season; Alpha=,100 Delta=,100 SERIES_G: Monthly passenger totals (in 1000's) SERIES_G	Smoothed Series	Resids	Seasonal Factors
141	508,0000	506,2786	1,7214	
142	461,0000	471,8748	-10,8748	
143	390,0000	441,3714	-51,3714	
144	432,0000	468,1717	-36,1717	
145		475,2827		
146		467,6489		
147		507,9464		
148		508,5692		
149		517,3071		
150		563,8587		
151		608,9954		
152		607,2675		
153		553,8059		
154		517,5418		

Рис. 18.39

На рис. 18.40 изображены графики исходного ряда, ряда прогнозов и ряда остатков. Видно что ряд остатков является стационарным. Это свидетельствует об адекватности построенной модели экспоненциального сглаживания.

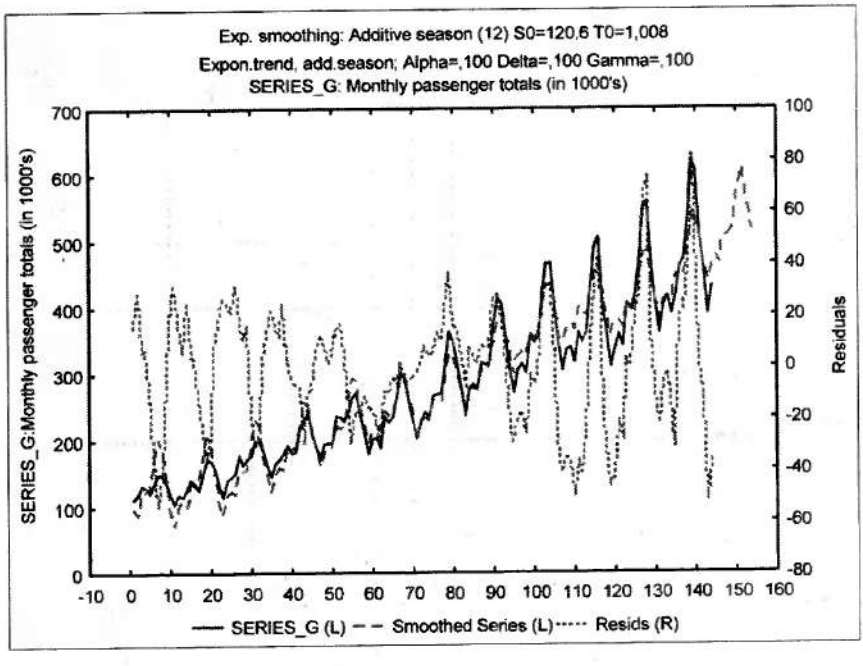


Рис. 18.40

Дополнительно убедиться в адекватности модели можно при помощи вкладки **Review series**.

На вкладке **Automatic Search** (автоматический поиск) реализована другая процедура поиска наилучших значений параметров. Задаются системой по умолчанию начальные значения параметров, число итераций (они могут быть изменены пользователем) и после нажатия кнопки **Automatic estimation** программа сразу строит графики временных рядов и показывает таблицы с соответствующими значениями.

Следует иметь в виду, что экспоненциальное сглаживание — наиболее простой способ построения прогнозов. Часто он дает быстрые эффективные результаты. Однако этот метод не позволяет строить доверительные интервалы и, следовательно, рассчитать риски при использовании прогнозов.

18.4. Сезонная декомпозиция

В модуле **Time Series/Forecasting** реализовано два вида сезонной декомпозиции: классическая сезонная декомпозиция **(Census I)** и так называемый **XII/Y2k (Census 1) — montly** (12-месячная сезонная корректировка), разработанный в Бюро переписей США и максимально приближенный к стандартам, применяемым в США. Отличие последнего от классической сезонной декомпозиции заключается в том, что в нем производится корректировка на различные дни недели и на различное количество дней в месяце (месячная корректировка), а также на месяцы с экстремальными наблюдениями (квартальная корректировка).

Основная идея сезонной декомпозиции проста. Как было замечено в начале этой главы, временной ряд x_t можно представить состоящим из четырех различных компонент: сезонной компоненты γ_t, тренда u_t, циклической компоненты c_t и случайной, нерегулярной компоненты ε_t. Разница между циклической и сезонной компонентами состоит в том, что последняя имеет регулярную (сезонную) периодичность, тогда как циклические факторы обычно обладают более длительным эффектом, который к тому же меняется от цикла к циклу. В методе сезонной декомпозиции тренд и циклическую компоненту обычно объединяют в одну тренд-циклическую компоненту (uc_t). Конкретные функциональные взаимосвязи между этими компонентами бывают самого разного вида. Однако можно выделить два основных способа, с помощью которых они взаимодействуют: аддитивно и мультипликативно.

Аддитивная модель имеет вид $x_t = uc_t + \gamma_t + \varepsilon_t$, мультипликативная модель — $x_t = uc_t \gamma_t \varepsilon_t$. Здесь x_t обозначает значение временного ряда в момент t. Если имеются какие-то априорные сведения о циклических факторах, влияющих на ряд (например, циклы деловой конъюнктуры), то можно использовать оценки для различных компонент для составления прогноза будущих значений ряда. Но для прогнозирования предпочтительнее экспоненциальное сглаживание, позволяющее учитывать сезонную составляющую и тренд.

Аддитивную (мультипликативную) модель желательно использовать, если значения ряда, соответствующие сезонным циклам, ведут себя аналогично членам арифметической (геометрической) прогрессии.

Если перейти к графикам временных рядов, то различие между этими двумя видами моделей будет проявляться так: в аддитивном случае ряд будет иметь постоянные сезонные колебания, величина которых не зависит от общего уровня значений ряда; в мультипликативном случае величина сезонных колебаний будет меняться в зависимости от общего уровня значений ряда.

В прогнозировании с помощью *ARIMA* сезонность учитывалась (бралась разность с лагом 12), но невозможно было проанализировать ее вид, понять, какое действие она оказывает на ряд. В методах сезонной декомпозиции можно строить графики сезонной компоненты, тренд-циклической и нерегулярной составляющей.

Обратим внимание еще и на то, что в диалоге *ARIMA* требуется минимум 8 полных сезонных циклов значений ряда (в нашем случае необходимо было иметь минимум $8 \times 12 = 96$ случаев), а для методов сезонной декомпозиции достаточно 5 полных сезонных циклов.

Осуществим сезонную декомпозицию для файла **Series G**. В стартовом окне модуля **Time Series/Forecasting** нажмите кнопку **Seasonal decomposition (Census I)**. Откроется окно диалога (рис. 18.41).

Опции, позволяющие задать модель декомпозиции, объединены в группу **Seasonal model** (сезонная модель):

- *Additive* (аддитивная);
- *Multiplicative* (мультипликативная).

В поле **Seasonal lag** (сезонный лаг) задается длина сезонного периода.

Опция *Centered moving averages (for even Seasonal lag only)* (центрированные скользящие средние для четного сезонного лага) позволяет пользователю при четном лаге выбрать одну из двух возможностей: брать скользящее среднее с одинаковыми весами или же так, чтобы первое и последнее наблюдение в окне имели неравные веса. Второй метод используется, если установлена галочка. Если же длина сезонного периода нечетна, то установка этой опции не влияет на вычисления.

Следующая группа опций *On OK append components to active work area* позволяет добавить в активное рабочее пространство следующие составляющие.

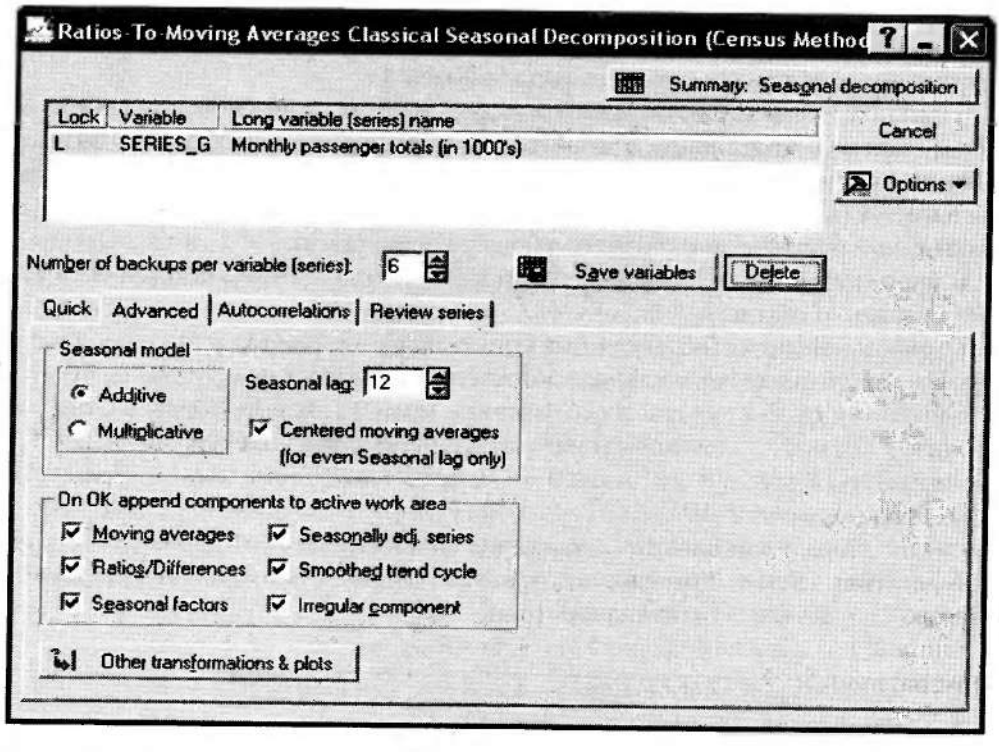


Рис. 18.41

Moving averages (скользящие средние). Сначала вычисляется скользящее среднее для временного ряда, при этом ширина окна берется равной периоду сезонности. Если период сезонности — четное число, пользователь может выбрать одну из двух возможностей: брать скользящее среднее с одинаковыми весами или же с неравными весами так, что первое и последнее наблюдения в окне имеют ополовиненные веса.

Ratios/Differences (отношения/разности). После взятия скользящих средних вся сезонная (т.е. внутри сезона) изменчивость будет исключена, и поэтому разность (в случае аддитивной модели) или отношение (для мультипликативной модели) между наблюдаемым и сглаженным рядом будет выделять сезонную составляющую (плюс нерегулярную компоненту). Более точно, ряд скользящих средних вычитается из наблюдаемого ряда (в аддитивной модели) или же значения наблюдаемого ряда делятся на значения скользящих средних (в мультипликативной модели).

Seasonal factors (сезонные факторы). На следующем шаге вычисляется сезонная составляющая как среднее (для аддитивных моделей) или медианное среднее (для мультипликативных моделей) всех значений ряда, соответствующих данному сезону.

Seasonal adj. series (ряд, скорректированный на сезонную составляющую). Исходный ряд можно скорректировать, вычитая из него (аддитивная модель)

или деля его значения (мультипликативная модель) на значения сезонной составляющей. Получающийся в результате ряд называется сезонной корректировкой ряда (из ряда убрана сезонная составляющая).

Smoothed trend cycle (сглаженная тренд-циклическая компонента). Напомним, что циклическая компонента отличается от сезонной компоненты тем, что продолжительность цикла, как правило, больше одного сезонного периода и разные циклы могут иметь разную продолжительность. Приближение для объединенной тренд-циклической компоненты можно получить, применяя к ряду с сезонной поправкой процедуру пятиточечного (центрированного) взвешенного скользящего среднего с весами 1, 2, 3, 2, 1.

Irregular components (нерегулярная составляющая). На последнем шаге выделяется случайная или нерегулярная компонента (погрешность) путем вычитания из ряда с сезонной поправкой (аддитивная модель) или делением этого ряда (мультипликативная модель) на тренд-циклическую компоненту.

Установите сезонный лаг, равный 12. Выделите все опции в группе опций **On OK append components to active work area.** Отметьте опцию **Additive**. Установите параметр **Namber of backups...**, равный 6. Нажмите кнопку **Summary: Seasonal decomposition** (итоги, сезонная декомпозиция). Появится таблица (рис. 18.42), в которой вычислены составляющие ряда.

Case	Seasonal Decomposition: Additive season (12); Centered means (Series_G) SERIES_G: Monthly passenger totals (in 1000's)						
	SERIES_G	Moving Averages	Diffrncs	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	112,0000			-24,7487	136,7487	143,2875	-6,5387
2	118,0000			-36,1881	154,1881	141,7260	12,4621
3	132,0000			-2,2412	134,2412	138,6031	-4,3620
4	129,0000			-8,0366	137,0366	131,5989	5,4377
5	121,0000			-4,5063	125,5063	118,6886	6,8178
6	135,0000			35,4028	99,5972	104,4840	-4,8868
7	148,0000	126,7917	21,2083	63,8308	84,1692	96,3380	-12,1688
8	148,0000	127,2500	20,7500	62,8232	85,1768	100,2298	-15,0530
9	136,0000	127,9583	8,0417	16,5202	119,4798	116,6490	2,8308
10	119,0000	128,5833	-9,5833	-20,6427	139,6427	133,8746	5,7681
11	104,0000	129,0000	-25,0000	-53,5934	157,5934	144,9482	12,6452
12	118,0000	129,7500	-11,7500	-28,6199	146,6199	148,4861	-1,8662
13	115,0000	131,2500	-16,2500	-24,7487	139,7487	148,6330	-8,8843
14	126,0000	133,0833	-7,0833	-36,1881	162,1881	149,1334	13,0547
15	141,0000	134,9167	6,0833	-2,2412	143,2412	145,4920	-2,2508

Рис. 18.42

Легко проверить, что результатом суммирования составляющих — тренд-цикла, сезонной и нерегулярной — является исходный временной ряд. Сумма скорректированного ряда и сезонной составляющей также равна исходному временному ряду.

