УРАЛЬСКАЯ АКАДЕМИЯ ГОСУДАРСТВЕННОЙ СЛУЖБЫ

Х. М. Биккин, А. В. Полтавец, С. Ю. Шашкин

КОМПЬЮТЕРНЫЙ АНАЛИЗ ДАННЫХ ДЛЯ МЕНЕДЖЕРОВ

Учебное пособие по курсу «Математические методы и компьютерные технологии в управлении»

Для студентов специальности
080500.068 «Менеджмент»

ЕКАТЕРИНБУРГ
2007

Рекомендовано к печати
учебно-методической комиссией УрАГС

Б603 **Х. М. Биккин, А. В. Полтавец С. Ю. Шашкин**
Компьютерный анализ данных для менеджеров:
Учеб. пособие. – Екатеринбург: УрАГС, 2007. – 304 с.

Рассмотрены основные математические модели и методы современного углубленного анализа данных, которые традиционно входят в направление, получившее в литературе название Data Mining. Изложение сопровождается большим числом примеров, позволяющих получить навыки практического применения современных программных пакетов для анализа данных социально-экономического содержания. Учебное пособие предназначено для студентов магистратуры и аспирантов специальности «Менеджмент», а также уже работающих специалистов, формируя навыки принятия управленческих решений на основании углубленного анализа данных. По каждой из обсуждаемых тем приводится большое число задач для самостоятельного решения. Данные для учебных примеров и задач содержатся на прилагаемом к изданию CD-диске.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ АВТОРОВ

Специалист в области государственного и муниципального управления должен ясно понимать, что благосостояние населения напрямую зависит от производительности труда работников и эффективности принятия управленческих решений. Осознание связи между доходом и производством помогает увидеть единственный реальный источник экономического благосостояния – увеличение производительности труда.

Жизненный уровень (доход) повышается при увеличении объема производства (выпуска нужных людям товаров). Объем выпуска продукции в расчете на одного работника определяет различия в заработках работников разных стран. Например, средний рабочий в Соединенных Штатах лучше обучен, применяет более производительные машины и пользуется преимуществами более эффективной экономической организации общества, чем такой же рабочий в России или Китае. Управленческий персонал в США принимает более эффективные управленческие решения. В результате этого производительность среднего американского рабочего существенно выше, нежели в России. Не будь этого, его заработки были бы не выше, чем в других странах. Ни профсоюзное, ни забастовочное движения не в состоянии увеличить производительность труда; скорее наоборот, победа профсоюзов зачастую достигается за счет уменьшения инвестиций, которые являются базой для увеличения производительности труда в будущем.

Таким образом, обеспечить высокую производительность труда можно только за счет:

- внедрения эффективных технологий;
- эффективного менеджмента;
- повышения образовательного уровня занятых в экономике;
- разумных инвестиций, обеспечивающих повышение производительности труда в ближайшей перспективе.

Одной из причин лидирующего положения Японии в области экономики является ориентация менеджмента в этой стране на увеличении производительности труда, а не на максимизацию прибыли в отличие от западных стран.

Многие склонны считать, что государство в состоянии обеспечить эффективное управление экономикой. В действительности это большое заблуждение. Государство – это всего лишь форма организации общества – институт власти, в рамках которого люди коллективно принимают решения и ведут определенную деятельность. Поэтому нет никакой гарантии, что политика, одобренная большинством избранных предста-

вителей народа, будет содействовать экономическому прогрессу. Наоборот, есть все основания опасаться, что всенародно избранные власти могут принимать решения, подрывающие общественное благосостояние, если большинство избирателей не сможет самоограничить свои намерения.

Учитывая вышесказанное, можно констатировать, что экономический прогресс в значительной степени определяется специалистами в области управления, которые должны наиболее эффективным способом организовать производство, обеспечить обучение персонала и использование новых технологий, правильно определить направление инвестиций.

Между тем менеджер любого уровня при принятии решений основывается лишь на доступной ему информации о предмете управления. Поэтому от качественных характеристик этой информации, таких как адекватность, полнота, достоверность, своевременность, непротиворечивость и т. п., непосредственно зависит эффективность его работы.

Основной парадигмой курса является представление о том, что анализ данных является составной частью процесса принятия управленческого решения в социальной сфере или сфере экономики.

Действительно, эффективное управление возможно только в том случае, когда имеются достоверные знания об объекте управления. Если эти данные ошибочны, то ошибочным будет и принятое решение.

Кроме этого аргумента, который представляется достаточно важным, есть еще по крайней мере один аргумент в пользу широкого применения анализа данных при принятии решений.

Очень часто опыт противопоставляют знаниям, отдавая явное предпочтение опыту. В действительности опыт зачастую сводится к неформализованным знаниям, которые трудно уложить в определенную логическую схему (эта схема может быть либо очень сложной, либо вообще отсутствовать). До недавнего времени реальный опыт в таких случаях был незаменим. Однако в конце прошлого столетия широкое распространение получили экспертные системы, которые в состоянии добыть неформализованные знания из данных. Если раньше анализ данных использовал в основном лишь методы математической статистики, которые лучше приспособлены для анализа количественной информации, то современные экспертные системы способны анализировать и качественную информацию. В первую очередь речь идет о нейросетевом моделировании и компьютерных программах, способных найти в данных все логические правила и взаимосвязи.

Это направление анализа данных получило в зарубежной литературе название Data Mining («раскопка данных») и в настоящее время широко используется наряду с традиционными статистическими методами анализа данных.

Целью курса является знакомство с современными методами углубленного анализа данных с использованием наиболее популярного в среде социологов пакета SPSS (*Statistical Package for Social Science*) и методами Data Mining с использованием пакетов для нейросетевого моделирования и выявления скрытых правил и закономерностей.

В связи с изучением статистических методов анализа данных закономерен вопрос, нужно ли знать сущность статистических методов для успешного применения статистических пакетов. Дать однозначный ответ на этот вопрос крайне затруднительно. С одной стороны, если пользователь не имеет представления о том, что может дать тот или иной метод статистического анализа, он просто не сможет воспользоваться результатами анализа и правильно их интерпретировать. С другой стороны, такой профессиональный пакет для анализа данных, как SPSS, содержит несколько сотен различных статистических процедур, даже беглый обзор которых в справочной системе SPSS занимает свыше 3 тыс. страниц. Правильный подход, на наш взгляд, состоит в том, чтобы понять принципы использования основных статистических методов, которые широко применяются на практике (их число не столь велико). Когда появится некоторый опыт в использовании статистического пакета для анализа данных, освоение новой статистической процедуры с использованием справочной системы SPSS уже не будет представлять большого труда.

Первая глава содержит лишь краткий обзор методов и проблем анализа данных и является по существу вводной.

В следующих четырех главах подробно рассмотрены алгоритмы интеллектуального анализа данных и критерии их применимости. Эти главы содержат также систему примеров с пошаговыми инструкциями выполнения наиболее сложных статистических процедур и подробной интерпретацией получаемых результатов. Каждая из этих глав завершается списком задач, решение которых способствует развитию навыков практического анализа данных.

Что касается задач, включенных в учебные материалы, то некоторые из них не являются оригинальными и были уже опубликованы в учебных изданиях, список которых приведен в конце книги. Мы не стали сопровождать такие задачи дополнительными ссылками.

ГЛАВА 1. СОВРЕМЕННЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ И ПРИНЯТИЯ РЕШЕНИЙ

1.1. Роль новых технологий в анализе данных

В настоящее время деятельность любого предприятия (коммерческого, государственного, медицинского, научного и т. д.) сопровождается регистрацией и записью на носители информации огромных объемов данных, касающихся деятельности предприятия. Возникает естественное желание использовать имеющуюся информацию для извлечения спрятанных в этих данных знаний, которые могут помочь оптимизировать управление технологическими процессами, улучшить деятельность организации, более точно узнать законы функционирования, присущие таким сложным объектам, как муниципальные или государственные органы управления, производственные или торговые предприятия.

Можно, конечно, действовать старым проверенным способом, т. е. использовать широко известные универсальные пакеты программ для статистического анализа. В этом случае потребуются опытные специалисты, которые, с одной стороны, владеют предметной областью и в состоянии генерировать продуктивные модели взаимосвязей в анализируемых данных, а с другой, способны эффективно использовать достаточно сложные универсальные пакеты. В связи с всевозрастающими объемами данных, которые требуют анализа, традиционный подход исчерпал свои возможности и, в принципе, не позволяет решать задачи анализа данных оперативно.

Хотелось бы автоматизировать процесс анализа и сделать его более объективным, а именно получить некоторую технологию, которая бы автоматически извлекала из данных новые нетривиальные знания в форме моделей, зависимостей, законов и т. д., гарантируя при этом их статистическую значимость.

Поэтому в конце XX века был предложен другой путь – путь создания специализированных пакетов программ, в которых используются практически те же алгоритмы статистической обработки данных, что и универсальные пакеты (никакими другими методами наука просто не располагает), но вся «кухня» статистического анализа скрыта от пользователя. Более того, такие программы способны сами генерировать возможные модели организации данных, проверять их на

имеющемся материале и предлагать пользователю наиболее значимую модель взаимосвязи данных.

Новейшая технология получила название KDD (*Knowledge discovery in databases* – дословно «обнаружение знаний в базах данных») – аналитический процесс исследования человеком большого объема информации с привлечением средств автоматизированного исследования данных с целью обнаружения скрытых в данных структур или зависимостей. Иногда эту технологию называют также Data Mining – раскопка данных.

KDD предполагает как минимум три этапа анализа данных. Первый этап включает предварительное осмысление и неполную формулировку задачи (в терминах целевых переменных), преобразование данных к формату, пригодному для автоматизированного анализа. На втором этапе средствами автоматического исследования данных производится обнаружение скрытых структур или зависимостей. На третьем этапе производится апробация обнаруженных моделей на новых, не использовавшихся для построения моделей данных и интерпретация человеком обнаруженных закономерностей.

KDD (или Data Mining) – это синтетическая область, впитавшая в себя последние достижения искусственного интеллекта, численных математических методов, статистики и эвристических подходов. Цель технологии – нахождение моделей и отношений, скрытых в базе данных, таких моделей, которые не могут быть найдены обычными методами. Следует отметить, что на плечи ЭВМ перекладываются не только рутинные операции (скажем, проверка статистической значимости гипотезы), но и операции, которые ранее было отнюдь не принято называть рутинными (выработка новой гипотезы). KDD позволяет увидеть такие взаимоотношения между данными, которые прежде даже не приходили в голову исследователю и применение которых может способствовать увеличению эффективности принимаемых решений или функционирования предприятия в целом.

Выявленная модель скорее всего не сможет претендовать на абсолютное знание, но она дает аналитику некоторое преимущество уже самим фактом обнаружения альтернативной статистически значимой модели. Наиболее важная цель KDD применительно к реальным системам – это улучшение понимания существа процессов. Технология KDD не заменяет аналитиков или менеджеров, а дает им современный, мощный инструмент для улучшения качества работы, которую они выполняют. И, ко-

нечно, технология нахождения нового знания в базе данных не может дать ответы на те вопросы, которые не были заданы исследователем.

В литературе, посвященной анализу баз данных, очень часто можно встретить и упоминание о технологии OLAP (*On-Line Analytical Processing*), т.е. технологии оперативной аналитической обработки данных в базах данных. Различие OLAP и KDD состоит в том, что при использовании технологии OLAP пользователь сам формирует модель – гипотезу об отношениях между данными – и после этого, используя серию запросов к базе данных, подтверждает или отклоняет эту гипотезу.

Средства OLAP обычно применяются на ранних стадиях процесса анализа данных. Они помогают лучшему пониманию структуры данных, что, в свою очередь, позволяет более эффективно добывать новые знания в базе данных.

Схема анализа данных с использованием технологии KDD приведена на рис. 1.1.

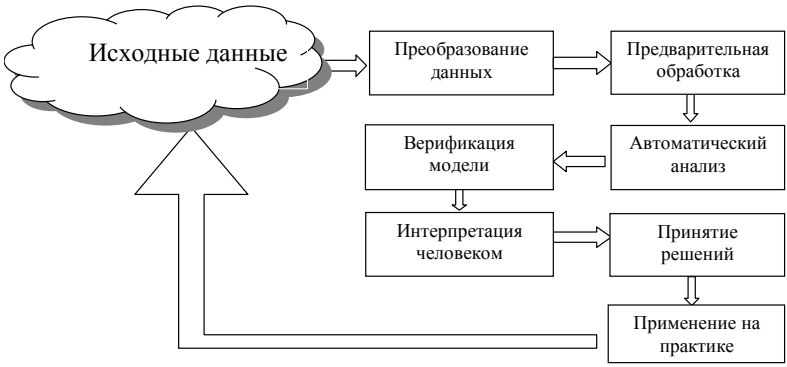


Рис.1.1. Схема анализа данных с использованием технологии KDD

Естественно возникает вопрос о том, насколько широко распространены первый (основанный на использовании универсальных пакетов) и второй (ориентированный на применение специализированных программ) подходы в анализе данных. В научном сообществе, где достаточно много квалифицированных специалистов, а результаты требуют строгой аргументации, шире распространены универсальные статистические пакеты. В бизнес-сообществе значительно больший интерес проявляется к специализированным программам, которые хорошо себя зарекомендовали при решении конкретных задач. Впрочем, такое разделение не является безусловно строгим. Современные универсальные пакеты содержат в своем составе специализи-

рованные модули, которые можно с полным основанием отнести к программным продуктам Data Mining.

Алгоритмы Data Mining в большинстве случаев являются программной реализацией хорошо известных методов многомерного статистического анализа. Универсальные статистические пакеты используют те же самые алгоритмы. Но в специализированных пакетах, ориентированных на применение технологии KDD, многие детали «кухни» статистического анализа скрыты от глаз пользователя, что, конечно, упрощает использование пакета в ущерб полноте исследования.

Можно выделить пять основных типов задач, для решения которых используются универсальные или специализированные статистические пакеты:

1) выявление взаимосвязей и прогнозирование;
2) классификация;
3) кластеризация;
4) выявление логических правил «если – условие – то результат»;
5) нахождение ассоциаций и последовательностей.

Естественно, возможна комбинация всех перечисленных задач в различных сочетаниях. Ниже мы рассмотрим основные идеи и алгоритмы, с помощью которых решаются перечисленные выше основные задачи анализа данных.

1.2. Основные методы анализа данных

Регрессионные методы

Выявление взаимосвязей и прогнозирование основаны на применении алгоритмов корреляционно-регрессионного анализа. Оставляя в стороне более простую задачу обнаружения корреляции в наборе данных, остановимся подробнее на проблеме построения регрессионных моделей.

Используемые в статистических пакетах регрессионные методы основаны на развитии традиционных статистических методик, в первую очередь – регрессионного анализа. Регрессия в этом контексте означает просто функциональную зависимость одной зависимой переменной y от некоторого набора других (независимых) переменных $x_1, x_2, \ldots, x_k$, которые часто называют регрессорами. Регрессионные методы применяются главным образом для обнаружения функциональных зависимостей в данных.

Пусть имеется набор значений результативной переменной y_i для n случаев ($i = 1, 2, \ldots, n$) и набор значений объясняющих переменных $x_{1i}, x_{2i}, \ldots, x_{ki}$ (предполагается, что значение результативной переменной y для каждого случая может быть объяснено действием факторов $x_1, x_2, \ldots, x_k$). Задача регрессионного анализа состоит в нахождении некоторой теоретической функции $y_i^T = f(x_{1i}, x_{2i}, \ldots, x_{ki})$, значения которой были бы наиболее близки к известным значениям переменной y_i для факторов $x_{1i}, x_{2i}, \ldots, x_{ki}$. При создании регрессионных моделей, как правило, предполагается, что уравнение связи зависимой и независимой переменных определяется линейной функцией. Большинство других моделей регрессии (степенная, экспоненциальная, гиперболическая) сводятся к линейной модели простой заменой переменных. Для линейной модели задача регрессионного анализа может быть сформулирована следующим образом: требуется подобрать коэффициенты b_i регрессионного уравнения

$$y_i^T = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \ldots + b_k \cdot x_{ki} \qquad (1.1)$$

таким образом, чтобы суммарный квадрат ошибки S был минимален:

$$S = \sum_{i=1}^{n} (y_i - y_i^T)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \to \min . \qquad (1.2)$$

Уравнения (1.1) и (1.2) позволяют получить замкнутую систему алгебраических уравнений для определения коэффициентов b_i; $i = 0, 1, 2, \ldots, k$. Метод получения регрессионных коэффициентов, основанный на минимизации суммы квадратов ошибок (1.2), является наиболее распространенным и известен в литературе как метод наименьших квадратов (МНК). Несмотря на кажущуюся простоту, построение регрессионных моделей по методу МНК является весьма сложной задачей. Рассмотрим на этом примере те проблемы, с которыми обычно сталкивается исследователь при анализе данных.

Выявление грубых ошибок

Данные могут содержать ошибочные записи, которые возникли при первичной регистрации или явились следствием ошибок при преобразовании данных. Ошибочные данные могут сильно исказить регрессионную модель. Поэтому их нужно выявить и исключить из анализа. Для этих целей используются алгоритмы *робастного (устойчивого) оценивания*. К сожалению, большинство статистических пакетов не

имеют процедур, позволяющих выявить ошибочные данные, и это приходится делать исследователю чаще всего вручную. Более подробно алгоритмы робастного оценивания будут рассмотрены в связи с изучением возможностей пакета SPSS для анализа данных.

Выбор формы уравнения регрессии

Рассмотрим проблему выбора формы уравнения регрессии. Если анализируемые данные измеряются в интервальной шкале (определение шкал измерения переменных будет приведено ниже), то возникает проблема определения функциональной взаимосвязи результативной и объясняющих переменных. Как уже указывалось, наряду с линейной формой регрессионного уравнения могут быть использованы и нелинейные (по объясняющим переменным) регрессионные уравнения. Большинство нелинейных моделей могут быть линеаризованы с помощью подходящей замены переменных.

Приведем вид регрессионных уравнений для некоторых моделей.

1. Линейная форма. Эта форма модели является основополагающей, и именно она до сих пор и рассматривалась выше. Тем не менее приведем математическую запись модели еще раз:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon. \qquad (1.3)$$

Обратим внимание на то, что свободный член в линейном уравнении регрессии очень часто не имеет экономической интерпретации.

2. Степенная форма уравнения множественной регрессии

$$y = a \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdot ... \cdot x_k^{\beta_k} \cdot \varepsilon\,. \qquad (1.4)$$

Степенная форма уравнения регрессии получила широкое распространение в связи с построением производственных функций. Примером может служить функция Кобба – Дугласа $Y = A \cdot K^{\beta_1} \cdot L^{\beta_2}$.

Степенная форма (1.4) сводится к линейной (1.3), если ввести новые переменные. Действительно, прологарифмируем уравнение (1.4):

$$\ln y = \ln a + \beta_1 \ln x_1 + \beta_2 \ln x_2 + ... + \beta_k \ln x_k + \ln \varepsilon\,. \qquad (1.5)$$

Для линеаризации достаточно ввести новые переменные, обозначив

$$\ln y = Z, \quad \ln x_1 = X_1, \quad \ln x_2 = X_2, \quad \ln x_k = X_k. \quad \ln \varepsilon = e, \quad \ln a = \beta_0\,.$$

В этих переменных получаем линейную регрессионную модель

$$Z = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k + e\,.$$

После того как регрессионные параметры $\beta_0, \beta_1, \beta_2, ..., \beta_k$ будут найдены, следует возвратиться к исходным переменным и вернуться к записи (1.4), учтя, что $a = e^{\beta_0}$.

Для линеаризации степенной модели можно использовать и десятичные логарифмы.

3. Экспоненциальная форма

$$y = e^{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon}. \quad (1.6)$$

Экспоненциальная форма используется, как правило, для моделирования временного изменения изучаемой величины.

Экспоненциальная форма сводится к линейной также простым логарифмированием. Вычисляя логарифм правой и левой частей уравнения (1.6), получаем $\ln y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon$. Обозначив $\ln y = Z$, снова получаем линейную модель.

4. Гиперболическая регрессия

$$y = \frac{1}{\beta_1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon}. \quad (1.7)$$

Гиперболическая форма регрессионной кривой используется при обратной связи признаков. Так, если y – объем плановых инвестиций, а x – норма банковского процента, то между ними существует связь, которая может быть выражена в форме $y = 1/(\beta_0 + \beta_1 \cdot x)$.

В случае гиперболической регрессионной модели линеаризация также производится простой заменой объясняемой переменной. Введем новую переменную $Z = 1/y$. Тогда вместо гиперболической получаем линейную зависимость для новой переменной:

$$\frac{1}{y} = Z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon.$$

5. Регрессия на классе произвольных функций. Наконец, если исследователя не устраивает ни один из вариантов рассмотренных выше моделей регрессии, то можно воспользоваться любыми другими функциями, которые с помощью замен переменных сводятся к линейным. Например, такая регрессионная модель

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot \frac{1}{x_2} + \cdot \beta_3 \cdot x_3^{0,5} + \beta_4 \cdot \ln x_4$$

с помощью замены переменных

$$z_1 = x_1, \quad z_2 = \frac{1}{x_2}, \quad z_3 = \sqrt{x_3}, \quad z_4 = \ln x_4$$

сводится к линейной модели.

Естественно, что после анализа линеаризованной модели нужно вернуться к исходным переменным.

Желательно после возвращения к исходным переменным проверить условие $M(\varepsilon) = 0$, которое просто сводится к тому, что

$$\sum_{i=1}^{n} (y_i - y_i^T) = \sum_{i=1}^{n} \varepsilon_i \approx 0 . \qquad (1.8)$$

Если это условие не выполняется, то, возможно, неправильно выбрана форма регрессионного уравнения или используется недостаточное количество регрессоров (объясняющих переменных). В любом случае невыполнимость условия (1.8) указывает на существенные ошибки в спецификации модели.

6. Логистическая регрессия. Если результативная переменная принадлежит к дихотомической шкале, то, как правило, речь идет о некотором событии, которое может произойти, а может и не произойти. Уравнение логистической регрессии позволяет рассчитать вероятность наступления события в зависимости от значений объясняющих переменных:

$$p_i = \frac{1}{1 + e^{-b_0 + b_1 \cdot x_{1i} + \ldots + b_k \cdot x_{ki} + \varepsilon_i}} . \qquad (1.9)$$

Задача построения логистической регрессионной модели очень похожа на задачу дискриминационного анализа, который позволяет отнести значение результативной переменной к одному из двух возможных классов.

Наконец, результативная переменная наряду с трендовой и случайной составляющими может иметь и сезонную составляющую, приводящую к возникновению периодических колебаний. В этом случае необходимо сначала выделить сезонные колебания и лишь затем строить регрессионную модель.

Обнаружение автокорреляции в данных

Важной предпосылкой построения качественной регрессионной модели по МНК является отсутствие автокорреляции. Случайные отклонения $\varepsilon_i = y_i - y_i^T$, $\varepsilon_j = y_j - y_j^T$ должны быть независимыми случайными величинами. Автокорреляция особенно часто возникает в рядах ди-

намики, представляющих собой данные, например, экономической деятельности предприятий, взятые в разные моменты времени. Для наличия автокорреляции в рядах динамики есть веские экономические аргументы. Действительно, если в текущий период t_n предприятие произвело продукции объемом $y(t_n)$ рублей, то объем производства $y(t_{n+1})$ в следующий временной промежуток в значительной степени определяется достигнутым уровнем производства.

В то же время часто обнаруживается, что значение уровня в некоторой точке временного ряда сильно коррелирует с несколькими предшествующими значениями. Корреляционную зависимость между последовательными уровнями ряда называют *автокорреляцией*. Количественно ее можно измерить при помощи вычисления коэффициента автокорреляции.

Автокорреляция первого порядка характеризует тесноту связи между соседними значениями временного ряда, автокорреляция второго порядка – между отстоящими друг от друга на два периода, и так далее.

Автокорреляция *n*-го порядка определяет степень взаимосвязи уровней ряда, разнесенных на *n* временных периодов.

Предполагая, что возникшая связь между значениями сохранится некоторое время в будущем, мы получаем механизм прогнозирования, основанный на построении регрессии точек ряда на самих себя, т. е. *авторегрессии*.

Авторегрессионные модели разных порядков – первого, второго, в общем случае n -го – можно описать уравнениями следующего вида:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot y_{i-1} + \varepsilon_i; \\ y_i &= \beta_0 + \beta_1 \cdot y_{i-1} + \beta_2 \cdot y_{i-2} + \varepsilon_i; \\ y_i &= \beta_0 + \beta_1 \cdot y_{i-1} + \ldots + \beta_n \cdot y_{i-n} + \varepsilon_i. \end{aligned} \qquad (1.10)$$

Таким образом, если обнаружена автокорреляция в исходном наборе данных, то нужно строить не регрессионную, а авторегрессионную модель. По этой причине желательно иметь простые критерии для своевременного обнаружения автокорреляции во входном наборе данных. Как правило, в состав статистических пакетов включается несколько таких процедур.

Отбор объясняющих переменных. Обнаружение мультиколлинеарности

Не существует метода, который бы сразу указывал, какие из объясняющих переменных следует включить в модель, а какие нет. По

этой причине следует пользоваться одной из возможных стратегий, например стратегией последовательного включения.

1. Вычислить коэффициенты корреляции между зависимой переменной и каждой из объясняющих переменных. Выделить фактор с наибольшим коэффициентом корреляции.

2. Построить уравнение регрессии, учитывая пока только этот один объясняющий фактор. Найти величину скорректированного коэффициента детерминации R^2.

3. Добавить в регрессионную модель следующий фактор, имеющий наибольшую корреляционную связь с зависимой переменной. Построить двухфакторную регрессионную модель и найти новое значение скорректированного коэффициента детерминации.

Если коэффициент детерминации увеличился незначительно, то добавление фактора не улучшает модель, а только затрудняет ее интерпретацию. В большинстве статистических пакетов предусмотрена возможность последовательного включения переменных в модель. Это позволяет хотя бы частично автоматизировать процесс отбора переменных в регрессионную модель.

Другой проблемой, которая возникает на стадии отбора объясняющих переменных, является проблема *мультиколлинеарности* (линейной зависимости) объясняющих переменных. При построении регрессионных моделей мультиколлинеарность обычно проявляется в том, что некоторые из регрессионных коэффициентов оказываются статистически незначимыми при высокой статистической значимости регрессионного уравнения в целом.

Для устранения или уменьшения мультиколлинеарности используется несколько методов. Самый простой из них (но далеко не всегда применимый) состоит том, что из регрессионной модели исключается одна или несколько объясняющих переменных, имеющих высокий коэффициент корреляции с другими переменными. При этом, какие из переменных следует оставить, а какие удалить, решается исходя из контекста задачи.

Другой прием предполагает предварительное использование факторного анализа с целью выявления скрытых (латентных или истинных) переменных. Если наличие мультиколлинеарности установлено и исключить из модели часть регрессоров не представляется возможным, следует перейти к новым переменным, которые являются главными компонентами вектора исходных объясняющих переменных.

Главные компоненты заведомо строятся как ортогональные вектора, что предотвращает появление мультиколлинеарности.

К сожалению, применение факторного анализа и введение латентных переменных (главных компонент) наталкивается на проблему интерпретации. Если содержательную интерпретацию переменных получить не удается, то построенное регрессионное уравнение теряет всякий смысл.

Обнаружение гетероскедастичности остатков

При практическом проведении регрессионного анализа следует обратить внимание на выполнимость предпосылок применимости МНК. В частности, следует убедиться в гомоскедастичности (равной дисперсии ошибки для каждого из наблюдений, т. е. $D(\varepsilon_i) = D(\varepsilon_j)$) исходного набора данных.

На первый взгляд не ясно, о какой дисперсии ошибок идет речь, если каждому значению x_i в исходном наборе данных соответствует одно значение зависимой переменной y_i и, следовательно, одна ошибка ε_i, $i = 1, 2, \ldots, n$. Дело в том, что данные выборочного наблюдения следует рассматривать как некую реализацию значений случайной величины. Если мы возьмем другую выборку, то получим новый ряд значений y_i при тех же самых значениях x_i.

Для обнаружения гетероскедастичности (неодинаковости разброса данных для различных точек наблюдения) можно использовать либо графический метод, либо один из возможных тестов.

В простейшем случае гетероскедастичность можно заметить непосредственно, построив поле корреляции переменных и линию тренда. Такой пример приведен на рис. 1.2.

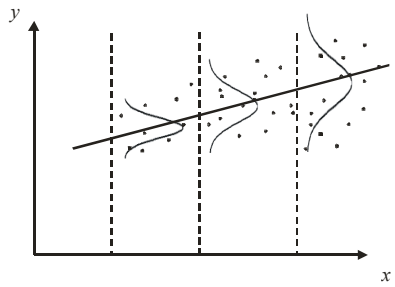


Рис. 1.2. Демонстрация гетероскедастичности исходного набора данных (показаны кривые плотности распределения ошибок)

На практике для графического обнаружения гетероскедастичности по оси абсцисс откладывают значение результативной переменной, получаемой из уравнения регрессии, а по оси ординат либо ошибку, либо квадрат ошибки для соответствующей точки.

При использовании статистических пакетов тесты проверки исходных данных на гомоскедастичность можно «заказать» непосредственно при построении регрессионной модели.

Наконец, если регрессионное уравнение строится по данным, представляющим собой некоторую выборку из генеральной совокупности, необходимо убедиться, что результативная переменная распределена по нормальному закону. Если закон распределения результативной переменной неизвестен, то невозможно оценить ошибку определения параметров регрессионного уравнения и ошибку предсказания результативного признака. Если ошибка неизвестна, то полученные результаты не могут использоваться для прогнозирования или принятия решения на их основе.

Выше мы рассмотрели некоторые проблемы, которые следует решить исследователю при построении регрессионной модели. Очевидно, что в полном объеме автоматизировать даже такую простую задачу, как построение регрессионной модели для произвольного набора данных, просто невозможно. Поэтому в универсальных пакетах статистического анализа пользователь сам должен провести необходимые исследования и выбрать наиболее подходящую модель.

В специализированных пакетах, реализующих технологию KDD, заложены некоторые алгоритмы выбора наиболее подходящей модели. Такие пакеты ориентированы на обработку данных определенной природы и могут на этих данных давать очень хорошие результаты.

Одним из статистических пакетов, который с полным основанием можно отнести к разряду технологии KDD, является пакет PolyAlalyst компании Megaputer Intelligence (www.megaputer.ru). При обработке исходных данных этот пакет позволяет обнаруживать многофакторные зависимости, которым придает затем вид функциональных выражений (класс функций в них практически произволен).

Классификация и кластеризация

Классификация и кластеризация по существу преследуют одну и ту же цель – разделение изучаемой совокупности объектов на группы «схожих» объектов, называемых *кластерами*. Различие состоит в том, что при классификации имеется целевая функция и некоторое число обу-

чающих примеров, на основании которых строится функция дискриминации, позволяющая классифицировать все остальные случаи.

Кластеризация в чем-то аналогична классификации, но отличается от нее тем, что для проведения анализа не требуется иметь выделенную целевую переменную. Отсутствуют также и обучающие примеры. Кластеризацию следует использовать на начальных этапах исследования, когда о данных мало что известно. Для этапа кластеризации характерно отсутствие информации о каких-либо различиях как между переменными, так и между случаями. Просто ищутся группы наиболее близких, похожих случаев (кластеризация объектов) или близкие по смыслу переменные (кластеризация переменных).

Методы автоматического разбиения на кластеры редко используются сами по себе, просто для получения групп схожих объектов. Анализ только начинается с разбиения на кластеры. Коли уж кластеры обнаружены, естественно использовать другие методы Data Mining, чтобы попытаться установить, что означает такое разбиение на кластеры, чем оно вызвано.

К недостаткам кластеризации следует отнести зависимость результатов от выбранного метода кластеризации. Кроме того, методы кластерного анализа не дают какого-либо способа для проверки достоверности разбиения на кластеры (проверки статистической гипотезы об адекватности разбиения).

Более подробно процедура и алгоритмы классификации и кластеризации описаны в главе 4, посвященной статистическому пакету SPSS.

Обнаружение логических закономерностей в данных

Очень часто бывает необходимо не только произвести классификацию объектов, но и выявить, на основании каких логических правил произведена эта классификация. Между тем стандартные методы производят классификацию на основании значений дискриминантной функции и совершенно не помогают выявлению правил принятия решений. В последние годы появилось достаточно большое число программ, в которых делается попытка выявить логические правила, содержащиеся в данных. Для иллюстрации сказанного рассмотрим пример, заимствованный из учебного пособия В. Дюка и А Самойленко[1].

[1] См. список литературы в конце учебного пособия.

На рис. 1.3 схематически изображены лица людей. Эти лица неким экспертом были разделены на два класса. Требуется выяснить закономерности, на основании которых произведено это разделение.

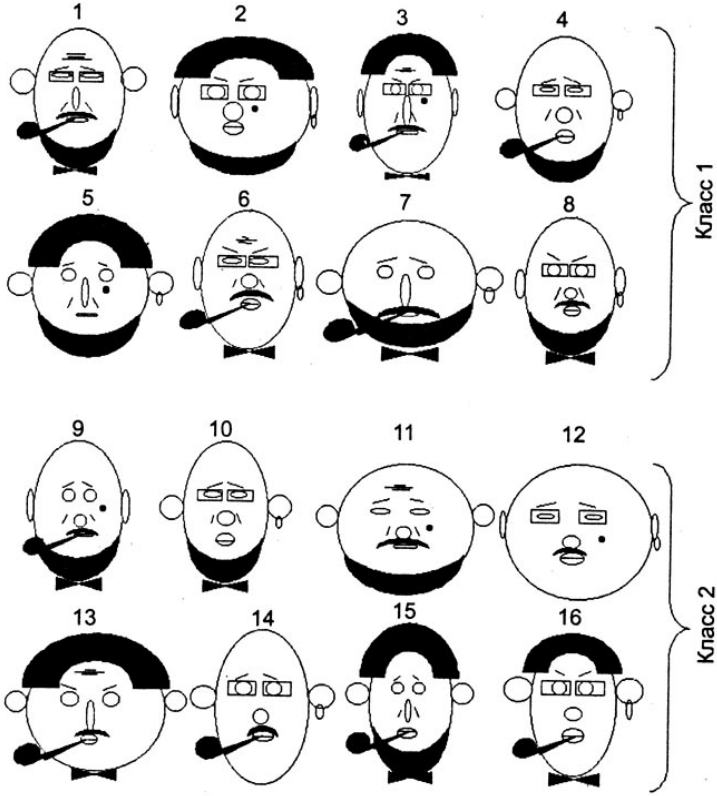


Рис. 1.3. Группировка изображений по двум классам

Нетрудно убедиться, что сформулировать с ходу правила группировки лиц по двум классам весьма затруднительно, хотя эти правила существуют. Можно попытаться решить эту задачу методами дискриминантного анализа. Здесь мы обсудим только основные моменты использования дискриминантного метода для решения этой задачи. Более подробно дискриминантный анализ обсуждается в главе 4.

Введем 16 дихотомных переменных для характеристики изображенных лиц (табл.1.1).

Таблица 1.1

Дихотомические переменные для характеристики изображений на рис. 1.3

Переменная	Характеризует	Значение	Кодируется числом	Значение	Кодируется числом
x_1	голову	круглая	1	овальная	0
x_2	уши	оттопыренные	1	прижатые	0
x_3	нос	круглый	1	длинный	0
x_4	глаза	круглые	1	узкие	0
x_5	лоб	с морщинами	1	без морщин	0
x_6	складку	носогубная	1	нет	0
x_7	губы	толстые	1	тонкие	0
x_8	волосы	есть	1	нет	0
x_9	усы	есть	1	нет	0
x_{10}	бороду	есть	1	нет	0
x_{11}	очки	есть	1	нет	0
x_{12}	родинку	есть	1	нет	0
x_{13}	бабочку	есть	1	нет	0
x_{14}	брови	подняты вверх	1	опущены	0
x_{15}	серьгу	есть	1	нет	0
x_{16}	трубку	есть	1	нет	0

Если подготовить в соответствии с рис. 1.3 таблицу числовых данных и произвести с помощью программы SPSS дискриминантный анализ, то обнаружим, что все случаи классифицированы правильно. Программа создала дискриминантную функцию

$$Z = -43{,}206 - 11{,}653 \cdot x_1 + 24{,}541 \cdot x_2 + 0{,}988 \cdot x_3 + 7{,}456 \cdot x_4 + \\ + 8{,}172 \cdot x_5 - 6{,}864 \cdot x_6 + 24{,}640 \cdot x_7 - 14{,}937 \cdot x_8 + 0{,}790 \cdot x_{10} - \\ - 6{,}765 \cdot x_{11} + 34{,}960 \cdot x_{12} + 14{,}418 \cdot x_{13}, \quad (1.11)$$

значения которой позволяют правильно классифицировать все объекты. Переменные $x_9, x_{14}, x_{15}, x_{16}$ были исключены программой из анализа из-за их линейной взаимосвязи с другими переменными.

Дело в том, что при проведении дискриминантного анализа для каждой из факторных переменных вычисляется *толерантность* – ме-

ра коллинеарности с другими переменными. Для определения толерантности переменной x_i строится регрессионная модель, в которой эта переменная считается результативной, а все оставшиеся факторные переменные x_j $(i \neq j)$ – регрессорами. Если значения переменной x_i удается выразить в виде линейной функции оставшихся переменных, то налицо линейная зависимость переменной x_i от других переменных, и ее следует исключить из дальнейшего анализа. Толерантность $T_i = 1 - R_i^2$, где R_i^2 – фактор детерминации для линейной регрессионной модели, когда результативной переменной является x_i, а факторными – все оставшиеся переменные x_j $(i \neq j)$. Переменная x_i исключается из дальнейшего анализа, если ее толерантность $T_i \leq 0{,}001$.

Дискриминантная функция имеет разные значения для объектов, принадлежащих к первому и второму классам. Среднее значение этой функции для объектов первого класса $\overline{Z}_1 = -7{,}876$, а для объектов второго класса – $\overline{Z}_2 = 7{,}876$ (эти средние значения в дискриминантном анализе называются центроидами групп). Имея дискриминантную функцию, мы можем отнести любой новый объект, который характеризуется теми же переменными, к первому или второму классу. Таким образом, задача классификации решена, но мы не установили логических правил, на основании которых эксперт разделил лица на два класса.

Логические правила дают возможность объяснять взаимосвязь явлений и необходимы для аргументированного прогнозирования. Выяснение логических правил взаимосвязи данных – это и есть обнаружение знаний, скрытых в данных. В любой науке новое знание возникает в результате анализа предварительно накопленных данных. Примеров такого подхода можно привести достаточно много. Достаточно вспомнить периодический закон, открытый Д. И. Менделеевым, который позволил обнаружить логическую причину схожести химических свойств совершенно разных на первый взгляд химических элементов.

За время развития теории анализа многомерных данных было предложено много различных методик поиска логических закономерностей в данных. Наиболее распространенным в настоящее время методом поиска закономерностей в данных является подход, основанный на построении дерева решений.

Дерево решений (*decision tree*) – создает иерархическую структуру классифицирующих правил типа «если... то...» (if-then), имеющих вид дерева. Для принятия решения, к какому классу отнести не-

который объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева начиная с его корня. Вопросы имеют вид «значение параметра А больше *x*?». Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный – то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана с наглядностью и простотой алгоритма. Но деревья решений принципиально не способны находить «лучшие» (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и «ухватывают» далеко не все закономерности. На рис. 1.4 приведено дерево решений, полученное с помощью пакета SPSS 13.

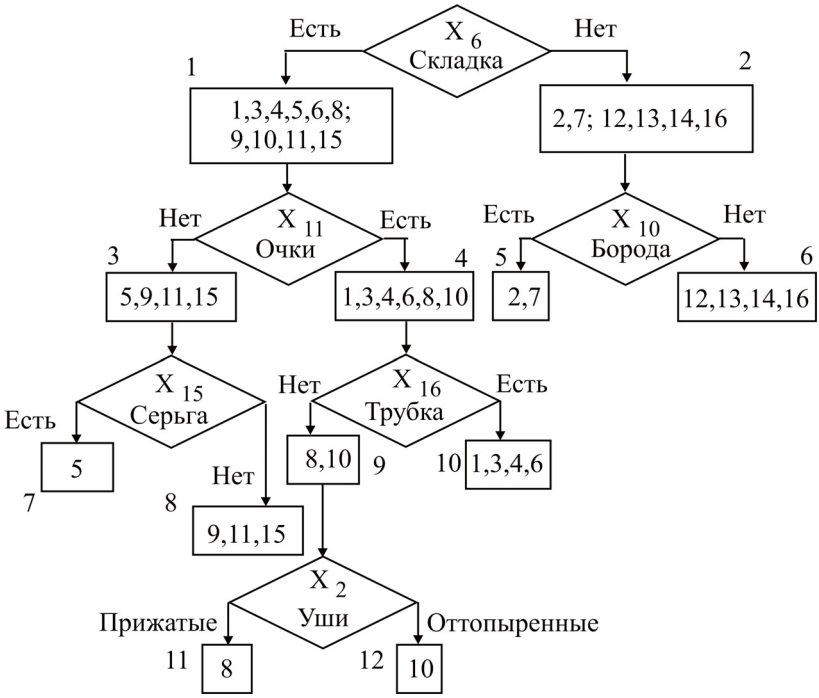


Рис. 1.4. Дерево решений для задачи классификации лиц

Логические правила, найденные программой, интерпретируются достаточно просто. Задачей построения дерева решений является расщепление исходного (корневого) узла на конечные (терминальные) узлы, в которых находятся либо только объекты одного класса, либо

деление которых уже невозможно по другим причинам (например, достигнут предел минимального числа объектов в родительском узле).