Для визуализации результатов декомпозиции перейдите на вкладку **Review series** и нажмите кнопку **Review multiple variables,** предварительно выставив нужные обозначения шкал *X* и *Y*. В появившемся окне можно выбрать исходный ряд и компоненты ряда, графики которых вас интересуют, например: *SERIES, Seasonal factors, Smoothed trend cycle*. Щелкните по кнопке **OK**, и программа построит соответствующие графики (рис. 18.43).

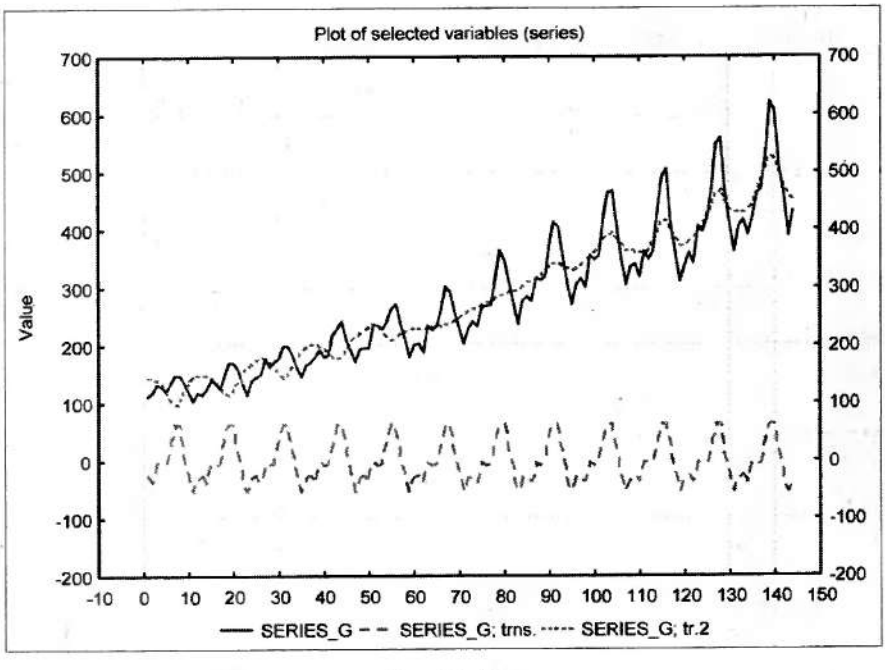


Рис. 18.43

18.5. XII-месячная сезонная корректировка

Основной метод сезонной декомпозиции, рассмотренный в 18.4, может быть усовершенствован различными способами. В отличие от многих методов моделирования временных рядов (в частности, АРПСС), которые основаны на определенной теоретической модели, метод **XII (Census II)** представляет собой просто результат многочисленных специально разработанных приемов и усовершенствований, которые доказали свою работоспособность в многолетней практике решения реальных задач. Некоторые из наиболее важных усовершенствований перечислены далее [6].

Поправка на число рабочих дней. В месяцах разное число дней и разное число рабочих дней (дней недели). Метод **XII (Census II)** дает возможность пользователю проверить, присутствует ли во временном ряду этот эффект числа рабочих дней, и если да, то внести соответствующие поправки.

Выбросы. Большинство реальных временных рядов содержит выбросы, т.е. резко выделяющиеся наблюдения, вызванные какими-то исключительными событиями.

Такие выбросы могут исказить оценки сезонной компоненты и тренда. В процедуре **XII (Census II)** предусмотрены корректировки на случай появления выбросов, основанные на использовании принципов статистического контроля: значения, выходящие за определенный диапазон (который определяется в терминах, кратных сигме, т.е. стандартному отклонению), могут быть преобразованы или вовсе пропущены, и только после этого будут вычисляться окончательные оценки параметров сезонности.

Последовательные уточнения. Корректировки, связанные с наличием выбросов и различным числом рабочих дней, можно производить многократно, чтобы последовательно получать для компонент оценки все лучшего качества. В методе XII делается несколько последовательных уточнений оценок для получения окончательных компонент тренд-цикличности и сезонности, нерегулярной составляющей и самого временного ряда с сезонными поправками.

Критерии и итоговые статистики. Помимо оценки основных компонент ряда, в системе могут быть вычислены различные сводные статистики. Например, можно сформировать таблицы дисперсионного анализа для проверки значимости фактора сезонной изменчивости, ряда и фактора рабочих дней (см. ранее). Процедура **XII** вычисляет также ежемесячные относительные изменения в случайной и тренд-циклической компонентах. С увеличением продолжительности временного промежутка, измеряемого в месяцах или в кварталах года (в случае квартального варианта метода **XII**), изменения в тренд-циклической компоненте, вообще говоря, будут нарастать, в то время как изменения случайной составляющей должны оставаться примерно на одном уровне. Средняя длина временного интервала, на котором изменения тренд-циклической компоненты становятся примерно равными изменениям случайной компоненты, называется месяцем (кварталом) циклического доминирования, или МЦД (соответственно КЦД). Например, если МЦД равно двум, то на сроках более двух месяцев тренд-циклическая компонента станет доминировать над флуктуациями нерегулярной (случайной) компоненты.

В качестве примера рассмотрим файл **Retail** из **Examples**, в котором приведены данные объемов розничных продаж с 1953 по 1964 г.

В стартовом окне модуля **Time Series Analysis** (рис. 18.1) нажмите кнопку **XII (Census 2) — monthly** и откройте окно диалога **XII/Y2k месячная сезонная корректировка** (рис. 18.44). Эта процедура осуществляет корректировку ряда на дни недели — им приписывают разные веса.

Опишем не встречавшиеся ранее кнопки и опции.

Вкладка **Advanced.** Группа опций **Dates (start of series)** (данные (начало рядов)) задает начало данных, а именно начальный месяц ряда (для этогонеобходимо или определить переменную дат, как это представлено на рис. 18.44, или ввести начальные значения месяца и года в поля месяца/года):

- *Variable* (переменная);
- *From — Month* (от — месяц);
- *Year* (год).

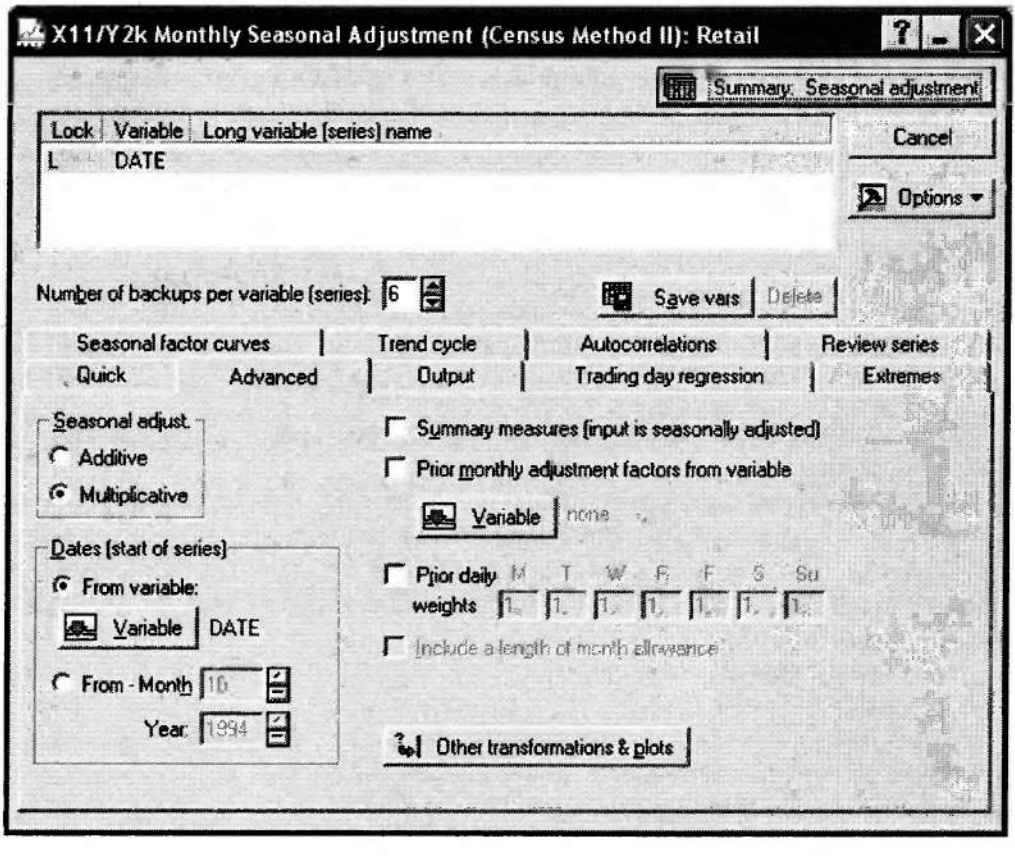


Рис. 18.44

Если нет переменной *Date*, программа присваивает первому наблюдению указанную дату.

Опция *Summary measures (input in seasonal adjusted)* включает итоговые измерения в сезонную поправку.

Опция *Prior monthly adjustment factors in var* (априорные месячные поправочные факторы в переменных) задает априорную (до проведения анализа) корректировку месячных факторов.

Следующие две опции позволяют ввести корректировку на дни недели и длительность месяца:

- *Prior daily weights* (априорные веса для дней недели);
- *Include a length of month allowance* (включая поправки на количество дней в месяце).

На вкладке **Trading-day regression** (регрессия рабочих дней) группа опций **Trading-day regression & adjustment of series** (регрессия рабочих дней и корректировка рядов) задает различные режимы корректировки ряда:

- *No adjustment* (нет корректировки): если выделена эта опция, то автоматической поправки на число рабочих дней не делается;
- *Compute only* (вычислить только): если выделена эта опция, то веса дней недели вычисляются и выводятся, но при этом ряд с их помощью не корректируется;
- *Compute & adjust series* (вычисленные и скорректированные ряды): если выделена эта опция, то веса дней недели вычисляются и используются для корректировки входного ряда, начиная с указанного месяца и года;
- *Compute & adjust conditionally* (вычислить и корректировать условно): если выделена эта опция, то веса дней недели вычисляются (начиная с указанного года) и используются для корректировки входного ряда, однако на шаге *C* (см. далее) веса рабочих дней используются только в том случае, если они объясняют статистически значимую вариацию.

На вкладке **Output** (вывод) задаются различные режимы вывода таблиц (*tables*) и диаграмм (*chart*).

На вкладке **Trend cycle** (тренд-цикл) пользователь указывает число точек, по которым будет сглаживаться тренд-циклическая компонента (автоматически, 9, 13, 23 точки). При помощи опции *Adjust trend-cycle for strikes (extremes)* (корректировка тренд-цикла для удаления экстремумов) задается «очистка» тренд-циклической компоненты от выбросов.

На вкладке **Seasonal factor curves** (сезонная составляющая) при помощи опции *Moving averages for seasonal factor curves* (скользящее среднее для сезонной составляющей) пользователь может задать процедуру сглаживания скользящим средним или оставить выбор этой процедуры за системой.

На вкладке **Extremes** опцией *Sigma limits for graduating extreme values* задаются сигма границы для градуировки экстремальных значений. Если регрессия рабочих дней используется, то при оценивании сезонной и тренд-циклической компонент значение в этом поле ввода определяет то, как будут обрабатываться выбросы нерегулярной компоненты. Значениям нерегулярной составляющей присваиваются разные веса в зависимости от того, как далеко они отклоняются от нуля (аддитивная модель) или единицы (мультипликативная модель).

Опция *Full weight* (полный вес) задает порог в терминах, кратных σ (сигма — оценка стандартного отклонения тренд-циклической компоненты), до достижения которого значению приписывается полный вес.

Опция *Zero weight* (нулевой вес) задает порог в терминах, кратных σ, при превышении которого значение получает нулевой вес. В промежутке между этими двумя пороговыми значениями вес убывает по линейному закону. То есть в вычислениях не будут использоваться те значения нерегулярной компоненты, которые отклоняются от 0 (аддитивная модель) или 1 (мультипликативная модель) больше, чем на $x \cdot s$ (здесь x — значение в этом поле ввода, а s — оценка стандартного отклонения числа рабочих дней).

После нажатия кнопки **Summary: Seasonal adjustment** программа выдаст каскад таблиц и графиков, количество которых определено на вкладке **Output**. После того как будет выдана последняя таблица результатов, программа автоматически добавит в активную рабочую область следующие компоненты ряда:

- сезонную составляющую (табл. *D10*);
- ряд с сезонной поправкой (*D11*);
- тренд-циклическую компоненту (*D12*);
- нерегулярную (случайную) составляющую (*D13*).

Буквенные обозначения таблиц приведены далее.

Чтобы зафиксировать копию, дважды щелкните мышью в колонке *Lock* (блок) напротив соответствующей переменной. Если выдача последовательности таблиц результатов прерывается, то ни одна из этих компонент не будет добавлена в активную рабочую область 10.

Опишем подробнее классификацию таблиц и графиков.

Последовательность вычислений, осуществляемых программой, отражена в системе результирующих таблиц, которые при этом выдаются.

Процедура корректировки разбивается на семь этапов, которые обычно обозначаются буквами *A*–*G*.

A. Априорная корректировка (помесячная сезонная корректировка). Перед тем как к временному ряду, содержащему ежемесячные значения, будет применяться какая-либо сезонная корректировка, могут быть произведены различные корректировки, заданные пользователем. Можно ввести еще один временной ряд, содержащий априорные корректирующие факторы; значения этого ряда будут вычитаться из исходного ряда (аддитивная модель) или же значения исходного ряда будут поделены на значения корректирующего ряда (мультипликативная модель). В случае мультипликативной модели пользователь может также определить свои поправочные коэффициенты (веса) на число рабочих дней. Эти веса будут использоваться для корректировки ежемесячных наблюдений так, чтобы учитывалось число рабочих дней в этом месяце.

B. Предварительное оценивание вариации числа рабочих дней (месячный вариант **XII**) и *весов*. На этом шаге вычисляются предварительные поправочные коэффициенты на число рабочих дней (только в месячном варианте **XII**) и веса, позволяющие уменьшить эффект выбросов.

C. Окончательное оценивание вариации числа рабочих дней и нерегулярных весов (месячный вариант **XII**). Поправки и веса, вычисленные в пункте *B*, используются для построения улучшенных оценок тренд-циклической и сезонной компонент. Эти улучшенные оценки служат для окончательного вычисления факторов числа рабочих дней (в месячном варианте **XII**) и весов.

D. Окончательное оценивание сезонных факторов, тренд-циклической, нерегулярной и сезонно скорректированной компонент ряда. Окончательные значения факторов рабочих дней и весов, вычисленные в пункте C, используются для вычисления окончательных оценок компонент ряда.

E. Модифицированные ряды: исходный, сезонно скорректированный и нерегулярный. Исходный и окончательный, сезонно скорректированный ряды, а также нерегулярная компонента модифицируются путем сглаживания

выбросов. Полученные в результате этого модифицированные ряды позволяют пользователю проверить устойчивость сезонной корректировки.

F. *Месяц (квартал) циклического доминирования* (МЦД, КЦД), *скользящее среднее и сводные показатели*. На этом этапе вычислений рассчитываются различные сводные характеристики (см. далее), позволяющие пользователю исследовать относительную важность разных компонент, среднюю флуктуацию от месяца к месяцу (от квартала к кварталу), среднее число идущих подряд изменений в одну сторону и др.

G. *Графики*. В завершение анализа программа строит различные графики, в обобщенном виде представляющие результаты анализа. Например, будет нарисован окончательный сезонно скорректированный ряд как функция времени.

На каждом из этапов *A*–*G* вычисляются различные таблицы результатов. Обычно все они нумеруются, им также приписывается буква, соответствующая этапу анализа. Например, табл. *B11* содержит предварительно сезонно скорректированный ряд; *C11* — это более точно сезонно скорректированный ряд, а *D11* — окончательный сезонно скорректированный ряд. Таблицы, помеченные звездочкой (*), недоступны (или неприменимы) при анализе квартальных показателей. Кроме того, в случае квартальной корректировки некоторые из описанных далее вычислений несколько видоизменяются. Так, например, для вычисления сезонных факторов вместо 12-периодного (т.е. 12-месячного) скользящего среднего используется 4-периодное (4-квартальное) скользящее среднее; предварительная тренд-циклическая компонента вычисляется по центрированному 4-периодному скользящему среднему, а окончательная оценка тренд-циклической компоненты вычисляется по 5-точечному среднему Хендерсона.

В соответствии со стандартом метода **XII**, принятым Бюро переписи США, в системе *STATISTICA* предусмотрены три степени подробности вывода: стандартный (17–27 таблиц), длинный (27–39 таблиц) и полный (44–59 таблиц). Имеется также возможность выводить только таблицы результатов, выбранные пользователем. В следующих далее описаниях таблиц буквы С (стандартный), Д (длинный) и П (полный) рядом с названием таблицы указывают, какие таблицы выводятся и/или распечатываются в соответствующем варианте вывода.

Для графиков предусмотрены два уровня подробности вывода: *стандартный* и *все*. Приведем обозначения и наименования таблиц.

*A1. Исходный ряд (С).

*A2. Априорные месячные поправки (С).

*A3. Исходный ряд, скорректированный с помощью априорных месячных поправок (С).

*А 4. Априорные поправки на рабочие дни (С).

B1. Ряд после априорной корректировки либо исходный ряд (С).

B2. Тренд-цикл (Д).

B3. Немодифицированные *S-I* разности или отношения (П).

B4. Значения для замены выбросов *S-I* разностей (отношений) (П).

B5. Сезонная составляющая (П).

B6. Сезонная корректировка ряда (П).

B7. Тренд-цикл (Д).

B8. Немодифицированные *S-I* разности (отношения) (П).

B9. Значения для замены выбросов *S-I* разностей (отношений) (П).

B10. Сезонная составляющая (Д).

B11. Сезонная корректировка ряда (П).

B12. (Не используется).

B13. Нерегулярная составляющая ряда (Д).

Табл. *B14–B16*, *B18* и *B19* — поправка на число рабочих дней. Эти таблицы доступны только при анализе ежемесячных данных. Число разных дней недели (понедельников, вторников и т.д.) колеблется от месяца к месяцу. Бывают ряды, в которых различия в числе рабочих дней в месяце могут давать заметный разброс ежемесячных показателей. Пользователь имеет возможность определить начальные веса для каждого дня недели (см. *A4*), и/или эти веса могут быть оценены по данным (пользователь также может сделать использование этих весов условным, т.е. только в тех случаях, когда они объясняют значительную часть дисперсии).

**B14.* Выбросы нерегулярной составляющей, исключенные из регрессии рабочих дней (Д).

**B15.* Предварительная регрессия рабочих дней (Д).

**B16.* Поправки на число рабочих дней, полученные из коэффициентов регрессии (П).

B17. Предварительные веса нерегулярной компоненты (Д).

**B18.* Поправки на число рабочих дней, полученные из комбинированных весов дней недели (П).

**B19.* Исходный ряд с поправками на рабочие дни и априорную вариацию (П).

C1. Исходный ряд, модифицированный с помощью предварительных весов, с поправкой на рабочие дни и априорную вариацию (Д).

C2. Тренд-цикл (П).

C3. (Не используется).

C4. Модифицированные *S-I* разности (отношения) (П).

C5. Сезонная составляющая (П).

C6. Сезонная корректировка ряда (П).

C7. Тренд-цикл (Д).

C8. (Не используется).

C9. Модифицированные *S-I* разности (отношения) (П).

C10. Сезонная составляющая (Д).

C11. Сезонная корректировка ряда (П).

C12. (Не используется).

C13. Нерегулярная составляющая (С).

Табл. *C14–C16*, *C18* и *C19*: поправка на число рабочих дней. Эти таблицы доступны только при анализе ежемесячных данных и если при этом требуется поправка на различное число рабочих дней. В этом случае поправки на число рабочих дней вычисляются по уточненным значениям сезонно скорректированных рядов аналогично тому, как это делалось в пункте *B* (см. *B14–B16*, *B18*, *B19*).

**C14*. Выбросы нерегулярной составляющей, исключенные из регрессии рабочих дней (С).

**C15*. Регрессия рабочих дней — окончательный вариант (С).

**C16*. Поправки на число рабочих дней, полученные из коэффициентов регрессии, — окончательный вариант (С).

C17. Окончательные веса нерегулярной компоненты (С).

**C18*. Поправки на число рабочих дней, полученные из комбинированных весов дней недели, — окончательный вариант (С).

**C19*. Исходный ряд с поправками на рабочие дни и априорную вариацию (С).

D1. Исходный ряд, модифицированный с помощью окончательных весов, с поправкой на рабочие дни и априорную вариацию (Д).

D2. Тренд-цикл (П).

D3. (Не используется).

D4. Модифицированные *S-I* разности (отношения) (П).

D5. Сезонная составляющая (П).

D6. Сезонная корректировка ряда (П).

D7. Тренд-цикл (Д).

D8. Немодифицированные *S-I* разности (отношения) — окончательный вариант (С).

D9. Окончательные значения для замены выбросов *S-I* разностей (отношений) (С).

D10. Сезонная составляющая — окончательный вариант (С).

D11. Сезонная корректировка ряда — окончательный вариант (С).

D12. Тренд-циклическая компонента — окончательный вариант (С).

D13. Нерегулярная составляющая — окончательный вариант (С).

E1. Модифицированный исходный ряд (С).

E2. Модифицированный ряд с сезонной поправкой (С).

E3. Модифицированная нерегулярная составляющая (С).

E4. Разности (отношения) годовых сумм (С).

E5. Разности (относительные изменения) исходного ряда (С).

E6. Разности (относительные изменения) окончательного варианта ряда с сезонной поправкой (С).

F1. МЦД (КЦД) — скользящее среднее (С).

F2. Сводные показатели (С).

G1. График (С).

G2. График (C).
G3. График (B).
G4. График (B).

В программе для файлов данных, в которых присутствует переменная даты, предусмотрена возможность проверки значимости различия в днях недели без корректировки ряда. Для этого на вкладке **Trading-day regression** выделите опцию *Compute only*, на вкладке **Advanced** произведите установки согласно рис. 18.44, на вкладке **Output** выберите опцию *Selected tables* (выбрать таблицы) и в открывшемся окне (рис. 18.45) в поле C выделите табл. *C15*. Щелкните по **OK** и, вернувшись в стартовое окно диалога, нажмите кнопку **Summary: Seasonal adjustment.**

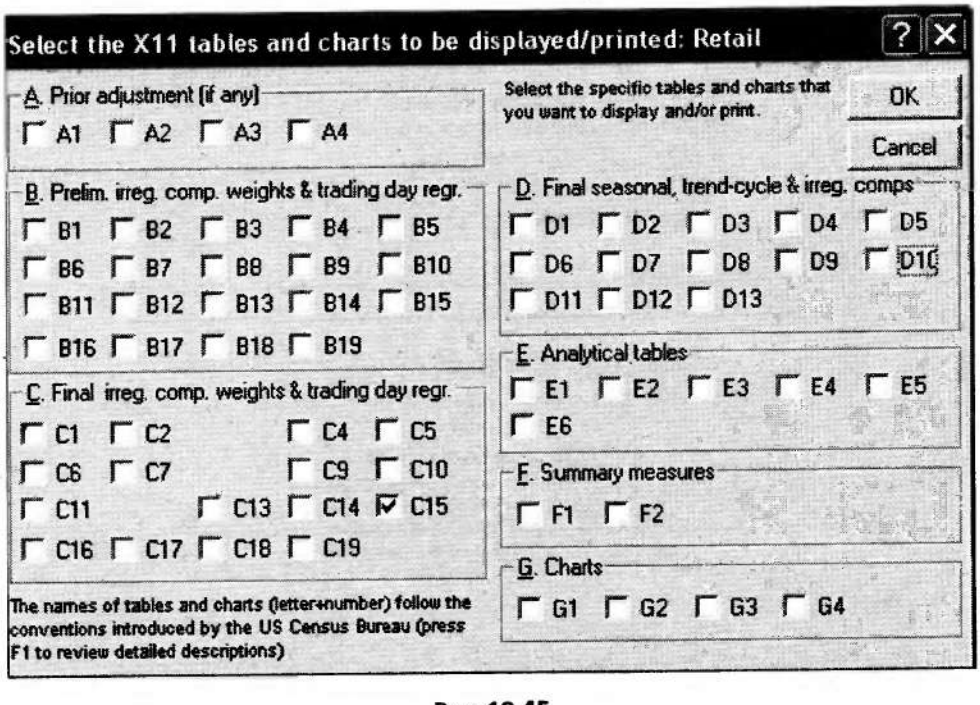


Рис. 18.45

Появится таблица с оценками весов (столбец *Combined Weight* на рис. 18.46) различных дней недели и таблица дисперсионного анализа (рис. 18.47), в которой приведены результаты проверки гипотезы о равенстве оценок весов априорным весам. Видно, что различие в днях недели высокозначимо (*уровень p* значимости *F-критерия* равен 0,000, что значительно меньше 0,05), т.е. имеет смысл осуществить корректировку модели. На вкладке **Trading-day regression** выделите опцию *Compute & adjust series*, на вкладке **Output** — опции *Standard*, щелкните по **OK (Seasonal adjustment).** Появятся различные таблицы и графики: значения сезонной составляющей (*D10*), оценки значений тренд-циклической составляющей (*D12*); графики ряда — ряд, скорректированный на сезонную составляющую, скорректированная тренд-циклическая составляющая (рис. 18.48).

C 15. Final trading day regression (Retail)
*Comb. wt. signif. different from 1/prior wt. at 1% level
SALES : U.S. total retail sales

Day	Combined Weight	Prior Weight	Regressn Coeff.	Std.Err. comb.wt.	t (1)	t prior wt
Monday	1,021304	1,000000	0,021304	0,052958	,402283	,402283
Tuesday	0,970008	1,000000	-0,029992	0,052632	-,56985	-,56985
Wednesd.	1,022603	1,000000	0,022603	0,052329	,431945	,431946
Thursday	0,980429	1,000000	-0,019571	0,054594	-,35848	-,35848
Friday	1,293735	1,000000	0,293735	0,054077	5,43175*	5,43175*
Saturday	1,215273	1,000000	0,215273	0,053545	4,02044*	4,02044*
Sunday	0,496647	1,000000	-0,503353	0,053888	-9,3407*	-9,3407*

Рис. 18.46

C 15. Final trading day regression (Retail)
Residual trading day variation present at 1% level
SALES : U.S. total retail sales

Source	Sum of Squares	Degrs.of Freedom	Mean Square	F	p
Regressn	11,27285	6	1,878809	37,72177	0,000000
Error	6,37530	128	0,049807		
Total	17,64815	134			

Рис. 18.47

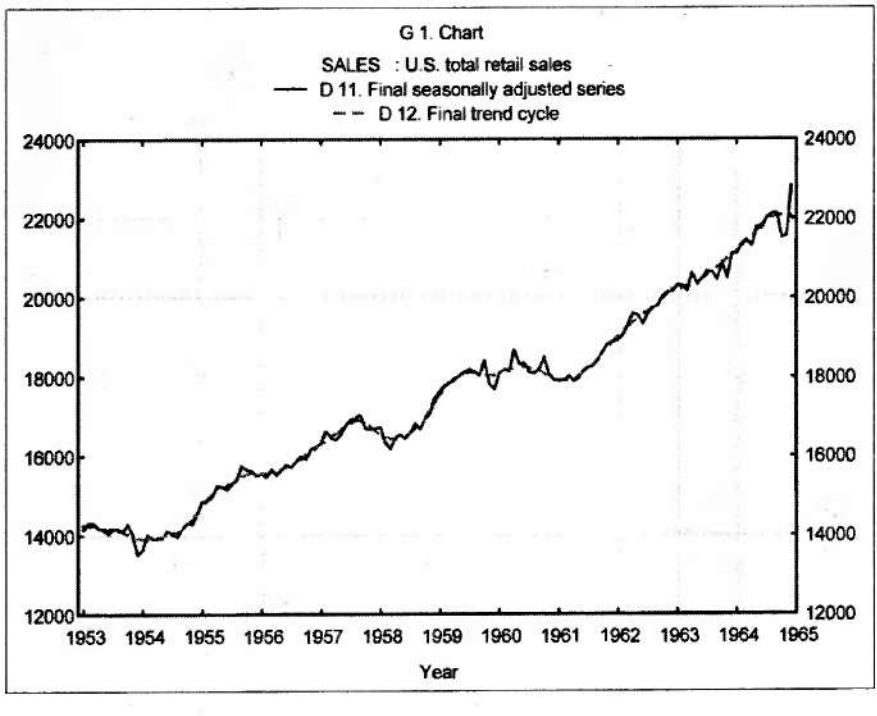


Рис. 18.48

Перейдите на вкладку **Review series**, последовательно высвечивая длинное имя переменных в информационном поле окна, нажмите кнопку **Plot** рядом с кнопкой **Review highlighted variable**. Программа построит графики исходного ряда (рис. 18.49), скорректированной сезонной составляющей (рис. 18.50), ряда, скорректированного на сезонную составляющую (рис. 18.51), скорректированной тренд-циклической составляющей (рис. 18.52), скорректированной нерегулярной составляющей (рис. 18.53).