Сначала корневой узел расщепляется на два по наличию признака «носогубная складка». Первый узел в левой ветви объединяет объекты 1, 3, 4, 5, 6, 8; 9, 10,11,15, а второй узел – объекты 2, 7; 12, 13, 14, 16.

Для второго узла правой ветви дерева решений можно записать всего лишь два логических условия, позволяющих провести окончательную классификацию:

1. Если нет складки и нет бороды, то объекты относятся ко второму классу. Этим условиям удовлетворяют объекты 12, 13, 14, 16.

2. Если нет складки, но есть борода, то объекты относятся к первому классу. Этим условиям удовлетворяют объекты 2 и 7.

Для узла 1 левой ветви дерева решений требуется еще несколько расщеплений, прежде чем будет произведена окончательная классификация.

Рассмотрим основные принципы построения дерева решений. Более подробно процедура построения дерева решений с использованием SPSS будет обсуждаться в главе 4. Здесь мы обсудим один из возможных подходов, который реализован в алгоритме ID3 (Interactive Dichotomizer). В улучшенном варианте он известен как алгоритм C4.5, предложенный Р. Куинленом (R. Quinlan).

При построении дерева решений сначала нужно выбрать переменную, обладающую наибольшей дискриминирующей силой. В нашем случае одинаковой дискриминирующей силой обладают сразу 7 переменных: $x_2, x_3, x_6, x_{10}, x_{11}, x_{14}, x_{15}$. Под дискриминирующей силой в данном случае понимается просто количество объектов, которые можно правильно отнести к первому или второму классу в соответствии со значением признака. Например, признак x_6 позволяет правильно классифицировать 6 объектов класса 1 (есть складка) и 4 объекта класса 2 (нет складки). Признак x_2 (форма ушей) обладает той же дискриминирующей силой, поскольку позволяет правильно классифицировать 6 объектов класса 2 и 4 объекта класса 1.

Если имеется несколько признаков с одинаковой дискриминирующей силой, то деревья решений получаются разными, в зависимости от того, какой признак выбран первым для классификации. Этот факт можно отнести к недостаткам алгоритма.

После того как все объекты расщеплены по значению первого дискриминирующего признака, возникнут несколько (в нашем случае два) новых множества объектов (узлов). В рассматриваемом случае это множества объектов 1, 3, 4, 5, 6, 8, 9, 10, 11, 15 в узле 1 (есть

складка) и 2, 7, 12, 13, 14, 16 в узле 2 (нет складки). Теперь тот же алгоритм следует применить к каждому из новых множеств (узлов) для выбора новой переменной с максимальной дискриминирующей силой (своей для каждого из узлов) и последующего расщепления каждого из них на две или более групп (по числу градаций дискриминирующей переменной). Этот процесс должен продолжаться до тех пор, пока возникающие множества не будут принадлежать к одному классу или пока не сработают ограничения на число объектов во вновь возникающих множествах. Дело в том, что в практически важных случаях нельзя устанавливать правила, если в множестве всего 1–2 объекта (как в случае дерева решений на рис. 1.4).

Приведенную выше методику построения дерева решений можно формализовать, если использовать понятие энтропии статистического распределения. *Энтропия распределения* показывает, насколько предсказуемыми являются значения результативной переменной. Энтропия распределения связана с информацией, измеряется в битах и определяется формулой

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i). \qquad (1.12)$$

В этой формуле p_i – вероятность встретить значение результативной переменной, относящееся к i-му классу, выбирая случайно некоторый объект из полной совокупности объектов в узле. Для рассматриваемого примера имеется всего два класса по 8 объектов в каждом. Поэтому вероятности $p_1 = p_2 = 1/2$ и $H = 1$.

На первом шаге необходимо выбрать переменную, обладающую наибольшей дискриминирующей силой. Рассмотрим дерево решений, изображенное на рис. 1.4. Построение дерева решений начинается с нулевого узла (корня), включающего в себя все объекты (на рис. 1.4 нулевой узел не изображен). Пусть в качестве первой дискриминирующей переменной взята переменная x_6. Тогда в результате расщепления корневого узла возникнет два дочерних узла (число дочерних узлов равно числу градаций дискриминирующей переменной). В первый узел попало 10 объектов, – объекты 1, 3, 4, 5, 6, 8, 9, 10, 11, 15 (на рисунке узлы отмечены прямоугольниками). Из них 6 объектов относятся к первому классу и 4 ко второму. Во второй узел попали 6 объектов: 2, 7, 12, 13, 14, 16. Из них 2 объекта принадлежат к первому классу и 4 ко второму.

Для каждого из вновь возникших дочерних узлов вычислим энтропию по формуле (1.12), рассматривая, естественно, только те объекты, которые попали в соответствующий узел:

$$H_1(x_6) = -6/10 \cdot \log_2(6/10) - 4/10 \cdot \log_2(4/10) = 0{,}971;$$
$$H_2(x_6) = -2/6 \cdot \log_2(2/6) - 4/6 \cdot \log_2(4/6) = 0{,}918. \qquad (1.13)$$

Оценку энтропии распределения после дискриминации по признаку x_6 найдем по формуле

$$H(x_6) = \sum_{i=1}^{2} \frac{n_i}{n} \cdot H_i(x_6) = 0{,}951,\ n_1 = 10,\ n_2 = 6,\ n = 16. \qquad (1.14)$$

Очевидно, что нашей конечной задачей является такое разделение множества объектов, когда в узлах остаются только объекты одного класса (вероятность появления этих объектов при случайном выборе равна 1) и нет объектов другого класса (вероятность появления этих объектов равна нулю). В этом случае энтропия распределения объектов в узле минимальна (равна нулю). И, наоборот, при полностью случайном распределении объектов в узле энтропия максимальна и равна 1. Поэтому в качестве дискриминирующей переменной следует выбрать такую переменную x_i, для которой значение энтропии (1.14) будет минимальным. Можно легко проверить, что для переменных $x_2, x_3, x_{10}, x_{11}, x_{14}, x_{15}$ мы получим ту же оценку энтропии, что и для переменной x_6. Если в качестве дискриминирующей переменной на первом шаге взять x_1, то получим большее значение энтропии $H(x_1) = 1$. Действительно, вместо (1.13), (1.14) имеем

$$H_1(x_1) = -3/6 \cdot \log_2(3/6) - 3/6 \cdot \log_2(3/6) = 1,$$
$$H_2(x_1) = -10 \cdot \log_2(5/10) - 5/10 \cdot \log_2(5/10) = 1;\ H(x_1) = 1.$$

Таким образом, на первом шаге, выполнив расщепление корневого узла по значениям переменной x_6, получаем два новых узла с номерами 1 и 2. На втором шаге описанный выше алгоритм применяется к узлу 1: ищется переменная, обладающая наибольшей дискриминирующей силой, и узел расщепляется на два новых узла с номерами 3 и 4. Затем процедура повторяется для узла 2, и он расщепляется на два новых – с номерами 4 и 5. Процедура расщепления завершается, если энтропия узла оказывается равной нулю или если будет достигнуто ограничение на минимальное число объектов, которое еще может содержать возникающий дочерний узел.

Достоинством этого алгоритма является то, что узел может расщепляться на число дочерних узлов, равное числу градаций выбранной дискриминирующей переменной. Алгоритм легко обобщается и на случай непрерывной переменной. Для этого достаточно непрерывную переменную разбить на интервалы и расщеплять родительский узел на дочерние в зависимости от того, в какой интервал попадает значение дискриминирующей переменной.

Как уже указывалось, не следует переоценивать возможности программ поиска правил в данных. Практически все известные программы поиска правил в данных отказываются найти правила в следующей тестовой задаче. Имеется 100 объектов, которые характеризуются двумя переменными x_1 и x_2, принимающими числовые значения от 0 до 9 (табл. 1.2). Результативная переменная может принимать всего лишь два значения – «крестик» или «нолик» в соответствии с правилами:

– если $(x_1 > 4)$ и $(x_2 < 5)$, тогда класс 1 – крестики;
– если $(x_1 < 5)$ и $(x_2 > 4)$, тогда класс 1 – крестики;
– если $(x_1 < 5)$ и $(x_2 < 5)$, тогда класс 2 – нолики;
– если $(x_1 > 4)$ и $(x_2 > 4)$, тогда класс 2 – нолики. (1.15)

Таблица 1.2

Данные тестового примера

x_2											
9	×	×	×	×	×	0	0	0	0	0	
8	×	×	×	×	×	0	0	0	0	0	
7	×	×	×	×	×	0	0	0	0	0	
6	×	×	×	×	×	0	0	0	0	0	
5	×	×	×	×	×	0	0	0	0	0	
4	0	0	0	0	0	×	×	×	×	×	
3	0	0	0	0	0	×	×	×	×	×	
2	0	0	0	0	0	×	×	×	×	×	
1	0	0	0	0	0	×	×	×	×	×	
0	0	0	0	0	0	×	×	×	×	×	
	0	1	2	3	4	5	6	7	8	9	x_1

Этот простейший тест оказывается «неподъемным» для специализированных программ поиска правил в данных SEE5 и WizWhy. Модуль построения деревьев решений SPSS 13 легко справляется с этой

задачей, если число интервалов разбиения интервальных независимых переменных установить равным двум.

Поиск ассоциативных правил

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий хлеб, приобретет и молоко с вероятностью 75%. Первый алгоритм поиска ассоциативных правил был разработан совсем недавно, в 1993 году, сотрудниками исследовательского центра IBM. Впервые задача поиска ассоциативных правил была использована для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют *анализом рыночной корзины* (*market basket analysis*).

Ассоциативным правилом называется импликация $X \Rightarrow Y$. Другими словами, целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов *X*, то на основании этого можно сделать вывод о том, что другой набор элементов *Y* также должен появиться в этой транзакции. Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил вида «если *X*, то *Y*», причем поддержка и достоверность этих правил должны быть выше некоторых, наперед определенных порогов, называемых соответственно минимальной поддержкой и минимальной достоверностью.

Транзакцией называется некоторая последовательность действий, представляющих единое целое (например, покупка человеком товаров в магазине).

Поддержкой ассоциативного правила «если *X*, то *Y*» называется доля транзакций во всем наборе данных, содержащих оба элемента *X* и *Y*. Обычно поддержка выражается в процентах.

Достоверностью правила называется вероятность встретить правило «если *X*, то *Y*» среди всех транзакций. Например, если среди всех транзакций в корзине покупателя хлеб и молоко встречается в 3 % случаев, а правило «если в корзине покупателя есть хлеб, то есть и молоко», выполняется в 75 % случаев, то говорят, что достоверность правила 75 % , а его поддержка 3 %.

Задача нахождения ассоциативных правил разбивается на две подзадачи:

1. Нахождение всех наборов элементов, которые удовлетворяют некоторому заданному заранее минимальному порогу поддержки. Такие наборы элементов называются *часто встречающимися*.

2. Генерация правил из наборов элементов, найденных согласно п.1. с достоверностью, удовлетворяющей минимальному порогу достоверности.

Хотя, как было сказано выше, задача поиска ассоциативных правил впервые применялась для анализа рыночной корзины, ассоциативные правила эффективно используются в сегментации покупателей по поведению при покупках, анализе предпочтений клиентов, планировании расположения товаров в супермаркетах, адресной рассылке. Более того, сфера применения этих алгоритмов не ограничивается лишь одной торговлей. Они успешно применяются в медицине для установления типичных симптомов заболевания, Web-администрировании для анализа посещений Web-страниц, статистических органах для анализа данных, (например, по результатам переписи населения) и т. д.

На практических занятиях процедура поиска ассоциативных правил будет рассмотрена на примерах с использованием программ Deductor и RulesWizard, созданных сотрудниками компании BaseGroup Labs (www.basegroup.ru).

1.3. Применение анализа данных в деятельности органов государственного управления

Приведенные выше сведения об основных направлениях анализа данных позволяют наметить перспективные области применения этих методик в сфере государственного и муниципального управления.

В большинстве случаев применение интеллектуального анализа данных дает возможность лишь усовершенствовать действующую и приносящую успех организационную схему. Чаще всего это происходит за счет небольших и постепенных изменений, а не революционных преобразований.

Иногда интеллектуальный анализ данных позволяет обнаружить и некоторые факты, радикально меняющие существующие взгляды. В качестве примера можно привести установленную учеными зависимость между развитием синдрома Рейе у детей (поражение митохондрий – внутриклеточных источников энергии – печени и головного мозга) и приемом аспирина при гриппе или ветряной оспе.

Новизна интеллектуального анализа информации заключается в расширении сферы применения указанных методов в управлении, ко-

торое стало возможно благодаря возросшей доступности данных и удешевлению вычислений. Вместо того чтобы полагаться только на интуицию, специалисты в области управления должны анализировать данные и выбирать лучшие стратегии принятия решений. Легко привести несколько примеров, где углубленный анализ данных уже нашел достаточно широкое применение.

• Мониторинг общественного мнения и анализ социально-экономической ситуации. Определение проблем, формирующих кризисную ситуацию. (Анализ данных необходим не только для выявления проблем региона, но и для определения причин их возникновения.)

• Анализ реакции населения на внедрение различных федеральных и региональных программ. Возможность корректировки программ для повышения их эффективности. Анализ экономического положения и уровня жизни населения.

• Предвыборные исследования и прогнозирование результатов выборов. Диагностика предвыборной ситуации, постоянный контроль за рейтингом кандидатов. Анализ эффективности политической рекламы кандидатов. Анализ основных проблем избирателей, необходимый для разработки направленной предвыборной программы кандидата.

• Анализ средств массовой информации. Выявление наиболее эффективных факторов влияния на мнения различных групп избирателей.

• Общественная безопасность и анализ преступности. (Анализ данных необходим не только для того, чтобы понять, какие типы преступлений совершаются и в каких районах они происходят, но и определить факторы, влияющие на количество и тяжесть преступлений.)
Отслеживание уровня рецидивизма. (Анализ данных нужен для обнаружения причин, по которым правонарушители снова совершают преступления.)

• Планирование школьных округов. Нахождение оптимального месторасположения новых школ в зависимости от условий района, демографической ситуации и других факторов. Отслеживание успеваемости учащихся, выявление факторов, способствующих повышению успеваемости. Контроль за уровнем выполнения обязательных программ и тестов.

• Трудоустройство. Анализ рынка труда. Определение состава и структуры рабочей силы. Анализ заявлений о приеме на работу – раз-

работка профилей претендентов. Отбор претендентов на престижные и конкурсные должности.

• Оценка соответствия размеров уплаченных налогов и имущества. Анализ мошенничеств. Выявление характеристик предприятий и физических лиц, имеющих предрасположенность к совершению мошенничеств. Создание образа надежного и ненадежного клиента в банковском деле.

• Отслеживание болезней и выявление взаимосвязи заболеваемости в регионе с другими показателями уровня жизни. Выявление причин заболеваний и увязка их с природно-климатическими условиями проживания на территории. Создание экспертных компьютерных систем для диагностики и распознавания заболеваний на ранней стадии, а также определения и выявление групп риска – людей, подверженных этим заболеваниям.

• Стратегическое планирование в сфере услуг. Анализ удовлетворенности клиентов, изучение изменений потребностей общества в предоставлении услуг. Профилирование населения. Создание более эффективных программ предложения, рассчитанных на определенные слои населения. Анализ затрат и выявление наиболее эффективных малозатратных программ.

Какую выгоду дает углубленный анализ данных?

С помощью аналитических методов можно получить ответ на серьезные вопросы. Например, простой суммарный отчет позволяет ответить на вопрос, в каком месяце был наибольший объем продаж, а используя углубленные аналитические методы анализа данных, можно ответить на вопрос, почему наибольший объем продаж был в прошлом месяце.

Особенно велико значение анализа данных в условиях жесткой конкуренции, когда даже небольшое увеличение прибыли может оказаться принципиальным в конкурентной борьбе. Нет статистики – нет результативного анализа, а без анализа нет интеллектуального бизнеса. Статистический анализ помогает превратить данные в знания. Например, обработка данных, считанных с ленты кассовых аппаратов супермаркета, позволяет выявить группы товаров, которые приобретаются вместе. Размещение таких товаров недалеко друг от друга на полках большого магазина способно ежемесячно увеличить прибыль на несколько процентов.

Использование статистики в области добычи знаний может существенно улучшить все аспекты работы организации. Например, используя

дискриминантный анализ, легко построить профиль проблемных клиентов банка, для которых риск невозвращения кредита оказывается слишком большим. Таким клиентам банк может либо просто отказать в выдаче кредита, либо выдать кредит на более жестких условиях.

Занимаясь поиском путей повышения вероятности возврата кредитов, аналитик может выявить и другие закономерности, которые, вероятно, он и не ожидал: например, установить, что степень риска невозврата кредитов банка зависит от уровня образования, возраста клиента или времени постоянного проживания на данной территории. После обнаружения таких неожиданных зависимостей необходимо оценить, как эти знания можно использовать в практической деятельности банка.

Интеллектуальный анализ данных – это не веяние моды, которое уйдет так же скоро, как и пришло. Подобные им количественные методики уже длительное время используются во многих отраслях экономики. Показательным примером может служить валютный и фондовый рынки, где тот, кто обладает лучшими математическими методами извлечения закономерностей из зашумленных, хаотических на первый взгляд данных курсов валют или ценных бумаг, может надеяться на большую норму прибыли за счет своих менее просвещенных собратьев.

Можно привести в пример и другие отрасли экономики, где интеллектуальный анализ данных уже давно стал рабочим инструментом менеджера: предсказание рынков, автоматический дилинг, оценка риска невозврата кредитов, предсказание банкротств, оценка стоимости недвижимости, выявление пере- и недооцененных компаний, автоматическое рейтингование, оптимизация портфелей ценных бумаг, оптимизация товарных и денежных потоков, автоматическое считывание чеков и форм, безопасность транзакций по пластиковым карточкам и т. д.

Таким образом, интеллектуальный анализ данных – это магистральный путь развития менеджмента в XXI в., и современный менеджер в полном объеме должен владеть этими методами. Отсутствие навыка анализа данных у специалиста в области управления увеличивает риск принятия им ложных решений, делает его неспособным воспринимать современные технологии управления.

Контрольные вопросы

1.1. Как ставится задача анализа данных с использованием технологии Data Mining?

1.2. В чем сходство и различие технологий OLAP и Data Mining?

1.3. Какова роль анализа данных в информационных технологиях поддержки принятия решений?

1.4. Какие основные типы задач решаются с применением технологии Data Mining?

1.5. В чем различие подходов при использовании универсальных и специализированных пакетов анализа данных?

1.6. Всегда ли возможно полностью доверять результатам анализа данных, выполненным с помощью специализированных пакетов? Перечислите проблемы, с которыми исследователь может столкнуться, например, при выполнении регрессионного анализа.

1.7. В чем сходство и различие задач классификации и кластеризации?

1.8. С помощью каких методов анализа можно выявить правила, содержащиеся в данных?

1.9. Позволяет ли методика построения дерева решений выявить все правила, содержащиеся в данных?

1.10. В чем сущность алгоритмов построения дерева решений?

1.11. Какой смысл имеет энтропия распределения? Как, используя энтропию распределения, можно выбрать переменную для расщепления узла дерева решений на два дочерних?

1.12. Дайте определение понятий «ассоциативное правило», «транзакция», «поддержка и достоверность ассоциативного правила».

1.13. Приведите примеры задач государственного и муниципального управления, в которых широко применяются технологии Data Mining.

ГЛАВА 2. АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ SPSS. ПЕРВЫЕ ШАГИ

Рассмотрим применение универсального статистического пакета SPSS, который является одним из лучших программных продуктов для анализа социологических и экономических данных. С его помощью легко могут решаться практически все виды задач анализа данных, которые обсуждались в предыдущей главе.

Многие из тех вопросов, которые будут обсуждаться в этой главе, изложены на страницах широко известного учебника А. Бююля и П. Цёфеля «SPSS: искусство обработки информации. Анализ данных и восстановление скрытых закономерностей». Эта книга является лучшим и наиболее полным учебным руководством по SPSS, несмотря на то, что за последние годы было издано достаточно много других книг. Большим достоинством этого руководства является то, что оно содержит файлы с исходными данными всех обсуждаемых примеров, что позволяет обучаться в интерактивном режиме (электронную версию книги и файлы примеров к ней можно найти в сети Internet).

Хотя учебной литературы по SPSS достаточно много, вся она рассчитана на специалистов, уже знакомых с принципами применения статистических методов для анализа данных, а не на студентов, столкнувшихся с этой проблемой впервые. В нашем учебном пособии также учитывается, что студенты имеют некоторую первичную подготовку по общей теории статистики и представляют цели и задачи аналитической статистики. Тем не менее, излагая материал, мы будем напоминать некоторые принципиальные моменты аналитической статистики, в частности, это касается основного метода оценки достоверности получаемых результатов – метода испытания статистических гипотез. Кроме того, в этой главе мы рассмотрим некоторые дополнительные примеры и приемы работы, которые не отражены в упомянутой книге А. Бююля. При таком изложении неизбежны некоторые повторы, которые, мы надеемся, будут полезны читателям, поскольку, как известно, задача преподавателя не только в том, чтобы познакомить студента с некоторым набором концепций и методов, но и в том, чтобы научить правильно к ним относиться.

2.1. Типы статистических шкал в SPSS

Практически все известные пакеты анализа данных оперируют исключительно данными, представленными в числовой форме. Поэтому, формируя данные, исследователь ставит в соответствие значениям переменной, имеющей содержательный смысл, числовые значения (например, мужской пол кодируется цифрой 1, женский – цифрой 2). Такое соответствие называется *шкалой измерения переменной*. В зависимости от свойств переменной выделяют номинативную, порядковую (ранговую), интервальную шкалы и шкалу отношений. Первые две шкалы являются неколичественными, а две последние – количественными. Возможность использования для анализа данных той или иной статистической процедуры зависит от шкалы измерения этих данных.

Номинативная (категориальная) *шкала* является самым «низким» уровнем измерения. В этом случае числовое значение приписывается переменным произвольно. Типичным примером переменной, которая измеряется в номинативной шкале, является пол. Например, в социологической анкете пол мужской кодируется цифрой 1, а женский пол – цифрой 2. В данном случае значения 1 и 2 не связаны между собой какими-либо отношениями. Бессмысленным бы было утверждение, что женский пол вдвое больше мужского. Другим примером переменной, измеряемой в номинативной шкале, может служить профессия. Например, при изучении профессионального состава работников цеха можно использовать следующее кодирование: профессия токаря закодирована цифрой 1; профессия слесаря – цифрой 2; профессия электрика – цифрой 3. Ясно, что переменные, измеренные в этой шкале, нельзя подвергать никаким арифметическим, алгебраическим или логическим операциям. Для переменных этого типа невозможно определить наименьшее и наибольшее значение, среднее значение, дисперсию, медиану и, как следствие, нельзя применять параметрическое тестирование (тестирование, основанное на использовании известных параметрах распределения). Поскольку для номинативных переменных нельзя определить понятие ранга, то невозможно определить и понятие ранговой корреляции. Единственный параметр статистического распределения, который здесь имеет смысл, – это мода распределения. В то же самое время переменные номинативного типа могут быть использованы как основание статистической группировки при проведении дисперсионного анализа, который позволяет установить взаимосвязь между переменными, измеренными, например, в номинативной и интервальной шкалах.

Исключением в некоторых ситуациях являются номинативные дихотомные (принимающие альтернативные значения) переменные. Значения этих переменных можно закодировать нулем и единицей. Например, при аудиторской проверке банковских счетов счет может быть оформлен верно или неверно. Верно оформленные счета («удача») кодируем цифрой 0, а неверно оформленные («неудача») – цифрой 1. Если для дихотомной случайной величины вероятность «удачи» p остается постоянной в n повторных испытаниях, то вероятность $P(x)$ выпадения x успехов в этой серии (выборке) определяется биноминальным распределением, которое при условии $n \cdot p > 5$, $n \cdot (1-p) > 5$ аппроксимируется нормальным распределением с математическим ожиданием «удачи» $M(x) = n \cdot p$ и дисперсией $D(x) = n \cdot p \cdot (1-p)$. Все это служит основанием того, что для дихотомных номинативных переменных возможно как интервальное оценивание, так и применение метода испытания статистических гипотез, например, для сравнения математических ожиданий для двух различных выборок.

Порядковая шкала применяется, если переменная выражает степень проявления какого-либо свойства, и ее значения могут быть упорядочены. Например, при обработке анкеты социологического опроса можно использовать порядковую шкалу для кодировании ответов на вопрос о том, представляется ли предлагаемая работа интересной: очень интересная – 3; интересная – 2; малоинтересная – 1; совершенно неинтересная – 0.

В этом случае между значениями переменных можно установить отношения порядка. Очевидно, что интересная работа более привлекательна, чем малоинтересная. Таким образом, вариационный ряд уже можно ранжировать, а значит, есть возможность определить медиану и моду распределения.

Для переменных, относящихся к порядковой шкале измерений, может исчисляться ранговый коэффициент корреляции, а для сравнения различных выборок могут применяться непараметрические тесты, формулы для которых оперируют рангами.

Интервальная шкала предполагает, что можно определить не только порядок значений, но и расстояние между значениями. Эта шкала, однако, такова, что не имеет смысла рассматривать, во сколько раз одно значение больше другого. Примером может служить шкала измерения температуры по Цельсию или Фаренгейту (принятая в США). Очевидно, что понятие разности температур можно опреде-

лить, и оно имеет смысл, а отношение температур – величина, лишенная всякого смысла. Действительно, если утром температура была $+1\ ^0C$, а днем поднялась до $+6\ ^0C$, то можно сказать, что она стала выше на 5 градусов, но нельзя сказать, что стало теплее в 6 раз.

Переменные интервальной шкалы могут обрабатываться любыми статистическими методами без ограничений.

Шкала отношений. Для переменных, измеренных в этой шкале, определены все арифметические и логические операции, которые можно производить с числовыми переменными. Например, мы можем смело заявить, что зарплата в 10 000 руб. вдвое выше зарплаты в 5 000 руб. К шкале отношений относятся и интервальные величины, которые имеют абсолютную нулевую точку (например, абсолютная температура, измеренная в шкале Кельвина). При статистическом анализе в SPSS переменные, относящиеся к интервальной шкале и шкале отношений, обычно не различаются (табл. 2.1).

Таблица 2.1

Типы шкал, применяемых для статистического анализа данных в SPSS

Тип шкалы	Отношения между значениями	Допустимые преобразования	Допустимые статистические расчеты	Наличие нуля и единиц измерения
Номинативная шкала	Отношения неравенства, различия	Установление соответствий	Доля, мода	Нет
Порядковая шкала	Больше, меньше, равно, не равно.	Ранжирование	Доля, мода, медиана	Нет
Интервальная шкала	Равенство, неравенство, больше на, меньше на	Арифметические операции, за исключением умножения и деления	Доля, мода, медиана, среднее арифметическое, дисперсия	Условный нуль, есть единицы измерения
Шкала отношений	Все отношения, допустимые в алгебре	Все арифметические преобразования	Доля, мода, медиана, среднее арифметическое, дисперсия,	Абсолютный нуль, есть единицы измерения

2.2. Подготовка данных для анализа

SPSS имеет встроенный редактор данных, который позволяет создавать, импортировать и экспортировать наборы данных. Данные

можно вставлять в окно редактора данных и через буфер обмена, например из электронных таблиц Excel. Данные можно сортировать в окне редактора данных, производить над переменными вычисления, группировать записи по значениям некоторой атрибутивной переменной, создавать новые переменные, решать вопрос о том, как система будет обрабатывать записи с пропущенными данными, и т. д. Таким образом, можно утверждать, что редактор данных является достаточно удобным средством подготовки данных для статистического анализа, хорошо взаимодействует с другими приложениями Windows.

Рассмотрим несколько примеров создания данных, проведения различных вычислений с использованием данных и импорта файлов данных. Эти примеры далеко не исчерпывают всех возможностей, которыми обладает SPSS для манипуляции данными. Для дальнейшего знакомства с этими возможностями следует обратиться к электронной версии учебника А. Бююля и П. Цёфеля.

Пример 2.1. Создание набора данных средствами SPSS

Построить график плотности нормального распределения, если среднее значение случайной величины $\bar{x} = 5$, а среднеквадратическое отклонение $\sigma = 2$.

Решение

Согласно известному правилу трех сигм 99,73 % всех значений случайной величины, распределенной по нормальному закону, будет находиться в интервале $\bar{x} - 3 \cdot \sigma \le x \le \bar{x} + 3 \cdot \sigma$. Поэтому для построения графика нужно выбрать интервал изменения переменной $-1 \le x \le 11$. Для построения графика плотности распределения нужно создать исходные данные, представляющие собой значения переменной x и функции $f(x)$, где $f(x)$ – значения плотности нормального распределения

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}. \qquad (2.1)$$

Для создания этих переменных запустим SPSS и переключимся на закладку Variable View (описание переменных). В поле Name наберем в первой строке *x*, во второй – *y*. Имена переменных могут быть набраны латиницей. Все остальные поля описания переменных можно пока не трогать. Переключимся на закладку Data View. Подготовим набор исход-

ных данных *x*. Нажимая клавишу Enter, перейдем к строке с номером 120 и введем с клавиатуры число «0». Во всех клетках первого и второго столбца появятся точки, которые символически отображают пропущенные значения переменных *x* и *y*. Для создания числовых значений вектора x выберем опции Transform/Compute, а затем в открывшемся окне определим переменную назначения *x* Target Variable (рис.2.1).

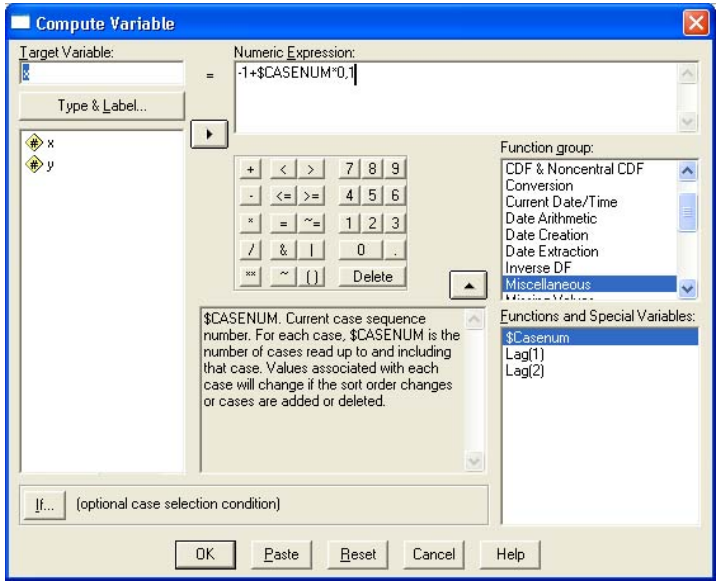


Рис. 2.1. Создание переменной, принимающей значения от –1 до 11 с шагом 0,1

В окне Numeric Expression наберем значение –1, среди доступных функций выберем закладку Miscellaneous (смешанные) функции и в окне Function and Special Variables выберем функцию $Casenum, которая возвращает значение номера строки. Полученное значение номера строки следует еще умножить на 0,1. В этом случае переменная *x* будет изменяться с шагом 0,1 в пределах от –0,9 до 11. Нажав кнопку OK, мы получим искомые значения *x*. Чтобы получить значения плотности распределения *y*, следует снова вызвать окно вычислений, выбрав Transform/Compute. Дальнейшие действия по созданию переменной *y* показаны на рис. 2.2.

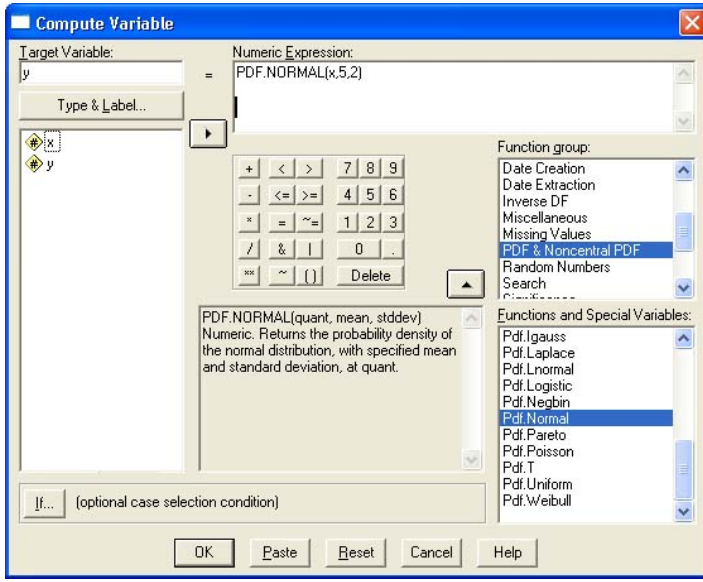


Рис. 2.2. Вычисление плотности нормального распределения величины x с параметрами $\bar{x} = 5$, $\sigma = 2$

В окне Target Variable набираем *y*, затем в окне Function group выбираем раздел PDF& Noncentral PDF, а в окне Function and Special Variables функцию Pdf.Normal, дважды щелкая мышкой по ее имени. В итоге в окне Numeric Expression должно появиться имя выбранной функции PDF.NORMAL(?,?,?), в которой вместо знаков вопросов нужно ввести параметры этой функции: значение переменной, среднее значение и среднеквадратичесое отклонение. Заканчивается создание переменной *y* нажатием клавиши OK. Построить простейший график в SPSS не составляет труда. Предлагаем читателям выполнить эту процедуру самостоятельно.

В рассмотренном выше примере все данные создавались непосредственно в редакторе данных. Такая ситуация практически никогда не встречается. Чаще приходится дополнять уже имеющиеся данные. В частности, весьма полезно производить нумерацию строк (случаев), если такой нумерации нет в исходных данных. В следующем примере мы обсудим более реалистичный случай, когда часть данных импортируется, а часть должна быть вычислена с использованием редактора данных.

Пример 2.2

Имеются данные о распределении студентов заочной формы обучения УрАГС по среднему экзаменационному баллу. Требуется выяснить, можно ли считать это распределение нормальным. Исходные данные в формате электронных таблиц Excel содержатся в файле Пример_2_2.xls.

Средняя оценка	Число студентов	Средняя оценка	Число студентов
2,25	1	3,75	249
2,5	2	4	234
2,75	1	4,25	189
3	39	4,5	183
3,25	98	4,75	78
3,5	222	5	33

Решение

Согласно центральной предельной теореме, если случайная величина Z получена сложением достаточно большого числа n других случайных величин x_i (обычно считается, что это число должно быть $n > 30$), то величина Z будет иметь распределение, близкое к нормальному, независимо от того, какому распределению подчиняются величины x_i. Поэтому есть все основания предполагать, что распределение средних экзаменационных оценок будет близким к нормальному.

Исходные данные из электронных таблиц Excel можно импортировать в SPSS через буфер обмена, предварительно описав эти переменные. Откроем окно редактора данных в SPSS, перейдем на закладку Variable View и в столбце Name в первой строке введем Ball, а в столбце Label (метка) опишем смысловое значение этой переменной – Средняя оценка. Аналогично во второй строке в столбце Name наберем – St_Numbe, а в столбце Label – Число студентов. При организации выдачи результатов SPSS может оперировать не именами переменных, а их смысловыми характеристиками (метками). Для того чтобы реализовать такое переключение, нужно выбрать опции Edit/Optons. В появившемся окне открыть закладку General и выбрать переключатель Display Labels (отображать метки).

После того как переменные определены, переключаемся в окно Data View, открываем лист Excel с данными, копируем их в буфер обмена (если первая строка содержит имена переменных, то ее копировать не следует) и вставляем данные в окно редактора данных SPSS, выделив первую ячейку в первой строке и первом столбце. Более правильно импортировать данные из базы данных Excel, используя встроенную в SPSS возможность импорта дан-

ных. Для этого следует выбрать опции File/Open Database /New Query. Дальше нужно следовать инструкциям мастера импорта.

После того как данные импортированы, найдем средний балл и стандартное отклонение среднего балла. Поскольку каждый балл имеет разный вес (разное число студентов имеют такой балл), то для вычисления среднего балла нужно указать процессору SPSS, что баллы нужно учитывать с весом. Для этого следует выбрать пункты меню Data/Weight Cases и в открывшемся окне указать, какая переменная определяет веса (рис. 2.3).

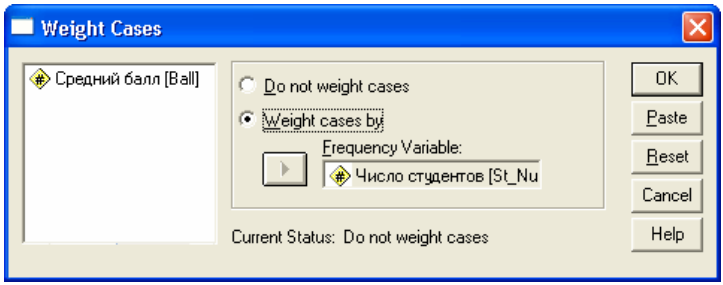


Рис. 2.3. Диалоговое окно, позволяющее определить частоты (веса) для переменной Средний балл

Теперь для нахождения среднего балла и стандартного отклонения достаточно запустить процедуру получения описательной (дескриптивной) статистики для переменной Средний балл, выбирая опции Analyze/Descriptive Statistics/ Descriptives. В открывшемся окне переменную Средний балл следует переместить в окно Variable(s) и, зайдя на закладку Option, отобрать нужные для отображения параметры. Затем следует вернуться в исходное окно и нажать OK. При этом в окне вывода появится таблица дескриптивной статистики с искомыми параметрами.

	N	Среднее	Стд. отклонение
Средний балл	1329	3,9537	0,47995
N валидных (целиком)	1329		

Теперь создадим новую переменную N_St_Numbe, значения которой должны содержать нормально распределенные значения числа студентов имеющих различные средние баллы. Для этого, перейдя на закладку Variable View, нужно сначала описать эту переменную. Вычисление значений переменной N_St_Numbe следует произвести точно так же, как вычисление значений переменной *y* в предыдущем примере. Различие состоит в том, что теперь нам нужна не плотность распределения, а число студентов, которые имеют средний бал в интервалах 2,25 – 2,5; 2,5 – 2,75; 2,75 – 3 и т. д. Поэтому для получения частот

распределения плотность распределения нужно умножить на полное число студентов 1329 и на ширину интервала группировки 0,25. Таким образом, в окне Numeric Expression (рис. 2.2) должно быть набрано выражение 1329*PDF.NORMAL(Ball, 3.9537, 0.4799)*0.25. Следует обратить внимание на то, что разделителем десятичных знаков в этом окне является точка, а не запятая.

Полученные таким образом данные позволяют построить графики эмпирического и теоретического распределения средних оценок студентов, но это будут два различных графика. Чтобы построить две кривых в одной координатной сетке, необходимо представить эти два распределения как две выборки из одной статистической совокупности, различающиеся значением группирующей переменной, в качестве которой возьмем переменную с именем Index. Скопируем значения переменной Ball в буфер обмена и вставим их ниже (начиная со строки 13). Затем скопируем значения переменной N_St_Numbe и вставим их, начиная с 13-й строки, в столбец St_Numbe. Переменная N_St_Numbe больше не нужна, и ее можно удалить.

Создадим новую перемену Index, используя методику, описанную в примере 2.1. По умолчанию SPSS создает переменные интервального (Scale) типа. Переменная Index должна иметь номинативный тип. Поэтому при создании этой переменной нужно в последнем столбце (Measure) выделить ячейку с описанием этой переменной по умолчанию и в выпадающем списке выбрать тип Nominal (номинативная шкала). Для переменных, измеренных в номинативной шкале, нужно определить смысл численных значений, которые принимает эта переменная. Поэтому выделим ячейку Values (значения), соответствующую этой переменной, откроем окно редактирования значений и наберем возможные значения переменной и их смысл. Пример того, как это можно сделать, показан на рис. 2.4. После набора числового и смыслового значений необходимо нажать кнопу Add (добавить). На этом описание переменной Index можно считать завершенным.

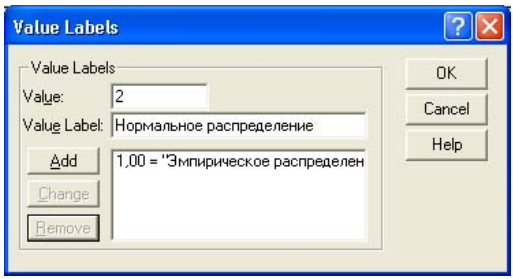


Рис. 2.4. Окно определения смысловых значений номинативной переменной

Числовые значения этой переменной (для первых 12 значений – 1, для остальных – 2) можно присвоить различными способами: просто ввести вручную; ввести вручную значение в одну из ячеек, скопировать в буфер обмена, а в остальные вставить, предварительно их выделив. Можно придумать и несколько других способов. Предоставляем читателям возможность поэкспериментировать и справиться с этой задачей самостоятельно.

После того как значения переменной Index определены, можно построить эмпирическое и модельное распределения в одной координатной сетке. Для этого выберем опции Graphs/Interactive/Line и в открывшемся окне с помощью мыши разместим переменные в окна, как показано на рис. 2.5.

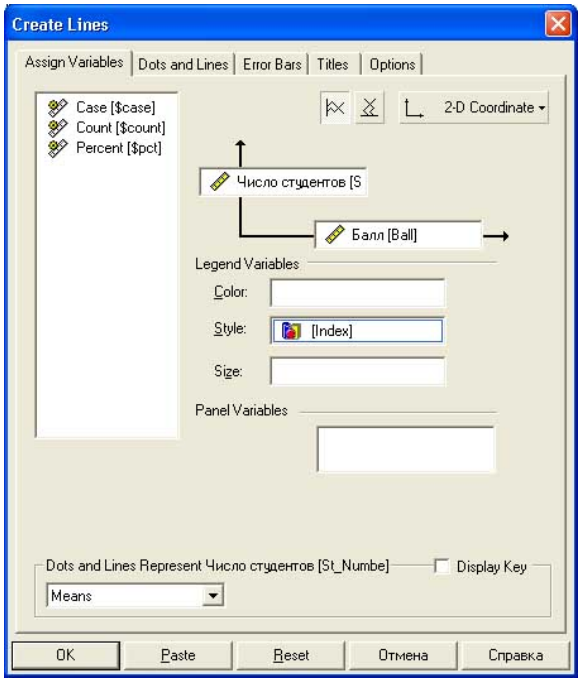


Рис. 2.5. Окно для размещения переменных при построении интерактивных графиков

Группирующую переменную можно было бы разместить в любое из окон: Color, Style, Size. Различие состоит лишь в способе оформления линий. В первом случае они будут различаться цветом, во втором – стилем линии, а в третьем – толщиной линии. Если группирующую переменную разместить в окно Panel Variables, то будет построена серия графиков от-

дельно для каждого значения группирующей переменной. Построенные графики можно дооформить. Это делается примерно так же, как и в Excel. Искомый график приведен на рис. 2.6.

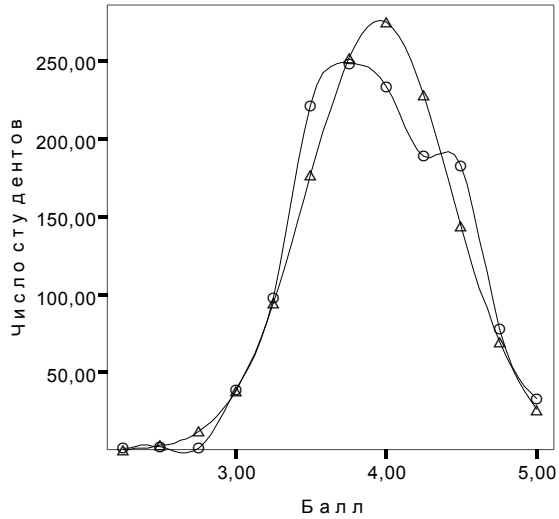


Рис. 2.6. График распределения студентов заочного отделения УрАГС по среднему баллу (кружками отмечено эмпирическое распределение, а треугольниками – нормальное)

Интерпретировать расхождения эмпирического и модельного распределений достаточно легко. Очевидно, что студента приходится отчислять, если средний балл меньше 3. Поэтому эмпирический график в этой области имеет провал. Дополнительный пик в области высоких баллов связан скорее всего с неоднородностью исходных данных. Дело в том, что средний балл юношей и девушек различается, и мы имеем дело на самом деле с двумя близкими по виду распределениями. Возможно также, что этот маленький пик связан с желанием преподавателей слегка завысить оценку и дотянуть ее до пятерки.

После завершения работы нужно сохранить файл данных и файл выходных данных. Сохранение данных не имеет каких-либо особенностей и не требует дальнейшего обсуждения. Желательно помнить, что файлы каждой задачи следует размещать в отдельной папке.

На этом мы завершим первое краткое знакомство со способами подготовки и манипуляции данными. По мере изложения материала мы еще не раз вернемся к этой теме.

2.3. Особенности применения выборочного метода

Выборочный метод зачастую является единственно возможным методом измерения социально-экономических показателей и в любом случае менее затратным. Однако его применение привносит в статистику огромное число проблем, большая часть которых и является предметом изучения аналитической статистики.

Как известно, выборочное наблюдение позволяет определить показатели, которыми характеризуется выборочная совокупность. Естественно встает вопрос: можно ли, зная выборочные показатели, дать оценку генеральных показателей? Под оценкой в данном случае понимается построение доверительного интервала для изучаемого показателя при заданной доверительной вероятности.

Следует ясно понимать, что большинство процедур, которые можно использовать для анализа данных, например в SPSS, связаны с определением степени достоверности получаемых результатов, оценкой применимости той или иной статистической процедуры для анализируемого набора данных. Причиной такого положения является то, что на практике, как правило, используется выборочный метод наблюдения, когда изучению подвергаются не все имеющиеся объекты (генеральная совокупность), а только отобранная по одной из специальных методик совокупность объектов (выборочная совокупность, или выборка), а результаты исследования требуется распространить на всю генеральную совокупность (выборочный метод в статистике).

В силу того, что применимость того или иного метода аналитической статистики для конкретного набора данных может оказаться под вопросом, в SPSS предлагается, как правило, целый набор процедур, имеющих или несколько различающуюся область применимости, или различную мощность критериев (понятие мощности критерия рассмотрено ниже). В большинстве случаев этот набор избыточен, и на практике можно ограничиться небольшим числом статистических методов, позволяющих решать стоящие перед исследователем задачи.

В следующем разделе мы рассмотрим статистические методы, позволяющие оценить ошибку, возникающую при использовании выборочного метода, и некоторые критерии оценки значимости получаемых результатов. Без интервальной оценки и оценки значимости любой результат, полученный с помощью выборочного метода, является случайной величиной и не может использоваться при принятии экономических и управленческих решений.

2.4. Универсальные законы распределения

Выше неоднократно отмечалось, что результаты статистического анализа в действительности являются случайными величинами и нуждаются в статистической оценке их достоверности. Обычно достоверность их проверяется с помощью метода, который принято называть методом испытания гипотез. Гипотезы, в свою очередь, проверяются с помощью некоторых специально подобранных статистических критериев, построенных с помощью универсальных распределений.

Рассмотрим несколько основных законов распределения, составляющих необходимый математический аппарат для построения в дальнейшем статистических критериев, применяемых для оценки справедливости найденных закономерностей.

Причиной, по которой рассматриваемые ниже распределения играют заметную роль в статистике, является универсальность. Для их построения не нужно задавать параметры, как для нормального распределения (напомним, что нормальное распределение задается двумя параметрами – средним значением и дисперсией).

Универсальные распределения однозначно определяются лишь величинами, которые обычно известны, или могут быть найдены из условия задачи, и поэтому служат основой для построения статистических критериев. Мы рассмотрим лишь некоторые, наиболее употребительные распределения и покажем, как они применяются при проверке статистических гипотез. Мы надеемся, что это поможет читателям понять, как используются другие многочисленные критерии проверки статистических гипотез в SPSS.

Распределение χ^2

Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение суммы квадратов k независимых случайных величин, распределенных по стандартному нормальному закону

$$\chi^2 = \sum_{i=1}^{k} Z_i^2,$$

где случайные величины Z_i $(i = 1, 2, \ldots, k)$ имеют стандартное нормальное распределение с математическим ожиданием, равным 0, и дисперсией, равной 1.

Пусть $x_1, x_2, \ldots, x_k$ – случайная выборка объема k из нормально распределенной генеральной совокупности со средним значением $\bar{x}_0$ и дисперсией $\sigma_0{}^2$. Приведем эти величины к стандартному виду путем преобразования $Z_i = \dfrac{x_i - \bar{x}_0}{\sigma_0}$. Очевидно, что величины Z_i подчиняются стандартному нормальному распределению, а величина

$$\chi^2 = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \left(\frac{x_i - \bar{x}_0}{\sigma_0} \right)^2 \qquad (2.2)$$

распределена по закону χ^2.

Функция плотности распределения хи-квадрат зависит лишь от одного параметра – числа степеней свободы *k*.

Числом степеней свободы k распределения называется число независимых значений случайной величины. Это число равно числу наблюдений (числу значений случайной величины) n за вычетом числа уравнений связи l, которые накладываются на эти наблюдения. Например, если величины Z_i связаны линейным соотношением $\sum_{i=1}^{n} Z_i = const$, то число степеней свободы k будет равным $n-1$.

Пример 2.3

Менеджер компании имеет бюджет 150 тыс. рублей на четыре проекта. Сколькими степенями свободы обладает распределение средств по четырем проектам?

Решение

Очевидно, что в данном случае только бюджеты трех проектов являются независимыми величинами. Как только бюджеты трех проектов распределены, то у менеджера не остается выбора, и четвертому проекту будет выделен лишь остаток средств. Таким образом, это распределение имеет три степени свободы.

На рис. 2.7 изображено χ^2 – распределение для различных значений числа степеней свободы. Это распределение легко получить, пользуясь приемами подготовки данных, описанными в примерах 2.1 и 2.2. Стандартная функция, возвращающая плотность распределения хи-квадрат, имеет название PDF.CHISQ(*x, k*), где *x* – аргумент, *k* – число степеней свободы.

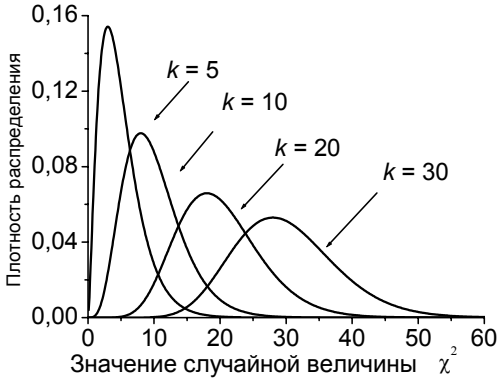


Рис. 2.7. График плотности распределения случайной величины x, подчиняющейся распределению χ^2 для различных степеней свободы

Как видно из рисунка, это распределение асимметрично, но асимметрия уменьшается с ростом числа степеней свободы.

При неограниченном увеличении числа степеней свободы распределение χ^2 приближается к нормальному распределению.

Математическое ожидание и дисперсия этого распределения равны числу степеней свободы и удвоенному числу степеней свободы соответственно: $M(\chi^2) = k, \quad D(\chi^2) = 2k.$

Распределение Стьюдента

Предположим, что было произведено большое число выборок $n > 30$ из нормально распределенной генеральной совокупности и для каждой из выборок получены выборочные средние $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$. Известно, что если объем выборок достаточно велик, то выборочные средние распределены по нормальному закону со средним значением, равным генеральной средней $\bar{x}_0$, и среднеквадратическим отклонением $s_{\bar{x}} = \sigma_0 / \sqrt{n}$. В этом случае величина

$$Z = \frac{\bar{x} - \bar{x}_0}{\sigma_0} \sqrt{n}$$

подчиняется стандартному нормальному закону.

В большинстве случаев значение генеральной дисперсии неизвестно. Поэтому естественно заменить генеральную дисперсию ее оцен-

кой по выборке $\sigma_0^2 = \sigma^2 \cdot n/(n-1)$. В результате получим новую случайную величину

$$t = \frac{(\bar{x} - \bar{x}_0)}{\sigma}\sqrt{n-1}\,. \qquad (2.3)$$

Величина t распределена уже не по нормальному закону, поскольку источником вариации здесь являются случайные величины $\bar{x}$ и σ. Открыл и подробно изучил это распределение английский статистик В. Госсет (Стьюдент) еще в 1908 году.

Распределение Стьюдента зависит от одного параметра – числа степеней свободы $k = n-1$. Одна степень свободы здесь теряется, поскольку n наблюдений связаны одним уравнением, задающим среднее значение :

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i\,.$$

Уже по способу построения ясно, что распределение Стьюдента очень походит на стандартное нормальное распределение. Ниже представлены графики плотности распределения Стьюдента для двух разных значений числа степеней свободы (рис. 2.8).

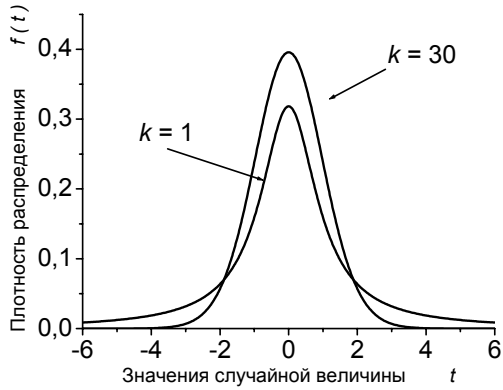


Рис. 2.8. Графики плотности распределения Стьюдента для двух различных значений числа степеней свободы

В SPSS распределение Стьюдента можно построить, используя функцию PDF.T(*x*,*k*), где x – случайная величина, k – число степеней свободы.

При значениях числа степеней свободы $k > 30$ график плотности распределения Стьюдента практически не изменяется и совпадает по виду с графиком стандартного нормального распределения.

Математическое ожидание и дисперсия случайной величины t соответственно равны $M(t) = 0;\ D(t) = k/(k-2),\ k > 2.$ В этой формуле k – число степеней свободы распределения Стьюдента.

Распределение Фишера – Снедекора

Часто возникает необходимость установить, являются ли дисперсии двух или более распределений равными. Для ответа на этот вопрос используется так называемое F-распределение.

Предположим, что существуют две генеральные совокупности, в каждой из которых случайная величина распределена нормально. Пусть для первой совокупности речь идет о случайной величине X с дисперсией $\sigma^2{}_{0x}$, а для второй совокупности – о случайной величине Y с дисперсией $\sigma^2{}_{0y}$. Из этих совокупностей извлечены две выборки объема n_x и n_y. Для каждой из выборок можно вычислить выборочные дисперсии $\sigma^2{}_x$ и $\sigma^2{}_y$. Тогда случайная величина

$$F = \frac{\sigma^2{}_x / \sigma^2{}_{0x}}{\sigma^2{}_y / \sigma^2{}_{0y}} \tag{2.4}$$

подчиняется распределению Фишера – Снедекора. В числителе и знаменателе этой формулы стоят величины, распределенные по закону χ^2, с числами степенями свободы $k_1 = n_x - 1$ и $k_2 = n_y - 1$. Формулу (2.4) поэтому можно переписать в виде

$$F = \frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}, \tag{2.5}$$

где $\chi^2(k_1)$ и $\chi^2(k_2)$ – распределения χ^2 с k_1 и k_2 степенями свободы.

F-распределение имеет асимметричную функцию плотности распределения и зависит от двух параметров – k_1 и k_2.

Графики этого распределения показаны на рис. 2.9. Построить такие распределения в SPSS достаточно просто, если воспользоваться

функцией PDF.F(x,k_1,k_2). Здесь x – случайная величина, k_1 и k_2 – числа степеней свободы.

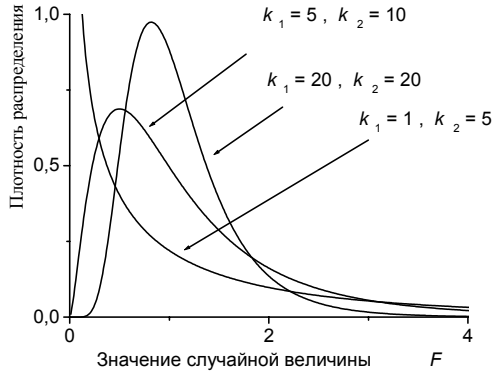


Рис. 2.9. Плотность распределения Фишера – Снедекора для различных значений параметров

2.5. Интервальная оценка выборочных параметров

Точечная оценка (оценка параметра одним числом) может быть близка к оцениваемому параметру, а может и сильно отличаться от него. Кроме того, точечная оценка не несет информации о точности процедуры оценивания. По этой причине при принятии решений следует использовать интервальную оценку параметров генеральной совокупности по данным выборочного наблюдения.

Оценка генеральной средней

Пусть имеется выборка объемом n и величина θ^* является статистической оценкой параметра θ. Величину $\Delta \geq \left|\theta - \theta^*\right|$ называют *предельной ошибкой выборки*.

Доверительной вероятностью оценки θ по θ^* называют вероятность γ, с которой выполняется неравенство $\left|\theta - \theta^*\right| < \Delta$. Иначе говоря,

$$\gamma = P(\left|\theta - \theta^*\right| < \Delta).$$

Доверительную вероятность выбирают достаточно большой: $\gamma = 0{,}95;\ 0{,}99$. Иногда вместо доверительной вероятности вводят понятие уровня значимости $\alpha = 1 - \gamma$, который равен вероятности того, что отклонение оцениваемого параметра от истинного значения оказалось больше предельной оценки Δ. В SPSS и ряде других статистических пакетов уровень значимости обозначается аббревиатурой Sig (*Significance value* – уровень значимости).

Можно дать следующее определение интервальной оценки.

Интервальной оценкой параметра θ называется числовой интервал $(\theta_1,\ \theta_2)$, который с заданной вероятностью γ накрывает неизвестное значение параметра θ. Важно отметить, что θ_1, и θ_2 определяются по выборочному наблюдению.

Построим доверительный интервал для генеральной средней в случае, когда генеральная дисперсия является неизвестной величиной. Если объем выборки достаточно велик, то для оценки генеральной дисперсии, как показано выше, можно использовать выборочную дисперсию, полагая $\sigma_0{}^2 = \sigma^2 \cdot n/(n-1)$.

Будем считать, что изучаемая случайная величина распределена в генеральной совокупности по нормальному закону. В этих условиях согласно формуле (2.3) случайная величина

$$t = \frac{\bar{x} - \bar{x}_0}{\sigma}\sqrt{n-1} \qquad (2.3^*)$$

подчиняется распределению Стьюдента с $n-1$ числом степеней свободы. Зададим доверительную вероятность γ и найдем интервал значений t_1, t_2, в который с заданной вероятностью γ попадет случайная величина, распределенная по закону Стьюдента .

Геометрический смысл поставленной задачи ясен из рис. 2.10. Нужно на графике распределения Стьюдента построить две симметричные относительно начала координат линии так, чтобы площадь под кривой между ними численно была равна γ. Напомним, что вероятность попадания случайной величины в некоторый интервал значений численно равна площади под кривой плотности распределения, оказавшейся внутри этого интервала.

Аналитически эту величину в SPSS можно найти, если воспользоваться обратной функцией распределения Стьюдента IDF.T(1 – 0.05/2, *k*), *k* – число степеней свободы. Эта функция возвращает вероятность того, что случайная величина примет значение меньшее, нежели вели-

чина t_2 на рис. 2.10. Все имеющиеся в SPSS обратные интегральные функции распределения сгруппированы в разделе с названием «Inverse DF» (обратные функции распределения).

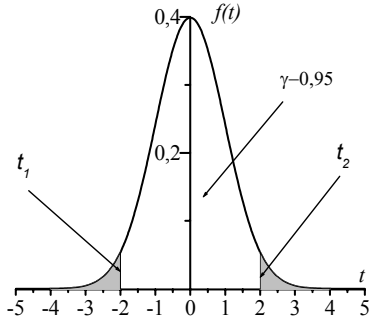


Рис. 2.10. Интервал значений t_1, t_2, внутри которого с вероятностью $\gamma = 0{,}95$ будет находиться случайная величина, распределенная по закону (2.3*)

Если взять число степеней свободы, например, равное 29, то случайная величина (2.3*) с вероятностью 0,95, будет принимать значения в интервале $-2{,}045 \le t \le 2{,}045$.

Для нахождения интервальной оценки среднего значения будем считать, что реализуется наихудшая ситуация, и случайная величина t принимает свои крайние значения $t_1 = -2{,}045$, а $t_2 = 2{,}045$. Тогда полуширину доверительного интервала можно оценить по формуле

$$\Delta = t_2 \cdot \sigma_0 / \sqrt{n}\,, \qquad (2.6)$$

а неизвестное значение генеральной средней с вероятностью $\gamma = 0{,}95$ будет находиться внутри интервала

$$\bar{x} - \Delta \le \bar{x}_0 \le \bar{x} + \Delta\,. \qquad (2.7)$$

В действительности при использовании статистического пакета SPSS для анализа данных нет необходимости производить эти рутинные вычисления. Чтобы получить интервальную оценку параметра, достаточно воспользоваться процедурой, возвращающей показатели дескриптивной статистики.

Пример 2.4

Имеются данные о средних денежных доходах и средних потребительских расходах (руб. в расчете на одного жителя) за август, сентябрь и октябрь 2005 года по 30 регионам РФ и данные о числе жителей в этих регионах (см. файл Пример_2_4.sav). Найти 95 %-й доверительный интер-

вал, в котором находятся средние доходы и потребительские расходы по месяцам в РФ в расчете на одного жителя.

Решение

Наиболее сложным в этом примере является правильный учет числа жителей, проживающих в различных областях. С одной стороны, полностью не учитывать различие в численности населения областей совершенно неверно, а с другой, если просто в качестве весового множителя взять число жителей, то средние значения будут вычислены правильно, а ошибка будет необоснованно занижена. Дело в том, что при таком взвешивании неявно предполагается, что мы располагаем числом наблюдений, равным суммарному числу жителей в этих 30 регионах. В действительности имеется всего лишь 30 наблюдений.

Для того чтобы правильно найти весовой коэффициент, найдем суммарное число жителей, активизировав получение описательной статистики для переменной N_people. Для этого выберем Analyze/Descriptive statistics/Descriptives и в открывшемся окне перенесем переменную N_people в правое окно (рис. 2.11). Открыв в этом окне закладку Options,

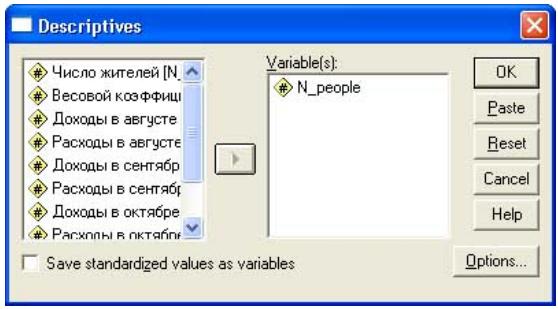


Рис. 2.11. Окно определения переменных, для которых нужно получить описательную статистику

потребуем вычисление суммы всех значений переменной, поставив в соответствующем окне галочку. Затем вернемся в исходное окно Descriptives, нажав кнопку Continue (продолжить), и завершим процедуру нажатием кнопки OK. Определим новую переменную Weight (как это делается, описано в примере 2.1). Затем откроем окно вычислений, используя пункты меню Transform/Compute, и переменной Weight присвоим значения, определяемые формулой Weight = N_people*30/ 58049500.

При таком способе определения весов, с одной стороны, будет учтено различие областей по численности, а с другой – сумма частот будет равна числу наблюдений (30).

Для получения правильной описательной статистики нужно указать процессору SPSS имя весовой переменной. В примере 2.2 подробно описана эта процедура (см. рис. 2.3). Теперь осталось запустить процедуру получения описательной статистики, выбрав Analyze/Descriptive statistics/Explore. В открывшемся окне (рис. 2.12) все переменные, для которых нужно получить описательную статистику, следует перенести в окно Dependent List.

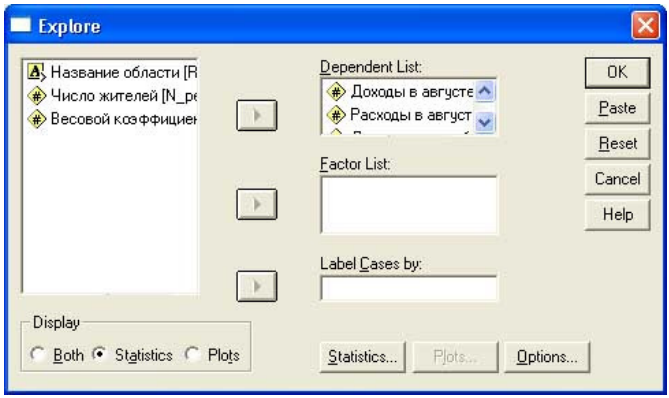


Рис. 2.12 Окно определения параметров описательной статистики

После нажатия клавиши OK будут выданы результаты в табличной форме. Мы приведем фрагмент этой таблицы только для одной переменной – Доходы в августе (таб. 2.2).

Полезно убедиться, что SPSS вычисляет доверительный интервал по приведенным выше формулам (2.6), (2.7). Действительно, стандартная ошибка для средней представляет собой оценку величины $\sigma_0/\sqrt{n}$. Значение статистики Стьюдента, как уже указывалось, можно вычислить с помощью обратной функции распределения Стьюдента IDF.T(1 – 0.05/2, 29), где число 29 – это число степеней свободы в данном примере. Эта функция возвращает так называемое критическое значение статистики Стьюдента, т. е. такое значение $t_{кр}$, которое определяет интервал $-t_{кр} \le t \le t_{кр}$, в который с вероятностью 95 % попадет значение случайной величины t, распределенной по закону Стьюдента (на рис. 2.10 t_2 и есть значение $t_{кр}$). Как уже указывалось, функция IDF.T(1 – 0.05/2, 29) возвращает значение 2,045.

Таблица 2.2

Фрагмент таблицы описательной статистики для переменной Доходы в августе *примера 2.4*

Показатели		Статистика	Стд. ошибка
Среднее		6861,492	400,305
95 % -й доверительный интервал для среднего	Нижняя граница	6042,777	
	Верхняя граница	7680,208	

Найдем верхнюю границу доверительного интервала, используя формулы (2.6), (2.7). Верхняя граница $= 6861{,}492 + 400{,}305 \cdot 2{,}045 = 7680{,}208$. Аналогично можно найти и нижнюю границу доверительного интервала.

Доверительный интервал для среднеквадратического отклонения

Пусть генеральная совокупность характеризуется нормальным распределением с параметрами $\bar{x}_0$ и ${\sigma_0}^2$, которые предполагаются неизвестными. По выборке найдены точечные оценки этих параметров:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{и} \quad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Требуется построить доверительный интервал, который с заданной доверительной вероятностью γ накроет неизвестное значение дисперсии в генеральной совокупности.

Для решения этой задачи рассмотрим случайную вспомогательную величину

$$\chi^2 = \frac{(n-1)\cdot\sigma^2}{{\sigma_0}^2}, \qquad (2.8)$$

которая распределена по закону χ^2 с $k = n-1$ степенями свободы (сравните это выражение с формулой (2.2). Для построения интервальной оценки генеральной дисперсии следует найти интервал ${\chi_1}^2$, ${\chi_2}^2$, внутрь которого с заданной вероятностью γ попадет значение случайной величины (2.8). На рис. 2.13 показано, как следует выбрать этот интервал.

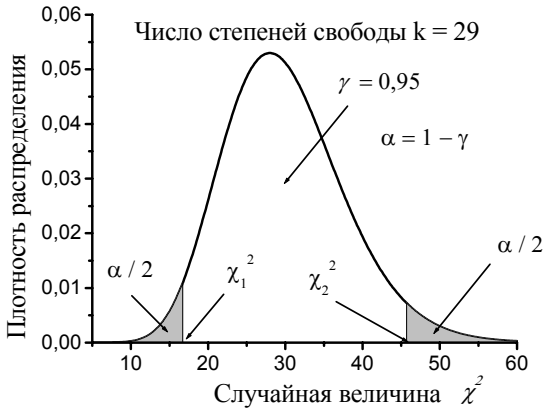


Рис. 2.13. Критические точки распределения χ^2
при доверительной вероятности $\gamma = 0{,}95$

Правую критическую точку $\chi_2{}^2$ выбираем так, чтобы вероятность того, что случайная величина χ^2 примет значение меньшее, нежели $\chi_2{}^2$, была равной $1-\alpha/2$ (эта вероятность равна площади под кривой распределения слева от $\chi_2{}^2$). Правую критическую точку в SPSS легко можно найти, используя функцию

$$\chi_2^2 = \text{IDF.CHISQ}(1-\alpha/2\,,\,k). \qquad (2.9)$$

В этой формуле k – число степеней свободы.

Аналогично находится и левая критическая точка. Нас интересует такая точка $\chi_1{}^2$ распределения χ^2, слева от которой площадь под кривой распределения оказалась бы равной $\alpha/2$. Поэтому

$$\chi_1^2 = \text{IDF.CHISQ}(\alpha/2\,,\,k). \qquad (2.10)$$

После того, как левая и правая критические точки распределения найдены, по формуле (2.8) находим нижнюю и верхнюю границы интервальной оценки для дисперсии и среднеквадратического отклонения в генеральной совокупности

$$\frac{(n-1)\cdot\sigma^2}{\chi_2{}^2} \le \sigma_0^2 \le \frac{(n-1)\cdot\sigma^2}{\chi_1{}^2}, \qquad \sigma\cdot\sqrt{\frac{n-1}{\chi_2{}^2}} \le \sigma_0 \le \sigma\cdot\sqrt{\frac{n-1}{\chi_1^2}}. \qquad (2.11)$$

Пример 2.5

Имеются данные о доходности акций компаний Microsoft, Intel и General Electric (GE) за период с августа 1997 по июль 2002 года (файл Пример_2_5.sav). Для этих активов можно определить среднюю доходность и риск (за меру риска обычно выбирается среднеквадратическое отклонение доходности). Поскольку средняя доходность и риск являются случайными величинами, для правильного выбора компании, в которую стоит вкладывать деньги, недостаточно ограничиться точечными оценками средней доходности и риска, а желательно иметь для этих величии 95 % -й доверительный интервал.

Определить интервальную оценку с 95 % -й доверительной вероятностью для доходности и риска активов компаний Microsoft, Intel и General Electric по данным наблюдения с августа 1997 по июль 2002 года.

Решение

Получить интервальную оценку для средней доходности можно просто, запустив процедуру получения описательной статистики выбором опций Analyze/Descriptive statistics/Explore.

К сожалению, нет аналогичной процедуры, с помощью которой можно было бы вычислить интервальную оценку для среднеквадратического отклонения. Поэтому эти вычисления придется выполнить вручную, применив приведенные выше формулы (2.9) – (2.11).

Для этого нужно создать новый файл данных, содержащий переменные: Comp – название компании; S – точечная оценка дисперсии (величины S следует скопировать из таблиц описательной статистики); xi1 и xi2 – левая и правая критические точки обратного распределения хи-квадрат; Smin, Smax – нижняя и верхняя границы 95 % -го доверительного интервала для среднеквадратического отклонения. Результаты вычисления приведены в табл. 2.3.

Таблица 2.3

Результаты расчета 95 % -го доверительного интервала для риска доходности акций компаний Microsoft, Intel и General Electric (G E)

Компания	Дисперсия доходности акций	χ_1^2	χ_2^2	Smin	Smax
Microsoft	0,021	39,66	82,12	1,76%	2,53%
Intel	0,023	39,66	82,12	1,98%	2,85%
G E	0,007	39,66	82,12	0,56%	0,81%

Подробно все результаты вычислений приведены в файле Пример_2_5a.sav.

2.6. Статистические гипотезы и методы их проверки

Большинство моделей требует тщательного анализа их состоятельности. Для этого необходимо проведение дополнительных расчетов, связанных с установлением выполнимости или невыполнимости тех или иных предпосылок модели, анализом качества найденных оценок, достоверностью полученных результатов.

Обычно эти расчеты проводятся по схеме статистической проверки гипотез. Поэтому знание основных принципов статистической проверки гипотез является обязательным для специалиста, который пытается обнаружить новые закономерности в данных.

Под *статистической гипотезой* понимают различного рода предположения о характере или параметрах распределения случайной величины, которые можно проверить, опираясь на результаты выборочного наблюдения.

Статистическая проверка гипотез носит вероятностный характер, и поэтому всегда существует риск совершить ошибку. Однако с помощью статистической теории можно оценить вероятность принятия ложного решения. Если эта вероятность мала, то решение можно считать статистически обоснованным.

Гипотезу, подлежащую проверке, обычно называют *нулевой гипотезой* и обозначают символом H_0. Наряду с нулевой гипотезой рассматривают *альтернативную (конкурирующую) гипотезу* (обозначается как H_1), которую придется принять, если будет отвергнута нулевая гипотеза. Например, в качестве нулевой гипотезы может быть выдвинуто предположение о равенстве нулю некоторого параметра θ в генеральной совокупности. Тогда альтернативной будет гипотеза о том, что $\theta \neq 0$ в генеральной совокупности.

Сущность проверки статистической гипотезы заключается в том, чтобы установить, согласуются или нет данные наблюдения и выдвинутая гипотеза. Ясно, что расхождения между результатами выборочного наблюдения и выдвинутой гипотезой будут практически всегда. Поэтому фактически решается вопрос о том, можно ли с определенным уровнем доверительной вероятности считать, что эти расхождения обусловлены действием случайных причин.

При проверке выборочные данные могут противоречить нулевой гипотезе H_0, и тогда она отклоняется, а принимается альтернативная гипотеза.

Статистическая проверка гипотез на основании выборочных данных неизбежно связана с риском принятия ложного решения. При этом ошибки могут быть двоякого рода.

Ошибка первого рода: проверяемая гипотеза H_0 является в действительности верной, но в результате статистической проверки принимается решение об отказе от нее (нулевая гипотеза отвергается).

Ошибка второго рода: нулевая гипотеза в действительности является ошибочной, но в результате статистической проверки она принимается.

Уровнем значимости называют вероятность совершить ошибку первого рода, т. е. отвергнуть гипотезу H_0 в результате статистических испытаний, когда она на самом деле верна. Уровень значимости обычно задают достаточно малым: $\alpha = 0{,}05;\ 0{,}01$.

Вероятность совершить ошибку второго рода обычно обозначают буквой β. Следует иметь в виду, что желательно обе эти вероятности сделать малыми. Это требование, однако, является противоречивым, поскольку уменьшение вероятности ошибки первого рода приводит к увеличению вероятности ошибки второго рода.

Продемонстрируем это положение на конкретном примере. Пусть нулевой гипотезой является утверждение, что в генеральной совокупности некоторая случайная величина распределена по нормальному закону и имеет среднее значение $\bar{x}_0$, равное нулю. В качестве альтернативной гипотезы H_1 выдвинем предположение, что эта случайная величина имеет среднее значение $\bar{x}_0 \neq 0$. Распределения случайной величины в условиях справедливости H_0 и в условиях справедливости H_1 будут различаться, и если дисперсия распределений не слишком мала, то распределения частично перекроются, как показано на рис. 2.14.

Из рисунка хорошо видно, что если уменьшать ошибку первого рода, увеличивая критическое значение параметра $x_{\text{кр}}$ (на рисунке $x_{\text{кр}} = 4$), то это приведет к увеличению вероятности совершить ошибку второго рода (увеличится β). Единственный способ уменьшить обе ошибки сразу – это увеличение объема выборки. В этом случае дисперсия нулевого и альтернативного распределений станет меньше, что приведет к уменьшению перекрытия этих распределений, и тогда обе ошибки станут меньше.

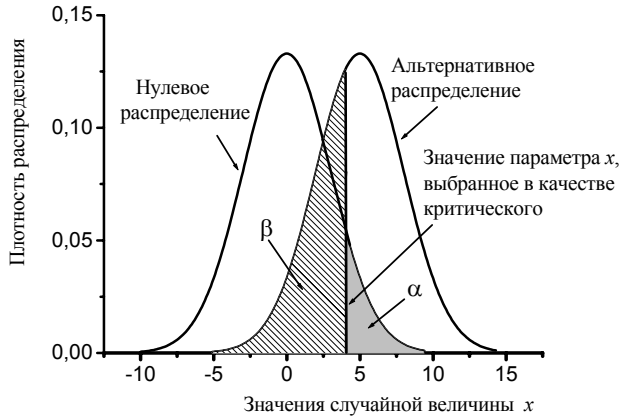


Рис. 2.14. Вероятности α и β, связанные с проверкой статистической гипотезы

При заданном уровне значимости α качество критерия для оценки статистической гипотезы измеряется вероятностью отвергнуть H_0, когда верна H_1 (или принять H_1, когда она верна). Эта вероятность называется мощностью критерия, обычно обозначается буквой π и равна вероятности не допустить ошибку второго рода: $\pi = 1 - \beta$. На рис. 2.14 мощность критерия $\pi = 1 - \beta$ равна площади под кривой альтернативного распределения справа от линии, определяющей значение параметра x, выбранного в качестве критического.

Статистическая проверка гипотез осуществляется на основании некоторых критериев. Для построения такого критерия необходимо:

- сформулировать нулевую гипотезу H_0;
- сформулировать альтернативную (конкурирующую) гипотезу H_1;
- подобрать случайную величину (часто называемую также статистикой), характер распределения которой был бы известен в условиях справедливости гипотезы H_0, и выбрать уровень значимости α, контролирующий допустимую ошибку первого рода;
- определить область допустимых значений и критическую область для изучаемого показателя (статистики);

- принять то или иное решение на основании сравнения наблюдаемого и критического значений показателя.

Для построения статистического критерии выбираются случайные величины, распределенные по законам Стьюдента (*t- критерий)*, Фишера – Снедекора (*F-критерий*), χ^2 (*хи-квадрат критерий)* либо какому-либо другому, априори известному закону распределения.

При испытании гипотезы следует принимать во внимание формулировку альтернативной гипотезы. То, какой выбрана гипотеза H_1, влияет на выбор критической области. Если проверяется гипотеза о равенстве нулю среднего значения в генеральной совокупности и выдвигается альтернативная гипотеза, что среднее нулю не равно, критическая область *t*-критерия Стьюдента выбирается как двусторонняя критическая область (рис. 2.15), поскольку нас не интересует, больше или меньше нуля будет среднее значение.

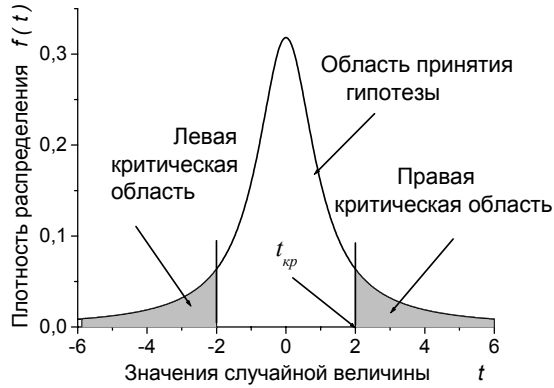


Рис.2.15. Двусторонняя критическая область для *t*-критерия Стьюдента

Если же в качестве альтернативной гипотезы выдвигается гипотеза о том, что в генеральной совокупности среднее больше нуля, то появление больших отрицательных значений случайной величины *t* не дает оснований принять альтернативную гипотезу. Поэтому критическая область в этом случае должна быть правосторонней (напоминаем, что площадь критической области численно равна уровню значимости α).

Поскольку понятие уровня значимости для статистической оценки гипотез является ключевым, приведем еще одно определение этого понятия.

Уровнем значимости называется такое малое значение вероятности попадания критерия в критическую область (при условии справедливости гипотезы H_0), что появление этого события можно расценивать как существенное расхождение выдвинутой гипотезы с результатом выборочного наблюдения. Попадание критерия в критическую область является основанием для отказа от гипотезы H_0 и принятия гипотезы H_1.

По своему прикладному содержанию статистические гипотезы можно разделить на несколько основных типов:

- о равенстве числовых характеристик генеральных совокупностей;
- о числовых значениях параметров;
- о виде закона распределения;
- об однородности выборок (т. е. принадлежности их одной и той же статистической совокупности).

Важно отметить, что принятие статистической гипотезы не дает логического доказательства ее верности. Принятие гипотезы следует рассматривать лишь как принятие весьма правдоподобного, не противоречащего опыту утверждения.

Пример 2.6

В рекламном проспекте компании Microsoft утверждается, что средняя доходность активов этой компании превышает 5 %. По данным 60 наблюдений за период с августа 1997 по июль 2002 года (файл Пример_2_5.sav) доходность акций этой компании составила 1,47 %, а среднеквадратическое отклонение доходности – 14,40 %. Не противоречат ли эти результаты утверждениям рекламного проспекта при уровне значимости $\alpha = 0{,}05$?

Решение

В качестве гипотезы H_0 примем предположение, что средняя доходность компании больше 5 %. В качестве альтернативной выдвигаем гипотезу, что средняя доходность акций компании Microsoft меньше 5 %.

Для проверки гипотезы H_0 используем t-критерий Стьюдента и рассмотрим случайную величину (2.3), выразив ее через исправленную выборочную дисперсию s:

$$t = \frac{(\bar{x} - \bar{x}_0)}{\sigma}\sqrt{n-1} = \frac{(\bar{x} - \bar{x}_0)}{s}\sqrt{n},$$

закон распределения которой представлен на рисунке 2.16. В данном случае критическая область должна быть левосторонней, поскольку большие положительные значения t не противоречат нулевой гипотезе.