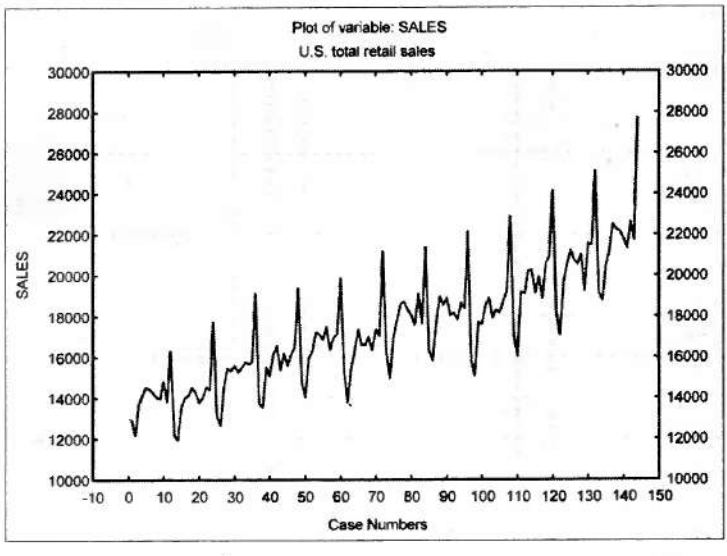


Рис. 18.49

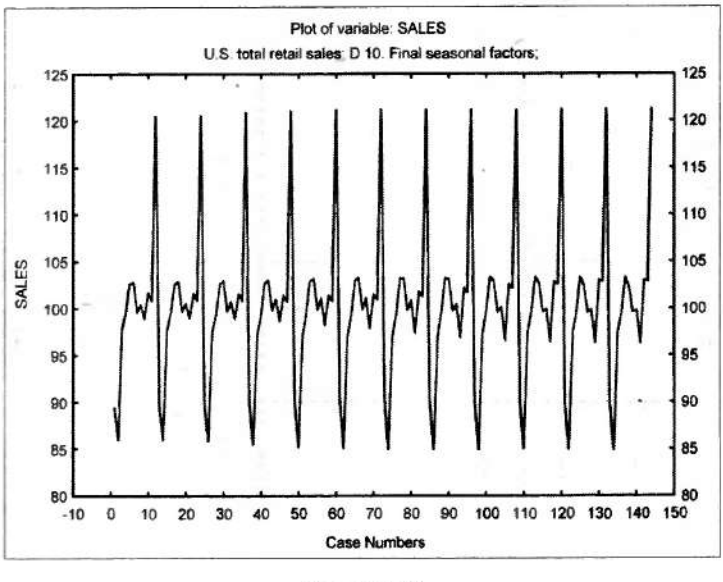


Рис. 18.50

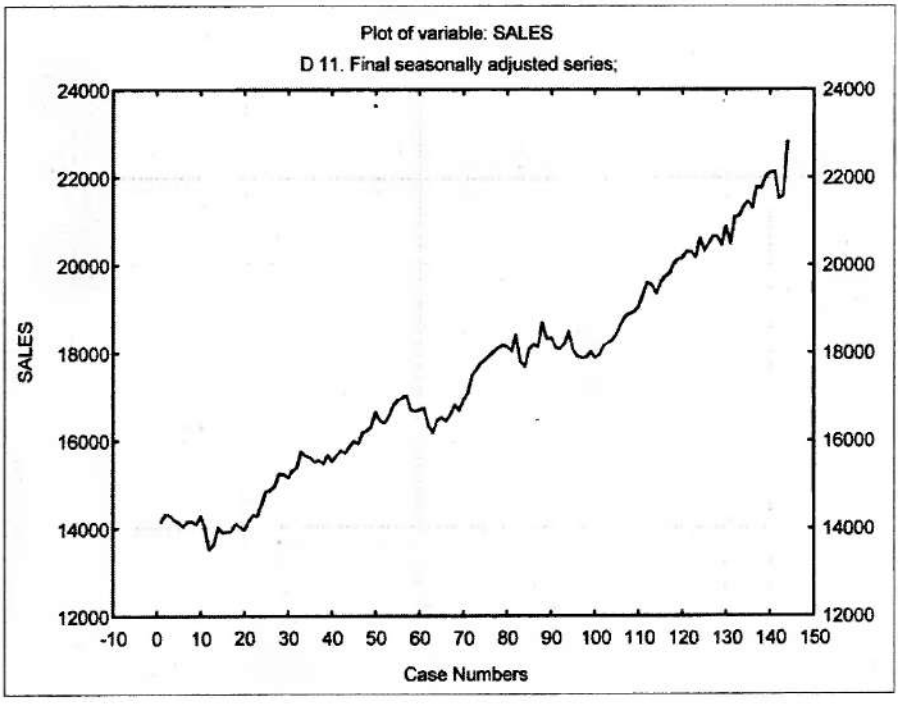


Рис. 18.51

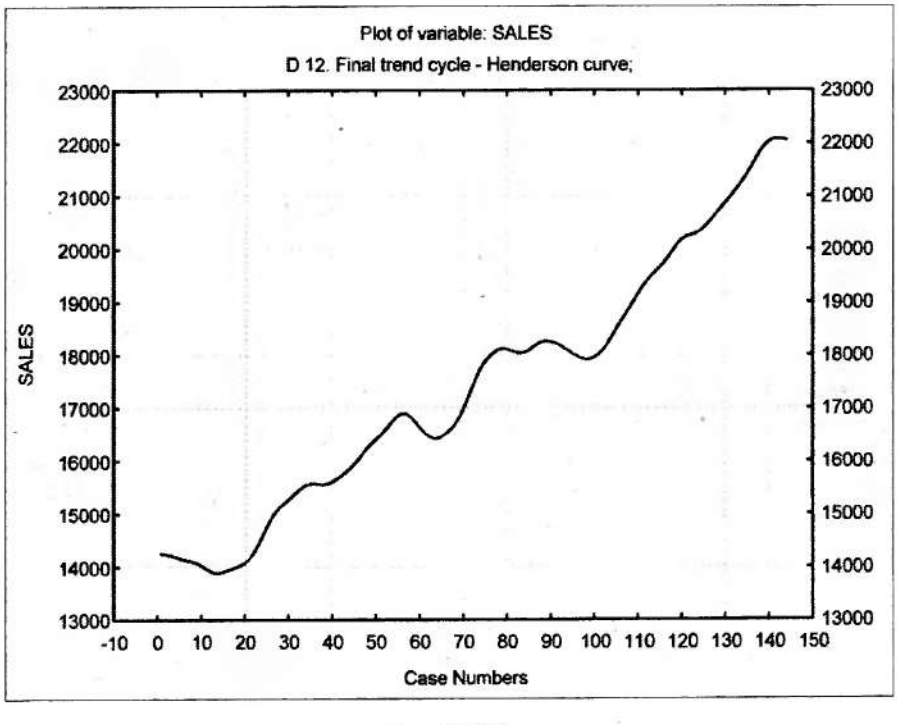


Рис. 18.52

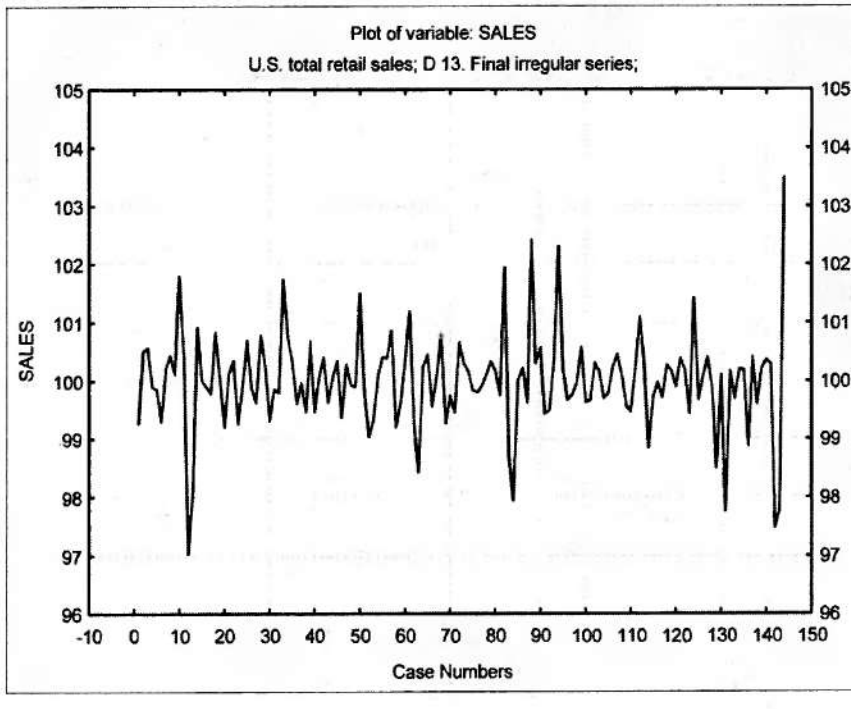


Рис. 18.53

Если в стартовом окне модуля **Time Series Analysis** (рис. 18.1) нажать кнопку **XII (Census 2) — quarterly**, откроется окно процедуры **XII/Y2k** —Квартальная сезонная корректировка. Эта процедура осуществляет корректировку ряда на месяцы с экстремальными наблюдениями.

Из стартового окна процедуры, изображенного на рис. 18.54, видно, что кнопки и опции диалога и последовательность шагов идентичны уже рассмотренным ранее в процедуре **XII/Y2k** (их значительно меньше), поэтому описывать их нет необходимости.

Заметим, что метод **XII (Census II)** оценивает значимость эффекта разного числа рабочих дней или выбросов и при наличии такого эффекта вносит в составляющие временного ряда корректировки. Чтобы учесть эти поправки при прогнозировании временного ряда, можно по скорректированным составляющим ряда составить новый скорректированный временной ряд (умножением или сложением составляющих ряда) и уже к нему применить процедуру **ARIMA** или **Exponential Smoothing&Forecasting**.

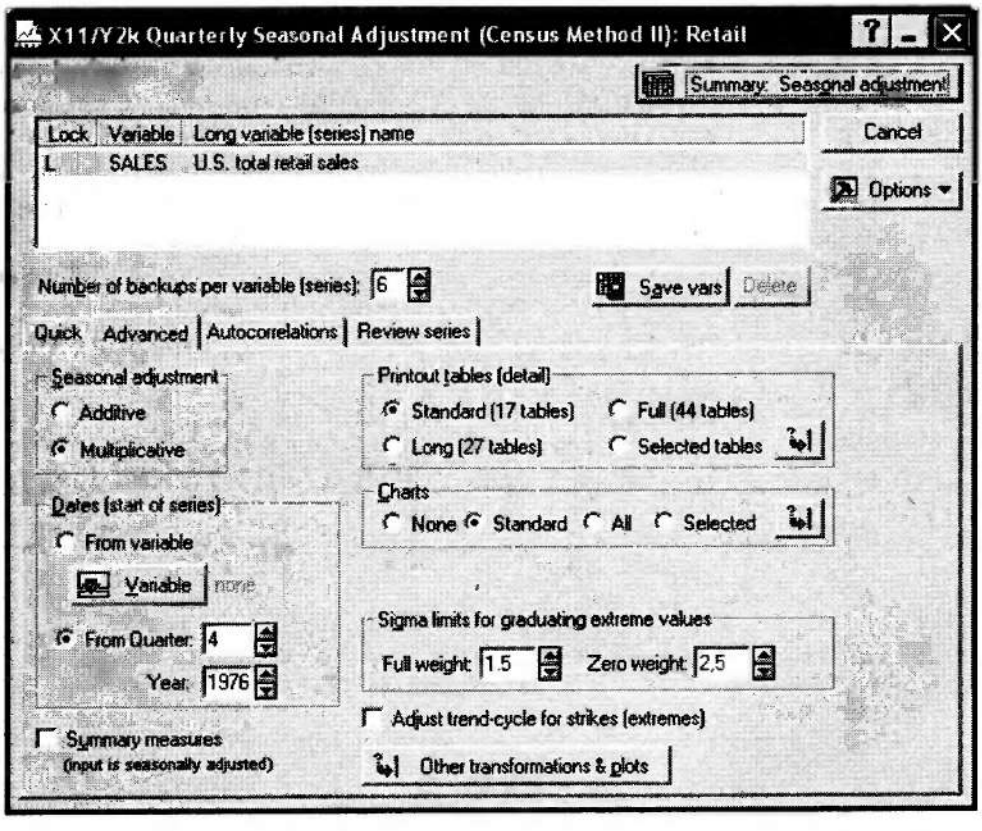


Рис. 18.54

18.6. Спектральный (Фурье) анализ

В спектральном анализе исследуются периодические модели данных. Цель анализа — разложить комплексные временные ряды с циклическими компонентами на несколько основных синусоидальных функций с определенной длиной волн. Термин «спектральный» — своеобразная метафора для описания природы этого анализа. В результате успешного анализа можно обнаружить всего несколько повторяющихся циклов различной длины в интересующих нас временных рядах, которые на первый взгляд выглядят как случайный шум.