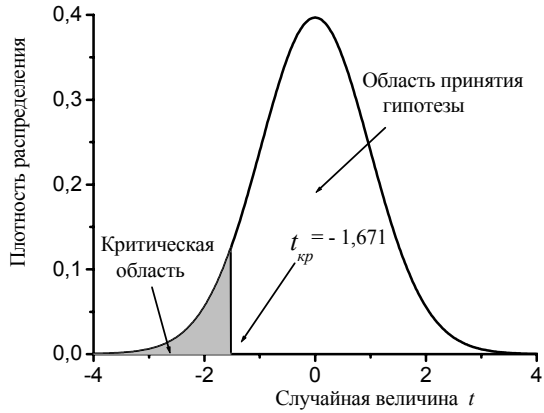


Рис. 2.16. Левосторонняя критическая область
для t-критерия Стьюдента (число степеней свободы $k = 59$, $\alpha = 0{,}05$)

Критическую точку левосторонней критической области при уровне значимости 0,05 найдем с помощью функции IDF.T(0.05,59), которая возвращает значение $t_{\text{кр}} = -1{,}671$. Эмпирическое значение критерия, найденное по выборочным данным:

$$t_{\text{эмп}} = \frac{(1{,}47 - 5) \cdot \sqrt{60}}{14{,}40} = -1{,}896 .$$

Поскольку эмпирическое значение критерия попадает в критическую область, то нулевую гипотезу придется отклонить. Это означает, что эмпирические данные дают основания принять альтернативную гипотезу и считать, что средний доход не превышает 5 %.

Используя стандартные процедуры SPSS, можно проверить только гипотезу, состоящую в том, что средняя доходность не равна 5 %. Для этого следует активизировать опции Analyze/Compare Means/One Sample T Test (Анализ/Сравнение средних/Т-тест для одной выборки).

В появившемся окне (рис. 2.17) нужно задать тестируемое значение, которое имеет смысл среднего в генеральной совокупности. В нашем случае – это 0,05.

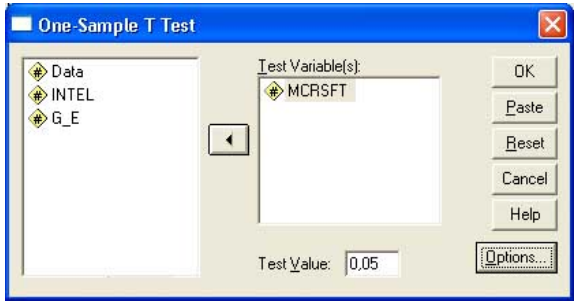


Рис. 2.17. Окно выбора параметров *t* -теста для одной выборки

Другие параметры этой процедуры можно не изменять. Результаты выполнения теста представлены в табл. 2.4

Таблица 2.4

Результаты t -теста для одной выборки

	Тестовое значение = 0,05					
	t	Ст.св.	Знч. (2-сторон)	Разность средних	95% -й доверительный интервал разности средних	
					нижняя граница	верхняя граница
MCRSFT	–1,896	59	0,063	–0,0352543	–0,072468	0,001959

По существу, этот тест содержит результаты проверки нулевой гипотезы о том, что среднее значение доходности равно 5 %. При уровне значимости 0,063 эту гипотезу можно принять, а при уровне значимости 0,05 ее следует отклонить. Следует обратить внимание на то, что в SPSS отсутствует возможность задать одностороннюю критическую область, и в нашем случае, чтобы ответить на поставленный вопрос, вычисления приходится делать вручную.

Другие примеры статистической проверки гипотез будут рассмотрены позже, по ходу изложения соответствующего материала.

2.7. Предварительный анализ данных

Хорошо известно, что целый ряд статистических процедур можно применять лишь в том случае, если исходный набор данных удовлетворяет условиям применимости того или иного статистического метода. Например, применять t-распределение Стьюдента для проверки гипотезы о равенстве двух выборочных средних можно лишь в том случае, если изучаемая величина распределена нормально (см. также пример 2.6). Если это условие не выполняется, то использовать t-статистику нельзя и следует применять один из непараметрических тестов. Критерии, используемые для проверки статистических гипотез, называются *непараметрическими*, если они не основываются на предположении об известном характере распределения случайной величины.

Поэтому, например, в примере 2.6, прежде чем использовать статистику Стьюдента для проверки гипотезы о том, что средняя доходность акций компании Microsoft превышает 5 %, следовало бы убедиться, что доходность акций распределена по нормальному закону.

Для целей предварительного знакомства со свойствами изучаемого статистического распределения используется вычисление показателей дескриптивной (описательной) статистики, таких как среднее значение, медиана, среднеквадратическое отклонение, асимметрия, эксцесс, строятся частотные распределения, диаграммы ветвей и листьев. Этим же целям служит построение гистограмм и коробчатых диаграмм распределения.

Предварительное изучение свойств анализируемого распределения является очень важным. В частности, таким образом можно заметить отклонения распределения случайной величины от нормального закона.

Изучая частотные характеристики распределения, можно обнаружить что выборка является нерепрезентативной. Например, если анализируется образовательный уровень населения России и частотный анализ показал, что в исследуемой выборке доля мужчин составляет 52 %, то сразу можно сказать, что эта выборка является нерепрезентативной, поскольку доля мужского населения в Российской Федерации составляет 45,5 %. Дальнейший анализ такой выборки смысла не имеет, и требуется провести коррекцию репрезентативности выборки. Такая коррекция реализуется с помощью взвешивания данных в окне редактора данных (пример такого рода будет рассмотрен на практических занятиях).

В связи с предварительным анализом данных рассмотрим еще одну проблему – обнаружение недостоверных данных и выбросов в исходной анализируемой информации. В простейшем случае выбросы (нетипичные значения изучаемой случайной величины) могут быть обна-

ружены при определении показателей описательной статистики и построении частотного распределения случайной величины. Если выбросы найдены и установлено, что эти значения ошибочны, то самый простой способ борьбы с ошибочными данными – отбросить их, если это возможно. На практике могут встретиться более сложные случаи, когда сразу нельзя утверждать, что отклоняющиеся от основного массива данные являются ошибками, или ситуации, когда имеется несколько схожих по величине значений, значительно отличающихся от основного массива данных. Как в этих случаях выявить наличие ошибочных данных и правильно провести оценку статистических показателей? Решению таких проблем посвящен специальный раздел статистики – робастное (устойчивое) оценивание.

Робастное оценивание – это методы статистического анализа, которые позволяют получить достаточно надежные оценки показателей статистической совокупности в условиях отсутствия данных о законе ее статистического распределения и наличия существенных отклонений в значении данных. У истоков развития методов робастного оценивания стояли американский статистик Д. Тьюки и швейцарский математик П. Хубер.

Пакет SPSS в своем составе имеет процедуры робастного оценивания (здесь они называются М-оценками). М-оценки Губера, Тьюки, Хампеля и Эндрюса можно получить на этапе предварительного исследования данных, если в меню выбрать Analyze/Descreptive Statistics /Explore и в открывшемся окне нажать клавишу Statistics.

Основная идея получения М-оценок состоит в том, что перед вычислением среднего значения разным случаям присваиваются веса. Чем дальше значение находится от среднего значения, тем меньше вес. Обычное среднее можно рассматривать как М-оценивание, когда все веса единичные. Используются М-оценки достаточно просто. Если М-оценки и обычные средние различаются, то это указывает на засоренность данных. В этом случае следует устранить из исходных данных недостоверные данные.

Критерий Колмогорова – Смирнова для проверки гипотезы о виде закона распределения

Для проверки гипотез о виде закона распределения очень часто используют критерий Колмогорова – Смирнова. Этот критерий очень прост для использования и в SPSS позволяет проверить гипотезу о принадлежности анализируемого распределения одному из распро-

страненных законов распределения – нормальному, Пуассона, однородному или экспоненциальному.

В качестве меры расхождения между эмпирическим и теоретическим распределениями принимается максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения $F(x)$

$$D = \max\left|F_n(x) - F(x)\right|. \qquad (2.12)$$

В качестве эмпирической функции распределения $F_n(x)$ используется просто функция накопленных частот, а предполагаемые параметры теоретической функции распределения рассчитываются по имеющимся выборочным данным.

Колмогоровым доказана теорема о том, что какому бы распределению $F(x)$ ни подчинялась случайная непрерывная величина x, при неограниченном числе наблюдений ($n \to \infty$) вероятность P того, что величина $D \cdot \sqrt{n}$ будет больше некоторого числа λ, определяется легко вычисляемой величиной

$$P(D \cdot \sqrt{n} \geq \lambda) = 1 - \sum_{k=-\infty}^{k=\infty} (-1)^k \cdot e^{-2 \cdot k^2 \cdot \lambda^2}. \qquad (2.13)$$

Чем больше величина λ, тем меньше оказывается величина (2.13). Величина $D \cdot \sqrt{n}$ называется *статистикой критерия Колмогорова – Смирнова*.

Для проверки, например, нормальности распределения случайной величины в выборке тест Колмогорова – Смирнова используется следующим образом.

- Выдвигается нулевая гипотеза о том, что распределение является нормальным.
- На основании имеющихся данных рассчитывается эмпирическое значение статистики Колмогорова – Смирнова, т. е. величина $\lambda_{\text{эмп}} = D \cdot \sqrt{n}$.
- Критическое значение статистики Колмогорова – Смирнова λ_α при заданном уровне значимости α определяется, исходя из выражения (2.13), т. е. $P(\lambda_\alpha) = \alpha$ (нужно подобрать такое значение λ в выражении для суммы формулы (2.13), чтобы, чтобы ее значение было равно доверительной вероятности $1 - \alpha$). Критическое значение статистики Колмогорова – Смирнова

ческое значение статистики Колмогорова – Смирнова можно взять и из таблиц критических точек для этой статистики.

- Если эмпирическое значение $\lambda_{эмп} > \lambda_{\alpha}$, то нулевую гипотезу придется отклонить и признать, что распределение не является нормальным. Если $\lambda_{эмп} < \lambda_{\alpha}$, то распределение является нормальным.

Пример 2.7

Для иллюстрации сказанного, используя данные примера 2.6, выясним, является ли распределение доходности акций фирмы Microsoft нормальным, как это неявно предполагалось выше.

Решение

В SPSS для проверки гипотезы о характере распределения в меню следует выбрать опции Analyze/Nonparametric Tests/1 Sample K-S (Анализ/Непараметрические тесты/Тест Колмогорова – Смирнова для одной выборки). В открывшемся окне следует переместить анализируемую переменную, в правое окно, как показано на рис. 2.18, отметить галочкой предполагаемый характер распределения и нажать кнопку OK.

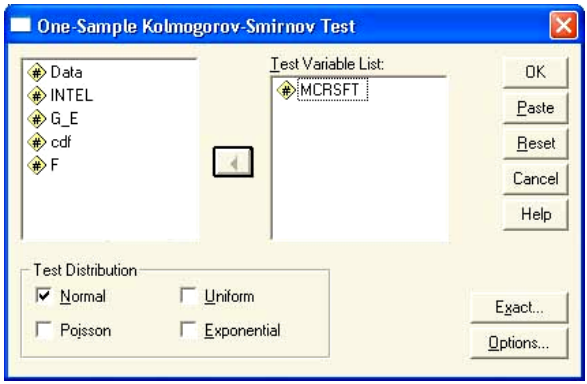


Рис. 2.18. Окно задания параметров проверки гипотезы о характере распределения

Таблица 2.5 представляет собой окно вывода для проверки гипотезы о характере закона распределения.

Таблица 2.5

Таблица вывода результатов критерия Колмогорова - Смирнова для одной выборки

N		60
Нормальные параметры [a, b]	среднее	0,014746
	стд. отклонение	0,1440551
Разности экстремумов (эмпирического и теоретического распределений)	модуль	0,081
	положительные	0,081
	отрицательные	–0,072
Статистика Колмогорова – Смирнова		0,626
Асимпт. знач. (двухсторонняя)		0,828

a) – сравнение с нормальным распределением; b) – оценивается по данным.

В этой таблице для нас представляют интерес только два параметра – Статистика Колмогорова – Смирнова и Асимптотическая значимость. Число 0,626 представляет собой эмпирическое значение статистики Колмогорова – Смирнова. Величина асимптотической значимости, равная 0,828, получается, если в формулу (2.13) подставить значение $\lambda = \lambda_{эмп} = 0{,}626$. Таким образом, значимость в данном случае представляет вероятность того, что для случайной величины, распределенной по нормальному закону, максимальное различие эмпирической и теоретической интегральных функций распределения окажется больше $\lambda_{эмп} = 0{,}626$. Поэтому нулевая гипотеза принимается, и распределение доходности акций фирмы Microsoft следует признать нормальным. Если бы величина значимости была меньше 0,05, то нулевую гипотезу пришлось бы отклонить, признав, что распределение доходности акций не подчиняется нормальному закону.

2.8. Выявление взаимосвязи явлений. Корреляционный анализ

В задачах корреляционного анализа требуется установить наличие взаимосвязи между изучаемыми явлениями (вычислить коэффициент корреляции и оценить его статистическую значимость). В SPSS имеется возможность вычислить коэффициенты корреляции Пирсона, Спирмена и Кендалла. Коэффициент корреляции по Пирсону следует использовать в том случае, когда изучаемые величины измеряются в интервальной шкале и имеют закон распределения, близкий к нормальному. Для переменных, измеренных в порядковой или в интервальной шкале, но имеющих распределение, сильно отличающееся от

нормального, следует вычислять ранговые коэффициенты корреляции по Спирмену или Кендаллу.

На практике иногда возникают ситуации, когда корреляционный анализ обнаруживает не поддающиеся логической интерпретации и противоречащие опыту взаимосвязи. Например, при маркетинговом исследовании была обнаружена логически необъяснимая, но статистически значимая взаимосвязь между количеством членов семьи и среднемесячным доходом на одного члена семьи. При более тщательном изучении было установлено, что взаимосвязь между количеством членов семьи и среднемесячным доходом на одного члена семьи объясняется третьей переменной – возрастом. Связанными (коррелирующими) являются в действительности пары переменных «возраст/уровень дохода» и «возраст/количество членов семьи».

Для выявления ложных корреляций в SPSS имеется процедура вычисления частных корреляций. При вычислении частной корреляции можно устранить влияние третьей переменной (в приведенном выше примере – возраста) и тем самым вычислить истинный коэффициент корреляции между изучаемыми переменными.

Пример 2.8

Используя данные маркетинговых исследований, содержащиеся в файле Пример_2_8.sav, выяснить, действительно ли существует взаимосвязь между среднемесячным доходом семьи, приходящимся на одного члена семьи, и количеством членов семьи.

Решение

Найдем коэффициенты парных корреляций для изучаемых переменных. Используя пункты меню Analyze/Correlate/Bivariate, получаем таблицу, из которой следует, что все переменные значимо коррелируют.

		Доход	Число членов семьи	Возраст
Доход	Корреляция Пирсона	1	0,626(**)	0,998(**)
	Знч.(2-сторон)		0,000	0,000
Число членов семьи	Корреляция Пирсона	0,626(**)	1	0,614(**)
	Знч.(2-сторон)	0,000		0,000
Возраст	Корреляция Пирсона	0,998(**)	0,614(**)	1
	Знч.(2-сторон)	0,000	0,000	

** – корреляция значима на уровне 0,01 (2-сторон.).

Метки переменных Доход, Число членов семьи и Возраст использованы для обозначения среднемесячного дохода на одного члена семьи, числа членов семьи и среднего возраста родителей соответственно. Поскольку есть основания считать, что корреляция переменных Доход и Число членов семьи обусловлена третьей переменной – Возраст, найдем частный коэффициент корреляции переменных Доход и Число членов семьи, исключив влияние переменной Возраст. Для этого используем опции Analyze/Correlate/Partial. В результате откроется окно, в котором нужно разместить переменные так, как показано на рис. 2.19.

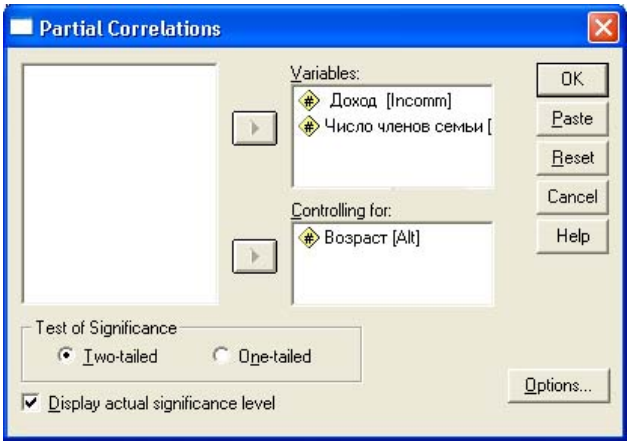


Рис. 2.19. Исключение влияния переменной Возраст на корреляцию переменных Доход и Число членов семьи

После нажатия кнопки OK получим табличку с результатами частной корреляции переменных

Контрольные переменные			Доход	Число членов семьи
Возраст	Доход	Корреляция	1,000	0,277
		Значимость (2-сторон.)	–	0,154
		ст.св.	0	26

Из приведенных результатов следует, что коэффициент частной корреляции переменных Доход и Число членов семьи составляет всего лишь 0,277, т.е. корреляция очень слабая и, кроме того, статистически незначима при уровне значимости 0,05.

2.9. Таблицы сопряженности и критерий хи-квадрат

Таблицы сопряженности (перекрестные распределения) служат для выявления зависимости между двумя и более переменными, которые измерены в номинативной или порядковой шкалах и имеют не очень большое число градаций. Для того чтобы понять, какой смысл имеют таблицы сопряженности, рассмотрим пример, в котором делается попытка выяснить, имеется ли взаимосвязь пола и психического состояния для студентов одного из вузов.

Пример 2.9

В файле Пример_2_9.sav содержатся данные о влиянии различных факторов (специальность, номер семестра, пол, возраст, успеваемость) на психическое состояние студентов. Требуется выяснить, зависит ли от пола психическое состояние, которое измерялось с помощью следующих градаций: крайне неустойчивое; неустойчивое; стабильное; очень стабильное.

Решение

Загрузим файл Пример_2_9.sav в редактор SPSS и затем с помошью выбора опций меню Analyze/Descriptive Statistics/Crosstabs (Анализ/Описательная статистика/Таблицы сопряженности). В результате откроется диалоговое окно, в котором перемену с меткой Пол нужно перенести в окно Row(s), а переменную с меткой Психическое состояние в окно Colum(s). После щелчка по кнопке OK будет создана таблица сопряженности (табл. 2.6).

Таблица 2.6

Таблица сопряженности для переменных «Пол» и «Психическое состояние»

Пол	Психическое состояние				Итого
	крайне неустойчивое	неустойчивое	стабильное	очень стабильное	
Женский	16 (7,88)	18	9	1	44
Мужской	3	22	32	5	62
Всего	*19*	*40*	*41*	*6*	106

В этой таблице на пересечении строк и столбцов стоят частоты f_{ij} (число случаев в исходном наборе данных, для которых имеется такое

сочетание признаков). Чтобы судить о наличии или отсутствии взаимосвязи признаков по таблице сопряженности обычно используется критерий χ^2. Смысл этого критерия состоит в том, что вначале вычисляются теоретические частоты f_{ij}^T сочетания признаков, исходя из предположения, что изучаемые факторы являются независимыми. В табл. 2.6 в скобках приведено теоретическое значение частоты f_{11}^T. Теоретические частоты определяют, сколько раз встречались бы искомые сочетания признаков, если бы между ними не было никакой связи и они были бы независимыми. Вычислим частоту появления признаков: «пол – женский», «психическое состояние – крайне неустойчивое» в предположении независимости этих факторов:

$$f_{11}^T = \frac{19}{106} \cdot \frac{44}{106} \cdot 106 = 7{,}88 \,. \qquad (2.14)$$

Первая дробь в формуле (2.14) $(19/106)$ равна вероятности того, что случайно взятый объект находится в крайне неустойчивом психическом состоянии, вторая определяет вероятность того, что случайно взятый объект будет иметь женский пол. Их произведение определяет вероятность появления сочетания этих признаков. Если вероятность умножить на число случаев (106), то мы найдем значение частоты f_{11}^T. Аналогично вычисляются и другие частоты. SPSS позволяет получить наряду с эмпирическими частотами и теоретические частоты, вычисленные в предположении независимости факторов. Для этого при построении таблиц сопряженности нужно на закладке Cells (Ячейки) в разделе Counts выбрать отображение ожидаемых (Expected) и наблюдаемых (Observed) частот.

Расчетное значение критерия χ^2 определяется формулой

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(f_{ij}^T - f_{ij}\right)^2}{f_{ij}^T}. \qquad (2.15)$$

В рассматриваемом примере $k = 2$, $m = 3$, по формуле (2.15) находим значение $\chi^2 = 22{,}455$. Этот результат и возвращает SPSS, если предусмотрительно позаботиться о выводе статистики хи-квадрат. Полученный результат еще не является окончательным. В действительности, для того чтобы установить наличие или отсутствие связи между признаками с помощью критерия χ^2, выдвигается гипотеза об отсутствии этой взаи-

мосвязи и вычисляется эмпирическое значение $\chi^2_{эмп}$ по формуле (2.15). Гипотеза об отсутствии взаимосвязи принимается, если $\chi^2_{эмп} < \chi^2_{кр}$. Пользователю SPSS нет необходимости вычислять критическое значение статистики, поскольку при выводе статистики хи-квадрат автоматически выводится и уровень значимости. В рассматриваем примере $\chi^2_{эмпир} = 22{,}455$, а значимость существенно меньше 0,001. Это означает, что исходную гипотезу об отсутствии взаимосвязи между полом и психическим состоянием студентов придется отклонить.

2.10. Сравнение выборочных средних

Очень часто возникает задача, в которой требуется, во-первых, вычислить выборочные средние, а во-вторых, выяснить, статистически значимо ли отличаются выборочные средние для двух разных выборок. Примеров такого рода постановок задач можно привести сколько угодно. Например, требуется проанализировать, значимо ли различаются для Свердловской и Челябинской областей процент голосов избирателей, отданных на выборах в Государственную Думу за представителей партии «Единая Россия».

Несмотря на простоту постановки задачи, единого алгоритма, позволяющего правильно ответить на поставленный вопрос, не существует. Приходится использовать разные статистические методы в зависимости от специфики анализируемого набора данных.

В общем случае в задачах такого типа требуется определить, анализируя исходные данные, статистическую значимость показателя или оценить, значимо ли различаются значения показателей, найденные по двум или нескольким выборкам. Решение последнего типа задач с применением *t* -критерия Стьюдента уже рассматривалось выше в примере 2.6. Сравнение выборочных средних – это класс наиболее простых задач, сводящихся обычно к стандартным вычислениям и испытанию некоторой статистической гипотезы. Тем не менее даже в этом случае SPSS предлагает достаточно большое число методик, имеющих различающиеся области применимости. Ниже приведена далеко не полная таблица критериев, используемых для оценки статистической значимости выборочных показателей или их различий для двух выборок (табл. 2.7).

Таблица 2.7

Критерии, используемые в SPSS для оценки статистической значимости выборочных показателей или их различий

Тип переменных	Тип критерия	Количество анализируемых выборок			
		Независимые		Зависимые	
		1	2	2	2 и более
Категориальные	Непараметрический	Критерий χ^2	Критерий χ^2	Критерий Мак Немура	Критерий Кохрана
Порядковые	Непараметрический	Критерий Колмогорова-Смирнова	Критерий Манна–Уитни	Критерий Вилкоксона	Критерий Фридмана
Интервальные	Параметрический: средние	*t*-критерий	*t*-критерий	t-критерий	Многофакторный анализ дисперсии
Интервальные	Непараметрический	Критерий Колмогорова-Смирнова	Критерий Манна–Уитни	Критерий Вилкоксона	Критерий Вилкоксона, критерий Фридмана

В приведенной таблице отражены лишь некоторые из тестов, позволяющих решить поставленную задачу статистической оценки выборочных средних или их различий. Как уже указывалось, при решении задачи о равенстве выборочных средних нельзя рекомендовать единый, универсальный подход. Поэтому мы ограничимся рассмотрением лишь нескольких типичных примеров, акцентируя внимание на принципах статистической проверки гипотез о равенстве выборочных средних.

Пример 2.10

Файл Пример_2_10.sav содержит данные о денежных доходах и потребительских расходах в расчете на душу населения в августе 2005 года. Используя эти данные, требуется определить: 1) можно ли считать одинаковыми доходы в расчете на одного человека в Южном и Сибирском федеральных округах? 2) можно ли считать, что денежные расходы граждан России совпадают с денежными доходами (это важно знать, чтобы убедиться, что неучтенные доходы невелики)?

Решение

Для примера в табл. 2.8 приведена часть статистического материала, относящаяся к Уральскому федеральному округу.

Таблица 2.8

Денежные доходы и потребительских расходы (руб.) в расчете на душу населения в августе 2005 года

Region	People	R_Number	Income	Costs	Index
Курганская область	992,1	5	4397,6	3212,2	0,323181
Свердловская область	4428,2	5	8542,1	6160,2	1,442504
Тюменская область	3307,5	5	13663,2	8913,8	1,077432
Челябинская область	3551,4	5	6706,4	4651,2	1,156883

Переменная People содержит число жителей области (тыс. чел.); R_Number – кодирует название федерального округа; Income и Costs – потребительские доходы и расходы в расчете на душу населения; переменная Index представляет собой весовой коэффициент, который учитывает различие областей региона по числу жителей. Значения этой переменной исчисляются следующим образом: определяется полное число жителей N_i в федеральном округе i; значение переменной Index для этого округа определяется по формуле

$$\text{Index} = \frac{\text{People}}{N_i} \cdot n_i .$$

В этой формуле n_i – число областей в федеральном округе i. Так, для Уральского федерального округа $N = 12279{,}2$, $n = 4$, таким образом, весовой коэффициент для Курганской области

$$\text{Index} = \frac{992{,}1}{12279{,}2} \cdot 4 = 0{,}323181.$$

Поскольку в задаче требуется выяснить, значимо ли различаются средние подушевые доходы для Южного и Сибирского федеральных округов, после загрузки данных в редактор SPSS произведем взвешивание данных, используя в качестве весов переменную Index. Как такое взвешивание производится, подробно описано в примере 2.2. Для того чтобы сравнить средние для двух федеральных округов, активизируем опции Analyze/Compare Means/Independent Samples T Test (Анализ/Сравнение средних/ Т-тест для независимых выборок). В результате откроется окно,

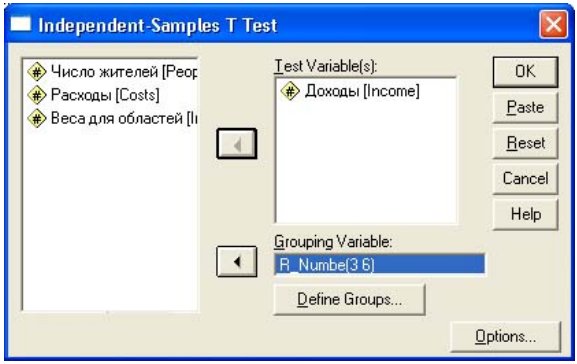


Рис. 2.20. Окно для настройки параметров сравнения средних с помощью T -теста для независимых выборок

в котором следует разместить переменные, как показано на рис. 2.20. Затем следует нажать кнопку Define Groups (Определить группы) и указать номера сравниваемых федеральных округов (в данном случае 3 и 6). Завершить работу с окном настройки параметров нажатием кнопки OK. В итоге получим таблицы отчета о сравнении средних доходов в Южном и Сибирском федеральных округах.

Таблица 2.9

Групповые статистики

Федеральный округ		n	Среднее	Стд. отклонение	Стд. ошибка среднего
Доходы	Южный	12	5430,98	1006,65	290,59
	Сибирский	12	6474,48	1293,44	373,38

Из таблицы 2.9 следует, что средние доходы на душу населения в Южном и Сибирском федеральных округах не равны, но значимо ли это различие? Ответ на этот вопрос содержится в таблицах 2.9 и 2.10, которые представляет собой часть информации, которую выдает SPSS при сравнении средних для двух выборок. Столбец Стандартное отклонение представляет собой корень квадратный из исправленной выборочной дисперсии σ, а столбец Стандартная ошибка среднего – значение $s = \sigma / \sqrt{n}$, где n – объем выборки. Поскольку данные о доходах измерены с использованием шкалы отношений и предполагается, что в каждой из выборок доходы распределены по закону, близкому к нормальному, для сравнения средних используется t -статистика Стьюдента.

В качестве нулевой гипотезы выдвинем предположение, что выборочные средние для изучаемых федеральных округов равны: $\bar{x}_3 = \bar{x}_6$. Для ответа на поставленный вопрос определим случайную величину

$$t = \frac{\bar{x}_3 - \bar{x}_6}{s_{(\bar{x}_3 - \bar{x}_6)}}. \qquad (2.16)$$

В знаменателе последней формулы $s_{(\bar{x}_3 - \bar{s}_6)}$ – среднеквадратическое отклонение разности выборочных средних (стандартная ошибка разности). При выдвинутой гипотезе обе выборки принадлежат к одной и той же генеральной совокупности, имеют одинаковое распределение, среднее значение и дисперсию. Оценим дисперсию разности средних. Поскольку дисперсия разности случайных величин равна сумме их дисперсий, получаем

$$s^2_{(\bar{x}_3 - \bar{x}_6)} = s^2_{\bar{x}_3} + s^2_{\bar{x}_6} = \frac{\sigma_0^2}{n_3} + \frac{\sigma_0^2}{n_6}. \qquad (2.17)$$

В нашем случае $n_1 = n_2 = 12$. Так как генеральная дисперсия неизвестна, оценим ее по выборочной дисперсии

$$\sigma_0^2 = \frac{\sigma_3^2 \cdot (n_3 - 1) + \sigma_6^2 \cdot (n_6 - 1)}{n_3 + n_6 - 2}. \qquad (2.18)$$

Формулы (2.16) – (2.18) и данные табл. 2.9 позволяют получить результаты, которые приводятся в таблице итоговой информации Т -теста на равенство средних в SPSS в предположении, что обе выборки взяты из одной генеральной совокупности (первая строчка в табл. 2.10).

Таблица 2.10

Таблица итоговой информации при анализе равенства выборочных средних в SPSS

Равенство дисперсий	*t*	Ст.св.	Знч. (2-сторон)	Разность Средних	Стд. ошибка разности
Предполагается	–2,205	22	0,038	–1043,49	473,14
Не предполагается	–2,205	20,749	0,039	–1043,49	473,14

Эти результаты говорят о том, что при уровне значимости 0,05 можно считать, что средние доходы в Южном и Сибирском федеральных округах в расчете на одного жителя оказались не одинаковыми.

В случае, когда равенство дисперсий в выборках не предполагается, стандартное отклонение разности средних определяется по формуле

$$s^2_{(\bar{x}_3-\bar{x}_6)} = s^2_{\bar{x}_3} + s^2_{\bar{x}_6} = \frac{\sigma^2_3}{n_3} + \frac{\sigma^2_6}{n_6}, \qquad (2.17^*)$$

где σ^2_3 и σ^2_6, как и раньше, – выборочные дисперсии. Существенным отличием является также формула для подсчета числа степеней свободы. Если предполагается равенство дисперсий, число степеней свободы определяется по формуле $k = n_3 + n_6 - 2$, а если дисперсии не равны, то

$$k = \frac{(\sigma^2_3/n_3 + \sigma^2_6/n_6)^2}{\frac{(\sigma^2_3/n_3)^2}{n_3 - 1} + \frac{(\sigma^2_6/n_6)^2}{n_6 - 1}}. \qquad (2.19)$$

Результат подсчета по формуле (2.19) дает значение $k = 20{,}749$, которое следует округлить до ближайшей целой величины – 21. Таким образом, если предположить, что выборки взяты не из одной генеральной совокупности, то все равно можно утверждать, что средние душевые доходы для жителей Южного и Сибирского федеральных округов не равны. Совпадение значений для t-статистики и стандартной ошибки разности в первой и второй строчках табл. 2.10 связано с равенством величин n_3 и n_6.

Обратимся ко второй части поставленной задачи – вопросу о равенстве средних доходов и расходов граждан России. В этом случае мы имеем дело с зависимыми выборками, и поэтому нужно активизировать опции Analyze/Compare Means/Paired Samples T Test. В результате откроется окно установки параметров теста на равенство средних для зависимых выборок (рис. 2.21). В левом окне следует выделить переменные с метками Доходы и Расходы и перенести их в правое окно.

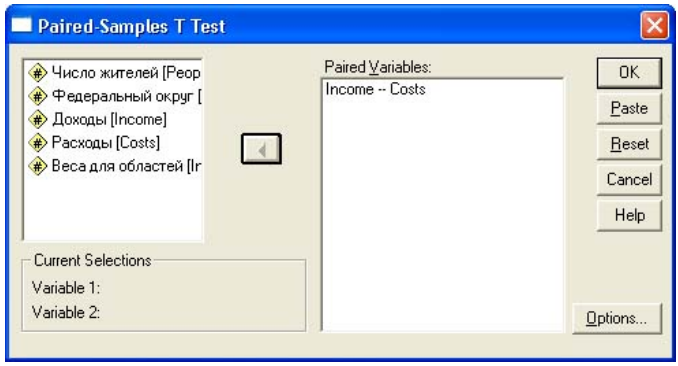


Рис. 2.21. Окно настройки параметров теста на равенство средних для зависимых выборок

После нажатия клавиши OK будут выведены на экран результаты, сгенерированные программой. Часть этих результатов представлена ниже.

	Среднее	N	Стд. отклонение	Стд. ошибка среднего
Доходы	7725,93	81	4520,04	502,23
Расходы	5656,25	81	3298,28	366,47

Здесь важно, что средние доходы оказались выше расходов. Вторая таблица позволяет сделать вывод о статистической значимости этой разницы, поскольку исходную гипотезу о равенстве расходов и доходов с вероятностью выше 99 % следует отклонить.

Разность средних	Стд. отклонение	Стд. ошибка среднего	95% -е доверительные границы разности средних		t	ст.св	Знч. (2-сторон)
			нижняя	верхняя			
2069,67	1476,38	164,04	1743,22	2396,13	12,617	80	0,000

В рассмотренном выше примере мы сравнивали средние выборочные с использованием параметрического t -критерия Стьюдента. Если случайная величина измерена в порядковой шкале, то в этом случае можно ввести лишь понятие медианы распределения, а математическое ожидание, дисперсию, которые нужны для применения t -критерия, определить не удается. В этом случае для сравнения двух выборок применяются непараметрические критерии, поскольку здесь используется не непосредственные значения случайных величин, а их ранги. Более того, эти критерии применимы и для того случая, когда случайные величины измерены в шкале отношений, но нет оснований считать, что их распределение близко к нормальному.

Наиболее известным из непараметрических критериев является критерий Манна – Уитни (Mann – Whitney), который является аналогом t -критерия Стьюдента и используется для сравнения медиан двух независимых выборок или сравнения средних двух независимых выборок, когда закон распределения случайных величин в анализируемых выборках не является нормальным.

Существует огромное число задач, которые сводятся к применению статистики Манна – Уитни. Например, для того чтобы выяснить, зависит ли усвоение изучаемого предмета в студенческой группе от пола, необходимо определить, значимо ли медианные балы для юно-

шей и девушек группы различаются между собой, т. е. провести тест Манна – Уитни.

Пример 2.11

Файл Пример_2_11.sav содержит данные (см. также таблицу ниже) о распределении балла юношей и девушек по итогам сдачи экзаменов за весь период обучения на заочном отделении УрАГС. Требуется выяснить, значимо ли различаются средние баллы юношей и девушек.

Балл	Число студ.	Пол	Балл	Число студ.	Пол
2,25	1	1	2,75	1	2
2,5	2	1	3	5	2
3	29	1	3,25	33	2
3,25	58	1	3,5	91	2
3,5	117	1	3,75	117	2
3,75	122	1	4	129	2
4	97	1	4,25	95	2
4,25	91	1	4,5	87	2
4,5	88	1	4,75	39	2
4,75	36	1	5	13	2
5	20	1			

Решение

Хотя представленные данные измерены в шкале отношений, применять t-статистику Стьюдента не вполне корректно, поскольку легко убедиться, используя критерий Колмогорова – Смирнова, что распределение среднего балла не является нормальным. Поэтому применим тест Манна – Уитни. Для этого загрузим файл Пример_2_11.sav в окно редактора данных SPSS и определим переменную Число студентов как весовую переменную (см. рис. 2.3). Для проведения теста Манна – Уитни активизируем пункт меню Analyze/Nonparametric tests/2 Independent Samples (Анализ /Непараметрические тесты /2 независимых выборки). В результате откроется окно настройки параметров, в котором следует сделать установки, изображенные на рис. 2.22. Следует особо подчеркнуть, что после того, как переменная Пол будет перенесена в окно с названием Grouping Variable (Группирующая переменная), необходимо щелкнуть по кнопке Define Groups (Определить группы) и выбрать значения номинативной переменной Пол, характеризующие сравниваемые выборки (в нашем случае это 1 и 2).

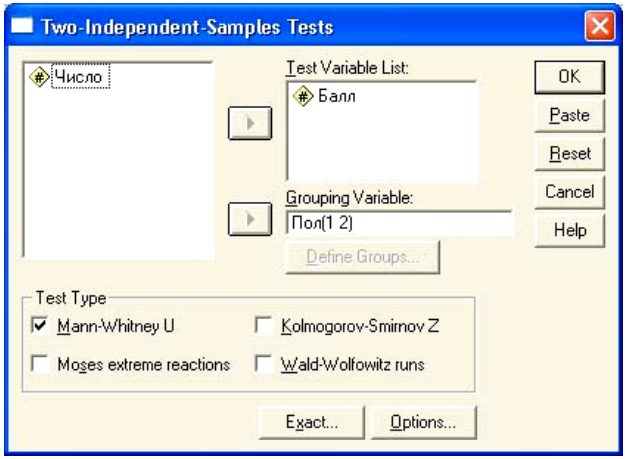


Рис. 2.22. Окно установки параметров теста Манна – Уитни

Все другие установки можно оставить без изменения. После нажатия кнопки OK появится окно, содержащее результаты теста, часть из которых приведена в табл 2.11 и 2.12.

Таблица 2.11

Ранги

Переменная	Пол	N	Средний ранг	Сумма рангов
Балл	Мужской	661	605,10	399971,00
	Женский	610	669,48	408385,00

Таблица 2.12

Статистика критерия

Статистика	Балл
U Манна – Уитни	181180,000
W Вилкоксона	399971,000
Z-статистика	–3,161
Асимпт. знч. (двухсторонняя)	0,002

Чтобы разобраться в результатах, нужно хотя бы кратко описать основные этапы расчета статистики Манна – Уитни U и стандартизованного значения Z -статистики. Чтобы применить критерий Манна – Уитни, необходимо заменить наблюдения, содержащиеся в двух выборках, имеющих объемы n и m, их рангами (если исходные данные не являют-

ся рангами изначально). Для этого объединим обе выборки, расположим данные в порядке возрастания и пронумеруем их (присвоим ранги). Для простоты будем предполагать, что все ранги различны. Обозначим R_i^1 ранг i-го элемента из первой выборки в общем вариационном ряду. Сумму рангов W_1 по первой выборке найдем, суммируя величины R_i^1:

$$W_1 = \sum_{i=1}^{l} R_i^1 . \qquad (2.20)$$

Аналогично можно найти и сумму рангов W_2 по второй выборке. Наименьшая из величин W_1 и W_2 называется статистикой Вилкоксона (Wilcoxon). Легко убедиться, что сумма рангов по первой и второй выборкам равна просто сумме $n = l + m$ первых натуральных чисел: $W_1 + W_2 = n(n+1)/2$. Статистика Манна – Уитни U линейно связана со статистикой Вилкоксона: $U = W - l(l+1)/2$, в чем легко убедиться, используя данные, приведенные в табл. 2.11 и 2.12 ($181180 = 399971 - 661 \cdot (661+1)/2$).

Далее выдвигается гипотеза, что медианы для обеих выборок равны, и альтернативная гипотеза, что медианы не равны. В качестве альтернативных могут выступать и другие гипотезы: например, медиана для первой выборки больше, чем медиана для второй выборки. Как уже указывалось, формулировка альтернативной гипотезы влияет на определение критической области.

В условиях справедливости выдвинутой нулевой гипотезы асимптотически при $n \to \infty$ величина (2.20) ведет себя как нормально распределенная случайная величина с математическим ожиданием $M(U) = 1/2 \cdot l \cdot (l + m + 1)$ и дисперсией $D(U) = 1/12 \cdot l \cdot m \cdot (l + m + 1)$.

Таким образом, можно ввести стандартизованную величину

$$Z = \frac{W - 1/2 \cdot l \cdot (l + m + 1)}{\sqrt{1/12 \cdot l \cdot m \cdot (l + m + 1)}}, \qquad (2.21)$$

которая приближенно подчиняется стандартному нормальному распределению. Сходимость к нормальному распределению достаточно быстрая, так что формула применима, если каждая из выборок содержат больше 8 членов. Вычисления по приближенной формуле (2.21) дают значение $Z = -3{,}12$, что несущественно отличается от точного результата, приведенного в табл. 2.10. Дальнейшие выводы делаются на основании Z-критерия. Поскольку значимость оказалась существенно меньше 0,05, то это означает, что гипотезу о равенстве средних баллов юношей и девушек за весь период следует отклонить.

Часто приходится сравнивать средние значения для одной и той же совокупности объектов. Такие выборки называются связанными. Например, с целью выяснить, зависит ли успеваемость студентов академической группы от номера семестра, сравнивают средние значения успеваемости после завершения первой и третьей экзаменационных сессий. Поскольку объекты исследования в обеих выборках одни и те же, мы имеем дело со связанными выборками.

В случае связанных выборок для определения статистической значимости получаемых результатов чаще других используется знаково-разностный тест Вилкоксона.

Для каждого случая определяется разница $d_i = X_i - Y_i$. Все ненулевые разности ранжируются по абсолютной величине, и им присваивается ранг. Для простоты опять предположим, что все величины d_i оказались различными. Затем вычисляется S_+ – сумма рангов положительных и S_- – сумма рангов отрицательных разностей. Если связь между X и Y отсутствует и распределение одинаково, то эти две суммы должны быть примерно равны. Статистика критерия Z строится как стандартное отклонение наименьшей суммы рангов от среднего значения:

$$Z = \frac{Min(S_+, S_-) - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}. \qquad (2.22)$$

В этой формуле n – число случаев с ненулевой разницей d_i. Величина Z приближенно подчиняется стандартному нормальному распределению, и поэтому дальнейший анализ статистической значимости получаемых результатов не представляет труда.

Пример 2.12

В файле Пример_2_12.sav содержатся данные о количестве правонарушений в расчете на 1000 жителей в областях РФ за период с января по май 2006, 2005 и 2004 годов. Требуется выяснить, значимо ли различается среднее число преступлений в 2005 и 2006 годах.

Решение

Для ответа на поставленный вопрос применим знаково-ранговый тест Вилкоксона. Для этого после загрузки файла данных в окно редактора активизируем опции Analyze/Nonparametric Tests/2 Related Samples (Анализ / Непараметрическое тестирование / 2 связанных выборки).

В результате появится окно, в котором тестируемую пару переменных следует перенести в правое окно с названием Test Pair (s) List. Все остальные параметры можно оставить по умолчанию. Результат выполнения теста отображается в двух таблицах (табл. 2. 13 и 2.14).

Таблица 2.13

Ранги

		N	Средний ранг	Сумма рангов
Число преступлений на 1000 чел в 2005 г - Число преступлений на 1000 чел в 2006 г.	Отрицательные ранги	73[a]	43,48	3174,00
	Положительные ранги	8[b]	18,38	147,00
	Всего	81		

a – число преступлений на 1000 чел в 2005 году < числа преступлений на 1000 чел в 2006 году;

b – число преступлений на 1000 чел в 2005 году > числа преступлений на 1000 чел в 2006 году;

Таблица 2.14

Статистика критерия знаковых рангов Вилкоксона

	Число преступлений на 1000 чел в 2005 г – число преступлений на 1000 чел в 2006 г.
Z	–7,126[a]
Асимпт. знч. (двухсторонняя)	,000

a – используются положительные ранги.

Поскольку проверяется гипотеза о равенстве средних (точнее, медиан распределений) и критерий попал в критическую область, то следует признать, что число преступлений на 1000 жителей в 2006 году значимо отличается от результатов 2005 года.

Сравнение нескольких выборочных средних. Дисперсионный анализ

При сравнении нескольких выборочных средних для независимых выборок используется либо дисперсионный анализ, либо его аналог – непараметрический ранговый критерий Крускала – Уоллеса (Kruskal – Wallis). Мы рассмотрим оба этих подхода, но вначале хотели бы обратить внимание еще раз на необходимость анализа условий применимости выбранной методики тестирования.

При проведении дисперсионного анализа требуется выполнение условия однородности дисперсии: дисперсия результативной величины не должна зависеть от значения факторной величины (номера точки наблюдения). Например, если изучается зависимость коэффициента интеллектуального развития IQ от типа общеобразовательной школы, то дисперсия величины IQ должна быть примерно одинаковой для учащихся школ с гуманитарным и естественно-научным уклоном.

Учитывая этот факт, в SPSS при проведении дисперсионного анализа предусмотрен тест на однородность дисперсии Левене (H. Levene, 1960). Если дисперсия окажется неоднородной, то доверять выводам дисперсионного анализа нельзя, поскольку нарушаются условия применимости метода.

Дисперсионный анализ представляет собой достаточно мощный и универсальный метод изучения взаимосвязи социально-экономических явлений. Поэтому представляется необходимым хотя бы кратко изложить суть этого метода.

Дисперсионный анализ представляет собой статистический метод, предназначенный для определения влияния различных факторов на результат эксперимента. Он был разработан Р. Фишером в 1918 году для обработки результатов агрономических опытов с целью выявления условий получения максимальной урожайности сельскохозяйственных культур.

По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный дисперсионный анализ.

Однофакторная дисперсионная модель имеет вид

$$y_{ij} = y_i + F_j + \varepsilon_{ij},$$

где y_{ij} – значение исследуемой переменной с учетом влияния фактора F_j, который учитывает изменение результативной переменной под влиянием j – уровня фактора ($j = 1, 2, \ldots, m$); y_i – значения изучаемой

величины в одной серии наблюдений при отсутствии влияния каких-либо факторов $(i = 1, 2, \ldots, n)$; ε_{ij} – случайное возмущение, вызванное действием неконтролируемых факторов. Под уровнем фактора понимается либо номер исследуемой партии, либо номер выборки.

Основная идея дисперсионного анализа состоит в разделении вариации результативного признака на часть, которая обусловлена действием факторного признака, и часть, обусловленную действием других причин. Если часть, обусловленная действием факторного признака, велика, то это и будет означать, что факторная и результативная величины взаимосвязаны.

В дисперсионном анализе принято не ограничиваться качественными выводами о наличии взаимосвязи факторной и результативной переменных. Обычно применяется стандартный для аналитической статистики прием испытания гипотез. Выдвигается гипотеза о равенстве средних значений во всех группах (т. е. об отсутствии влияния факторного признака на результат), которая затем проверяется с помощью критерия (статистики) Фишера.

Критерий Фишера строится следующим образом. Если данные каждой группы представляют собой случайную выборку из нормально распределенной генеральной совокупности, то можно оценить дисперсию в генеральной совокупности двумя способами: 1) разделить межгрупповую вариацию на число степеней свободы для межгрупповой вариации; 2) разделить вариацию, обусловленную действием других причин, на число степеней свободы для внутригрупповой вариации. Первый способ оценки дает значение $s^2_{\text{факт}}$, второй – $s^2_{\text{групп}}$.

$$s^2_{\text{факт}} = \frac{\sum_{j=1}^{m} n_j \cdot (\bar{y}_j - \bar{y})^2}{m-1}; \qquad s^2_{\text{групп}} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n-m}. \quad (2.23)$$

В формуле (2.23) $\bar{y}_j$ – среднее значение признака для j-й группы, $\bar{y}$ – общее среднее. В условиях справедливости нулевой гипотезы случайные величины $s^2_{\text{факт}}$ и $s^2_{\text{групп}}$ распределены по закону χ^2 с $m-1$ и $n-m$ степенями свободы соответственно. Поэтому величина

$$F = \frac{s^2_{\text{факт}}}{s^2_{\text{групп}}} \quad (2.24)$$

является случайной величиной, распределенной по закону Фишера – Снедекора со степенями свободы $m-1$ и $n-m$.

Для проверки нулевой гипотезы эмпирическое значение критерия Фишера, даваемое формулой (2.24), следует сравнить с критическим значением статистики, которое определяется по таблицам критических точек распределения Фишера. В SPSS критическое значение можно найти, используя функцию IDF.F(P, m-1, n-m), зависящую от трех аргументов: P – доверительной вероятности и чисел степеней свободы $m-1$ и $n-m$. Если критическое значение статистики Фишера больше эмпирического значения, то нулевую гипотезу (гипотезу о равенстве средних во всех группах) можно принять. Если же эмпирическое значение статистики будет больше критического, то принимается альтернативная гипотеза об отсутствии равенства средних в группах.

Для применимости дисперсионного анализа должны выполняться следующие предпосылки:

- Случайная ошибка должна подчиняться нормальному закону распределения с математическим ожиданием $M(\varepsilon_{ij},)=0$. Возмущения ε_{ij} и ε_{kl} для несовпадающих i, j, и k, l должны быть независимы.
- Дисперсия возмущений ε_{ij} должна быть одинаковой для разных групп

Если дисперсия изучаемой величины в разных группах сильно различается, то использование критерия Фишера для проверки равенства средних в группах оказывается неправомерным.

Пример 2.13

Имеются данные о годовых затратах туристических фирм на рекламу (млн руб.) и ежегодном количестве туристов, которых обслуживают эти фирмы (см. табл. 2.15 или файл Пример_2_13.sav на электронном диске). С 95 % -й доверительной вероятностью требуется выяснить, влияют ли затраты на рекламу на увеличение числа туристов, обслуживаемых фирмой.

Для ответа на поставленный вопрос следует использовать однофакторный дисперсионный анализ. Однако, как указывалось выше, условием применимости дисперсионного анализа является равенство дисперсии факторной величины в различных группах. Проверить, удовлетворяют ли исходные данные этому условию, можно с помощью теста Левене.

При проведении теста Левене выдвигается нулевая гипотеза о том, что дисперсия для всех k групп одинакова и $\sigma_1=\sigma_2=\ldots=\sigma_k$ (в рассматри-

ваемом случае имеем пять групп туристических фирм, различающихся объемом средств, отпускаемых на рекламу). Альтернативной гипотезой является предположение о неравенстве дисперсий для любых двух групп ($\sigma_i \neq \sigma_j,\ i, j \leq k$).

Таблица 2.15

Исходные статистические данные примера 2.13

Номера фирм	Затраты на рекламу, млн руб.	Количество туристов, чел.	Номера фирм	Затраты на рекламу, млн руб.	Количество туристов, чел.
1	8	800	11	10	920
2	8	850	12	10	1060
3	8	720	13	10	950
4	9	850	14	11	900
5	9	800	15	11	1200
6	9	880	15	11	1150
7	9	950	17	11	1000
8	9	820	18	12	1200
9	10	900	19	12	1100
10	10	1000	20	12	1000

Статистический критерий Левене для проверки нулевой гипотезы очень напоминает критерий Фишера (2.24) с той лишь разницей, что для L -статистики Левене используются не значения результативного признака y_{ij} , а значения $z_{ij} = \left| y_{ij} - \bar{y}_j \right|$ – отклонения результативного признака от среднего в группах:

$$L = \frac{n-m}{m-1} \cdot \frac{\sum_{j=1}^{m} n_j \cdot (\bar{z}_j - \bar{z})^2}{\sum_{j=1}^{m} \sum_{i=1}^{n_j} (z_{ij} - \bar{z}_j)^2}. \qquad (2.25)$$

В этой формуле $\bar{z}$ – общее среднее отклонение, $\bar{z}_j$ – среднее значение отклонения в группе j .

Критерий (2.25) используется точно так же, как и критерий Фишера (2.24). Находим эмпирическое значение статистике Левене по формуле (2.25) и затем сравниваем его с критическим значением статистики Фи-

шера $F(\alpha, m-1, n-m)$, которое находится либо по таблицам, либо с использованием функции IDF.F$(1-\alpha, m-1, n-m)$ в SPSS. Если критическое значение статистики Левене больше эмпирического при заданном уровне значимости, то гипотеза принимается и дисперсию в группах можно считать одинаковой, а применение дисперсионного анализа оправданным.

Вернемся к анализу примера 2.13. Для проверки наличия взаимосвязи между вложением денег в рекламу и числом обслуженных туристов загрузим файл Пример_2_13.sav в окно редактора SPSS и проведем дисперсионный анализ, активизировав опции Analyse/Compare Means/One-Way Anova (Анализ/Сравнение средних/Однофакторный дисперсионный анализ). В появившемся окне следует переменную Количество туристов перенести в окно Dependet List (Список зависимых), а переменную Затраты на рекламу в окно Factor (рис. 2.23).

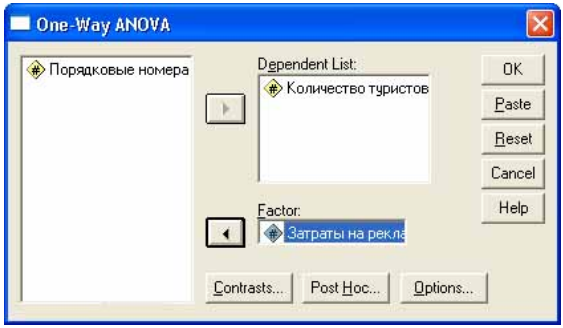


Рис. 2.23. Окно выбора зависимых и факторных переменных однофакторного дисперсионного анализа

Для того чтобы проверить однородность дисперсии, необходимо нажать кнопку Options, в появившемся окне заказать выполнение теста на однородность дисперсии и нажать кнопку OK. В окне результатов интерес для нас представляют таблица итогов проверки входных данных на однородность дисперсии (табл. 2.16) и таблица итогов однофакторного дисперсионного анализа (табл. 2.17).

Таблица 2.16

Итоги теста Левене на однородность дисперсии

Статистика Левене	Ст.св.1	Ст.св.2	Знач.
2,169	4	15	,122

Из приведенных данных следует, что эмпирическое значение статистики Левене L = 2,169. Значимость 0,122 означает вероятность совер-

шить ошибку первого рода, т. е. если мы отклоним гипотезу о равенстве дисперсий в группах, то с вероятностью 0,122 это решение будет ошибочным. Если мы хотим отклонить гипотезу о равенстве дисперсий в группах с высокой степенью достоверности $P = 0{,}95$, то нужно сравнить критическое значение статистики Фишера IDF.F(0,95,4,15) = 3,06 с эмпирическим значением статистики Левене L = 2,169. Поскольку эмпирическое значение статистики меньше критического, мы не можем отклонить гипотезу о равенстве дисперсий в группах с достоверностью 95 %.

Таблица 2.17

Итоги однофакторного дисперсионного анализа в SPSS

	Сумма квадратов отклонений	Ст.св.	Средний квадрат	F	Знач.
Между группами	236580,00	4	59145,00	7,648	0,001
Внутри групп	115995,00	15	7733,00		
Итого	352575,00	19			

В таблице 2.17 показаны промежуточные результаты вычисления статистики Фишера по формулам (2.23) – (2.24). Для нас представляет интерес лишь окончательный результат: эмпирическое значение статистики Фишера в нашем примере $F_{\text{эмп}} = 7{,}648$ и значимость $\alpha = 0{,}001$. Поскольку проверялась исходная гипотеза об отсутствии взаимосвязи между количеством туристов, которых обслужила фирма, и уровнем расходов на рекламу и эмпирическое значение $F_{\text{эмп}} = 7{,}648$ при уровне значимости $\alpha = 0{,}05$ оказалось существенно больше критического значения $F_{\text{кр}} = 3{,}06$, то нулевую гипотезу придется отклонить и принять альтернативную – о наличии взаимосвязи между факторным и результативным признаками.

Ранговый критерий Крускала – Уоллеса

Ранговый критерий Крускала – Уоллеса позволяет сравнить медианы нескольких ($k > 2$) независимых выборок и является обобщением рангового теста Вилкоксона для двух выборок. С другой стороны, этот тест является непараметрической альтернативой F - критерию Фишера в однофакторном дисперсионном анализе и позволяет сравнивать данные, измеренные как в интервальной, так и порядковой шкалах.

Ранговый критерий Крускала – Уоллеса применяется для проверки гипотезы о том, что медианы M_i ($i = 1, 2, \ldots, k$) в k выборках совпадают между собой (выборки взяты из генеральных совокупностей,

имеющих одинаковые медианы). Относительно закона распределения случайной величины в генеральных совокупностях, из которых взяты выборки, не делается каких-либо предположений, за исключением предположения о непрерывности распределения.

Для того чтобы применить тест Крускала – Уоллеса, необходимо заменить наблюдения в выборках их объединенными рангами. При этом наименьшему значению соответствует первый ранг, следующему по величине – второй и т. д. Наибольшему значению, таким образом, будет соответствовать ранг $n = n_1 + n_2 + \ldots + n_k$, где n_i – число наблюдений в i - й выборке. Если некоторые значения повторяются, им присваиваются средние значения их рангов. В дальнейшем ради упрощения изложения мы будем предполагать, что все наблюдения различны.

На следующем шаге вычисляется сумма рангов R_i для каждой из выборок ($i = 1, 2, \ldots, k$). Тест Крускала – Уоллеса определяет, насколько эти суммы рангов отличаются по величине. Ясно, что если различие мало, то выборки очень похожи и можно считать, что они взяты из одной генеральной совокупности.

Показано, что если все выборки взяты из одной генеральной совокупности (в условиях справедливости нулевой гипотезы), можно, используя вычисленные суммы рангов, определить H -статистику Крускала – Уоллеса

$$H = \frac{12}{n \cdot (n-1)} \cdot \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3 \cdot (n-1), \qquad (2.26)$$

которая при достаточно большом объеме выборок (практически должно выполняться условие $n_i > 5, \quad i = 1, 2, \ldots, k$) приближенно подчиняется статистике χ^2 с $k-1$ степенью свободы. Таким образом, при заданном уровне значимости, например $\alpha = 0{,}05$, нужно найти критическое значение статистики $\chi^2_{\text{кр}}$ с $k-1$ степенью свободы и сравнить это значение с критерием H (2.26). Если $H < \chi^2_{\text{кр}}$, то гипотезу о том, что все выборки взяты из одной генеральной совокупности, следует принять, а если выполняется обратное неравенство $H > \chi^2_{\text{кр}}$, то нулевую гипотезу следует отклонить и принять альтернативную, которая состоит в том, что хотя бы одна из выборок взята из генеральной совокупности с другим значением медианы.

Пример 2.14

Используя набор данных примера 2.12 с 95 % -й доверительной вероятностью, выяснить, можно ли считать, что по числу преступлений в расчете на 1000 жителей Приволжский, Сибирский и Дальневосточный федеральные округа значимо не различаются.

Решение

Применение теста Крускала – Уоллеса избавляет нас от проблемы предварительного изучения свойств статистического распределения случайной величины. Даже если выполняются условия применимости F -критерия Фишера, мощность критерия Крускала – Уоллеса не хуже, и поэтому применение этого теста вполне оправданно.

После загрузки файла Пример_2_14.sav в редактор данных SPSS для проведения теста Крускала – Уоллеса выберем пункт меню Analyze/ Nonparametric Tests/ K Independent Samples (Анализ / Непараметрические тесты / К независимых выборок). В появившемся окне задания параметров необходимо переменные разместить так, как показано на рисунке 2.24, а затем с помощью кнопки Define Range задать минимальное и максимальное значения группирующей переменной (в нашем случае 4 и 7).

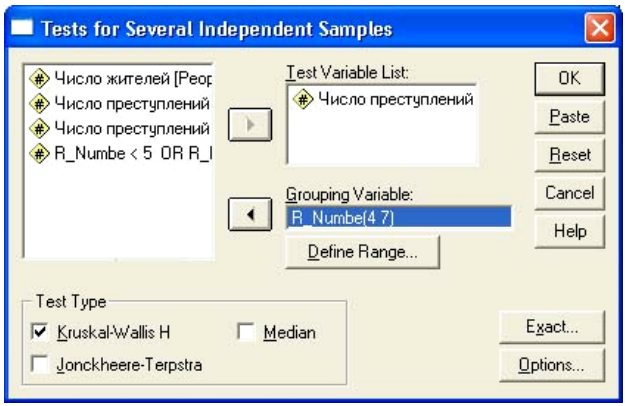


Рис. 2.24. Окно установки параметров теста для нескольких независимых выборок

Мы исключили из анализа данные по Уральскому федеральному округу в связи с тем, что здесь объем выборки составляет всего 4 объекта. Для достижения этой цели при отборе объектов, участвующих в анализе, использовался фильтр, позволяющий отобрать объекты, для которых значение переменной R_Numbe не равно 5 (R_Numbe < 5 OR R_Numbe > 5).

Результаты анализа представлены в табл. 2.18 и 2.19.

Таблица 2.18

Число преступлений на 1000 чел в 2006 году. Ранги

Федеральный округ	N	Средний ранг
Приволжский	14	15,71
Сибирский	12	19,33
Дальневосточный	9	19,78
Всего	*35*	

Таблица 2.19

Статистика критерия Крускала –Уоллеса

	Число преступлений на 1000 чел в 2006 г.
Хи-квадрат	1,171
Степени свободы	2
Асимптотическая значимость	0,557

Данные в табл. 2.18 носят информационный характер. Наибольший интерес представляют данные табл. 2.19. Значение статистики Крускала – Уоллеса (хи-квадрат) равно всего лишь 1,171 при уровне значимости 0, 557. Поскольку нулевой гипотезой являлось предположение о том, что выборки взяты из одной генеральной совокупности и критерий находится в области принятия нулевой гипотезы, то нет оснований утверждать с 95% -й вероятностью, что по числу преступлений на 1000 человек Приволжский, Сибирский и Дальневосточный федеральные округа различаются между собой.

Контрольные вопросы

2.1. Какие типы шкал используются для анализа данных в SPSS?

2.2. Приведите примеры данных, для измерения которых используются номинативная, порядковая и интервальная шкалы.

2.3. Перечислите основные операции, которые допускают данные, измеренные в номинативной, порядковой и интервальной шкалах.

2.4. Назовите способы импорта данных из программы Excel в SPSS.

2.5. В чем состоит сущность выборочного метода исследования?

2.6. Перечислите основные виды распределений, которые широко используются для построения статистических критериев в SPSS.

2.7. Сформулируйте концепцию интервального оценивания, широко применяемую для оценки значения показателя в генеральной совокупности по выборочным данным.

2.8. Дайте определение понятий «доверительная вероятность», «уровень значимости», «число степеней свободы».

2.9. Перечислите основные свойства нормального и стандартного нормального распределений.

2.10. В чем состоит сущность метода статистического испытания гипотез при анализе выборочных данных?

2.11. Чем различаются ошибки первого и второго рода? Какие существуют способы для одновременного уменьшения этих ошибок?

2.12. Дайте определение понятий «критическая область» и «область принятия решений».

2.13. Как можно проверить предположение о нормальности распределения во входном наборе данных?

2.14. Какими способами можно установить взаимосвязь переменных? Какие методы оценки корреляций имеются в SPSS?

2.15. В чем состоит смысл таблиц сопряженности и как на их основе можно установить взаимосвязь признаков?

2.16. В чем заключается смысл t -критерия Стьюдента, F -критерия Фишера и критерия χ^2 Пирсона?

2.17. В каких случаях для сравнения выборочных средних используется t -критерий Стьюдента?

2.18. Чем параметрический критерий отличается от непараметрического? Приведите примеры параметрических и непараметрических критериев.

2.19. В чем состоит сущность метода дисперсионного анализа?

2.20. Перечислите основные непараметрические критерии и опишите методологию проверки статистических гипотез о параметрах выборочных показателей на их основе.

Задачи и упражнения

2.1. В файле Задача_2_1. xls содержатся данные о распределении среднего экзаменационного балла юношей и девушек, обучавшихся на заочном отделении УрАГС. Значение «1» номинативной переменной Пол соответствует юношам, «2» – девушкам. Импортируйте данные в SPSS, ис-

пользуя стандартный интерфейс импорта баз данных. Найдите средний балл и дисперсию среднего балла для юношей и девушек, используя метод получения дескриптивной статистики. Постройте графики распределений баллов для юношей и девушек в одной системе координат. Постройте графики распределения баллов девушек (юношей) и предполагаемого нормального распределения в одной координатной системе. При выполнении этих заданий следует обратиться к примерам 2.1 и 2.2.

2.2. В файле Задача_2_2. xls содержатся данные тестовых исследований 38 учащихся 10 -х классов школ с физико-математическим уклоном (значение переменной Категория для них выбрана равной 1) и 38 учащихся 10-х классов школ с гуманитарным уклоном (значение переменной Категория для них выбрана равной 2). Пол учащихся кодируется цифрами: 1 – юноши и 2 – девушки. Первый тест характеризует способность к рассуждению и условно назван «Здравомыслие». Второй тест характеризует способности находить аналогии. Третий тест характеризует способность к обобщению. Четвертый тест оценивает развитость навыков численного счета. Наконец, последний тест направлен на оценку воображения. В результате теста каждому из респондентов выставлялась некоторая сумма баллов, причем, чем выше балл, тем качественнее был данный ответ. Импортируйте тестовые данные в редактор данных SPSS. Используя дескриптивную статистику, исследуйте, зависят ли результаты тестов от категории (пола) учащихся. Постройте коробчатые диаграммы для каждого из тестов и для каждой категории учащихся. Какой вывод можно сделать в результате проведенного анализа?

2.3. Файл Задача_2_3.sav содержит данные об итогах тестирования учащихся трех выпускных классов, а также данные о среднем балле за 10-й, 11-й классы, внешкольных увлечениях (хобби) и профиле вуза, выбранного учеником для поступления. Подробная информация о смысле и характере переменных содержится в самом файле и становится доступной после загрузки данных в окне Variable View редактора данных SPSS. Создайте новую переменную Средний балл, представляющую собой средний балл за 10-й и 11-й классы. Определите средний балл для юношей и девушек. Построив коробчатые диаграммы, определите, есть ли взаимосвязь между средним баллом и типом вуза, в который собирается поступить абитуриент. Выясните, существует ли взаимосвязь между средним баллом и внешкольными увлечениями. Одинаковым ли будет этот результат, если рассмотреть эту взаимосвязь отдельно для юношей и девушек?

2.4. Используя числовые данные предыдущей задачи, постройте распределение школьников в 10-м классе по среднему баллу с шагом 0,2 балла. Найдите среднее значение и среднеквадратическое отклонение этого распреде-

ления. Постройте в одних координатных осях найденное и предполагаемое нормальное распределение. Проведите те же вычисления для среднего балла за 11-й класс и среднего балла за два последних года обучения.

2.5. В файле Задача_2_5. xls содержатся данные о денежных доходах и потребительских расходах в расчете на душу населения в 2005 году в августе, сентябре и октябре в областях РФ, а также численность жителей этих областей. Используя опции Transform/Compute, найдите средние доходы и расходы за три месяца, определив новые переменные. Найдите средние месячные доходы и расходы, приходящиеся на одного жителя семи федеральных округов РФ и постройте по этим данным столбчатые диаграммы. Обратите внимание на правильный учет числа жителей, проживающих в различных областях, входящих в федеральный округ, определив весовой коэффициент.

2.6. Файл Задача_2_6.xls содержит данные об обеспеченности жильем жителей различных областей РФ. В этом файле также приведены данные о числе жителей в этих областях и принадлежность их к одному из федеральных округов РФ. Предполагая, что обеспеченность жильем подчиняется нормальному распределению, найдите 95 % -й доверительный интервал для общей площади, приходящейся на одного человека в различных федеральных округах РФ. Обратите внимание на необходимость предварительной подготовки данных до импорта их в редактор данных SPSS.

2.7. Имеются данные об ежегодной доходности по акциям, облигациям и казначейским векселям за период с 1972 по 2001 год (файл Задача_2_7.xls). Постройте 95 % -е доверительные интервалы доходности и риска для акций, облигаций и казначейских векселей (под риском доходности активов обычно понимают среднеквадратическое отклонение доходности).

2.8. В файле Задача_2_8.xls приведены официальные результаты (процент голосов, набранных различными партиями) в ходе выборов в Государственную Думу в 2003 году и итоги Exit poll, а также разность между официальными результатами и данными Exit pool. Используя t - критерий Стьюдента, выясните, можно ли считать с 95 % -й вероятностью, что средняя ошибка подсчета голосов Exit pool равна нулю. Найдите среднюю ошибку и М-оценки средней ошибки.

2.9. Используя данные предыдущей задачи, определите, используя тест Колмогорова – Смирнова, можно ли считать, что разность между официальными результатами и данными Exit pool выборов в Государственную Думу в 2003 году подчиняется нормальному распределению.

2.10. Используя данные, содержащиеся в файле Задача_2_1.xls, выясните, можно ли считать, что распределение среднего балла юношей и девушек, обучающихся на заочном отделении УрАГС, является нормальным. Для

ответа на поставленный вопрос постройте гистограмму распределения и выполните тест Колмогорова – Смирнова.

2.11. Используя тест Колмогорова – Смирнова, выясните, подчиняется ли доходность акций, облигаций и казначейских векселей нормальному распределению (данные находятся в файле Задача_2_7.xls). Проверьте, подчиняется ли доходность этих активов однородному распределению.

2.12. В файле Задача_2_12.sav приведен рейтинг стран по уровню коррумпированности чиновников и уровню демократичности (прозрачности) экономики. Данные относятся к 2003 году. Используя коэффициент ранговой корреляции Спирмена, выясните, существует ли тесная взаимосвязь между уровнем коррупции в стране и степенью прозрачности экономики.

2.13. Известно, что доходность различных активов взаимосвязана. Но активы в разной степени чувствительны к колебаниям цен на рынке. Имеются данные о доходности акций компаний Microsoft, Intel и General Electric за период с августа 1997 по июль 2002 года (файл Задача_2_13.sav). Определите по этим данным, какие из активов наиболее тесно связаны между собой, используя коэффициенты корреляции Пирсона, Спирмена и Кендалла. Оцените статистическую значимость взаимосвязи при уровне значимости 0,05.

2.14. Используя данные, содержащиеся в файле Задача_2_7.xls , выясните, имеется ли статистически значимая (при уровне значимости 0,05) взаимосвязь между доходностью акций, облигаций и казначейских векселей.

2.15. Имеются данные о годовом доходе (тыс. долларов США), опыте работы в промышленности и поле сотрудников (мужской – 0, женский – 1) некоторой фирмы (файл Задача_2_15.xls). Найдите коэффициенты парной корреляции всех переменных, а также коэффициент частной корреляции переменных Оклад – Опыт при исключении влияния переменной Пол. Интерпретируйте полученные результаты.

2.16. В средствах массовой информации США было опубликовано сообщение, что по данным статистики, чем больше муниципалитет вкладывает денег в фонд бесплатной медицины, тем больше регистрируется заболевших. В действительности есть другая переменная, которая сильно коррелирует и с объемом фонда, и с числом заболевших. Этой переменной является число обратившихся за медицинской помощью. Используя данные файла Задача_2_16.sav, покажите, что частная корреляция переменных Фонд и Число_больных невелика, если исключить влияние переменной Число_обращений.

2.17. Имеются данные о среднем балле студентов заочного отделения УрАГС и их возрасте (файл Задача_2_17.xls). Используя градации успевае-

мости Очень высокая, Высокая, Средняя и Низкая и градации возраста студентов Пожилой, Средний, Молодой, определите с помощью критерия хи-квадрат с 95 % -й доверительной вероятностью, есть ли взаимосвязь между успеваемостью и возрастом студентов.

2.18. Имеются данные о занимаемой должности студентов заочного отделения УрАГС и их среднем балле за период обучения в вузе (файл Задача_2_18.xls). Выборка ограничена только такими должностями, которые достаточно часто повторяются. Построив таблицу сопряженности с помощью критерия хи-квадрат, выясните, есть ли статистически значимая взаимосвязь между должностью студентов и их успеваемостью.

2.19. В файле Задача_2_19.xls имеются данные о числе преступлений на 1000 жителей по областям РФ за 2004, 2005 и 2006 годы. Оценивая уровень преступности градациями Высокая – число преступлений на 1000 жителей больше 12, Средняя – число преступлений на 1000 жителей колеблется в интервале от 8 до 12, и Низкая – число преступлений на 1000 жителей меньше 8, постройте таблицу сопряженности «год – преступность» и с помощью критерия хи-квадрат установите, значимо ли изменяется уровень преступности по годам.

2.20. Используя данные, содержащиеся в файле Задача_2_6.xls, выясните, можно ли считать с 95 % -й доверительной вероятностью, что обеспеченность жильем в Центральном федеральном округе выше, чем в Уральском. Аналогичным образом сравните другие пары федеральных округов. Проверьте применимость статистики Стьюдента для анализа этого набора данных.

2.21. Используя непараметрический тест Манна – Уитни, выясните с 95 % -й доверительной вероятностью, различается ли величина прожиточного минимума для трудоспособного населения, пенсионеров и детей в различных федеральных округах РФ. Данные о прожиточном минимуме по областям РФ, а также численность жителей приведены в файле Задача_2_21.xls. Пары сравниваемых округов можно выбирать по своему усмотрению. Следует обратить внимание на необходимость подготовки данных для импорта их в SPSS.

2.22. Используя непараметрический критерий Вилкоксона для двух повторных выборок, определите с 95 % -й доверительной вероятностью, изменился ли уровень миграции населения РФ (отдельно приезд и выезд) в страны дальнего зарубежья и в страны ближнего зарубежья в 2005 году по сравнению с 2004 годом. Аналогичные исследования проведите, взяв другие временные сроки. Данные о миграции содержатся в файле Задача_2_22.xls. Подготовьте данные на рабочем листе книги Excel, а затем загрузите их в редактор SPSS.

2.23. Менеджер крупной сети супермаркета захотел выяснить, зависит ли качество обслуживания клиентов торговыми агентами от сроков их обучения. Он разделил проходящих обучение агентов на три группы по срокам обучения, выделив группы с длительность обучения 1 неделя, 2 недели и 3 недели. Качество работы оценивалось в баллах. Данные представлены в файле Задача_2_23. sav. Выясните, можно ли применить дисперсионный анализ для ответа на поставленный вопрос. Постройте для этих данных коробчатую диаграмму и проведите тест Левене на однородность дисперсии в группах. Интерпретируйте результаты.

2.24. Муниципалитет решил провести конкурс проектов на реконструкцию исторической части города. На конкурс было выставлено 4 проекта, которым были присвоены номера с первого по четвертый. Каждый проект оценивали 9 экспертов по четырем номинациям. В каждой из номинаций проекты оценивались по семибалльной шкале, причем, чем выше качество проекта, тем больше баллов он набирал. Итоги деятельности экспертов по оценке проектов представлены в файле Задача_2_24.xls. Используя дисперсионный анализ, выясните, значимо ли различаются проекты между собой при уровне значимости 0,05. Проверьте также выполнение условия однородности дисперсий для анализируемых данных. Если проекты различаются значимо, то, используя контрасты, определите проект, который является победителем конкурса.

2.25. Менеджер рекламного агентства решил выяснить влияние рекламы на восприятие качества продукции. Был организован эксперимент, в котором использовалось пять видов рекламы одного и того же товара. В рекламе 1 свойства товара были сильно занижены; в рекламе 2 – занижены умеренно; в рекламе 3 – слегка завышены; в рекламе 4 – сильно завышены и в рекламе 5 свойства товара описывались объективно. Была произведена случайная выборка 30 респондентов, которые были разбиты на пять групп по шесть человек. Каждой группе вручался один из вариантов рекламного проспекта и образцы одного и того же товара. Респондентам предстояло, изучив рекламу, оценить качество товара по трем параметрам: внешний вид, долговечность, потребительские качества. Каждый показатель оценивался по семибалльной шкале, и результат суммировался. Результаты оценки качества товара респондентами представлен в файле Задача_2_25.xls. Используя дисперсионный анализ, 1) выясните, существует ли статистически значимая разница между средними рейтингами товара для разных групп; 2) определите, для какой группы средний рейтинг товара значимо отличается от рейтинга других групп; 3) используя тест Левене, проверьте данные на однородность дисперсии в разных группах.

2.26. Используя ранговый критерий Крускала – Уоллеса с 95 % -й доверительной вероятностью, выясните, можно ли считать, что по числу престу-

плений в расчете на 1000 жителей Центральный, Северо-Западный и Южный федеральные округа значимо не различаются. Данные содержатся в файле Задача_2.26.xls.

2.27. Менеджер учебного центра решил выяснить, зависит ли скорость реакции работников на сборочном конвейере от программы обучения. Из 25 вновь нанятых работников было сформировано три группы, которые обучались по разным программам. После обучения работники прошли испытание, в ходе которого измерялась скорость их реакции (1– самая быстрая, 25 – самая медленная) при работе на сборочном конвейере. Результаты испытаний приведены в файле Задача_2_27.xls. С помощью рангового критерия Крускала – Уоллеса выясните, существует ли статистически значимая разница между медианами скорости реакции работников, прошедших подготовку по разным программам, если уровень значимости равен 0,01.

2.28. В файле Задача_2_28.xls содержатся данные о приросте/убыли населения в областях РФ за период с 1990 по 2005 год (чел). Используя эти данные, найдите ответы на следующие вопросы: 1) различаются ли тенденции прироста /убыли населения в областях РФ в 1990 и 2000 годах? Аналогичным образом проанализируйте данные за 2004 и 2005 годы; 2) одинакова ли тенденция прироста/убыли населения на 100 000 жителей в областях, входящих в различные федеральных округа в 2005 году? Самостоятельно выберите 3 – 4 федеральных округа и проведите анализ только для этих округов. Данные о числе жителей в различных областях РФ содержатся в файле Задача_2_28а.xls. Самостоятельно подготовьте данные в нужном формате на листе рабочей книги Excel, а затем импортируйте данные в SPSS. Выполните анализ, используя различные процедуры SPSS, и затем сравните полученные результаты. Получив результаты, сделайте статистически обоснованный вывод о процессах прироста/убыли населения в областях РФ.

ГЛАВА 3. РЕГРЕССИОННЫЙ АНАЛИЗ. СТАТИСТИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ И ПРИНЯТИЕ РЕШЕНИЙ

3.1. Однофакторная и многофакторная регрессии в SPSS

Как уже указывалось в главе 1, задачей регрессионного анализа является построение математической модели взаимосвязи явлений и оценка ее статистической значимости на основании имеющихся данных. Если построенная модель статистически значима, то она позволяет строить прогноз развития явлений.

Прогнозирование с использованием однофакторной или многофакторной регрессий начнем с рассмотрения линейных моделей или моделей, сводящихся к линейным, простым преобразованием переменных (модели линейные по существу). Для построения регрессионных уравнений в этом случае SPSS использует стандартный метод наименьших квадратов (МНК), о котором уже упоминалось в главе 1.

При практическом применении регрессионного анализа приходится сталкиваться с двумя проблемами: во-первых, это проверка применимости метода МНК, а во-вторых – оценка статистической значимости построенной модели. В обоих случаях исключительно важную роль играет анализ остатков (не объясненная регрессионной моделью часть вариации результативной переменной).

Пусть имеется набор значений результативной переменной y_i для n случаев ($i = 1, 2, \ldots, n$) и набор значений объясняющих переменных $x_{1i}, x_{2i}, \ldots, x_{ki}$ (предполагается, что значение результативной переменной y для каждого случая может быть объяснено действием факторов $x_1, x_2, \ldots, x_k$). Для линейной регрессионной модели очевидно можно записать

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_k \cdot x_{ki} + \varepsilon_i, \qquad (3.1)$$

где b_j $(j = 0, 1, \ldots, k)$ – регрессионные коэффициенты; ε_i представляет собой ошибку модели, которая описывает суммарное действие не учтенных моделью факторов. Таким образом, остаток ε_i – это разница между эмпирическим значением результативной переменной y_i и предсказанным значением $y_i^T = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_k \cdot x_{ki}$:

$$\varepsilon_i = y_i - y_i^T .$$

Используя приведенное выше определение ошибки ε_i, сформулируем условия применимости метода МНК.

1. Математическое ожидание ошибки ε_i должно быть равно нулю для всех наблюдений: $M(\varepsilon_i) = 0, i = 1, 2, \ldots, n$. На практике проверка этого условия сводится к нахождению суммы ошибок для всех наблюдений. Если эта сумма близка к нулю, то можно считать, что ошибка действительно является случайной величиной. Если сумма ошибок велика, то это указывает на существенные недостатки модели (либо неправильно выбрана форма регрессионного уравнения, либо в модели недостаточно объясняющих переменных).

2. Случайная ошибка должна иметь постоянную дисперсию (гомоскедастичность исходного набора данных): $D(\varepsilon_i) = \sigma^2$ для любых $i = 1, 2, \ldots, n$. Хотя разработано достаточно большое количество тестов для проверки входного набора данных на гетероскедастичность, в SPSS предлагаются только визуальные (графические) методы. Подробнее мы познакомимся с ними при рассмотрении примеров.

3. Отсутствие автокорреляции. Случайные отклонения $\varepsilon_i, \varepsilon_j$ являются независимыми случайными величинами. Автокорреляция особенно часто возникает в рядах динамики, представляющих собой данные, например, экономической деятельности предприятий, взятые в разные моменты времени. Для выявления автокорреляции во входном наборе данных в SPSS используется тест Дарбина – Уотсона. При построении регрессионной модели можно заказать проведение этого теста, однако интерпретировать результат аналитик должен самостоятельно.

4. Модель является линейной относительно параметров $b_0, b_1, \ldots, b_k$. Строго говоря, линейность по объясняющим переменным не требуется, поскольку, например, регрессионную модель $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$ можно представить в виде $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$, где $x = x_1$, $x^2 = x_2$.