В отличие от АРПСС или метода экспоненциального сглаживания, цель спектрального анализа — распознать сезонные колебания различной длины, в то время как в предшествующих типах анализа длина сезонных компонент обычно известна заранее (или предполагается) и затем включается в некоторые теоретические модели скользящего среднего или автокорреляции. Эта процедура позволяет провести спектральный анализ стационарных временных рядов, построить периодограмму, определить оценки спектральной плотности с разнообразными, в том числе и определяемыми пользователем спектральными окнами. Методы спектрального анализа имеют большое значение для опре-

деления скрытых периодичностей в данных. Также они могут быть рассмотрены в связи с другими методами обработки временных рядов, например, для проверки адекватности модели АРПСС (спектральный анализ остатков). Стандартные методы предварительной обработки ряда включают косинус-сглаживание, вычитание среднего и удаление тренда. Результаты обычного спектрального анализа содержат коэффициенты частоты, периода, коэффициенты при синусах и косинусах, периодограмму и оценку спектральной плотности.

Откройте файл данных **Series_G**. В стартовом окне модуля **Time Series Analysis** (рис. 18.1) нажмите кнопку **Spectral (Fourier) Analysis** (спектральный (Фурье) анализ). Программа откроет окно диалога **Fourier (Spectral) Analysis** (рис. 18.55).

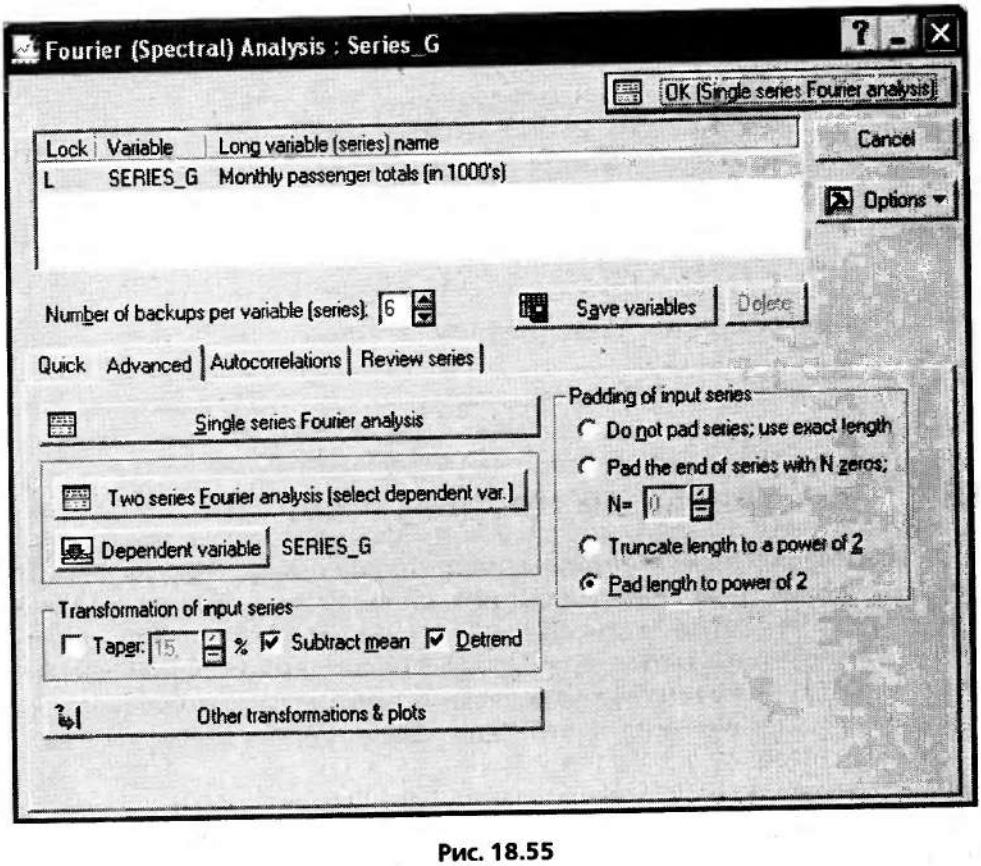


Рис. 18.55

Из функциональных кнопок наиболее важны следующие:

- **Single series Fourier analysis** (анализ Фурье одного ряда);
- **Two series Fourier analysis** (анализ Фурье двух рядов).

Если начальные установки произведены, эти кнопки запустят процесс вычисления.

Группа опций **Transformation of input series** (преобразования исходного ряда) дает возможность выполнить различные преобразования ряда перед построением оценок спектральной плотности:

- *Taper* (сглаживание на концах); эта процедура рекомендуется для устранения ложных пиков в периодограмме;
- *Subtract mean* (вычитание среднего); из значений ряда вычисляется выборочное среднее;
- *Detrend* (вычитание тренда); из ряда вычитается линейный тренд.

Группа опций **Padding of input series** (добавление нулей к концу исходного ряда) предназначена для того, чтобы увеличить число частот в периодограмме и использовать эффективный алгоритм быстрого преобразования Фурье при анализе длинных временных рядов. Рассмотрим их подробнее.

Do not pad series; use exact length (не добавлять нули, использовать точную длину ряда). Добавление нулей в ряд не производится и формально пользоваться алгоритмом быстрого преобразования Фурье нельзя.

Pad the end of the series N zeros (добавить *N* нулей к концу ряда). Число нулей задается в поле рядом с опцией.

Truncate length to a power 2 (урезать длину ряда до того, что она станет равной степени 2). Длина ряда уменьшается, пока не станет равна ближайшему числу, являющемуся степенью 2. Значения в конце ряда отбрасываются.

Pad length to a power of 2 (увеличить длину ряда до ближайшей степени 2). К концу исходного ряда добавляется необходимое число нулей. Как и при использовании опции *Truncate length to a power 2*, длина ряда теперь является степенью 2, и, следовательно, можно применять алгоритм быстрого преобразования Фурье.

Произведите установки согласно рис. 18.55 и нажмите кнопку **Single series Fourier analysis.** Программа выполнит спектральный анализ высвеченного ряда. На экране появится окно результатов **Single Series Fourier (Spectral) Analysis Results** (рис. 18.56).

Кнопка **Summary** позволяет в компактном виде представить итоговый обзор основных статистик спектрального анализа. Эта кнопка выведет итоговую таблицу результатов с частотами, периодами, коэффициентами при косинусах и синусах, значения периодограммы, оценки спектральной плотности (вычисленные в соответствии с выбором в рамке **Data windows for spectral density estimation** (см. далее) и веса, используемые для получения оценок спектральной плотности. Опишем подробнее величины, выводимые по команде этой кнопки.

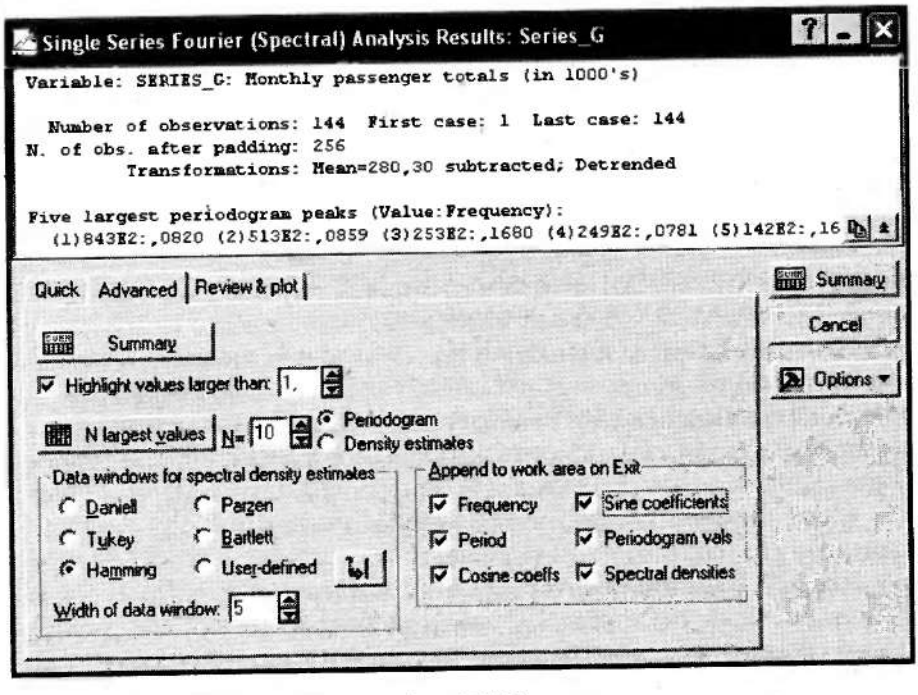


Рис. 18.56

Значения временного ряда вычисляются по формуле

$$x_i = a_0/2 + \sum_k \{a_k \cdot \cos(2\pi \cdot f_k(t-1)) + b_k \cdot \sin(2\pi \cdot f_k(t-1))\}.$$

Частота f_k определяется как число циклов в единицу времени. В модуле **Time Series/Forecasting** за единицу времени берется одно наблюдение (т.е. частота выражается как часть цикла на одно наблюдение), последовательные частоты вычисляются как k/n (от $k = 0$ до $n/2$), где n — число наблюдений ряда. Например, частота 0,0833 означает, что каждое наблюдение составляет 0,0833 от целого цикла, или что 12 наблюдений составляют один цикл (0,0833·12 = 1). Таким образом, если ряд содержит месячные данные за несколько лет, соответствующая периодичность определяет годичный цикл.

Период $1/f_k$ есть число, обратное частоте, т.е. это число наблюдений в полном цикле соответствующей частоты.

Косинус-коэффициенты a_k — это коэффициенты регрессии; они показывают степень корреляции функций косинусов с данными на соответствующих частотах. Синус-коэффициенты b_k интерпретируются аналогично косинус-коэффициентам.

Значения периодограммы вычисляются как суммы квадратов коэффициентов при синусах и косинусах для каждой частоты ($n/2$ раза). Значения периодограммы – дисперсия (сумма квадратов) данных соответствующей частоты или периода.

Оценки спектральной плотности вычисляются путем сглаживания значений периодограммы и выбираются в рамке **Data windows for spectral density**

estimation (окна для оценки спектральной плотности). Сглаживая периодограмму, можно определить основные частотные области (или спектральные плотности), которые вносят значительный вклад в циклическое поведение ряда.

Веса используются в окне сглаживания для получения оценок спектральной плотности. Различные сглаживающие окна описаны далее (окна для оценок спектральной плотности). Заметим, что веса нормируются так, чтобы их сумма была равна 1.

Если нажать кнопку **N largest values...** (*N* наибольших значений...), то программа построит таблицу только для *N* (которое определено в окне редактирования справа) наибольших значений периодограммы и спектральной плотности, в зависимости от установки кнопки справа. Эта таблица является выборкой значений из итоговой таблицы основных статистик спектрального анализа, которая доступна после нажатия кнопки **Summary.** По этой таблице легко определить период сезонного цикла (лаг). Заметим, что для этой таблицы результатов можно построить графики коэффициентов при синусах/косинусах значений периодограммы (или лог-периодограмм) или оценок спектральной плотности (или лог-плотности) по частоте или по периоду.

Группа опций **Data windows for spectral density estimation** (окна для оценки спектральной плотности) определяет спектральные окна, в которых реализованы различные способы определения весов для сглаживания скользящим средним. На выбор предлагаются следующие спектральные окна: Даниэля, Тьюки, Хэминга, Парзена, Барлетта. Можно также задать собственное окно с помощью опции *User-defined* (определенное пользователем) в этой же группе опций. Ширина окна задается в поле **Width of data window** (ширина окна).

Поля рамки **Append to work area on Exit** (добавить в активную рабочую область) определяют, какие статистики будут добавлены в активную рабочую область при выходе из этого диалога (рис. 18.55, 18.56). Эти статистики описаны ранее для опции **Summary**. Если число текущих доступных (неиспользуемых и незапертых) копий в активной рабочей области меньше, чем число рядов, которые должны быть добавлены, то параметр *Number of backups per variable* (число резервов для переменных) будет увеличен, насколько это возможно.

На вкладке **Review series** программа предоставит пользователю различную графическую иллюстрацию результатов анализа: графики спектральной плотности, периодограммы, их логарифмов, косинус- и синус- коэффициенты. Графики могут быть построены по частоте, периоду или лог-периоду, в зависимости от выделенного поля в рамке **Plot by**.

Если наблюдения во временном ряду независимы друг от друга (т.е. нет периодичности) и подчинены нормальному распределению, такой временной ряд может быть белым шумом. Если исходный ряд — белый шум, соответствующие значения периодограммы будут иметь экспоненциальное распределение. Таким образом, путем проверки на экспоненциальность значений периодограммы можно узнать, отличается ли исходный ряд от белого шума. Кнопка **Histogram of periodogram** построит гистограмму значений периодограммы и подгонит экспоненциальное распределение на гистограме. Также пользователь может запро-

сить вычислить *d-статистику* Колмогорова-Смирнова. К сожалению, значение статистики программа выводит без уровня значимости p. На рис. 18.57 изображен график гистограммы периодограммы, построенный процедурой **Fitting Continuous**, с оценкой возможности аппроксимации экспоненциальным законом распределения по критериям χ^2 и Колмогорова-Смирнова.