5. Отсутствие мультиколлинеарности. Между объясняющими переменными не должно быть сильной линейной взаимосвязи (иначе говоря, никакая объясняющая переменная не может быть представлена в виде линейной комбинации других переменных). В SPSS степень коллинеарности для каждой переменной i $(i = 1, 2, \ldots, k)$ контролируется вычислением толерантности $T_i = 1 - R_i^2$, где R_i^2 – коэффициент детермина-

ции для регрессионной модели, в которой i-я переменная выступает как результативная, а остальные $k-1$ переменные – как факторные. Наряду с толерантностью в SPSS приводится связанная с ней величина – коэффициент инфляции VIF (*Variance Inflationary Factor*), который равен величине, обратной толерантности

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}. \qquad (3.2)$$

Обычно считается, что если для некоторой переменной $\mathrm{VIF} \geq 5$, то эта переменная сильно связана с другими и ее следует исключить из регрессионной модели.

6. Случайные ошибки ε_i распределены по стандартному нормальному закону. Это последнее условие уже не связано с применимостью МНК, однако оно необходимо, поскольку позволяет получить статистические оценки значимости регрессионного уравнения и регрессионных коэффициентов.

В действительности следует иметь в виду, что если условия 1 – 5 применимости метода МНК выполняются, то найденные регрессионные коэффициенты являются наилучшими. Невыполнение некоторых из этих условий может и не означать полного краха регрессионной модели, но в значительной степени ставит ее под сомнение. В дальнейшем мы еще вернемся к этому вопросу при рассмотрении конкретных примеров.

В табл. 3.1 представлены основные характеристики переменных, которые могут использоваться в регрессионном анализе

Таблица 3.1

Основные характеристики переменных, участвующих в регрессионном анализе

Зависимые переменные		Независимые переменные	
количество	тип	количество	тип
Одна	Интервальные, порядковые	Любое	Интервальные, порядковые, дихотомические

Если условия применимости регрессионной модели выполняются, то на следующем шаге следует оценить качество модели в целом и статистическую значимость регрессионных коэффициентов.

Рассмотрим вначале способы оценки качества регрессионного уравнения в целом. Для этих целей в статистике используется так на-

зываемый фактор детерминации R^2, который представляет собой долю дисперсии результативной переменной, которая объясняется регрессионной моделью:

$$R^2 = \frac{Q_R}{Q_T} = \frac{\sum_{i=1}^{n}(y_i^T - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (3.3)$$

В этой формуле величина Q_R представляет собой сумму квадратов отклонений, объясняемую регрессионной моделью, Q_T – полную сумму квадратов отклонений. Их разность образует сумму квадратов отклонений Q_E, не объясненную регрессионной моделью: $Q_E = Q_T - Q_R$. Чем больше доля дисперсии, объясняемая моделью, тем выше качество регрессионного уравнения. Поэтому, чем больше значение R^2 приближается к единице, тем выше качество модели.

Чтобы избежать необоснованного завышения качества регрессионной модели при добавлении новой объясняющей переменной, вводится нормированный фактор детерминации. Очевидно, что формулу (3,3) можно представить в виде

$$R^2 = 1 - \frac{Q_E}{Q_T}. \qquad (3.4)$$

Нормированный (скорректированный) фактор детерминации использует не просто отношение Q_E / Q_T, а отношение этих величин, приходящихся на одну степень свободы

$$R^2{}_{норм} = 1 - \frac{Q_E/(n-m-1)}{Q_T/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-m-1}. \qquad (3.5)$$

В этой формуле m – число предикторов (объясняющих переменных), n – число наблюдений. Для оценки качества многофакторной регрессионной модели предпочтительнее использовать нормированный фактор детерминации. При построении регрессионного уравнения в SPSS фактор детерминации и нормированный фактор детерминации автоматически вычисляются, и их можно найти в таблице выдачи результатов.

Фактор детерминации дает лишь качественную оценку регрессионной модели. Количественно значимость уравнения множественной регрессии в целом оценивается с помощью F -критерия Фишера. По

существу, в этом случае проверяется гипотеза H_0 об одновременном равенстве нулю всех коэффициентов при объясняющих переменных:

$$\beta_0, = \beta_1 = \beta_2 = \ldots = \beta_k = 0 .$$

Если эта гипотеза не отвергается, то делается заключение о том, что совокупное влияние объясняющих переменных на результативную переменную является статистически незначимым.

Проверка гипотезы H_0 производится на основании дисперсионного анализа (сравнения объясненной и остаточной дисперсии). Строится статистический критерий

$$F = \frac{Q_R \cdot (n-m-1)}{Q_E \cdot m}. \qquad (3.6)$$

Буквой *m* здесь обозначено число факторов, включенных в модель. Если $F_{\text{эмп}} > F_{\text{кр}}$, то нулевая гипотеза отклоняется и регрессионное уравнение признается статистически обоснованным. Величина $F_{\text{эмп}}$ вычисляется по формуле (3.6), $F_{\text{кр}}$ – это критическое значение статистики Фишера для распределения с числами степеней свободы m и $n-m-1$. В SPSS таблица итоговых результатов содержит эмпирическое значение статистики Фишера, вычисляемое по формуле (3.6), и соответствующий этому значению уровень значимости, который численно равен площади под кривой распределения Фишера справа от значения случайной величины $F = F_{\text{эмп}}$.

В линейной регрессии обычно оценивается значимость не только уравнения в целом, но и отдельных его регрессионных коэффициентов. Для оценки статистической значимости коэффициентов регрессии используются случайные величины

$$t_{b_i} = \frac{|b_i - \beta_i|}{\Delta b_i},$$

где β_i – регрессионные коэффициенты в генеральной совокупности, Δb_i – оценки величин $\sqrt{\sigma^2_{\beta_i}}$, характеризующих дисперсию регрессионных коэффициентов в генеральной совокупности. В качестве нулевой гипотезы выдвинем предположение, что регрессионные коэффициенты в генеральной совокупности равны нулю: $\beta_i = 0$.

В условиях справедливости выдвинутой гипотезы случайные величины t_{b_i} подчиняются распределению Стьюдента. Поэтому для проверки гипотезы нужно вычислить эмпирические значения t_{b_i}:

$$t_{b_i} = \frac{|b_i|}{\Delta b_i}, \tag{3.7}$$

и затем сравнить их с критическим значением статистики Стьюдента $t_{\text{кр}}$ при заданном уровне значимости и числе степеней свободы $n-2$.

В SPSS нет необходимости вычислять эмпирическое значение статистики Стьюдента для каждого из регрессионных коэффициентов по формуле (3.7), поскольку в итоговой таблице отчета по модели содержатся как эмпирические значения статистики Стьюдента, вычисленные по формуле (3.7), так и соответствующий этим значениям уровень значимости.

Как известно, в случае многофакторной регрессии существует проблема отбора объясняющих переменных. Всегда встает вопрос, улучшает или нет регрессионную модель включение еще одной переменной. Для решения этой проблемы SPSS предлагает пошаговый метод включения переменных, позволяющий отобрать переменные, объясняющие большую часть дисперсии результативной переменной.

Закончив краткое обсуждения основных проблем регрессионного анализа, обратимся к примерам реализации регрессионных моделей в SPSS начиная с простейших задач однофакторной регрессии.

Пример 3.1

Менеджер крупной школы бизнеса с балльно-рейтинговой системой оценки знаний учащихся хотел бы выяснить, есть ли статистически значимая зависимость межу итоговым баллом, который студенты набирают к концу обучения, и баллом, которым оцениваются знания студентов по базовому курсу «Современный менеджмент». В файле Пример_3_1.xls приведены данные о 20 студентах, окончивших школу бизнеса накануне.

а) Построить корреляционную диаграмму (диаграмму разброса) и линию тренда на этой диаграмме.

б) Построить линейную регрессионную модель, оценить статистическую значимость модели в целом и регрессионных коэффициентов. Сохранить предсказываемые моделью значения результативной переменной и не объясняемые моделью остатки. Убедиться, что остатки подчиняются распределению, близкому к нормальному, а сумма остатков близка к нулю.

в) Построить график зависимости не объясняемых регрессионной моделью остатков от предсказываемых моделью значений результативной

переменной и на основании этого дать заключение о наличии гетероскедастичности во входном наборе данных.

Решение

Загрузим данные в редактор данных SPSS. Для построения корреляционной диаграммы выбираем опции меню Graphs/Scatter/Dot/Simple Scatter/Define (Графики/Диаграммы рассеяния/Точки/Простые диаграммы/Определить). В результате откроется окно, в котором переменную Итог следует поместить в окошко Y Axis (Ось Y), переменную Менеджмент в окошко X Axis (Ось X) и нажать кнопу OK. В результате в окне выдачи результатов появится корреляционная диаграмма (точки), изображенная на рис. 3.1.

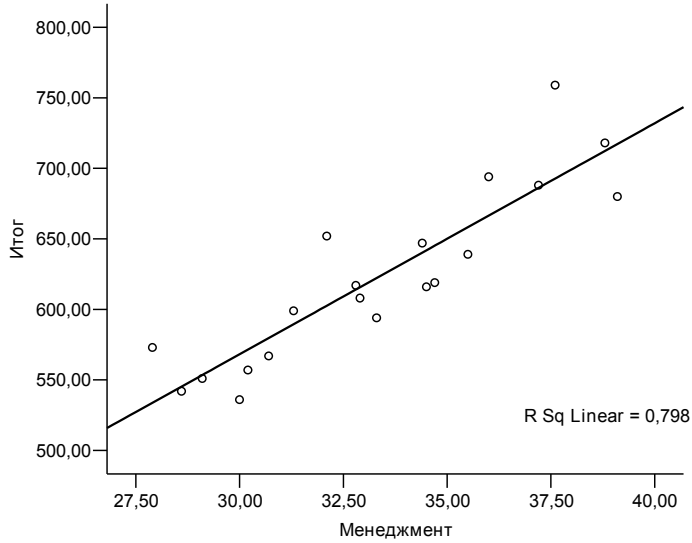


Рис. 3.1. Корреляционная диаграмма и линия тренда

Для того чтобы на диаграмме появилась линия тренда, необходимо дважды щелкнуть левой клавишей мыши в поле диаграммы (при этом откроется окно редактора диаграмм), затем в меню редактора диаграмм выбрать опцию Elements/Fit Line at Total (Элементы/Подгонка линии).

Из рисунка следует, что для предлагаемого набора данных может оказаться подходящей линейная форма регрессионной линии. Построим линейную регрессионную модель по входному набору данных. Результативной переменной выбираем переменную Итог, а объясняющей пере-

менной – набранный балл по курсу «Современный менеджмент» (переменная Менеджмент). Для построения регрессионной модели выбираем опции Analyze/Regression/Linear (Анализ/Регрессия/Линейная). В итоге откроется окно, в котором переменную Итог следует поместить в окошко, озаглавленное Depend (зависимая переменная), а переменную Менеджмент в окно Independent (независимая переменная) (рис. 3.2).

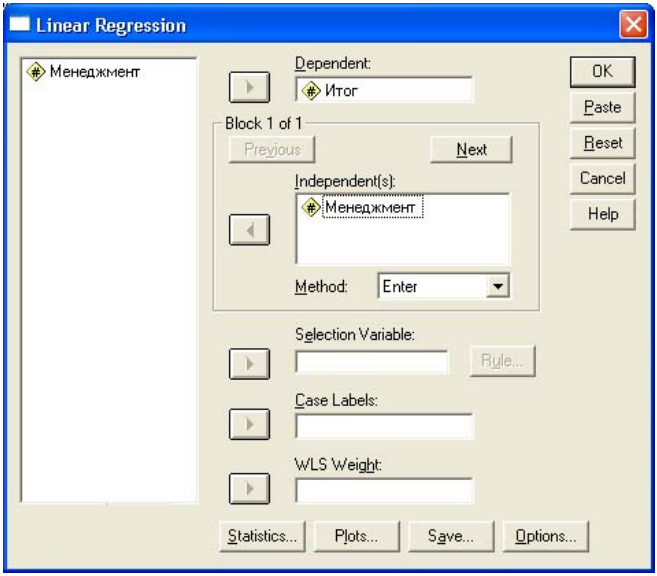


Рис.3.2. Окно установки параметров линейной регрессионной модели

Затем нужно нажать кнопку Statistics (Статистика) и в открывшемся окне поставить дополнительно галочки в окошках Confidence interval (Доверительные интервалы) и Durbin-Watson. Это обеспечит нам вывод доверительных интервалов для регрессионных коэффициентов и вывод статистики Дарбина – Уотсона, которая позволяет исключить или подтвердить наличие автокорреляции в данных. Нажав кнопку Continue, вернемся в окно задания параметров линейной регрессионной модели.

Для обеспечения вывода графической информации нажмем кнопку Plots и поставим галочку в окошке Histogram (Гистограмма), что обеспечит нам вывод информации о распределении остатков и величине их среднего значения. Для проверки входного набора данных на гомоскедастичность переменную *ZRESID поместим в окно Y, а переменную *ZPRED – в окно X. Переменная *ZRESID представляет собой стандартизированные остатки (величина остатка ε_i, деленная на оценку стандартного отклонения остатков

$\Delta\varepsilon = \sqrt{\sigma_\varepsilon^2}$. Переменная *ZPRED представляет собой стандартизированное предсказанное значение результативной переменной:

$$*\,\mathrm{ZPRED} = \frac{Y^T - \overline{Y^T}}{\sigma_{Y^T}}.$$

Для продолжения работы следует вернуться в окно установки параметров линейной регрессии и нажать здесь клавишу OK. В результате появится окно выдачи результатов. Начнем анализ с гистограммы распределения остатков, на которую наложена кривая нормального распределения (рис. 3.3).

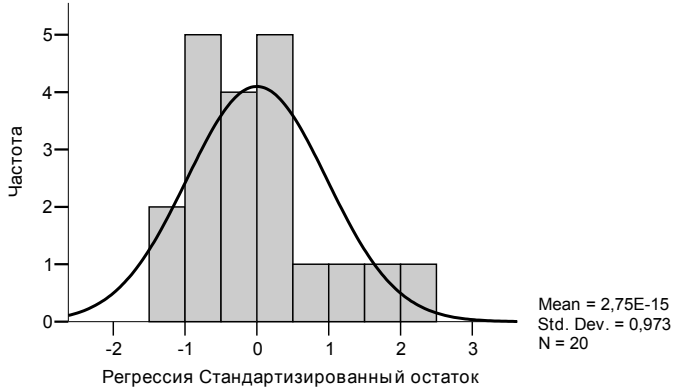


Рис. 3.3. Гистограмма распределения остатков с наложенной кривой нормального распределения

Для оценки качества модели необходимо убедиться в применимости метода МНК для анализируемого набора данных. Из рисунка следует, что остаток является случайной величиной с распределением, близким к нормальному, и средним значением, практически равным нулю.

Гомоскедастичность набора данных можно подтвердить, изучая график стандартизированных остатков в зависимости от стандартизированных значений результативной величины. Если на такой диаграмме разброс точек по оси ординат не имеет тенденции к увеличению/ уменьшению при смещении по оси абсцисс (разброс одинаков в начале и в конце таблицы, упорядоченной по величине результативной переменной), как на рисунке 3.4, то набор данных можно признать гомоскедастичным.

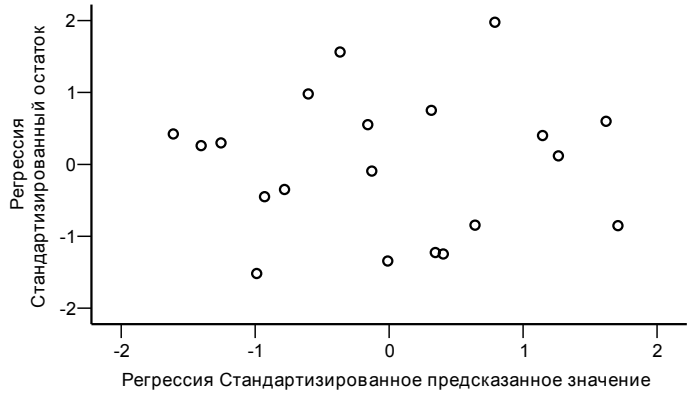


Рис. 3.4. Диаграмма рассеяния, позволяющая обнаружить гомоскедастичность набора данных

Одним из условий применимости МНК является независимость значений результативной переменной. Выполнение этого условия проверяется с использованием статистики Дарбина –Уотсона, значение которой содержится в итоговой сводке по модели (табл.3.2).

Таблица 3.2

Итоговая сводка по регрессионной модели

R	R -квадрат	Скорректированный R -квадрат	Стд. ошибка оценки	Статистика Дарбина – Уотсона
0,893	0,798	0,787	28,58693	1,642

Как следует из табл. 3.2, значение статистики Дарбина – Уотсона $d = 1{,}642$. Несложные вычисления показывают, что статистика Дарбина – Уотсона просто связана с коэффициентом автокорреляции первого порядка: $d \approx 2(1-r)$, где r – коэффициент автокорреляции первого порядка для остатков регрессионной модели. Если корреляции нет, то $d = 2$. Если корреляция полная, то $d = 0$. Если корреляция полная и отрицательная, то $d = 4$. Хотя тест Дарбина – Уотсона не является в полном смысле этого слова статистическим тестом, тем не менее для него разработаны специальные таблицы, которые для заданного уровня значимости α, числа наблюдений n и количества объясняющих переменных m дают два числа $d_{\text{верх}}$ и $d_{\text{нижн}}$. В рассматриваемом нами примере примем уровень значимости $\alpha = 0{,}05$, число точек наблюдения – 20, число объясняющих переменных – 1. По таблицам статистики Дарбина – Уотсона (см., напри-

мер: Бородич С. А. Эконометрика, прил. 6) $d_{\text{нижн}} = 1{,}201$ и $d_{\text{верхн}} = 1{,}411$. Теперь для проверки гипотезы об отсутствии автокорреляции следует обратиться к диаграмме на рис. 3.5.

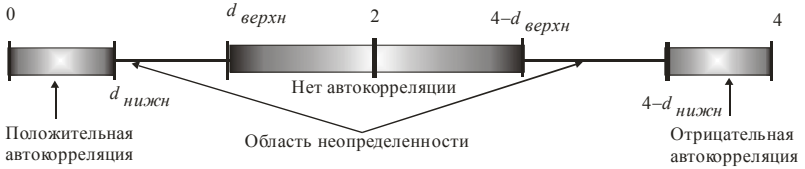


Рис. 3.5. Интерпретация значений величины d -критерия Дарбина – Уотсона

Как следует из приведенного рисунка, критерий попал в область отсутствия автокорреляции, и поэтому значения результативной переменной можно считать независимыми.

После того как мы установили обоснованность применения метода МНК для анализа данных этого примера, можно проанализировать качество построенной регрессионной модели. Общее качество регрессионной модели определяется с помощью фактора $R^2 = 0{,}798$. Поскольку максимальное значение $R^2 = 1$, то можно утверждать, что качество регрессионной модели весьма высокое (около 80 % дисперсии результативной переменной предлагаемая модель объясняет). Скорректированный или нормированный фактор детерминации в данном случае использовать нет большого смысла, поскольку в задаче всего одна объясняющая переменная. Величина R представляет собой просто коэффициент корреляции факторной и результативной переменных (фактор детерминации равен квадрату коэффициента корреляции).

В табл. 3.3. представлены данные дисперсионного анализа, из которых

Таблица 3.3

Дисперсионный анализ

	Сумма квадратов	Ст.св.	Средний квадрат	*F*	Знач.
Регрессия	58047,378	1	58047,378	71,031	0,000
Остаток	14709,822	18	817,212		
Итого	*72757,200*	*19*			

следует статистическая значимость модели в целом при уровне значимости 0,05. Регрессионные коэффициенты, их интервальная оценка и уровень статистической значимости приведены в табл. 3.4.

Таблица 3.4

Регрессионные коэффициенты

Нестандартизованные коэффициенты		Стандартизованные коэффициенты	*t*	Знач.	95%-й доверительный интервал для B	
B	Стд. ошибка	Бета			нижняя граница	верхняя граница
76,72	65,108		1,18	0,25	–60,07	213,50
16,38	1,94	0,89	8,43	1,1E-07	12,30	20,46

Как следует из таблицы, построенная регрессионная модель имеет вид

$$y_i^T = 76{,}72 + 16{,}38 \cdot x_i \,. \qquad (3.8)$$

Регрессионный коэффициент b_0 (свободный член) статистически незначим при уровне значимости 0,05, а регрессионный коэффициент b_1 – значим. Фактически статистическая незначимость коэффициента b_0 означает, что его можно принять равным нулю, поскольку нулевое значение входит в 95%-й доверительный интервал для b_0. Для регрессионной модели более важна значимость коэффициента b_1, поскольку именно он содержит информацию о линейной взаимосвязи факторной и результативной переменных.

Завершая обсуждение этой простой задачи, следует сказать, что рассчитанные по формуле (3.8) значения результативной переменной, их интервальную оценку, а также величины остатков (ошибки) можно сохранить в редакторе данных, если воспользоваться кнопкой Save в окне задания параметров линейной регрессионной модели (см. рис. 3.2).

Пример 3.2

В файле Пример_3_2.xls содержатся данные о стоимости квартир (долл. США) в зависимости от числа комнат, общей площади, жилой площади, площади кухни, типа дома и т. д. (всего используется 11 характеристик жилплощади). Требуется построить статистически значимую регрессионную модель, позволяющую правильно определять стоимость одно- и двухкомнатных квартир, используя 5 – 6 наиболее важных характеристик жилья.

Решение

В случае многофакторной модели, точно так же, как и в предыдущем случае, следует убедиться, что выполняются условия применимости ме-

тода МНК. Новыми по сравнению с однофакторной моделью являются часто возникающая проблема мультиколлинеарности и проблема отбора объясняющих факторов. При рассмотрении этого примера мы на них и сконцентрируем свое внимание.

Поместим переменную Цена в окно зависимых (Dependent) переменных. Переменную Число комнат в окно Selection Variable (переменная для отбора) и, нажав кнопочку Rule, установим отбираемые значения: less then or equal to 2 (меньше или равно 2). Оставшиеся переменные поместим в окно независимых переменных и выберем метод их пошагового включения в модель (Stepwise). Никогда не следует сразу включать в модель все объясняющие переменные. Более правильным является пошаговый метод включения переменных в модель или метод их пошагового удаления. При этом генерируется несколько моделей сразу, и пользователь может отобрать наиболее приемлемую из них.

Чтобы обеспечить диагностику коллинеарности при включении переменных в модель, в окне выбора параметров линейной регрессии (см. рис. 3.2) следует нажать кнопку Statistics (статистика) и в появившемся окне поставить галочку в окошке Collinearity diagnostics (Диагностика коллинеарности).

Таблица 3.5

Отчет о параметрах многофакторной регрессионной модели

Переменные	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	*t*	Знач.	Толерантность
	B	Стд. ошибка	Бета			
Константа	8330,25	1978,94		4,21	0,00	
Общая_площадь	1253,18	32,25	0,59	38,86	0,00	0,78
Индекс_места	–269,75	8,80	–0,44	–30,65	0,00	0,85
Площадь_кухни	1298,43	179,89	0,11	7,22	0,00	0,74
Тип_дома	–2847,20	484,70	–0,08	–5,87	0,00	0,86
Телефон	3465,80	817,56	0,06	4,24	0,00	0,95
Этаж_квартиры	195,33	73,46	0,04	2,66	0,01	0,89

Как следует из табл. 3.5, все объясняющие переменные, включенные в модель, имеют высокую статистическую значимость и толерантность.

Переменная Жилая_площадь не включена в модель из-за линейной взаимосвязи с другими переменными, на что указывает очень низкое значение толерантности для этой переменной: $T = 0{,}107$ (табл. 3.6).

Таблица 3.6

Отчет о переменных, не включенных в модель

Переменные	Бета	*t*	Знач.	Толерантность
Жилая_площадь	–0,031	–0,751	0,453	0,107
Балкон_лоджия	0,012	,869	0,385	0,885

Переменная Балкон_лоджия дает очень малый вклад в значение результативного фактора. Это можно обнаружить, анализируя стандартизированные значения регрессионных коэффициентов, значения которых приведены в столбце с названием «Бета». Смысл перехода к стандартизированным переменным состоит в том, что в стандартизованной форме все коэффициенты сопоставимы между собой, и по величине коэффициента можно сразу сказать, следует ли переменную включать в модель. Стандартизованный коэффициент для переменной Балкон_лоджия составляет всего 0,012. Поэтому влияние этой переменной на цену почти в четыре раза меньше, нежели влияние переменной Этаж_квартиры.

Достаточно высокое качество регрессионной модели подтверждает и анализ регрессионных остатков (рис. 3.6).

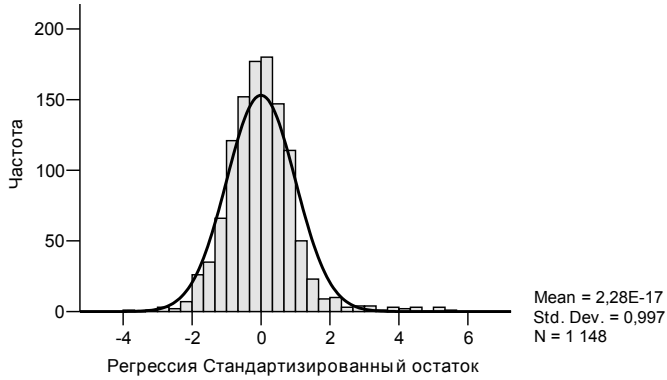


Рис. 3.6. Гистограмма распределения остатков многофакторной модели

Средний остаток фактически равен нулю, а распределение остатков с высокой степенью воспроизводит нормальное распределение. Это означает, что все наиболее важные факторы, определяющие цену квартир, в модель включены и имеющиеся ошибки являются случайными.

Важно отметить, что если бы мы включили в модель все переменные, используя метод включения переменных Enter, то качество регрессионной модели оказалось бы хуже из-за статистической незначимости некоторых регрессионных коэффициентов и трудно интерпретируемой из-за проблемы взаимосвязи (мультиколлинеарности) переменных.

3.2. Нелинейные регрессионные модели в SPSS

Регрессия одно- и многофакторная совсем не обязательно должна быть линейной. В общем случае нелинейные регрессионные модели могут быть сведены к двум основным типам:

1) линеаризуемые модели;

2) нелинеаризуемые модели.

Линеаризуемые модели можно, в свою очередь, также разбить на два класса:

1) модели нелинейные по переменным;

2) модели нелинейные по параметрам.

В SPSS есть возможность построения нелинейных моделей любого типа, хотя чаще всего приходится сталкиваться с моделями, допускающими линеаризацию. Линеаризация по существу сводится к подходящей замене переменных (примеры моделей, допускающих линеаризацию, приведены в главе 1). Здесь важно отметить, что после нахождения параметров линеаризованной модели необходимо вернуться к исходным переменным и обязательно проанализировать поведение остатков, в частности проверить выполнение условия (1.8).

Для построения простых типов нелинейных моделей однофакторной регрессии в SPSS предусмотрена еще одна возможность, которая названа Curve Estimation (Оценка криволинейности). Здесь сразу можно сгенерировать несколько моделей, различающихся формой регрессионной кривой, и затем выбрать наилучшую из построенных моделей.

Для определения регрессионных коэффициентов в моделях, не сводящихся к линейным, в SPSS используется метод итераций, подбирающий регрессионные параметры с использованием алгоритма минимизации суммы квадратов остатков. Для успешной работы этого алгоритма необходимо задать более или менее правильно найденные стартовые значения параметров модели.

Процедуру построения нелинейных регрессионных моделей рассмотрим на нескольких простых примерах.

Пример 3.3

Менеджер небольшой компании решил выяснить, как связаны затраты на рекламирование продукции с ее сбытом. В файле Пример_3_3.sav приведены данные о месячных затратах на рекламу (тыс. долл. США) и месячных объемах продаж (млн долл. США). Требуется установить закон взаимосвязи затрат на рекламу и объема продаж.

Решение

После загрузки файла данных в редактор SPSS активизируем опции меню Analyse/Regression/Curve Estimation. В результате откроется окно установки параметров оценки криволинейности, изображенное на рис. 3.7, в котором отметим галочкой актуальные для решаемой задачи установки.

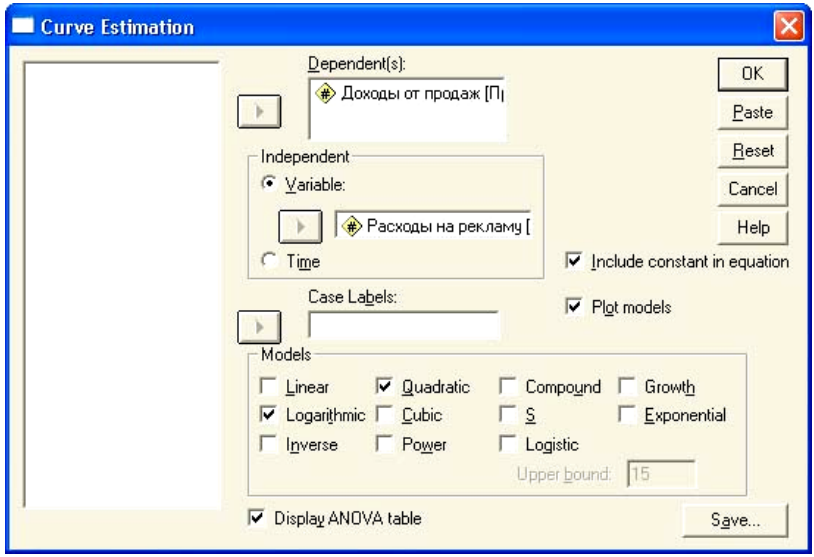


Рис.3.7. Окно установки параметров для подбора вида регрессионной кривой в SPSS

Для того чтобы правильно воспользоваться предоставленной возможностью подбора регрессионной модели с помощью этого окна, необходимо иметь ясное понимание, какие формы регрессионной модели здесь можно

выбрать. В табл. 3.7 приведены названия моделей и соответствующие уравнения взаимосвязи факторной и результативной переменных.

Таблица 3.7

Типы моделей, которые можно выбрать в окне Curve Estimation

Модель	Уравнение взаимосвязи
Линейная (Linear)	$y = b_0 + b_1 \cdot x$
Логарифмическая (Logarithmic)	$y = b_0 + b_1 \cdot \ln x$
Обратная (Inverse)	$y = b_0 + b_1 / x$
Квадратичная (Quadratic)	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2$
Кубическая (Cubic)	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$
Степенная (Pover)	$b_0 \cdot x^{b_1}$
Показательная	$b_0 \cdot b_1^x$
S -кривая	$y = \exp(b_0 + b_1 / x)$
Логистическая (Logistic)	$y = 1/(1/U + b_0 \cdot b_1^x)^*$
Кривая роста (Growth)	$y = \exp(b_0 + b_1 \cdot x)$
Экспоненциальная (Exponential)	$y = b_0 \cdot \exp(b_1 \cdot x)$

* Число U выбирается пользователем и должно превышать максимальное значение зависимой переменной

Для анализируемого набора данных выберем квадратичную модель. Для этого есть достаточно веские экономические основания. В экономике хорошо известен закон убывающей эффективности ресурсов производства. Поэтому подходящими могут оказаться логарифмическая, квадратичная и степенная модели. Эти модели дают примерно одинаковые значения скорректированного значении $R^2 \approx 0{,}9$, но сумма остатков для степенной модели недопустимо высока. По этой причине степенную модель следует сразу отвергнуть. Логарифмическая и квадратичная модели дают одинаковое значение средней ошибки, но скорректированное значение R^2 несколько выше для квадратичной модели. По этой причине для анализируемого набора данных выбрана квадратичная модель. Следует отметить, что с точки зрения экономической интерпретации более привлекательной является логарифмическая модель, поскольку квадратичная модель предсказывает уменьшение объемов продаж при дальнейшем уве-

личении расходов на рекламу, что скорее всего не соответствует действительности. На рис. 3.8 приведена диаграмма рассеяния, квадратичная (штрихпунктирная линия) и логарифмическая (сплошная линия) модели зависимости объема продаж от объема средств, выделяемых на рекламу. Исходные данные объемов продаж изображены светлыми кружками.

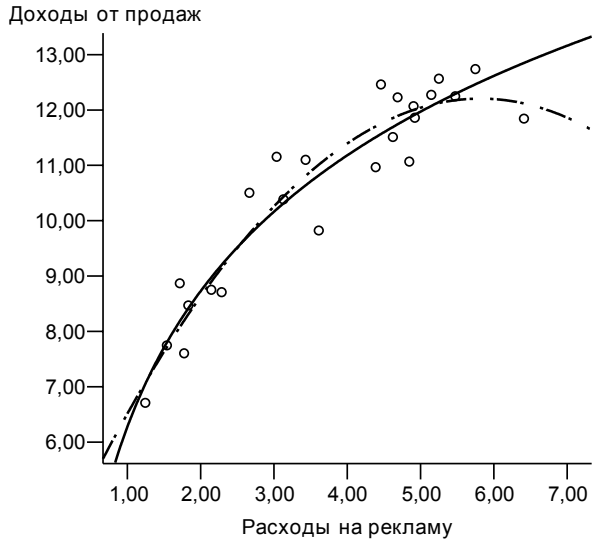


Рис. 3.8. Квадратичная и логарифмическая модели, объясняющие зависимость объема продаж от объема средств, выделяемых на рекламу

В табл. 3.8 приведены данные о регрессионных коэффициентах модели.

Таблица 3.8

Отчет о регрессионных коэффициентах квадратичной модели

Коэффициент	Коэффициенты			*t*	Sig.
	нестандартизованные		стандартизованные		
	Значение коэффициента	Станд. ошибка	Значение коэффициента		
b_1	2,854	0,453	2,440	6,302	0,000
b_2	–0,245	0,061	–1,547	–3,996	0,001
b_0	3,903	0,739		5,280	0,000

Из приведенных данных следует, что все регрессионные коэффициенты являются статистически значимыми и модель может использо-

ваться для принятия экономически обоснованных решений на этом предприятии.

Рассмотрим общий подход построения произвольной нелинейной модели в SPSS, которая не может быть линеаризована с помощью преобразования переменных. Для пояснения деталей такого построения рассмотрим пример.

Пример 3.4

Интернет-провайдер решил выяснить влияние появления вируса в компьютерной сети на долю зараженного вирусом трафика. Для этого отслеживалась доля почтовых отправлений, содержащих вирус. В файле Пример_3.4.sav содержатся данные о доле инфицированных почтовых отправлений в сети в зависимости от времени (числа часов), прошедших от начала заражения. Требуется построить регрессионную модель зависимости доли зараженных почтовых отправлений от времени, прошедшего от начала заражения.

Решение

Чтобы иметь представление о характере взаимосвязи изучаемых переменных, построим поле корреляции для изучаемого набора данных (рис. 3.8). Как строится такая диаграмма, подробно описано в примере 3.1.

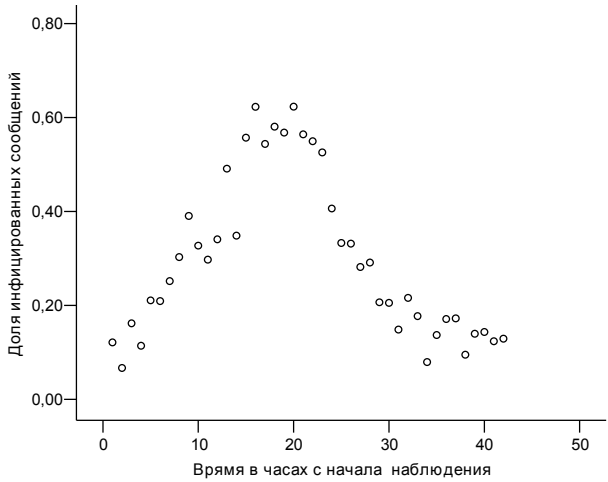


Рис. 3.8. Диаграмма рассеяния для данных примера

Из графика видно, что регрессионная модель должна описываться кривой с максимумом. Нетрудно убедиться, что, например, квадратичная

кривая дает невысокое качество модели ($R^2 \approx 0{,}63$). Вообще говоря, заранее неизвестно, можно ли описать этот набор данных какой-либо одной кривой. Вполне возможно, что вирус вначале беспрепятственно размножался, что приводило к быстрому росту доли инфицированных почтовых сообщений, но по истечении примерно 20 часов пользователи обнаружили вирус и начали с ним эффективно бороться, что привело к быстрому падению числа зараженных сообщений. Поэтому используем для первого участка кривой ($x \le 19$) логистическую модель:

$$y_i^T = \frac{b_0}{1 + b_1 \cdot \exp(-b_2 \cdot x_i)}, \tag{3.9}$$

а для $x \ge 20$ – экспоненциальную модель:

$$y_i^T = a_0 + a_1 \cdot \exp(a_2 \cdot x_i). \tag{3.10}$$

Модели (3.9), (3.10) не могут быть линеаризованы преобразованием переменных. Как уже указывалось выше, в SPSS в этом случае используется прямой численный метод минимизации суммы квадратов отклонений (1.2).

Для правильной работы этого метода необходимо задать более или менее правильные стартовые значения (оценки) регрессионных коэффициентов b_0, b_1, b_2. Эти оценки легко получить, используя входной набор данных. Очевидно, что величина b_0 должна быть больше самого большого значения результативной переменной. Поэтому в качестве стартового значения возьмем $b_0 = 0{,}7$. Для оценки коэффициента b_1 возьмем самое маленькое значение $x_1 = 1$. В этом случае значение экспоненты в формуле (3.9) можно принять близким к единице. Тогда

$$b_1 \approx \frac{b_0}{y_1} - 1 = 4{,}8.$$

Коэффициент b_2 теперь можно оценить по формуле (3.9), подставляя стартовые значения коэффициентов $b_0 = 0{,}7$ и $b_1 = 4{,}8$ и, например, значения $x_{15} = 15$ и $y_{15} = 0{,}56$ получаем $b_2 \approx 0{,}2$.

Аналогично оцениваем и численные значения коэффициентов a_0, a_1, a_2. Будем исходить из того, что коэффициент $a_2 < 0$ (кривая должна описывать спад). Стартовое значение коэффициента a_0 можно положить равным нулю. Тогда коэффициент a_1 оценим как максимальное значения результативной переменной, т. е. $a_1 = 0{,}62$. Коэффициент a_2 найдем, используя одно из известных значений факторной и результа-

тивной переменных, оценки коэффициентов a_0, a_1 и формулу (3.10). Вычисления дают для этого коэффициента значение $a_2 \approx -0{,}04$.

После того как предварительные оценки регрессионных коэффициентов сделаны, открываем окно задания параметров нелинейной модели, используя пункты меню Analyse/Regression/Nonlinear. В результате откроется окно (рис. 3.9), в котором нужно результативную переменную

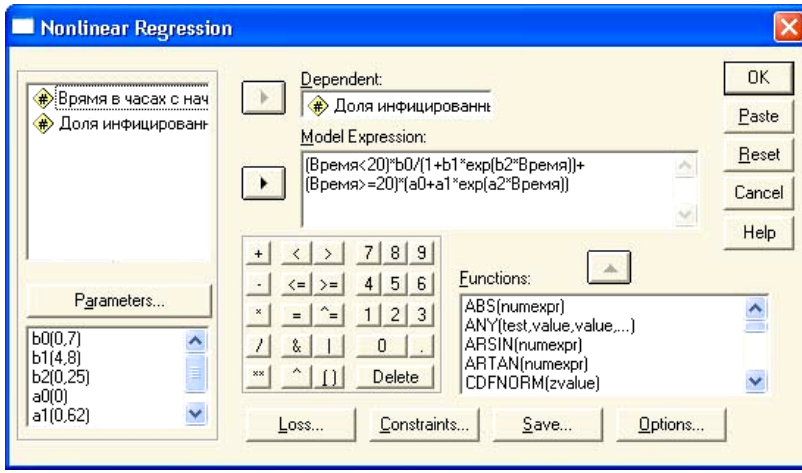


Рис. 3.9. Окно задания параметров нелинейной регрессионной модели

поместить в окно Dependent, а в окне Model Expression (Выражение для модели) с помощью клавиатуры набрать формулу, изображенную на рис. 3.9. После того как формула записана, необходимо задать стартовые значения параметров. Для этого нужно нажать кнопку Параметры и в открывшемся окне последовательно ввести имена и значения параметров, не забывая нажимать кнопку Add для сохранения введенных значений (рис. 3.10).

После того как все параметры введены, следует вернуться в окно установки параметров нелинейной регрессии и с помощью кнопки Save (Сохранить) обеспечить вывод предсказанных значений, поставив галочки в окнах Predicted values и Residuals (Остатки). Запускается процедура поиска регрессионных параметров нажатием кнопки OK в окне установки параметров нелинейной регрессии. В результате будет выдан достаточно подробный отчет о работе программы, который содержит информацию о численных значениях коэффициентов на всех шагах итерации. Этот отчет для нас не представляет интереса, и мы его не приводим. В табл. 3.9 приведен только отчет о значениях регрессионных коэффициентов.

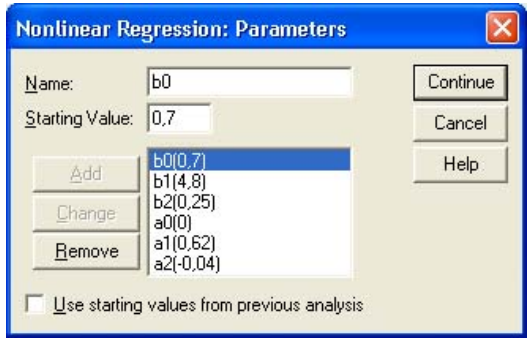


Рис. 3.10. Окно задания стартовых значений параметров нелинейной регрессии

Таблица 3.9

Отчет о регрессионных коэффициентах нелинейной модели

Параметр	Оценка	Станд. ошибка	95% -й доверительный интервал	
			нижняя граница	верхняя граница
b0	0,734	0,127	0,477	0,991
b1	7,428	1,375	4,638	10,217
b2	–0,184	0,040	–0,265	–0,103
a0	0,091	0,030	0,030	0,153
a1	11,471	6,205	–1,114	24,056
a2	–0,150	0,027	–0,205	–0,095

В табл. 3.10 приведены данные дисперсионного анализа для построенной модели.

Таблица 3.10

Данные дисперсионного анализа для нелинейной модели

Источник вариации	Сумма квадратов	Число степеней свободы	Средняя сумма квадратов
Регрессия Q_R	4,884	6	0,814
Остатки Q_E	0,082	36	0,002

Фактор детерминации R^2 может быть вычислен с использованием данных табл. 3.10 по формуле (3.3). Вычисления дают $R^2 = 0{,}98$, что говорит о высоком качестве модели.

Как проверить наличие нелинейных зависимостей и взаимодействие переменных при построении регрессионного уравнения? Существует достаточно простой подход, состоящий в том, что в качестве объясняющей переменной добавляется новая переменная, являющаяся некоторой степенью исходной. Если регрессионные коэффициенты для новой переменной получаются значимыми, то включение новой объясняющей переменной может быть оправданным.

Точно так же определяется и необходимость учета взаимодействия переменных. В регрессионное уравнение вводится новая объясняющая переменная, которая, например, равна произведению двух исходных. Если в построенной модели регрессионные коэффициенты получаются значимыми, то это может свидетельствовать о необходимости учета взаимодействия переменных. Конечно, построенная модель должна иметь разумную содержательную интерпретацию. Если такой интерпретации добиться не удается, то использование математической модели взаимосвязи явлений остается под вопросом.

Пример 3.5

В файле Пример_3_5.xls содержатся данные о годовом окладе работников некоторой фирмы (тыс. долл. США), опыте работы в промышленности и поле (мужской, женский). Переменная Пол закодирована таким образом, что ее значение равно 1 для женщин и 0 для мужчин. Требуется спрогнозировать оклад в зависимости от пола и опыта работы, а также ответить на вопрос о наличии или отсутствии дискриминации работников этого предприятия по половому признаку.

Решение

После загрузки данных в окно редактора SPSS создадим новые переменные с именами Опыт2 и Опыт_Пол. Значения первой переменной Опыт2 – это просто значения переменной Опыт, возведенные в квадрат. Смысл введения новой переменной состоит в том, что заработная плата с опытом растет до определенных пределов. При дальнейшем увеличении опыта заработная плата будет падать из-за снижения производительности труда с возрастом. Появление квадратичного члена как раз позволяет учесть это явление, если он будет входить в регрессионное уравнение с отрицательным знаком. Переменная Опыт_Пол представляет собой произве-

дение переменных Опыт и Пол. Включение этой переменной в регрессионное уравнение позволяет выяснить, как влияет опыт на заработную плату у женщин. Если регрессионный коэффициент при этой переменной окажется отрицательным, то это будет свидетельствовать о наличии дополнительного фактора снижения заработной платы у женщин.

Как можно присвоить значения новым переменным, подробно обсуждалось в примере 2.1.

После того как новым переменным присвоены значения, построим линейную многофакторную модель с предикторами Опыт, Пол, Опыт2 и Опыт_Пол. Выберем пункты меню Analyze/Regression/Linear (Анализ /Регрессия/Линейная). В результате откроется окно установки параметров линейной регрессионной модели (см. рис. 3.2), в котором переменную Оклад нужно поместить в окно зависимых переменных (Dependent), а остальные переменные – в окно независимых переменных. В качестве способа включения переменных выберем метод Backward (Обратный), при котором сначала все переменные включаются в модель, а затем поочередно подвергаются анализу, при котором некоторые из переменных могут быть исключены из модели. Все остальные параметры модели можно не изменять и просто нажать кнопку OK. В результате появится окно выдачи результатов, в котором для нас представляют интерес таблица данных дисперсионного анализа (табл. 3.11) и таблица регрессионных коэффициентов (табл. 3.12).

Таблица 3.11

Сводка для модели

Модель	R	R -квадрат	Скорректированный R -квадрат	Стд. ошибка оценки	Статистика Дарбина – Уотсона
1	0,931	0,866	0,860	4,53028	
2	0,930	0,866	0,861	4,51399	1,871

Модель 1 – предикторы: константа, Опыт2, Пол, Опыт_Пол, Опыт.
Модель 2 – предикторы: константа, Опыт2, Опыт_Пол, Опыт.

Из двух представленных в таблице моделей программой была отобрана вторая модель, содержащая меньшее число переменных, но дающая большее значение скорректированного фактора R^2. Следует обратить внимание на статистику Дарбина – Уотсона. Значение 1, 871 показывает, что автокорреляция во входном наборе данных отсутствует. Довольно высокое значение скорректированного значения $R^2 = 0{,}861$ означает, что

регрессионная модель объясняет свыше 86 % изменчивости результативного признака и может считаться вполне успешной.

Обсудим теперь данные, приведенные в табл. 3.12. Статистика коллинеарности указывает, что переменные Опыт и Опыт2 мультиколлинеарны, что и следовало ожидать, поскольку эти переменные находятся в простой функциональной связи. Это, однако, не сказывается существенно на значимости регрессионных коэффициентов.

Таблица 3.12

Регрессионные коэффициенты и их значимость

Переменные модели	Нестандартизованные коэффициенты		*t*	Знач.	Статистики коллинеарности
	B	Стд. ошибка			Толерантность
Константа	59,057	1,455	40,578	0,000	
Опыт	0,781	0,355	2,203	0,030	0,059
Опыт_Пол	–2,073	0,088	–23,622	0,000	0,882
Опыт2	–0,034	0,017	–1,921	0,058	0,061

Можно считать, что все регрессионные коэффициенты значимы. Даже наихудший уровень значимости, равный 0,058 для регрессионного коэффициента Опыт2, дает основания считать модель вполне приемлемой.

Переменная Пол исключена программой из регрессионной модели как раз из-за того, что она оказалась незначимой.

Построим уравнение регрессионной кривой для предсказания уровня заработной платы в зависимости от опыта, пола и стажа. Поскольку переменная Пол была исключена из модели, можно сказать, что прямой дискриминации женщин по уровню заработной платы в этой организации нет. В то же время переменная Опыт_Пол имеет вполне значимый отрицательный регрессионный коэффициент, равный –2,073. По существу это означает, что с каждым проработанным годом разница в годовой заработной плате между мужчинами и женщинами на этом предприятии возрастает на 2,073 тыс. долларов, что дает основание полагать наличие дискриминации по половому признаку на этом предприятии.

Запишем уравнение, определяющее уровень заработной платы в зависимости от факторных переменных

$$y = 59{,}057 + 0{,}781 \cdot \text{Опыт} - 2{,}073 \cdot \text{Пол\_Опыт} - 0{,}034 \cdot \text{Опыт2}\,.$$

Приведенное уравнение предсказывает любопытную закономерность. Пока опыт меньше 23 лет, годовая заработная плата растет с опытом, а затем начинает падать.

3.3. Логистическая регрессия

Очень часто на практике возникает задача классификации объектов по известным значениям признаков. Легко можно привести несколько примеров постановок задач такого рода. Например, может быть поставлена задача отбора налогоплательщиков, которые весьма вероятно уклоняются от выплаты всех налогов, или больных, которым по совокупности диагностических признаков показано оперативное вмешательство. В общем случае для решения задач отбора (классификации) используется дискриминантный анализ, но эту задачу можно решить, используя логистическую регрессию, в которой строится значение результативной дихотомической переменной в зависимости от значений факторных переменных.

Будем считать, что событие определяется дихотомической переменной (0 – не произошло событие, 1 – произошло). Для построения модели предсказания можно было бы построить линейное регрессионное уравнение с зависимой дихотомической переменной y, но оно не будет адекватно поставленной задаче, так как в классическом уравнении регрессии предполагается, что y – непрерывная переменная. Поэтому вводится логистическая регрессия.

С помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических переменных от независимых переменных, измеренных в любой шкале. Регрессия в таком случае рассчитывает вероятность p наступления события в зависимости от значений независимых переменных $x_{1i}, x_{2i}, \ldots, x_{ki}$:

$$p = \frac{\exp(z)}{1+\exp(z)},\ z = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k . \qquad (3.11)$$

Если для p получится значение, меньшее 0,5, то можно предположить, что событие не наступит; в противном случае предполагается наступление события.

В SPSS реализованы три возможности построения логистической регрессионной модели в зависимости от того, какие значения принимает результативная переменная. Если результативная переменная является дихотомической и принимает всего лишь два возможных значения (0, 1), то модель называется бинарной логистической. Если результативная переменная может принимать несколько различных значений (более двух), то SPSS позволяет построить мультиноминальную регрессионную модель. Наконец, если результативная переменная из-

мерена в порядковой шкале, то следует строить модель порядковой логистической регрессии. Основные характеристики переменных, участвующих в логистическом анализе, приведены в табл. 3.13.

Таблица 3.13

Основные характеристики переменных в моделях логистической регрессии

Модель регрессии	Зависимые переменные		Независимые переменные	
	Число	Тип	Число	Тип
Бинарная	Одна	Дихотомическая	Любое	Любой
Мультиноминальная	Одна	Номинальная	Любое	Номинальная, порядковая
Порядковая	Одна	Порядковая	Любое	Номинальная, порядковая

Чтобы понять выводимые в SPSS результаты построения логистической модели, рассмотрим основные идеи, используемые при построении логистической модели, на простейшем примере, когда имеется лишь один объясняющий фактор и в формуле (3.11) можно положить, что $z = b_0 + b_1 \cdot x$. Казалось бы, что логистическая модель может быть сведена к линейной при использовании преобразования

$$g(x) = \ln(\frac{p(x)}{1 - p(x)}) = b_0 + b_1 \cdot x . \qquad (3.12)$$

В действительности это, однако, не так. Дело в том, что МНК основан на предположении, что для линейной модели $y = b_0 + b_1 \cdot x + \varepsilon$, где случайная ошибка ε распределена по нормальному закону. В логистической модели $y = p(x) + \varepsilon$ и ошибка ε может принимать два возможных значения: если $y = 1$, то $\varepsilon = 1 - p(x)$ с вероятностью $p(x)$, определяемой формулой (3.11), и если $y = 0$, то $\varepsilon = -p(x)$ с вероятностью $1 - p(x)$. Таким образом, случайная ошибка в логистической модели имеет распределение, для которого среднее значение равно нулю, а дисперсия равна $p(x) \cdot (1 - p(x))$. Поэтому, для логистической регрессии нарушается одно из условий применимости метода МНК, а именно постоянство дисперсии ошибки для различных наблюдений (условие гомоскедастичности).

По этой причине для определения регрессионных коэффициентов для логистической регрессионной модели используется метод *мак-*

симального правдоподобия. Этот метод можно использовать и для определения регрессионных коэффициентов линейной модели, где он дает точно такие же результаты, как и метод МНК. Для лучшего понимания метода максимального правдоподобия рассмотрим вначале случай линейной регрессии.

Будем рассматривать величины y_i как независимые нормально распределенные случайные величины с математическим ожиданием $M(y_i) = \alpha + \beta \cdot x_i$ и постоянной дисперсией σ_y^2. Следовательно, плотность распределения случайной величины y_i

$$f(y_i) = \frac{1}{\sigma_y \sqrt{2\pi}} \cdot e^{\frac{-(y_i - \alpha - \beta \cdot x_i)^2}{2\sigma_y^2}}. \qquad (3.13)$$

Плотность вероятности появления случайных величин $y_1, y_2, \ldots, y_n$ (в выборке объема n) задается произведением функций (3.13)

$$f(y_1, y_2 \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sigma_y \sqrt{2\pi}} \cdot e^{\frac{-(y_i - \alpha - \beta \cdot x_i)^2}{2\sigma_y^2}} = \frac{1}{\left(\sigma_y \sqrt{2\pi}\right)^n} \cdot e^{\frac{-\sum_{i=1}^{n} (y_i - \alpha - \beta \cdot x_i)^2}{2\sigma_y^2}}.$$

Записанная выше функция в математической статистике называется *функцией максимального правдоподобия*. Функция правдоподобия полезна тем, что она позволяет получить оценку параметров α, β и σ_y^2. Согласно методу максимального правдоподобия, в качестве параметров α, β и σ_y^2 следует брать такие значения, которые максимизируют $f(y_1, y_2, \ldots, y_n)$. Очевидно, что при заданных значениях $x_1, x_2, \ldots, x_n$ объясняющей переменной максимум этой функции достигается тогда, когда минимально значение суммы, стоящей в показателе экспоненциальной функции. Отсюда получаем условие

$$\sum_{i=1}^{n} (y_i - \alpha - \beta \cdot x_i)^2 \to \min,$$

которое полностью эквивалентно условию (1.2) МНК.

Построим теперь функцию максимального правдоподобия для логистической регрессионной модели. Пусть значения результативной переменной кодируются как «0» или «1». Тогда выражение (3.11)

для некоторых значений b_0, b_1 дает вероятность того, что $y_i = 1$ при заданном x_i (для простоты, как это уже упоминалось, рассматриваем случай одной объясняющей переменной). Тогда величина $1 - p(x_i)$ определяет вероятность, что $y_i = 0$ при заданном x_i.

Таким образом, для тех пар x_i, y_i, в которых $y_i = 1$, вклад в функцию максимального правдоподобия будет $p(x_i)$, а для пар,в которых $y_i = 0$, этот вклад будет $1 - p(x_i)$. Очевидно, что для произвольной пары x_i, y_i вклад в функцию максимального правдоподобия можно записать в виде

$$f(y_i) = p(x_i)^{y_i} \cdot [1 - p(x_i)]^{1-y_i}. \qquad (3.14)$$

Таким образом, если имеется набор эмпирических точек x_i, y_i, $i = 1, 2, \ldots, n$, и все наблюдения независимы, то функция максимального правдоподобия будет

$$L(b_0, b_1) = \prod_{i=1}^{n} f(y_i). \qquad (3.15)$$

Неизвестные параметры логистической регрессии b_0, b_1 могут быть найдены из условия максимальности функции (3.15). Выкладки значительно упрощаются, если предварительно вычислить натуральный логарифм функции правдоподобия

$$\ln L(b_0, b_1) = \sum_{i=1}^{n} \{y_i \cdot \ln p(x_i) + (1 - y_i) \cdot \ln[1 - p(x)]\}. \qquad (3.16)$$

Регрессионные параметры найдем из решения системы уравнений

$$\begin{aligned} \frac{d \ln L(b_0, b_1)}{db_0} &= \sum_{i=1}^{n} (y_i - p(x_i)) = 0; \\ \frac{d \ln L(b_0, b_1)}{db_1} &= \sum_{i=1}^{n} x_i \cdot (y_i - p(x_i)) = 0. \end{aligned} \qquad (3.17)$$

В общем случае многофакторной логистической модели все сделанные выше выкладки сохраняют свою силу, и в результате мы получим для определения регрессионных коэффициентов систему уравнений, аналогичную системе (3.17).

Для оценки качества регрессионной модели также широко используется функция правдоподобия. Обычно используется прямой метод включения переменных в модель, при котором на первом шаге вклю-

чается в модель только константа b_0, на втором шаге включается первая объясняющая переменная и т. д. При этом анализируется статистика G (в SPSS она имеет название –2Log Правдоподобие), которая вычисляется по формуле

$$G = -2 \cdot \ln\left(\frac{\text{Правдоподобие без переменной}}{\text{Правдоподобие с включенной новой переменной}}\right). \qquad (3.18)$$

В числителе стоит функция правдоподобия предыдущего шага, а в знаменателе функция правдоподобия следующего шага, когда в модель включена новая переменная. Если включенная переменная действительно улучшает модель, то величина G будет положительной. Значимость включения переменной в модель можно оценить, используя тот факт, что величина G подчиняется распределению хи-квадрат с числом степеней свободы, равным 1.

Общее качество модели определяется с помощью факторов R^2_{CS} Кокса и Снелла

$$R^2_{CS} = 1 - \left(\frac{L(b_0)}{L(b_0, b_1, \ldots, b_k)}\right)^{2/n} \qquad (3.19)$$

или фактора R^2_N Нейджелкерка

$$R^2_N = \frac{R^2_{CS}}{1 - \{L(b_0)\}^{2/n}}. \qquad (3.20)$$

В формулах (3.19), (3.20) n – это число наблюдений. На практике считается, что R^2_N Нейджелкерка дает более достоверную информацию о значимости модели.

Значимость регрессионных коэффициентов модели в SPSS оценивается с помощью критерия Вальда W:

$$W = \frac{b_i}{\Delta b_i}, \quad i = 0, 1, \ldots, k, \qquad (3.21)$$

где Δb_i – стандартная ошибка для i-го регрессионного коэффициента. Статистика Вальда подчиняется стандартному нормальному распределению, и поэтому значимость регрессионных коэффициентов оценивается так же, как и в случае линейной регрессии.

Другие детали построения регрессионных логистических моделей в SPSS мы обсудим при рассмотрении конкретных примеров.

Пример 3.6

В файле Пример_3_6.sav собраны кредитные истории 700 клиентов банка и данные еще о 150 клиентах, которые намерены обратиться в банк за кредитом. Информация о клиентах банка содержит такие данные:

- возраст (числовая переменная);
- образование (номинативная переменная);
- стаж коммерческой деятельности;
- число полных лет постоянного места проживания;
- годовой доход (в тыс. долларов США);
- долги (в процентах к годовому доходу);
- долги по кредитной карте (в тыс. долларов США);
- другие долги (в тыс. долларов США);
- для клиентов, уже бравших кредит, данные о том, являлся ли он должником банка.

Используя 70 % -ю случайную выборку клиентов, уже бравших кредит в банке, создать регрессионную логистическую модель, позволяющую предсказать невозвращение кредита клиентом. Используя оставшиеся 30 % клиентов (проверочная совокупность), выяснить степень пригодности построенной модели для предсказания случаев невозвращения кредита клиентами банка. Определить, какие переменные могут быть без ущерба исключены из модели.

Решение

После загрузки данных в SPSS отберем случайным образом 70 % объектов для создания регрессионной модели. Для этого, используя опции меню Transform/Compute (Преобразование/Вычислить), создадим новую переменную Выборка, присвоив ей значения случайной величины, распределенной по закону Бернулли. Для этих целей воспользуемся функцией RV.Bernoulli (0.7) из набора стандартных функций SPSS, предназначенных для генерации случайных величин. По закону Бернулли распределена случайная величина, которая может принимать значения либо 0, либо 1 с некоторой заданной вероятностью. Например, вероятность выпадения «орла» либо «решки» подчиняется закону Бернулли и может быть смоделирована в SPSS функцией RV.Bernoulli (0.5). Поскольку нам нужно выбрать только те объекты, для которых уже известно, своевременно возвратил клиент кредит или нет, то переменной Выборка мы будем присваивать значения только в том случае, если переменной Должник присвоено какое-либо значение. Для этого нужно нажать кнопку с надписью if (см. рис. 3.11) и в открывшемся окне набрать условие Должник >= 0 .

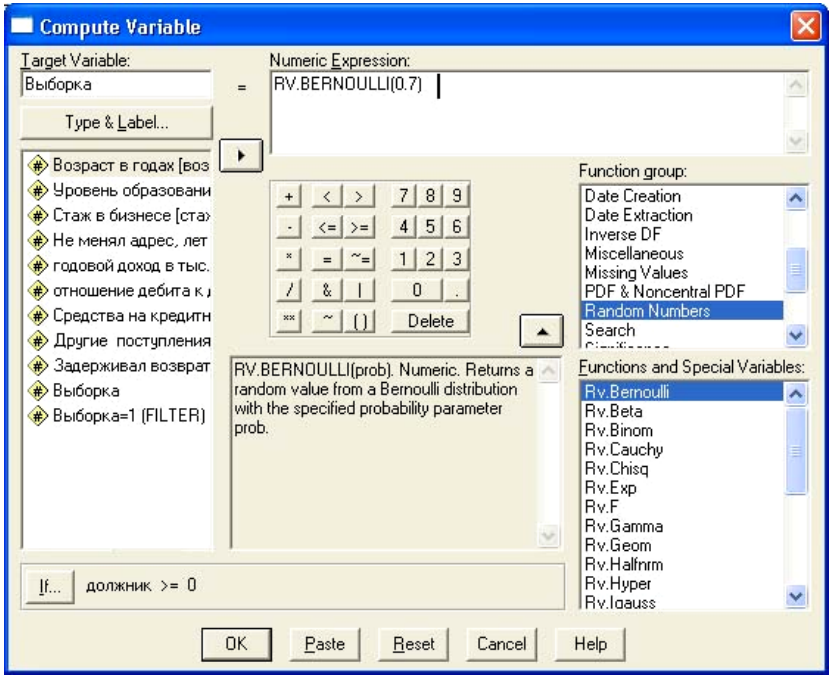


Рис. 3.11. Создание переменной, позволяющей произвести случайную выборку 70 % случаев

После того как новая переменная создана, необходимо отобрать для создания регрессионной модели только случаи, для которых переменная Выборка = 1. Для этих целей выбираем пункты меню Data /Select Cases (Данные /Отбор случаев) и в появившемся окне добавляем условие Выборка =1. В результате этих действий программа создаст автоматически еще одну переменную с именем filter_$, значения которой она и будет использовать для отбора данных.

После этих подготовительных шагов можно приступать к созданию регрессионной модели. Для открытия окна установки параметров логистической регрессионной модели выберем пункты меню Analyze /Regression / Binary Logistic. В появившемся новом окне переменную Должник (метка Задерживал возврат) следует поместить в окно Dependent (Зависимая переменная), остальные факторные переменные – в окно Covariates. В данном контексте термин Covariates следует перевести как «влияющие или факторные переменные». Отметим, что при построении логистической регрессионной модели мы можем отобрать случаи, участвующие в создании моде-

ли, и в окне выбора параметров модели (рис. 3.12), а не с помощью пункта меню Data /Select Cases, как указывалось выше.

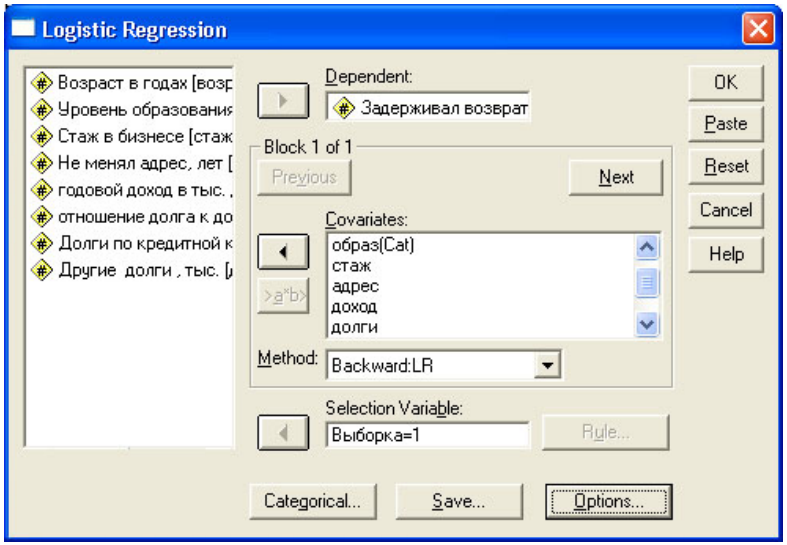


Рис. 3.12. Окно установки параметров логистической регрессии

Для этого переменную Выборка нужно поместить в окно Selection Variable (Переменная отбора) и, нажав клавишу Rule (Правило), задать правила отбора переменных (Выборка = 1).

Выберем в окне Method способ включения переменных в модель. При выборе метода Forward LR переменные будут включаться в модель по очереди. При этом необязательно все переменные будут включены в модель. Если значимость коэффициента новой переменной будет ниже определенного уровня, то переменная не будет включена в модель.

Если среди предикторов есть переменные, измеренные в номинативной шкале с более чем двумя градациями, то следует нажать кнопку Categorical. В результате откроется окно, в котором все номинальные переменные следует переместить в окошко с названием Categorical variables (Категориальные переменные). Это необходимо сделать потому, что если категориальная переменная имеет m градаций, то SPSS для анализа значений этой переменной создает $m-1$ дихотомическую переменную. В рассматриваемом примере переменная Образ (Образование) является номинальной и имеет пять уровней градации.

Для сохранения предсказанных значений нужно нажать кнопку Save (Сохранить) и в открывшемся окне поставить галочку в поле Probabilites (Вероятности). Все остальные установки можно оставить по умолчанию. После нажатия клавиши OK программа откроет окно выдачи результатов логистической регрессии. Это окно содержит избыточную информацию, большую часть которой на первых порах можно опустить.

Рассмотрим таблицу, которая в окне SPSS называется «Объединенные тесты для коэффициентов модели» (табл. 3.14). Здесь содержится отчет о том, как изменялись характеристики модели на каждом шаге включения очередной переменной в модель. Для пользователя представляет интерес, как правило, последний шаг, на котором модель была окончательно сформирована (в нашем случае шаг 4).

Таблица 3.14

Объединенные тесты для коэффициентов модели (шаг 4)

	Хи-квадрат	Ст.св.	Знач.
Шаг	10,368	1	0,001
Блок	172,968	4	0,000
Модель	172,968	4	0,000

Значение хи-квадрат в строке «Шаг» содержит значение функции G (3.18). Фактически эта величина определяет изменение функции правдоподобия при включении дополнительной переменной на шаге 4. Величина хи-квадрат в строке «Блок» определяет изменение логарифма функции правдоподобия в текущей модели по сравнению с предыдущим блоком, когда в модель была включена только константа. Наконец, в строке «Модель» значение хи-квадрат равно изменению логарифма функции правдоподобия для конечной модели по сравнению с начальной моделью. Высокий уровень значимости для включения переменных в модель говорит о том, что включенные в модель переменные действительно являются объясняющими.

Как указывалось выше, качество регрессионной модели можно оценить с помощью критерия R_N^2 Нейджелкерка. Как следует из отчета по модели (здесь мы не приводим эту таблицу), $R_N^2 = 0{,}426$, что говорит о том, что качество модели оставляет желать лучшего, но вполне достаточно для использования модели с целью прогноза перспектив невозвращения долга кредитором. В табл. 3.15 приведены отчет о параметрах, включенных в модель на последнем шаге, и их значимость.

Таблица 3.15

Отчет о переменных, включенных в модель на шаге 4

Переменная	Значение коэффициента	Стд. ошибка	Вальд	Ст.св.	Знч.
Стаж	–0,227	0,032	50,493	1	0,000
Адрес	–0,069	0,022	9,584	1	0,002
Долги	0,105	0,023	20,240	1	0,000
Кредит_карта	0,509	0,101	25,295	1	0,000
Константа	–0,962	0,299	10,318	1	0,001

Как следует из приведенной таблицы, не вошли в итоговую модель переменные, характеризующие возраст, образование, годовой доход и другие поступления денежных средств клиента. Если включить все переменные в модель, то это не улучшит ее качество, а лишь приведет к статистической незначимости регрессионных коэффициентов.

Данные о проценте правильных предсказаний для выборки, взятой для создания модели, и проверочной выборки приведены в табл. 3.16.

Таблица 3.16

Процент правильных предсказаний

Наблюденное		Предсказанное на основе модели					
		Отобранные наблюдения			Неотобранные наблюдения		
		Задерживал возврат кредита ранее		Процент корректных	Задерживал возврат кредита ранее		Процент корректных
		Нет	Да		Нет	Да	
Задерживал возврат кредита ранее	Нет	331,0	29,0	91,9	141,0	16,0	89,8
	Да	71,0	65,0	47,8	19,0	28,0	59,6
Общий процент				79,8			82,8

Из приведенных данных следует, что процент правильных предсказаний клиентов, возвращающих кредит, достаточно высок и составляет около 90 % для проверочной выборки. Значительно хуже обстоит дело с предсказанием невозвращений кредита. Здесь процент правильных предсказаний составляет 59,6 %. Поэтому при практическом использовании модели следует ориентироваться на предсказании возвращений и не анализировать вероятность невозвращений.

3.4. Анализ рядов динамики в SPSS

Общие сведения о временных рядах

При анализе данных в сфере управления очень важной является задача прогнозирования значений временных рядов. Под *временным рядом* (*рядом динамики*) понимается совокупность наблюдений некоторого признака (случайной величины) в последовательные моменты времени.

На первый взгляд может показаться, что задача построения регрессионной модели для временного ряда ничем не отличается от рассмотренных выше задач построения регрессионных моделей по результатам выборочного наблюдения. В действительности это не так. Временные ряды описывают один и тот же объект, взятый в разные моменты времени, и поэтому рядом стоящие значения уровней ряда очень часто бывают коррелированными (автокорреляция). Для этого есть веские экономические основания: выпуск продукции, например, в следующем году, скорее всего будет зависеть от уровня объема производства текущего года.

Таким образом, если при анализе выборочного наблюдения предполагалось, что $y_1, y_2, ..., y_n$ можно рассматривать как независимые случайные величины, то значения временного ряда y_t $(t = 1, 2, \ldots, n)$ не являются статистически независимыми. Для многих явлений их современное состояние функционально определяется предшествующими состояниями системы, в большей степени недавними, в гораздо меньшей – далеко отстоящими от заданного по временному ряду. Например, потребление электроэнергии в городе подвержено сезонным колебаниям, и поэтому будут сильно коррелировать данные, относящиеся к одному и тому же месяцу для разных лет. При построении моделей рядов динамики следует учитывать эту специфику.

Пусть исследуется показатель y. Его значение в текущий момент (период) обозначим y_t, а в предшествующие моменты времени – y_{t-1}, y_{t-2} и т. д. Значение объясняющей переменной в различные моменты времени обозначим аналогично: x_t, $x_{t-1}, \ldots, x_{t-k}$.

При построении временных рядов динамики используются не только текущие значения переменных, но и их значения в предыдущие моменты времени. Естественно, что объясняющей переменной может служить и время t.

Переменные, влияние которых характеризуется некоторым запаздыванием, называются *лаговыми переменными*.

Обычно модели динамики делят на три больших класса.

1. Модели с лаговыми переменными. Примером является модель

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_1 \cdot x_{t-1} + \ldots + \beta_k \cdot x_{t-k} + \varepsilon_t . \qquad (3.22)$$

2. Авторегрессионые модели – это модели, уравнения которых в качестве лаговых переменных могут включать и значение зависимых переменных. Примером является модель

$$y_t = \alpha + \beta \cdot x_t + \gamma \cdot y_{t-1} + \varepsilon_t . \qquad (3.23)$$

3. Модели, содержащие циклические колебания различной природы. Достаточно часто встречаются модели, содержащие сезонные колебания. Например, продажа мороженого в розничной торговле будет ежегодно демонстрировать летние максимумы и зимние спады.

Каждый уровень (значение) временного ряда формируется под действием большого числа факторов, которые можно разделить на четыре группы:

- факторы, формирующие тенденции ряда (тренд);
- факторы, формирующие сезонные колебания, отражающие повторяемость экономических процессов в течение не очень длительного периода;
- факторы, отражающие повторяемость экономических процессов в течение длительных периодов;
- случайные факторы.

Естественно предположить, что все четыре компоненты (трендовая, сезонная, циклическая и случайная) будут формировать наблюдаемое значение случайной величины y.

Поэтому в общем случае временной ряд можно представить либо в виде аддитивной модели

$$y_t = u_t + \nu_t + c_t + \varepsilon_t \ , \qquad (3.24)$$

либо в виде мультипликативной модели

$$y_t = u_t \cdot \nu_t \cdot c_t \cdot \varepsilon_t . \qquad (3.25)$$

В этих формулах u_t – трендовая, ν_t – сезонная, c_t – циклическая и ε_t – случайная составляющие временного ряда. Важно подчеркнуть, что в отличие от ε_t величины u_t, ν_t, c_t случайными не являются,

формируя детерминированную (закономерную) составляющую временного ряда.

Сглаживание временных рядов

Как уже указывалось, одной из главных задач анализа временных рядов является выделение трендовой составляющей, позволяющей выявить устойчивую тенденцию развития уровней временного ряда. Если для выявления тренда строить регрессионную модель непосредственно, используя первичные данные, то построенная модель окажется скорее всего несостоятельной, поскольку, например, сезонная составляющая будет рассматриваться как ошибка, и поэтому фактор детерминации R^2 модели окажется слишком малым. Простейшим способом устранения этой проблемы является процедура сглаживания уровней временного ряда. Существует несколько способов сглаживания, простейшим из которых является метод скользящей средней. Например, если у нас имеются данные квартального потребления электроэнергии, демонстрирующие периодичность с циклом в один год (четыре квартала), то разумно для сглаживания временного ряда использовать четырехзвенную скользящую среднюю

$$y_{i-2}^{ср} = \frac{1}{4}(y_{i-3} + y_{i-2} + y_{i-1} + y_i),\ i = 4, 5, \ldots, n\,. \qquad (3.26)$$

При таком сглаживании несколько первых значений уровня ряда теряются.

В SPSS используется более совершенный метод экспоненциального сглаживания временных рядов. Алгоритм этого метода зависит от модели временного ряда, и поэтому мы будем знакомиться с ними, используя простые примеры.

Пример 3.7

Имеются следующие данные об общем объеме розничного товарооборота вновь открывшегося магазина по месяцам в 2006 г., млн руб. (файл Пример_3_7.xls).

t	1	2	3	4	5	6	7	8	9	10	11	12
Y	22,8	24,9	31,0	29,5	30,5	35,6	36,4	42,6	45,1	47,3	51,0	53,4

Требуется произвести экспоненциальное сглаживание уровней временного ряда.

Решение

При анализе временных рядов очень важно иметь правильное представление о структуре временного ряда. Поэтому анализ следует начинать с построения графика исходных данных. Для этого выбираем пункты меню Graph/Sequence и в открывшемся окне с названием Sequence Charts (График последовательности) перемещаем переменную Товарооборот (Y) в окно Variables (Переменные), а переменную Время в окно Time Axis Labels (Значения по временной оси). Все остальные установки в этом окне можно оставить по умолчанию. После нажатия кнопки OK в окне Sequence Charts будет построена диаграмма, приведенная на рис. 3.13.

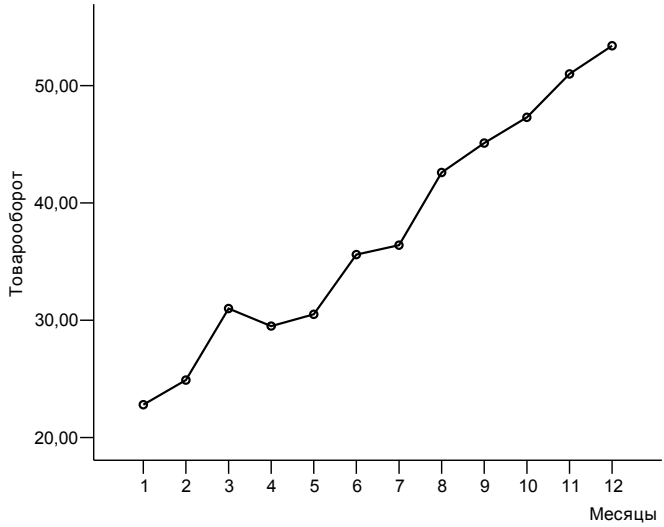


Рис. 3.13. Товарооборот магазина по месяцам

Из приведенного графика следует, что имеется устойчивая тенденция развития уровней ряда (тренд). Поэтому при выборе алгоритма экспоненциального сглаживания следует выбрать тот, который предполагает наличие тренда. Чтобы познакомиться со всеми возможностями, которые имеются в SPSS для сглаживания временных рядов, дополним исходные данные временными переменными (если этого не сделать, кнопки, запускающие некоторые из алгоритмов сглаживания, будут неактивными). Для этого в меню выберем пункты Data/Define Dates (Данные/Определить даты) и в открывшемся окне установим значения параметров, которые выбраны на рис. 3.14.

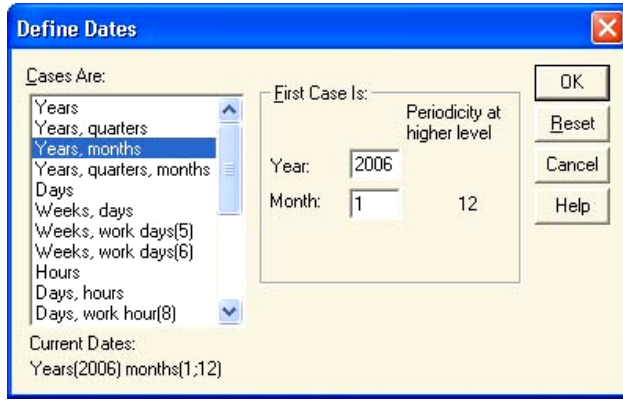


Рис. 3.14 Окно определения дат в SPSS

После нажатия кнопки OK в окне редактора SPSS появятся новые переменные: YEAR_, MONTH_, DATE_.

Изучим возможные модели временного сглаживания, которые предлагаются в SPSS. Выберем пункты меню Analyze/Time Series/Exponential Smoothing (Временные ряды/Экспоненциальное сглаживание). В результате откроется окно, изображенное на рис. 3.15.

Рис. 3.15. Окно выбора модели экспоненциального сглаживания

Модель Simple

В этой модели предполагается, что нет ни трендовой, ни сезонной составляющих, и временной ряд имеет следующую структуру:

$$y_i = b + \varepsilon_i , \qquad (3.27)$$

где b – некоторая константа, ε_i – случайная ошибка. Обозначим $y_1, y_2, \ldots, y_n$ – уровни исходного временного ряда, $\bar{y}$ – среднее значение этого ряда, $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_n$ – значения уровней ряда, получившиеся после сглаживания, $\varepsilon_i = y_i - \widetilde{y}_{i-1}$ – остаток (разность уровней исходного и сглаженного ряда в предыдущий временной период). Алгоритм сглаживания при отсутствии трендовой и сезонной составляющих имеет вид

$$\widetilde{y}_1 = \bar{y}, \quad \varepsilon_i = y_i - \widetilde{y}_{i-1}, \quad \widetilde{y}_i = \widetilde{y}_{i-1} + \alpha \cdot \varepsilon_i, \quad i = 2, 3, \ldots, n. \qquad (3.28)$$

Алгоритм (3.28) позволяет получить последовательно все уровни сглаженного временного ряда. Действительно, выполняя простые вычисления, получаем $\widetilde{y}_1 = \bar{y}$, $\widetilde{y}_2 = \widetilde{y}_1 + \alpha \cdot (y_2 - \widetilde{y}_1)$, $\widetilde{y}_3 = \widetilde{y}_2 + \alpha \cdot (y_3 - \widetilde{y}_2)$ и т. д.

В формуле (3.28) значение коэффициента α изменяется в пределах $0 \le \alpha \le 1$ и может быть выбрано пользователем произвольно. Существует, однако, возможность предоставить SPSS выбрать наиболее подходящее значение этого коэффициента.

Модель Holt

В этой модели предполагается наличие трендовой и случайной составляющих, а структура временного ряда имеет форму

$$y_t = b_0 + b_1 \cdot t + \varepsilon_t. \qquad (3.29)$$

Уровни сглаженного ряда определяются последовательно, начиная с первого, в соответствии с алгоритмом

$$T_1 = \frac{y_n - y_1}{n-1}, \; S_1 = y_1 - 0{,}5 \cdot T_1;$$
$$T_i = T_{i-1} + \alpha \cdot \gamma \cdot \varepsilon_i; \quad S_i = S_{i-1} + T_{i-1} + \alpha \cdot \varepsilon_i;$$
$$\widetilde{y}_i = S_i + T_i; \quad \varepsilon_i = y_i - \widetilde{y}_{i-1}, \quad i = 2, 3, \ldots, n. \qquad (3.30)$$

В формуле (3.30) коэффициент γ также изменяется в пределах от 0 до 1 и может выбираться либо пользователем, либо наиболее оптимальное значение этого параметра может подобрать SPSS, исходя из принципа минимизации суммы квадратов отклонений между истинными и сглаженными уровнями ряда. Значения трендовой составляющей T_i представляют собой тангенс угла наклона участка линии тренда к оси абсцисс.

Рассматриваемые в примере 3.7 данные как раз следует сглаживать с помощью модели Holt. Поэтому оставим на время обсуждение других алгоритмов сглаживания и завершим рассмотрение примера.