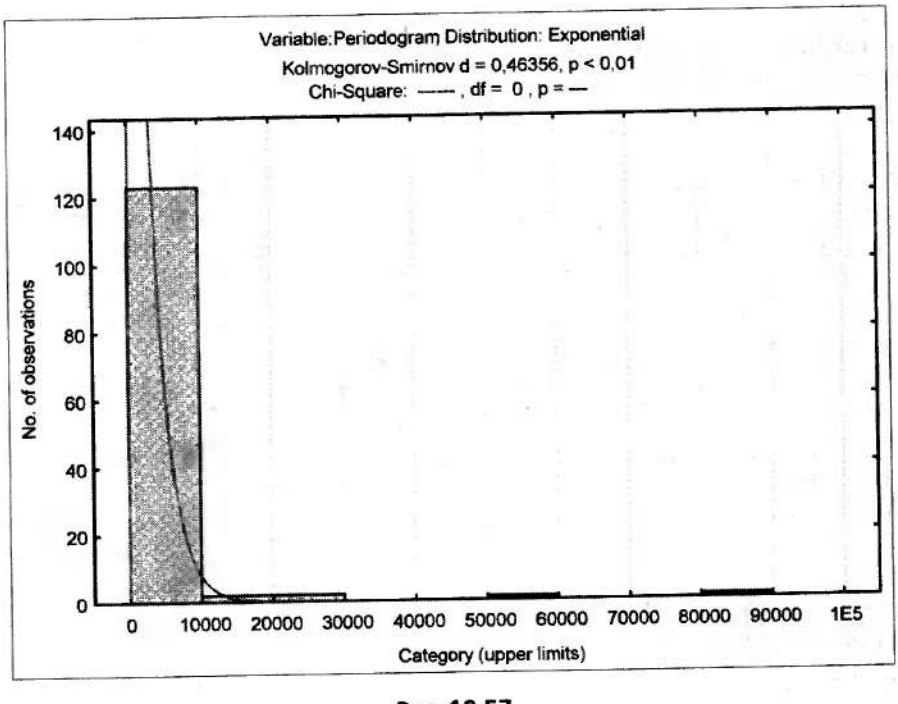


Рис. 18.57

Из значений оценок упомянутых критериев следует, что гистограмма периодограммы не соответствует экспоненциальному распределению, а значит, и анализируемый ряд не является белым шумом и в нем присутствуют периодичности. К тому же значения ряда не подчинены нормальному распределению — p значительно меньше, чем 0,05 (рис. 18.58).

Если выделить опцию **Period** и нажать кнопку **Periodogram**, программа построит график периодограммы, изображенный на рис. 18.5, по которому также можно судить о наличии или отсутствии регулярных циклов, определить период сезонного цикла (лаг). Если выделить опцию **Period** и нажать кнопку **Spectral density**, программа построит график спектральной плотности (рис.18.59). Сравнив графики периодограммы на рис. 18.5 и построенный график спектральной плотности, легко увидеть разницу между ними — спектральная плотность является результатом сглаживания периодограммы.

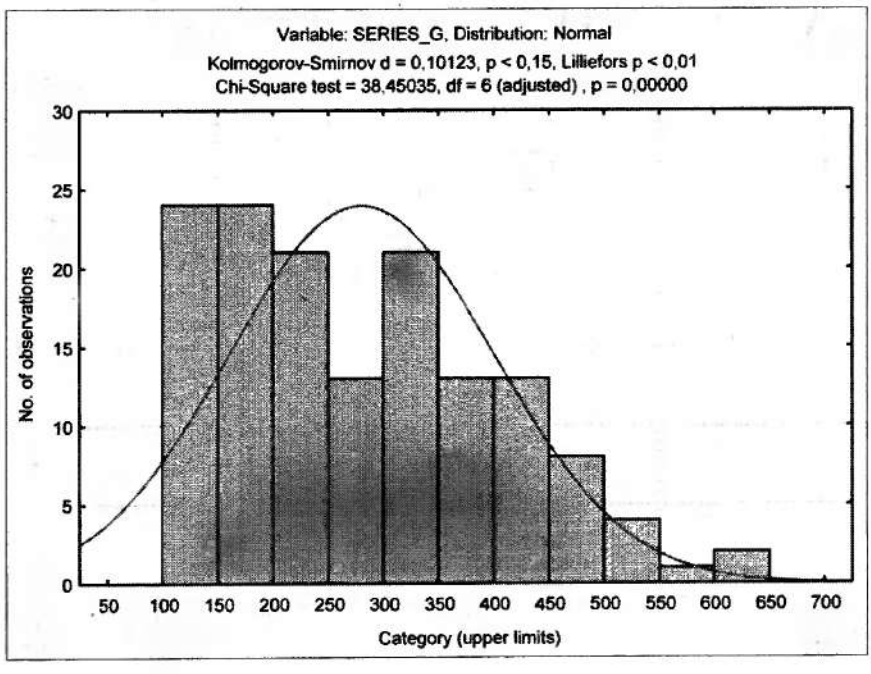


Рис. 18.58

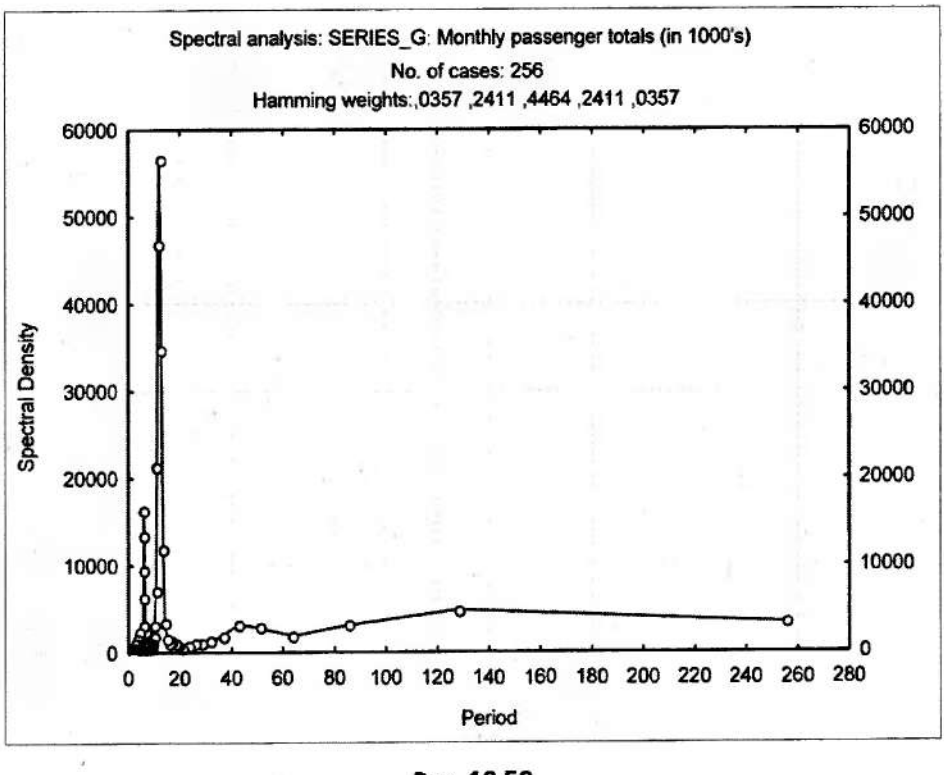


Рис. 18.59

18.7. Анализ распределенных лагов

Анализ распределенных лагов — это специальный метод оценки запаздывающей зависимости между рядами [6]. Он позволяет построить регрессию одного ряда на другой [21], если требуется предсказать значения одного ряда на основе значений другого, например, зависимого ряда на основе измерений со сдвигом независимого ряда. Такого рода зависимости с запаздыванием особенно часто возникают в эконометрике. Например, доход от инвестиций в новое оборудование отчетливо проявится не сразу, а только через определенное время. Более высокий доход изменяет выбор жилья людьми, однако эта зависимость тоже проявляется с запаздыванием. Подобные задачи возникают в страховании, где временной ряд клиентов и ряд денежных поступлений сдвинуты относительно друг друга.

Пусть Y — зависимая переменная, а X — независимая. Эти переменные измеряются несколько раз в течение определенного отрезка времени. Простейший способ описать зависимость между этими переменными дает линейное уравнение

$$y_t = b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \ldots + b_k x_{t-k}, t = 1, 2, \ldots$$

В этом уравнении значение зависимой переменной в момент t — линейная функция переменной X, измеренной в моменты $t-1, t-2$ и т.д. Будем рассматривать это уравнение как специальный случай уравнения линейной регрессии. Если коэффициент b_k переменной с определенным запаздыванием (лагом) значим, то можно заключить, что переменная Y предсказывается (или объясняется) с запаздыванием.

Иногда соседние значения X сильно коррелируют. В самом крайнем случае это приводит к тому, что корреляционная матрица не будет обратимой и коэффициенты b_k не могут быть вычислены. В менее экстремальных ситуациях вычисления этих коэффициентов и их стандартные ошибки становятся ненадежными из-за вычислительных ошибок (ошибок округления).

Алмон [21] предложил специальную процедуру, которая в данном случае уменьшает мультиколлинеарность. Пусть каждый неизвестный коэффициент b_k записан в виде

$$b_k = a_0 + a_1 i + a_2 i^2 + \ldots + a_q i^q, q < k.$$

Тогда во многих случаях (в частности, чтобы избежать мультиколлинеарности) легче оценить коэффициенты a_q, чем непосредственно коэффициенты b_k. Такой метод оценивания коэффициентов называется полиномиальной аппроксимацией. Общая проблема полиномиальной аппроксимации состоит в том, что длина лага и степень полинома не известны заранее. Последствия неправильного определения (спецификации) этих параметров потенциально серьезны [6].

Откройте файл **Teatchers** из **Examples → Datasets**. В файле приведены данные о количестве учащихся, учителей и их заработной плате в США, регистрируемые через каждые 10 лет с 1900 по 1980 г.

Нажмите кнопку **Distributed Lags Analysis** (анализ распределенных лагов) в стартовом окне модуля **Time Series/Forecasting**. Задайте имена переменных для анализа: *CHILDREN* (учащийся), *TEATCHER* (учитель), *SALARY* (заработная

плата). Откроется стартовое окно процедуры **Distributed Lags** (рис.18.60). В группе опций **Method** производится выбор метода оценивания коэффициентов регрессии. Опция *Unconstrained Polynomial lags* (обычные лаги) задает оценивание без использования полиномиальных лагов. Опция *Almon polynomial lags* определяет коэффициенты полиномиальной регрессии для вычисления коэффициентов b_k.

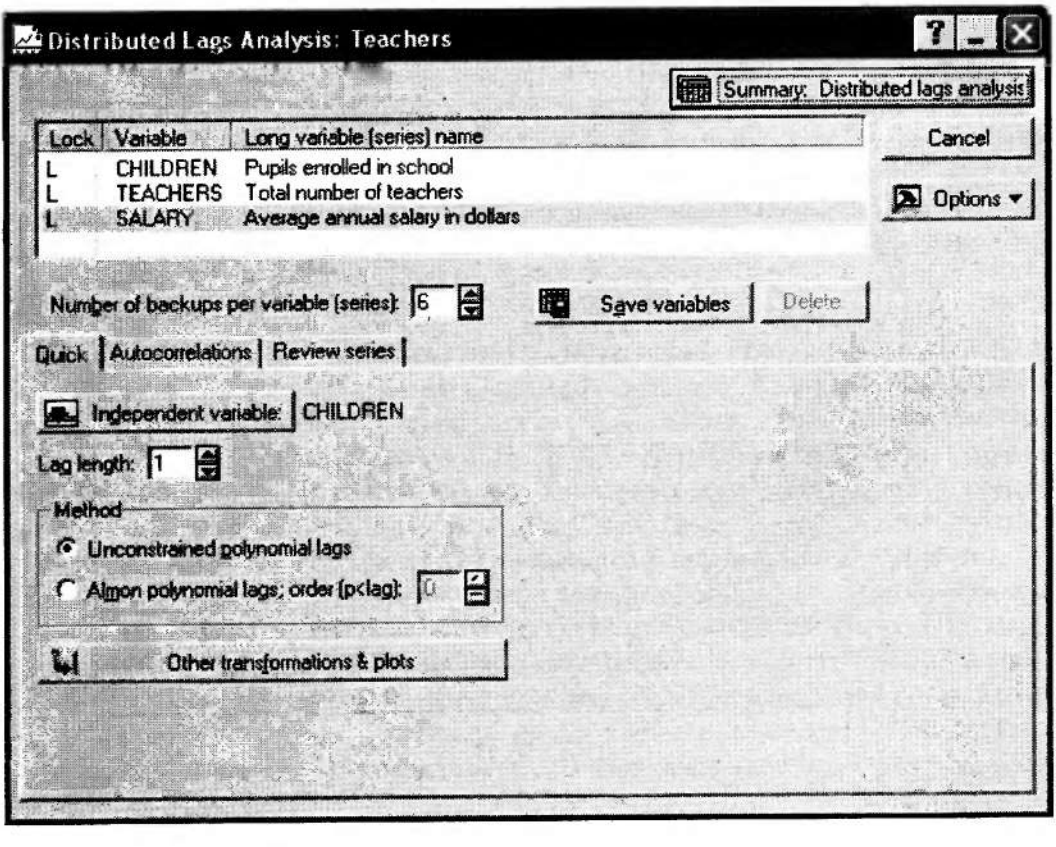


Рис. 18.60

Функциональная кнопка **Independent variable** (независимая переменная) позволяет выбрать независимую переменную, поле **Lag length** (длина лага) — задать величину сдвига одного ряда относительно другого.

На вкладке **Review series** группа кнопок **Review and plot variables** (просмотреть переменные и построить графики) позволяет всесторонне просмотреть данные в электронных таблицах и построить графики, в частности, графики из двух списков переменных в различных шкалах.

В активном рабочем окне высветите переменную *SALARY*, нажмите кнопку **Independent variable** и выберите независимую переменную — *CHILDREN*. Выделите опцию *Unconstrained Polynomial lags*. Установите значение **Lag length** — 2 и нажмите кнопку **Summary: Distributed Lags Analysis.** Появится таблица

(рис. 18.61), в которой приведены оценки коэффициентов регрессии, стандартные ошибки, значения *t-критерия*, соответствующие уровню значимости *p*. Большие значения *p* (все значительно больше 0,05) свидетельствуют о статистической незначимости оценок. О несостоятельности оценок говорят и большие значения стандартных ошибок. Воспользуйтесь процедурой Алмона для вычисления коэффициентов регрессии. Выделите опцию *Almon polynomial lags*, установите значение **order** – 1 и нажмите кнопку **Summary: Distributed Lags Analysis**. Из появившейся таблицы (рис. 18.62) видно, что Lag = 1 соответствует незначительная стандартная ошибка и *t-критерий* статистически значим на уровне 0,1.

	Polyn. Distr. Lags; Regression Coefficients (Teachers) Indep: CHILDREN: Pupils enroll Dep: SALARY : Average annu Lag: 2 R=,9174 R-square=,8416 N:7			
Lag	Regressn Coeff.	Standard Error	t(4)	p
0	-0,00021727357(	0,000281701358	-0,7712904616:	0,483568282472
1	0,00089564510(	0,000489451859	1,8298941627(	0,141244911195
2	-0,00051225584(	0,000503821740	-1,0167402632(	0,366773969256

Рис. 18.61

	Almon Polyn. Distr.Lags; Regression Coefficients (Teachers) Indep: CHILDREN: Pupils enroll Dep: SALARY : Average annu Lag: 2 Polyn. order: 1 R=,8560 R-square=,7327 N:7			
Lag	Regressn Coeff.	Standard Error	t(4)	p
0	0,000008026892	0,000286737126	0,027993903765	0,979007999982
1	0,000085559661	0,000035972522	2,378472656005	0,076115626976
2	0,000163092431	0,000344687020	0,473160928771	0,660763459634

Рис. 18.62

Таким образом, со статистической достоверностью можно утверждать, что зависимость между рядами *SALARY* и *CHILDREN* имеет вид

$SALARY(t) = 0{,}000086 \cdot CHILDREN(t-1), t = 2, 3, \ldots$

Заметим, что лучший результат можно получить, если установить значение **Lag length** –1.

Перейдите на вкладку **Review series** и нажмите кнопку **Plot two var lists with different scales** (графики двух переменных с разными шкалами), появится график (рис. 18.63), на котором левая вертикальная ось соответствует шкале измерения переменной *CHILDREN*, правая *SALARY*.

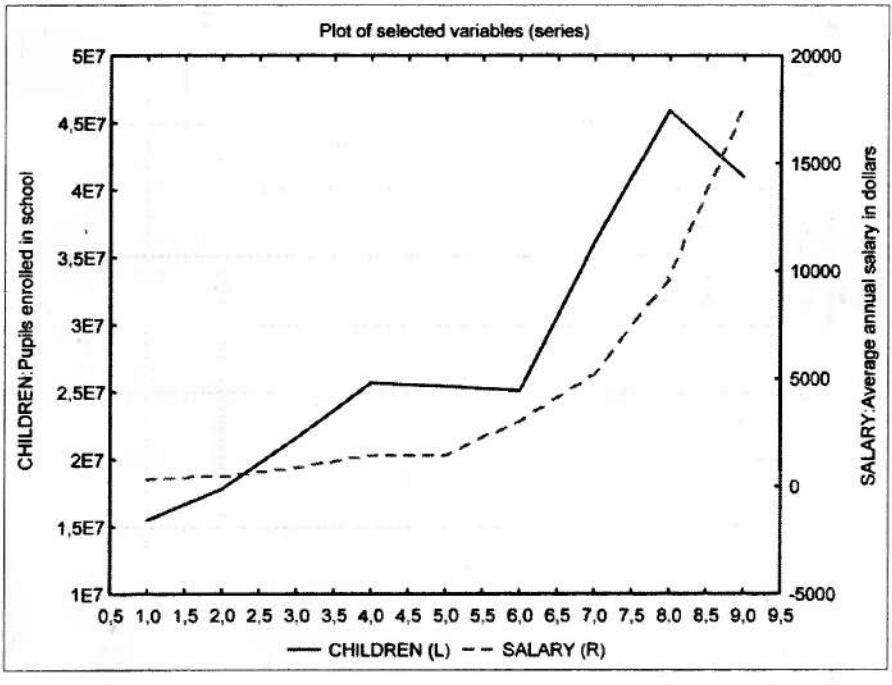

Plot of selected variables (series)
CHILDREN:Pupils enrolled in school
SALARY:Average annual salary in dollars
5E7
4,5E7
4E7
3,5E7
3E7
2,5E7
2E7
1,5E7
1E7
20000
15000
10000
5000
0
-5000
0,5 1,0 1,5 2,0 2,5 3,0 3,5 4,0 4,5 5,0 5,5 6,0 6,5 7,0 7,5 8,0 8,5 9,0 9,5
CHILDREN (L)
SALARY (R)

Рис. 18.63

Глава 19
Создание макросов

Очень часто при статистической обработке однотипных наборов данных приходится периодически многократно выполнять одну и ту же серию операций. Записав макрос, содержащий необходимую последовательность нужных операций, которая может занимать, к примеру, несколько часов путешествий по меню, установку нужных параметров и т.д., при следующей порции данных можно сократить это время до нескольких секунд.

Поэтому создание макросов — полезная и зачастую необходимая процедура, которая присутствует во многих программных продуктах, в том числе и в программе *STATISTICA*. Основное ее назначение — автоматизация обработки данных и соответственно значительная экономия времени.

При помощи макросов также можно разрабатывать приложения с собственным интерфейсом. Код, созданный таким образом, может быть легко интегрирован в большинство сред программирования [6]. Макросы создаются при помощи языка *STATISTICA VISUAL BASIC (SVB)*. Существует несколько методов создания макросов *SVB*.

1. Автоматическая запись макроса. Каждый раз при выполнении процедур из меню *Statistics* или *Graphs, SVB* записывает в фоновом режиме программный код, соответствующий всем спецификациям процедур и параметрам вывода. Этот код может впоследствии многократно выполняться и редактироваться. В процессе редактирования можно изменять настройки процедур анализа, используемые переменные и их спецификации, файлы данных, добавлять элементы пользовательского интерфейса и т.д.
2. Макросы могут быть написаны с нуля с помощью профессиональной среды разработчика *SVB*. Данная среда представляет собой удобный редактор

программного кода с мощным отладчиком. Кроме того, имеется наглядный мастер создания диалогов, а также множество других удобных функций для эффективного написания макросов.

3. *SVB* макросы могут создаваться на основе уже готовых программ на *VISUAL*, написанных в других приложениях (например, *MICROSOFT EXCEL*), путем добавления встроенных процедур и функций *STATISTICA*.

Как простые, так и очень сложные *SVB* программы и макросы, включающие расширенный интерфейс пользователя и работу с файлами, могут быть запущены прямо в *STATISTICA*. Но так как *SVB* удовлетворяют промышленным стандартам совместимости, можно использовать функции *SVB* в других приложениях (например, *MICROSOFT EXCEL, MICROSOFT WORD* или установленная отдельно среда языка программирования *VISUAL BASIC*). Однако когда вы запустите приложение SVB или попытаетесь вызвать функции *STATISTICA* из другого приложения, все вызовы к специфичным функциям *STATISTICA* (в отличие от функций *MS VISUAL BASIC*) будут выполнены, только если соответствующие библиотеки *STATISTICA* есть на компьютере, на котором эти вызовы производятся. Таким образом, пользователь программы должен иметь лицензию на использование соответствующих библиотек функций *STATISTICA*. Библиотека функций *STATISTICA* (более 10 000 процедур) открыта для использования не только из *VISUAL BASIC* (как встроенного, так и стандартного), но и из любого другого совместимого языка программирования, например, из *C/C ++, JAVA* или *DELPHI*.

В *STATISTICA* предусмотрено три категории макросов, которые могут быть автоматически написаны. Для активации этих макросов в меню **Tools** на панели инструментов выделите команду **Macro**.

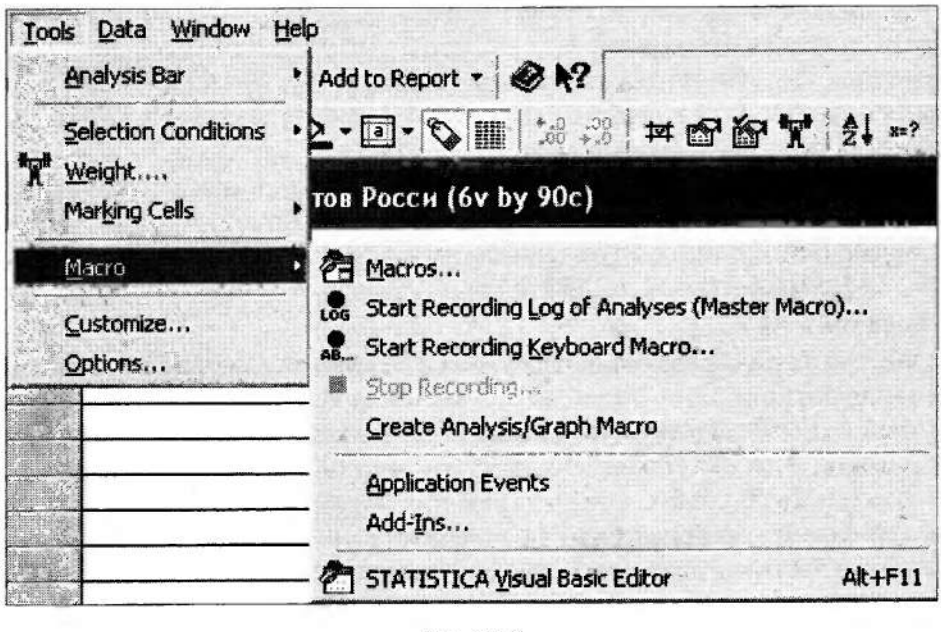


Рис. 19.1

Analysis/Graph Macro (макрос анализа/графика) — макросы создаваемые для конкретных типов анализа из меню **Statistics** и **Graphs**. В макрос записываются все настройки, параметры, присущие данному типу анализа, а также переменные, над которыми он проводится. После выбора модуля или процедуры из указанных меню в фоновом режиме осуществляется запись всех выполняемых действий: выбор переменных, изменение параметров и др. В любой момент можно перенести записанную информацию (код макроса *VISUAL BASIC*) в окно редактора макросов *VISUAL BASIC*. Заметим, что команда **Create Analysis/Graph Macro** (создать макрос) доступна либо из меню **Tools** (сервис), если предварительно открыто окно соответствующего модуля, либо через кнопку **Options** в этом же окне, либо, если окно модуля свернуто, через быстрое меню нажатием правой кнопки мыши на кнопке свернутого модуля.

Log of Analyses (Master Macro) (мастер-макрос (журнал)) — макросы, содержащие любую последовательность модулей из меню **Statistics** или **Graphs**. В мастер-макрос записывается последовательность проведенных анализов с указанными для них параметрами и переменными от момента включения записи макроса до ее отключения. Такая запись объединяет различные модули, выбранные в меню **Statistics** или **Graphs**. В отличие от простого **Analysis Macro**, запись **Master Macros** может быть приостановлена и возобновлена. Запись мастера-макроса начинается при нажатии кнопки записи и приостанавливается нажатием кнопки останова. Все действия, совершенные между этими событиями, записываются в соответствующей последовательности: выбор файлов данных, операции преобразования переменных, выбор элементов и др.

Key board Macro (клавиатурный макрос) — макросы, содержащие последовательности нажатия клавиш во время проведения анализа. При остановке записи в окне редактора *SVB* откроется простая программа, содержащая одну команду **SendKeys** с символами, которые соответствуют клавишам, нажатым при проведении анализа. Данный тип макроса довольно прост — он не записывает контекст, в котором происходило нажатие клавиш (т.е. команды, которые при этом выбирались), но данное свойство может быть полезно для решения определенных задач.

Диалог **Macros** (рис. 19.1) используется для редактирования, удаления или запуска существующих макросов, а также для создания новых макросов.

Все три категории макросов имеют одинаковый синтаксис и могут быть впоследствии модифицированы, но из-за различий в способах создания каждый из них имеет свои преимущества.

Рассмотрим пример создания **Analysis/Graph Macro** для модуля «Факторный анализ».

Откройте файл данных **Factor** из библиотеки **Eamples** и произведите те же манипуляции и в той же последовательности с модулем **Factor Analysis**, которые описаны в параграфе 14.2.

Далее в окне **Factor Analysis Results** нажмите кнопку **Options** и в открывшемся меню выберите команду **Create Macro...** (рис. 19.2). В появившемся окне **New Macro**, в поле **Name** укажите имя макроса или оставьте без изменения появившееся имя *Macro 1* (рис. 19.3).

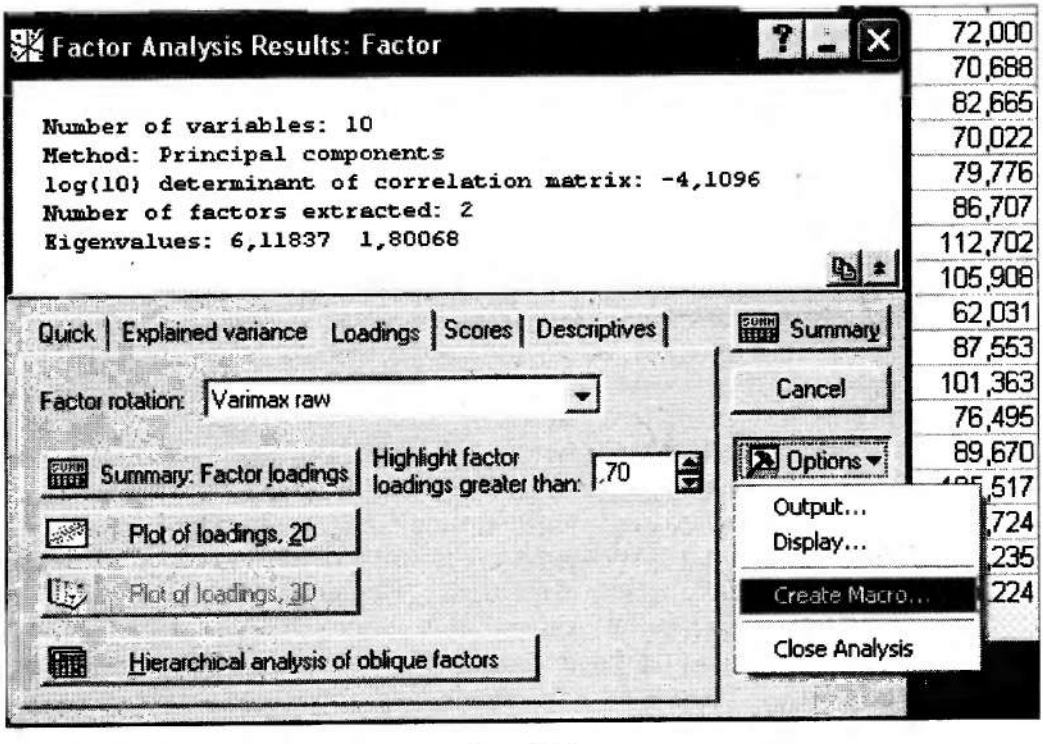


Рис. 19.2

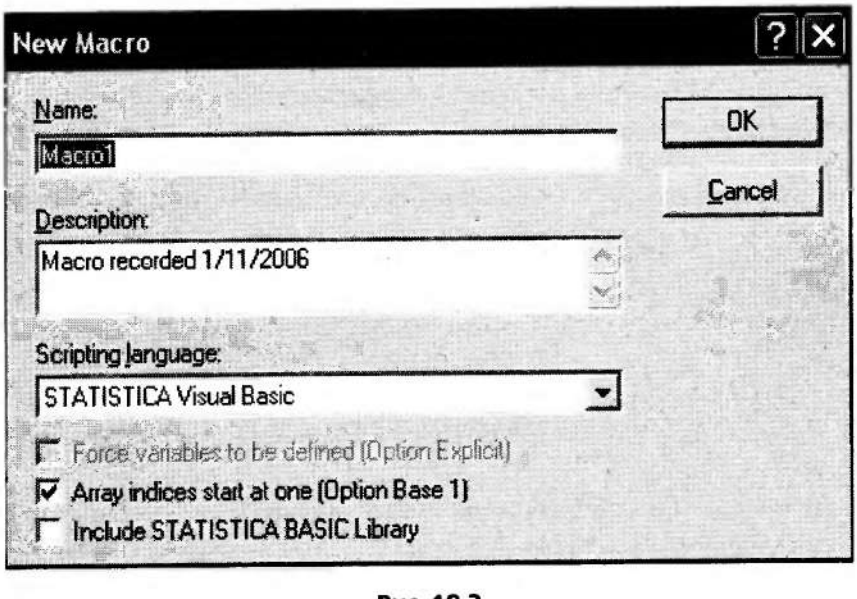


Рис. 19.3

В поле **Description** записана дата создания макроса, здесь же можно записать необходимую дополнительную информацию. В поле **Scriting language** указан язык программирования. Опция *Array indeces start at one (Ootion Base 1)* требует,

чтобы элементы массива начинались с первого. Опция *Include STATISTICA BASIC Library* требует подключения библиотеки *STATISTICA BASIC*.

Нажмите **OK**. Откроется макрос, на рис. 19.4 приведен фрагмент записанного макроса.

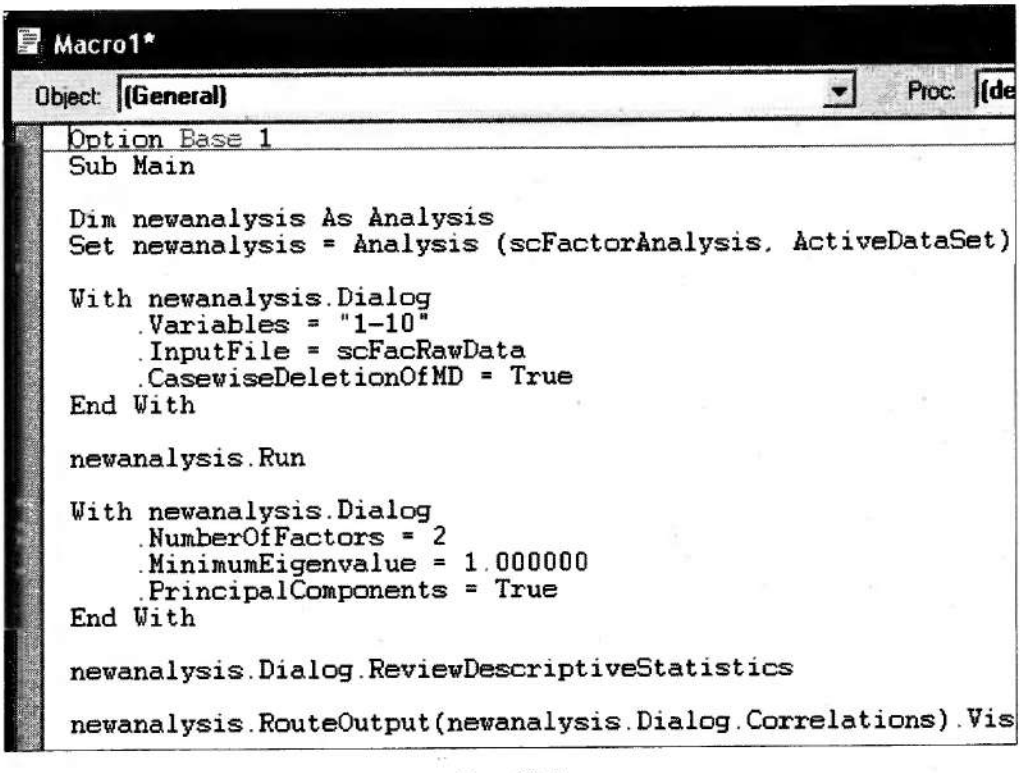


Рис. 19.4

Сохраните макрос через меню **File->Save...** Сохраненным макросом можно воспользоваться при анализе аналогичных данных. Однако количество переменных должно остаться неизменным, в противном случае нужно изменить их число в тексте макроса (*Variables* = и *Dialog. ResultsVariables* =).

Для того чтобы воспользоваться ранее созданным макросом, надо открыть таблицу с данными, выбрать файл макроса, например, через меню **File->Open...** и открыть его. На панели инструментов появится кнопка для запуска макроса (рис. 19.5). Нажмите эту кнопку. Программа произведет все необходимые расчеты, построит таблицы и графики в соответствии с командами, записанными в макросе.

Аналогичным образом можно создать и использовать **Analysis Macro** для любых модулей и процедур из меню **Statistics** и **Graphs**.

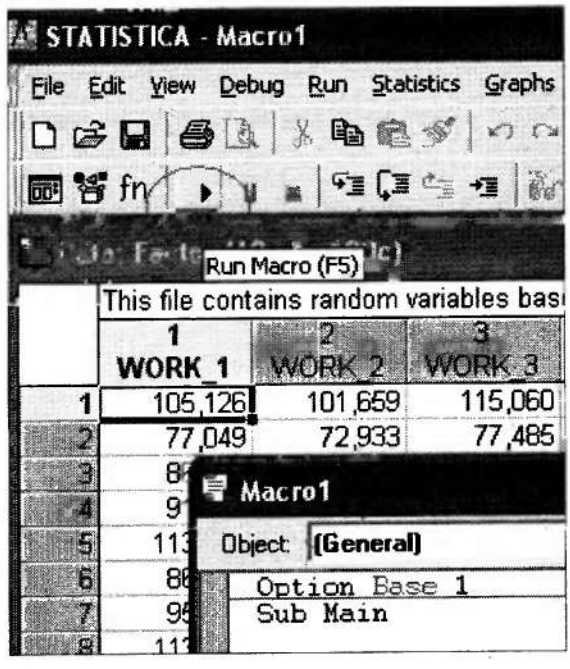


Рис. 19.5

Рассмотрим пример создания **Log of Analyses (Master Macro)**. Как уже отмечалось, данный тип макросов позволяет объединить различные виды анализов из меню **Statistics** и **Graphs** в цепочку. Кроме самих анализов в данные макросы записывается большинство действий пользователя, включая открытие файлов, манипуляции с данными и т.д. *STATISTICA* записывает все параметры и переменные каждого анализа и объединяет их в той последовательности, в которой они проводились. Запись **Master Macros** можно начать в любой момент из меню **Tools** главного меню, выделив команду **Macro**. В появившемся меню (рис. 19.1) выберите команду **Start Recording Log of Analyses (Master Macro)**. В верхнем левом углу файла данных появится небольшое окно (рис. 19.6) с двумя кнопками. С этого момента начинается запись макроса.

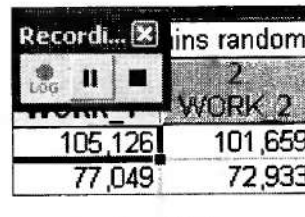


Рис. 19.6

Если необходимо временно приостановить запись макроса, а затем вновь восстановить, надо нажать на кнопку  – **Pauses/resumes Macro Recording**.

Если потребуется остановить запись макроса, нужно нажать на кнопку  — **Stop Recording**, откроется окно **New Macro** (рис. 19.3). Остальные действия по открытию, сохранению и использованию макроса идентичны изложеным для **Analysis/Graph Macro**.

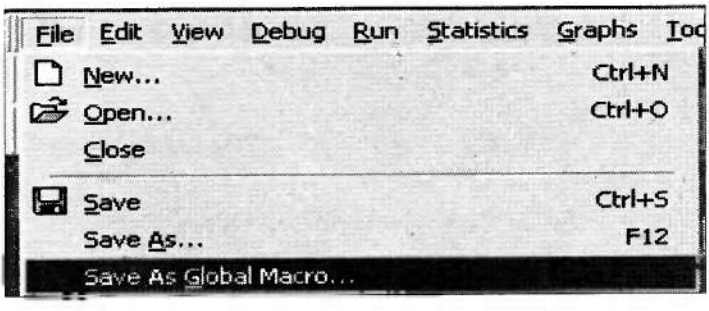


Рис. 19.7

Если некотрые макросы используются достаточно часто, то есть смысл сохранить их как **Global Macros** через меню **File**. В этом случае они сохраняются в корневой папке *STATISTICA* и всегда доступны через меню.

Заключение

Мы живем в эпоху стремительного роста научных знаний и интенсивного проникновения информационных технологий во все области человеческой деятельности. Пакеты прикладных программ (*ППП*) являются неотъемлемой частью современных информационных технологий. Особое место среди *ППП* занимают статистические пакеты по обработке данных. Модули *ППП STATISTICA 6.0* охватывают практически все разделы прикладной статистики.

Многие модули имеют сложную структуру, содержат большое число вкладок, состоящих из множества процедур и опций, описать подробно каждую не представляется технически возможным и целесообразным. Основное достоинство использования *ППП* — это возможность экспериментировать при реализации тех или иных статистических методов. Только экспериментируя, можно накопить необходимый опыт и приобрести практические навыки и знания работы с программой *STATISTICA*. Поэтому изложенный материал позволит пользователю самостоятельно в полном объеме разобраться с рассмотренными в книге методами статистического анализа.

Буду признателен читателям за замечания и пожелания по содержанию и оформлению книги. Адреса для сообщений: г. Краснодар, 350040, Ставропольская 142, Кубанский государственный университет, факультет прикладной математики, Халафяну А.А., или khaliphyan@kubannet.ru

Библиографические ссылки

Приведенные издания могут использоваться в качестве рекомендуемой литературы.

1. *Елисеева И.И., Князевский В.С., Ниворожкина Л.И., Морозова З.А.* Теория статистики с основами теории вероятностей. М.: Юнити, 2001.
2. *Боровиков В.П., Боровиков И.П.* STATISTICA. Статистический анализ и обработка данных в среде Windows. М.: Филинъ, 1997.
3. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.
4. *Дюк В., Самойленко А.* Data Mining. СПб.: Питер, 2001.
5. *Перегудов Ф.И., Тарасенко Ф.П.* Введение в системный анализ. М.: Высшая школа, 1989.
6. *STATISTICA* (Версия 6.1). Электронное руководство.
7. Вероятность и математическая статистика: Энциклопедия. Под ред. *Ю.В. Прохорова. М.*: Большая Российская энциклопедия, 2003.
8. *Алексахин С.В., Бадлин А.В., Николаев А.Б., Строганов В.Ю.* Прикладной статистический нализ. М.: ПРИОР, 2001.
9. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы: Учебник. М.: Финансы и статистика, 2000.
10. *Боровиков В.П. STATISTICA.* Искусство анализа данных на компьютере. Для профессионалов. СПб.: Питер, 2001.
11. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. М.: ЮНИТИ-ДАНА, 2001.
12. *Реброва О.Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М.: Медиа Сфера, 2003.

13. *Холлендер М., Вульф Д.* Непараметрические методы статистики. М.: Мир, 1983.
14. *Боровиков В.П.* Популярное введение в программу STATISTICA. М.: Компьютер-Пресс, 1998.
15. *Ширяев А.Н.* Вероятность, статистика, случайные процессы. М.: МГУ, 1973.
16. StatSoft.Inc.(2001). Электронный учебник по статистике. М. StatSoft.WEB: http://www. StatSoft.ru/home/textbook/default.htm.
17. *Елисеева И.И., Курышева С.В., Костеева Т.В., Бабаева И.В., Михайлов Б.А.* Эконометрика. М.: Финансы и статистика, 2001.
18. *Блехман И.И., Мышкис А.Д.* Механика и прикладная математика. М.: Наука, 1983.
19. *Львовский Е.Н.* Статистические методы построения эмпирических формул. М.:Высшая школа, 1988.
20. *Томас Р.* Количественный анализ хозяйственных операций и управленческих решений. М.: Дело и Сервис, 2003.
21. *Боровиков В.П., Ивченко Г.И.* Прогнозирование в системе STATISTICA в среде WINDOWS. М.: Финансы и статистика, 1999.