Поместим переменную Товарооборот (см. рис. 3.15) в окно Variables (Переменные), переключатель выбора модели установим в положение Holt и откроем окно установки параметров модели, щелкнув мышкой на

кнопке Parameters (Параметры). В результате откроется окно, изображенное на рис.3.16, в котором пользователю предоставляется возможность выбрать значения параметров α и γ [см. формулы (3.30)] и установить начальные значения параметров S_1 и T_1, которые должны быть предварительно вычислены по формулам (3.30).

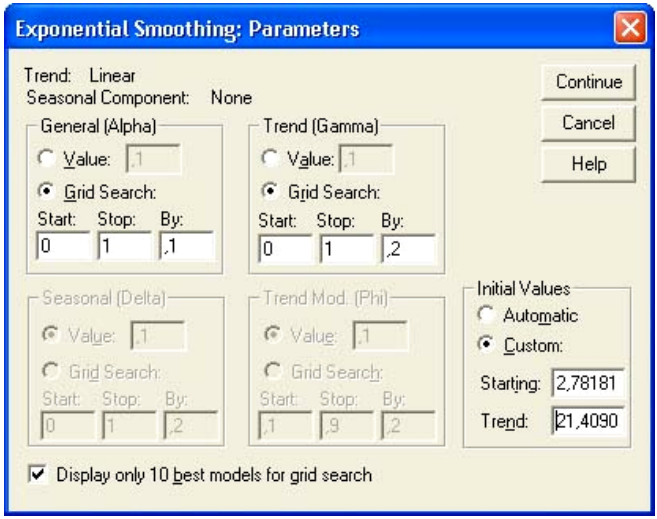


Рис. 3.16. Окно установки параметров модели Holt

Если установить переключатели в положение Grid Search (Поиск на сетке), то будет осуществлен перебор всех возможных значений параметров α и γ в интервале от 0 до 1 с шагом, который можно выбрать в окошке с названием By (в данном контексте – посредством выбора шага). В итоге будут отображены 10 лучших результатов, дающих наименьшую сумму квадратов отклонений истинных и сглаженных уровней ряда.

Значения переключателя в окне Initial Values (Начальные значения) можно оставить и в положении Automatic (Автоматически). При этом, следуя формулам (3.30), будут взяты в качестве начальных данных как раз те значения, которые изображены на рис. 3.16.

Для продолжения вычислений достаточно нажать кнопку Continue (Продолжить). Запуск процедуры сглаживания можно запустить, нажав кнопку OK (см. рис. 3.15). В результате появится окно выдачи результатов, в котором для нас представляют интерес только подобранные SPSS значения параметров $\alpha = 0{,}4$ и $\gamma = 0{,}0$. После завершения вычислений в

окне редактора отображаются уровни сглаженного ряда (переменная FIT_1) и разность исходных и сглаженных значений уровня ряда.

Представляет интерес построить уровни исходного и сглаженного ряда. Это проще всего сделать, если на панели инструментов нажать кнопку Dialog Recall (Возврат к диалогу) и в открывшемся окне выбрать диалог Sequence Charts (График последовательности). В открывшемся окне переменную FIT_1 следует переместить в окно Variables и нажать OK. В результате будет построена диаграмма, приведенная на рисунке 3.17.

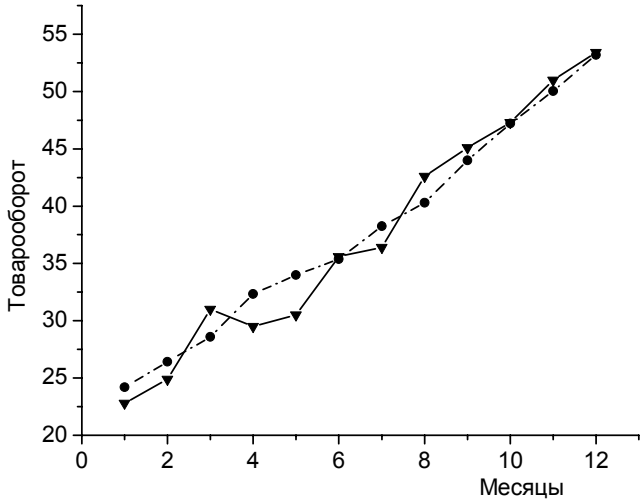


Рис. 3.17. Исходный и сглаженный временные ряды товарооборота:
треугольники – исходные, кружки – сглаженные данные

Для прогноза значений временного ряда нужно найти регрессионные коэффициенты. Для этого нужно воспользоваться стандартными методами построения линейного регрессионного уравнения. Важно отметить, что после проведения процедуры экспоненциального сглаживания качество регрессионного уравнения получается заметно выше. Собственно, в этом и состоит основной смысл экспоненциального сглаживания данных.

Наиболее полно возможности SPSS по сглаживанию данных представлены на закладке Custom (по выбору пользователя). Это окно представлено на рис. 3.18.

Рис.3.18. Окно выбора вариантов модели
экспоненциального сглаживания

Для того чтобы были доступны модели сглаживания с использованием сезонных компонент, необходимо определить переменную, содержащую даты, с помощью пунктов меню Data/Define Dates (Данные/Определить даты), как это уже указывалось выше.

Выделение трендовой и сезонной составляющих. Предсказание уровней ряда

Процедура сглаживания еще не решает задачи прогнозирования уровней ряда в ближайшей перспективе. Для этого нужна декомпозиция ряда динамики на трендовую, сезонную и случайную составляющие в соответствии с аддитивной (3.24) и мультипликативной (3.25) моделями временного ряда. Обсудим проблемы декомпозиции временного ряда на простом примере.

Пример 3.8

Рассмотрим данные потребления электроэнергии (млрд кВт· ч) жителями региона за 16 кварталов 2000 – 2003 гг. Статистические данные приведены в таблице (см. также файл Пример_3_8.xls).

Квартал	Объем потребления электроэнергии, млрд кВт · ч	Квартал	Объем потребления электроэнергии, млрд кВт · ч
1	6	9	8
2	4,4	10	5,6
3	5	11	6,4
4	9	12	11
5	7,2	13	9
6	4,8	14	6,6
7	6	15	7
8	10	16	10,8

Требуется выделить сезонную трендовую и случайную составляющие и предсказать потребление электроэнергии в 18 -м квартале.

Решение

После загрузки данных в SPSS полезно получить общее представление о поведении временного ряда. Для этого нужно полностью повторить процедуру построения графика последовательности, описанную в начале примера 3.7. Результат такого построения приведен на рис. 3.19.

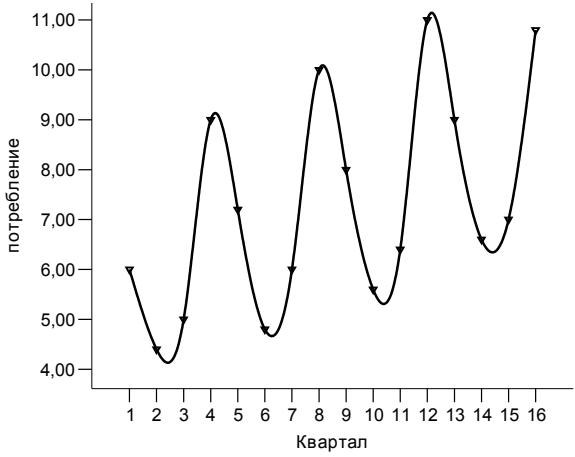


Рис. 3.19. Временной ряд, демонстрирующий сезонную периодичность и устойчивую тенденцию роста

С помощью пунктов меню Data/Define Dates (Данные/Определить даты) определим новые временные переменные в полной аналогии с тем, как это было описано в примере 3.7. Поскольку в рассматриваемом случае мы имеем дело с квартальными данными, то в окне определения формата даты (см. рис. 3.14) следует выбрать закладку Years, quarters (Года, кварталы). В результате в редакторе данных появятся три новых переменных: YEAR_ QUARTER_, DATE_ (Год, Квартал, Дата).

Для разделения уровней временного ряда на сезонную, трендовую и случайную составляющие выберем пункты меню Analyze/Time Series /Seasonal Decomposition (Анализ/Временные ряды/Сезонная декомпозиция). В результате откроется окно, в котором переменную Потребление нужно переместить в окошко Variables (Переменные) и переключатель на панели Model перевести в положение Additive (Аддитивная). Все остальные параметры можно оставить неизменными. После нажатия кнопки

OK появится окно выдачи результатов, которое на сей раз не представляет большого интереса, поскольку все необходимые результаты будут отображены в окне редактора в виде четырех вновь сгенерированных переменных (SAF_1, SAS_1, STC_1, ERR_1).

В первую очередь следует обратить внимание на сгенерированную SPSS переменную с именем SAF_1 (SAF – это аббревиатура термина *Seasonal adjustment factors* – сезонные подгоночные факторы). Как следует из названия, переменная содержит сезонные факторы. В анализируемом примере сезонные факторы составляют

Квартал	Сезонный фактор
1	0,58333
2	–1,98333
3	–1,30000
4	2,70000

Переменная SAS_1 (*Seasonally adjusted series*) представляет собой исходный ряд с устраненной сезонной составляющей. Значения этой переменной получаются простым вычитанием сезонной составляющей из уровней исходного ряда. Эти значения можно использовать для построения уравнения тренда.

Переменная STC_1 (*Smoothed trend-cycle component*) представляет собой сглаженный ряд значений переменной SAS_1. Дополнительное сглаживание необходимо, если кроме сезонной составляющей имеется еще и циклическая составляющая. Сглаживание производится по специальной методике, описание которой можно найти в файле Season.pdf, (папка «Алгоритмы SPSS» на электронном диске, прилагаемом к учебному пособию). Если кроме сезонной есть еще и циклическая переменная, то для отыскания уравнения тренда лучше использовать переменную STC_1. Даже если и нет сезонной составляющей, сглаженный ряд дает более качественную регрессионную модель.

Наконец, переменная ERR_1 представляет собой случайную ошибку, которая определяется просто как разность переменных SAS_1 и STC_1.

Для определения уравнения тренда запустим процедуру линейного регрессионного анализа и выберем в качестве зависимой переменной сглаженный ряд STC_1, а в качестве факторной переменной – переменную Квартал. Полученное регрессионное уравнение

$$y_i^T = 5{,}752 + 0{,}181 \cdot x_i \qquad (3.31)$$

имеет достаточно высокий фактор детерминации $R^2 = 0{,}962$ и значимые при уровне значимости $\alpha = 0{,}01$ регрессионные коэффициенты. Это

позволяет использовать уравнение для прогнозирования уровня потребления электроэнергии в 18-м квартале.

Прогнозное потребление электроэнергии в 18-м квартале (2-й квартал от начала года) складывается из трендовой составляющей для 18-го квартала и сезонной составляющей для 2-го квартала

$$y_{18}^{T} = 5{,}752 + 0{,}181 \cdot 18 - 1{,}983 = 7{,}027 .$$

Аналогично можно найти прогнозные значения для 19-го и 20-го кварталов.

Авторегрессионные модели временного ряда. Устранение автокорреляции остатков

В SPSS предлагается несколько методик построения однофакторных и многофакторных регрессионных моделей при наличии автокорреляции, включая наиболее универсальный метод ARIMA (AutoRegressive Integrated Moving Average) – метод авторегрессионного интегрированного скользящего среднего, предложенный Боксом и Дженкинсом в 1994 г. Это довольно сложный комплексный метод, состоящий из нескольких различных процедур. Мы не будем изучать в полном объеме все методы анализа временных рядов, имеющиеся в SPSS, и рассмотрим всего лишь два простых, но достаточно поучительных примера.

Как уже указывалось, в рядах динамики последовательные уровни ряда очень часто оказываются взаимосвязанными. Это проявляется, в частности, в том, что взаимосвязанными оказываются и ошибки.

Рассмотрим простейшую модель

$$y_t = b_0 + b_1 \cdot x_t + \varepsilon_t , \qquad (3.32)$$

описывающую взаимосвязь факторной переменной x и результативной переменной y в момент времени t. Очевидно, что в момент времени $t-1$ эта взаимосвязь будет иметь вид

$$y_{t-1} = b_0 + b_1 \cdot x_{t-1} + \varepsilon_{t-1} . \qquad (3.33)$$

Если имеется взаимосвязь между уровнями ряда y_t и y_{t-1}, то это значит, что и ε_t и ε_{t-1} являются взаимосвязанными. Это нарушает одно из основных положений применимости метода МНК (ошибки ε должны быть независимыми случайными величинами), и поэтому пользоваться стандартными методами регрессионного анализа некорректно. Можно попытаться уменьшить степень взаимосвязи уровней ряда с помощью простого преобразования. Если случайные ошибки подвержены автокорреляции первого порядка, то для них справедлива линейная взаимосвязь

$$\varepsilon_t = \rho \cdot \varepsilon_{t-1} + v_t , \qquad (3.34)$$

где v_t – случайные отклонения, а величина ρ является некоторым, пока не известным коэффициентом.

Вычтем из уравнения (3.32) уравнение (3.33), умноженное на ρ:

$$y_t - \rho \cdot y_{t-1} = b_0 \cdot (1-\rho) + b_1 \cdot (x_t - \rho \cdot x_{t-1}) + (\varepsilon_t - \rho \cdot \varepsilon_{t-1}). \qquad (3.35)$$

Если определить новые переменные $y_t^* = y_t - \rho \cdot y_{t-1}$ и $x_t^* = x_t - \rho \cdot x_{t-1}$, $b_0^* = b_0 \cdot (1-\rho)$, то мы с учетом выражения (3.34) получим стандартную регрессионную модель, для нахождения коэффициентов которой применим метод МНК:

$$y_t^* = b_0^* + b_1 \cdot x_t^* + v_t. \qquad (3,36)$$

Преобразования, приводящие нас от модели (3.32) к модели (3.36), называются авторегрессионной схемой первого порядка. Это преобразование позволяет устранить автокорреляцию в остатках, но коэффициент ρ в формуле (3.35) остался пока неопределенным, и следует найти методы его оценки. Простой способ оценки коэффициента основан на статистике Дарбина – Уотсона $\rho \approx 1 - DW/2$, где DW – статистика Дарбина – Уотсона. В SPSS для определения коэффициента ρ используется итерационный процесс Кохрана – Оркатта (Kochrane – Orcutt). Суть его в том, что строится некоторый итерационный процесс определения ρ, который прекращается, когда разница между двумя последовательными значениями окажется меньше некоторой заданной величины.

Метод Прайса – Винстена (Prais – Winsten) является развитием метода Кохрана – Оркатта и обычно дает более приемлемые результаты. Мы не будем вдаваться в анализ деталей этих алгоритмов, отсылая интересующихся читателей к документации, в которой содержится описание алгоритмов SPSS (см. файл AREG.pdf в папке «Алгоритмы SPSS» на электронном диске, прилагаемом к учебному пособию). Наконец, в том случае, если в данных много пропущенных значений и описанные выше методы не дают достоверных результатов для регрессионных коэффициентов, можно воспользоваться методом максимального правдоподобия (*Exact Maximum Likelihood*), который является более затратным по вычислительным ресурсам и фактически использует вычислительные алгоритмы ARIMA, которые кратко мы обсудим ниже.

Хотя выше мы обсуждали лишь однофакторную модель, авторегрессионное преобразование может быть обобщено на случай произвольного числа переменных.

Пример 3.9

Менеджер магазина, торгующего бытовой радиоэлектроникой, желает выяснить, от каких факторов рекламной кампании зависит недельный объем продаж. В файле Пример_3_9.sav содержатся данные о недельных продажах (тыс. руб.), числе рекламных роликов, которые появляются за неделю на радио, числе посетителей рекламного сайта магазина за неделю и числе рекламных досок, на которых еженедельно размещается информация о товарах магазина. На основании этих данных требуется выяснить, какой вид рекламы положительно сказывается на объеме продаж. Каким будет относительный прирост объема продаж, если относительный прирост числа рекламных роликов на радио составит 50 %?

Решение

Попробуем вначале просто построить линейную регрессионную модель зависимости объемов продаж от трех объясняющих параметров – числа листовок, размещенных на рекламных щитах; числа обращений на сайт в неделю; рекламы на радио (роликов в неделю). В таблице ниже приведены некоторые данные построенной модели.

R	*R*-квадрат	Скорректированный *R*-квадрат	Стд. ошибка оценки	Статистика Дарбина –Уотсона
0,554(a)	0,307	0,264	329,67616	0,773

Обращает на себя внимание малое значение скорректированного фактора детерминации, всего 0,264, которое указывает на неудовлетворительное качество модели. Возможная причина этого состоит в том, что существует взаимосвязь уровней ряда, на что указывает значение статистики Дарбина – Уотсона, равное 0,773 (при отсутствии автокорреляции это значение должно быть близким к 2).

Наличие автокорреляции легко установить, если при построении линейной регрессионной модели задать программе вывод нестандартизированных остатков в виде значений новой переменной. Для этого нужно в окне установки параметров линейной регрессионной модели (см. рис. 3.2) нажать кнопку Save (Сохранить) и в появившемся окне поставить галочку Residuals (Остатки) Unstandardized (Нестандартизованные). В этом случае после завершения работы программы по построению линейной регрессионной модели в окне редактора SPSS появится новая переменная RES_1, значения которой и будут содержать ошибку. Для определения коэффициентов автокорреляции различных порядков следует выбрать пункты меню Graph/Time Series/Autocorrelations и в открывшемся окне

переменную RES_1 переместить в правое окошко, как показано на рис. 3.20. Все остальные параметры можно оставить без изменения.

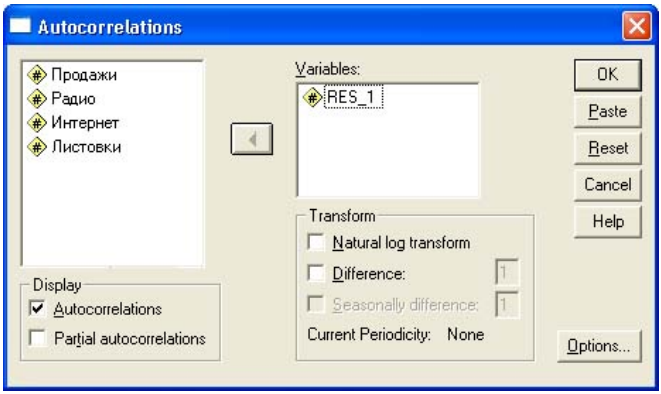


Рис. 3.20. Окно выбора параметров построения корреляционной диаграммы

Если теперь нажать кнопку OK в окне выбора параметров построения автокорреляционной диаграммы, то в окне выдачи результатов появится диаграмма автокорреляции остатков с 1-го до 16 -го порядка. Как следует из рис. 3.21,

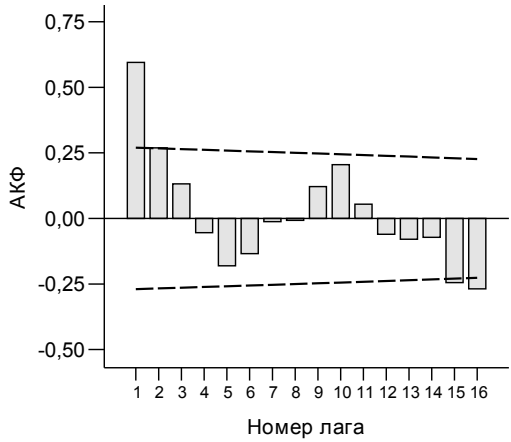


Рис. 3.21. Автокорреляционная диаграмма остатков

наиболее сильной является автокорреляция с лагом 1 (т. е. наибольшая связь имеется между уровнями ряда, сдвинутыми на один временной ин-

тервал). Величина коэффициента автокорреляции первого порядка существенно превосходит верхнюю границу доверительного интервала, отмеченную на рис. 3.21 пунктирной линией, и является значимой, в то время как автокорреляция с лагом больше единицы является несущественной. Проведенный анализ позволяет применить методику (3.35), (3,36) для устранения автокорреляции в остатках.

Чтобы выполнить авторегрессионное преобразование и получить модель, свободную от автокорреляции, выберем пункты меню Analyze/Time Series/Autoregression и в появившемся окне результативную и факторные переменные разместим так, как показано на рис. 3.22.

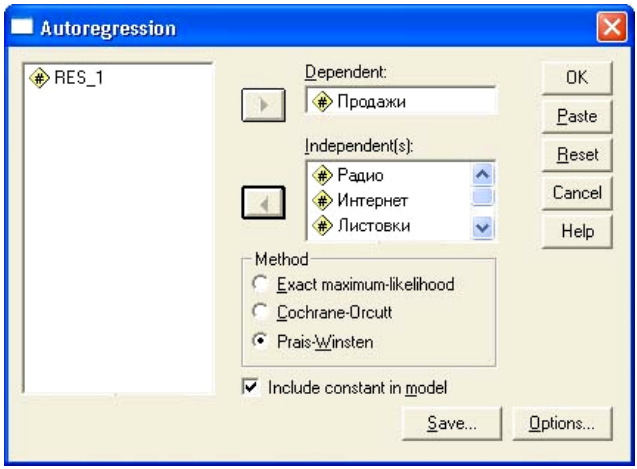


Рис. 3.22. Окно установки параметров для построения авторегрессионной модели

Поскольку пропущенных значений нет, выбираем метод Прайса – Винстена. Все остальные параметры построения модели можно не изменять. После запуска процедуры построения модели кнопкой OK появится сводка результатов, содержащая сведения о параметрах модели.

R	*R*-квадрат	Скорректированный *R* - квадрат	Стд. ошибка оценки	Статистика Дарбина – Уотсона
0,689	0,475	0,430	260,272	1,756

Как следует из приведенной ниже таблицы, качество регрессионной модели заметно улучшилось, а авторегрессия в остатках оказалась устраненной, о чем говорит значение статистики Дарбина – Уотсона, близкое к 2. Значения регрессионных коэффициентов модели также приведены в таблице.

Переменная	Регрессионные коэффициенты			
	Значения	Ст. ошибка	*t*	Значимость
Реклама на радио (роликов в неделю)	202,825	51,564	3,933	0,000
Число обращений на сайт в неделю	1,369	,653	2,097	0,041
Число листовок, размещенных на рекламных щитах	0,783	1,754	0,447	0,657
Константа	–524,839	257,346	–2,039	0,047

Как явствует из приведенных результатов, незначимым при уровне значимости 0,05 является только коэффициент при переменной, задающей число рекламных щитов, на которых размещалась реклама о товарах в магазине. Следовательно, эту переменную следует исключить из модели и построить модель с двумя регрессорами. Параметры двухфакторной модели приведены в табл. 3.17.

Таблица 3.17

Параметры авторегрессионной модели

Переменная	B	Ст. ошибка	*t*	Значимость
Реклама на радио (роликов в неделю) (b_1)	211,023	47,710	4,423	0,000
Число обращений на сайт в неделю (b_2)	1,435	0,628	2,285	0,027
Константа (b_0)	–487,092	241,353	–2,018	0,049

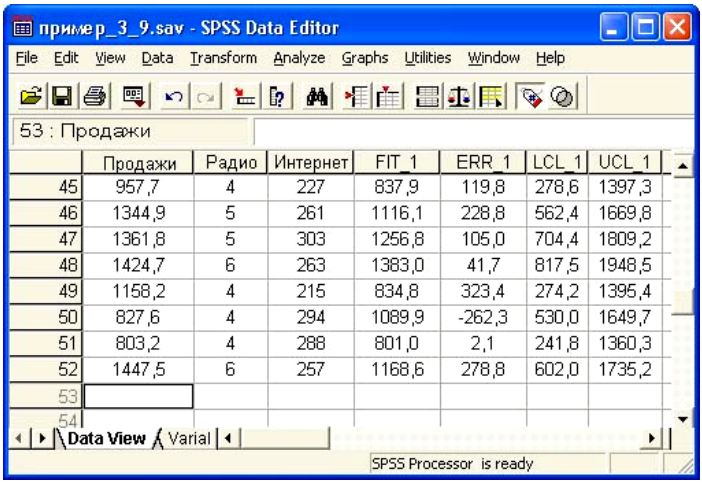
Найдем, каким будет относительный прирост объема продаж, если относительный прирост числа рекламных роликов на радио составит 50 %. Для этого необходимо знать, как рассчитываются значения уровней ряда с помощью найденных параметров авторегрессионной модели. В SPSS прогнозные значения авторегрессионной модели y_i^T вычисляются в соответствии с несколько модифицированной формулой (3.35)

$$\tilde{y}_1^T = y_1^T; \qquad y_i^T = b_0 + b_1 \cdot x_i + b_2 \cdot x_i; \;\; i = 1, 2, \ldots, n;$$

$$\tilde{y}_i^T = y_i^T + \rho \cdot (y_{i-1} - y_{i-1}^T); \quad i = 2, 3, \ldots, n, \qquad (3.37)$$

где $\tilde{y}_i^T$ – прогнозные значения зависимой переменной; $\rho = 0{,}631$ – автокорреляционный коэффициент, найденный на последнем шаге итерации; b_0, b_1, b_2 – регрессионные коэффициенты модели (см. табл. 3.17).

В SPSS прогнозные значения результативной переменной $\tilde{y}_i^T$, ошибка $\varepsilon_i = y_i - \tilde{y}_i^T$ и верхняя и нижняя границы 95% -го доверительного интервала для прогноза генерируются программой при создании авторегрессионной модели и размещаются в окне редактора данных (переменные FIT_1, ERR_1, UCL_1 и LCL_1). На рис. 3.23 изображен фрагмент входных данных (переменные Продажи, Радио, Интернет) и выходных данных (переменные FIT_1, ERR_1, UCL_1 и LCL_1).



	Продажи	Радио	Интернет	FIT_1	ERR_1	LCL_1	UCL_1
45	957,7	4	227	837,9	119,8	278,6	1397,3
46	1344,9	5	261	1116,1	228,8	562,4	1669,8
47	1361,8	5	303	1256,8	105,0	704,4	1809,2
48	1424,7	6	263	1383,0	41,7	817,5	1948,5
49	1158,2	4	215	834,8	323,4	274,2	1395,4
50	827,6	4	294	1089,9	-262,3	530,0	1649,7
51	803,2	4	288	801,0	2,1	241,8	1360,3
52	1447,5	6	257	1168,6	278,8	602,0	1735,2
53							
54							

Рис. 3.23. Окно редактора данных SPSS с фрагментом входных и выходных данных

Как следует из рисунка, объем продаж в последнюю неделю составил 1447, 5 тыс. руб., а рекламных роликов на радио было 6. При относительном увеличении на 50 % число рекламных роликов должно быть равно 9. Найдем прогнозное значение объема продаж в этом случае. По формуле (3.37) вычислим $\tilde{y}_i^T$ последней недели при значении параметров $x_1 = 9,\ x_2 = 257$:

$$\tilde{y}_{52}^T = -487{,}092 + 211{,}023 \cdot 9 + 1{,}435 \cdot 257 + 0{,}631 \cdot (803{,}2 - 770{,}4) = 1801{,}695.$$

Таким образом, при относительном приросте числа рекламных роликов на радио 50 % прирост объема продаж составит 354,235 тыс. рублей.

Пример 3.10

Файл Пример_3_10.xls содержит данные о валовом внутреннем продукте нескольких европейских стран за 1977 – 2000 гг. Все величины ВВП выражены в сопоставимых ценах 1990 г. и указаны в миллиардах единиц национальных валют. Требуется, используя данные 1977 – 1998 гг., создать модель временного ряда. Используя данные 1999 – 2000 гг. протестировать созданную модель, найдя относительную ошибку предсказаний. Предсказать уровень ВВП в этих странах в 2001 г.

Решение

При построении прогнозных значений уровней временного ряда важно иметь правильные представления об общих свойствах временного ряда. Поэтому начинать исследование следует с построения графика ВВП в зависимости от года (рис. 3.24). Дальнейшие исследования будем проводить на примере одной страны – Австрии.

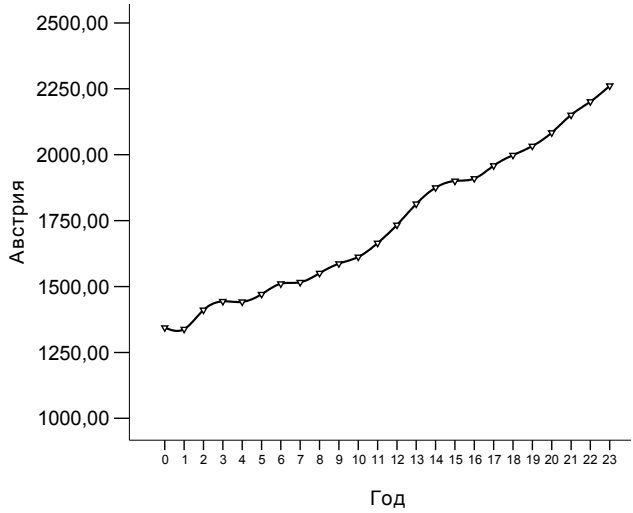
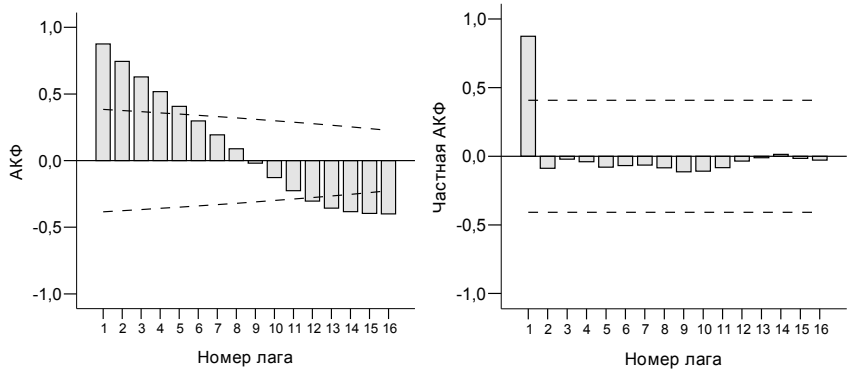


Рис. 3.24. Зависимость ВВП Австрии с 1977 по 2000 г. (млрд шиллингов) (1977 г. принят за начало отсчета)

Из графика следует, что временной ряд имеет очевидный тренд. Ясно, что объясняющей переменной не может являться год, и поэтому строить простую линейную регрессионную модель не имеет никакого смысла. В этом легко убедиться, если построить функцию автокорреляции и функцию частной автокорреляции, изображенные на рис. 3.25. В частной АКФ устраняется зависимость между промежуточными наблюдениями (наблю-

дениями внутри лага). Другими словами, частная автокорреляция на данном лаге аналогична обычной автокорреляции, за исключением того, что при вычислении из нее удаляется влияние автокорреляций с меньшими лагами. На лаге 1 (когда нет промежуточных элементов внутри лага) частная автокорреляция равна, очевидно, обычной автокорреляции. Частная автокорреляционная функция дает более достоверную информацию, поскольку она очищена от влияния корреляций более низкого порядка.



а *б*

Рис. 3.25. Коэффициенты автокорреляции (*а*) и частной автокорреляции (*б*) для 16 последовательных лагов

Из рисунка следует, что для построения правильной модели изучаемого ряда достаточно использовать авторегрессионную модель первого порядка

$$y_t = m + b_1 \cdot (y_{t-1} - m) + v_t , \qquad (3.38)$$

где m – некоторая константа (обычно это среднее значение уровней исходного ряда), b_1 – регрессионный коэффициент, v_t – случайная ошибка, имеющая свойства белого шума (среднее значение v_t равно нулю, ошибки для разных значений t независимы). Преобразование (3.38) обычно называют авторегрессионным преобразованием первого порядка AR(1). Если частная автокорреляционная функция указывает на наличие автокорреляции порядка p, то потребуется и авторегрессионное преобразование порядка p

$$y_t = m + b_1 \cdot (y_{t-1} - m) + b_2 \cdot (y_{t-2} - m) + \ldots + b_p \cdot (y_{t-p} - m) + v_t . \qquad (3.39)$$

Это преобразование обычно обозначается как AR(p) или ARIMA(p,0,0).

Преобразования (3.38), (3.39) применимы только для стационарных рядов, в которых отсутствует изменение средних во времени. В нашем

случае, как следует из рис. 3.24, имеется ярко выраженная трендовая зависимость, и преобразование (3.38) не может обеспечить построение адекватной модели. Для сведения ряда к стационарному можно подвергнуть исходные данные еще одному преобразованию. Если временной ряд имеет вид

$$y_t = m + y_{t-1} + \varepsilon_t, \tag{3.40}$$

то, вычитая из левой и правой частей (3.40) y_{t-1}, получим новый ряд $z_t = y_t - y_{t-1}$, который уже будет стационарным:

$$z_t = m + \varepsilon_t. \tag{3.41}$$

Если ряд (3.41) не является стационарным, то процесс взятия разностей можно продолжить, выполнив преобразование

$$y_t = m + y_{t-1} + y_{t-2} + \ldots + y_{t-d} + \varepsilon_t. \tag{3.42}$$

Модель временного ряда (3.42) обозначается как $I(d)$ – интегрированная модель временного ряда порядка d или ARIMA$(0,d,0)$.

Наконец, возможна модель временного ряда, при которой уровни ряда определяются некоторой константой и значением ошибок в прошлые моменты времени:

$$y_t = m + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \ldots + \varepsilon_{t-q}. \tag{3.43}$$

Модель временного ряда (3.43) носит название модели скользящей средней порядка q (Moving Average) и обозначается как Ma(q) или ARIMA(0,0,q). Модель временного ряда (3.43), задаваемая процедурой скользящего среднего, не имеет ничего общего с процедурой сглаживания временных рядов.

Общая модель временного ряда, которая включает в себя авторегрессионное преобразование порядка p, усреднение методом скользящей средней порядка q и переход от исходных величин к приращениям порядка d, называется моделью ARIMA(p,d,q).

В рассматриваемом нами случае достаточно воспользоваться простейшей моделью временного ряда ARIMA(0,1,0). Для того чтобы исключить данные 1999 и 2000 гг. из анализа при построении модели, выберем опции Data/Select Cases (Данные/Выбор случаев) и в открывшемся окне введем условие Год < 1999. В результате данные 1999 и 2000 гг. будут исключены из анализа при построении модели временного ряда. Выбрав пункты меню Analyze/TimeSeries/ARIMA (Анализ/Временные ряды /ARIMA), откроем окно выбора параметров регрессионной модели ARIMA. В этом окне следует выбрать параметры, отмеченные на рис. 3.26. Все остальные параметры влияют лишь на степень полноты выводимой информации, и их можно оставить выбранными по умолчанию.

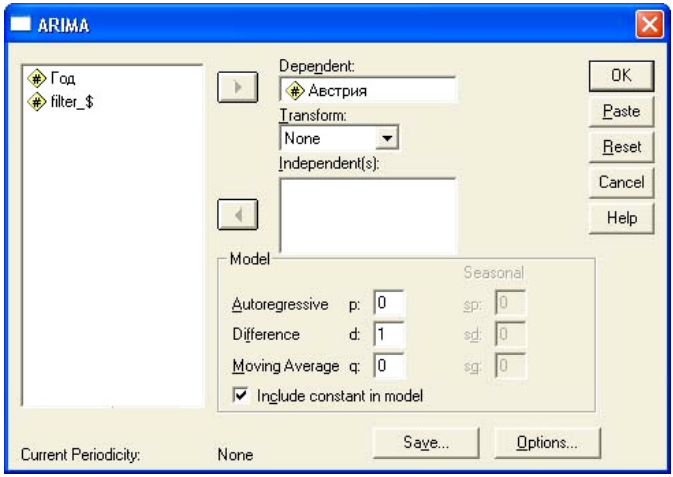


Рис. 3.26. Окно выбора параметров модели ARIMA

Запустить программу создания модели следует нажатием кнопки OK. В результате в окне вывода появится сводка о параметрах модели, а в окне редактора данных будут сгенерированы значения пяти новых переменных: FIT_1, ERR_1, UCL_1, LCL_1 и SEP_1. Смысл четырех первых переменных обсуждался в предыдущем примере, SEP_1 – это стандартная ошибка предсказаний. В окне вывода результатов интерес для нас представляет лишь табличка оценки параметров модели. Как следует из выражения (3.40), в модели ARIMA(0,1,0) всего лишь один оцениваемый параметр – константа m (табл. 3.18).

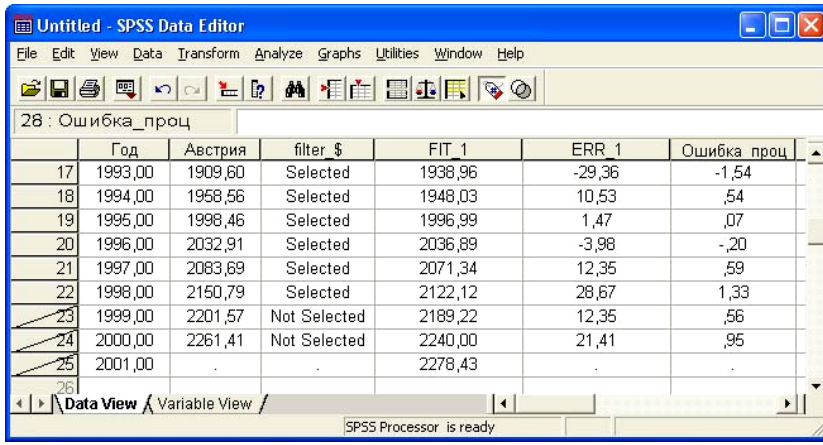
Таблица 3.18

Итоговый отчет о параметрах модели ARIMA(0,1,0)

Параметр	Оценка	Станд. откл.	t	Значимость
m	38,429	5,261	7,304	0,000

Для нахождения численного значения этого параметра строится ряд $z_t = y_t - y_{t-1} = m + \varepsilon_t$. Величина константы m – это просто среднее значения уровней ряда z_t. Предсказанные значения уровней ряда строятся в соответствии с формулой (3.40) $y_t^T = y_{t-1} + m$ (напомним, что предсказанные моделью значения временного ряда – это значения переменной FIT_1). Качество построенной модели является достаточно высоким. Параметр m яв-

ляется значимым, относительная ошибка прогноза на тестовой выборке (1999 и 2000 гг.) не превышает 1% (рис. 3.27).



	Год	Австрия	filter_$	FIT_1	ERR_1	Ошибка_проц
17	1993,00	1909,60	Selected	1938,96	-29,36	-1,54
18	1994,00	1958,56	Selected	1948,03	10,53	,54
19	1995,00	1998,46	Selected	1996,99	1,47	,07
20	1996,00	2032,91	Selected	2036,89	-3,98	-,20
21	1997,00	2083,69	Selected	2071,34	12,35	,59
22	1998,00	2150,79	Selected	2122,12	28,67	1,33
23	1999,00	2201,57	Not Selected	2189,22	12,35	,56
24	2000,00	2261,41	Not Selected	2240,00	21,41	,95
25	2001,00	.	.	2278,43	.	.
26						

Рис. 3.27. Предсказанные моделью ARIMA(0,1,0) (переменная FIT_1) объемы ВВП Австрии и их абсолютная и относительная ошибки

На рис. 3.28 приведены исходные данные (помечены треугольниками) и результаты моделирования ВВП Австрии.

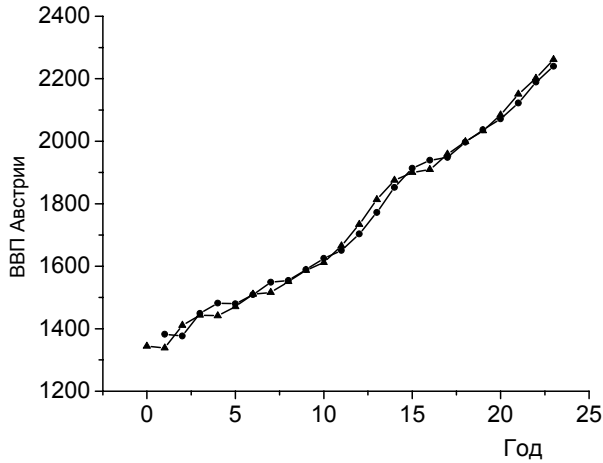


Рис. 3.28. Исходные данные и результаты моделирования ВВП Австрии по годам (1977 г. принят за начало отсчета)

Результаты, представленные на рис. 3.28, показывают, что модель достаточно точно описывает структуру анализируемых данных, и дальнейшее усложнение модели не оправдано.

Этим примером мы завершим рассмотрение главы, посвященной построению регрессионных моделей в SPSS. Материал для дальнейшего изучения можно найти в литературе, список которой приведен в конце учебного пособия.

Контрольные вопросы

3.1. Что понимается под моделью множественной линейной регрессии?

3.2. Перечислите основные посылки МНК.

3.3. Какой смысл имеют коэффициенты регрессионного уравнения?

3.4. Какие преимущества имеет стандартизованная модель регрессии?

3.5. В чем смысл коэффициента детерминации? Какие значения он может принимать?

3.6. Чем нормированный коэффициент детерминации отличается от обычного?

3.7. Если число объясняющих переменных увеличилось, то обязательно ли увеличится и значение нормированного коэффициента детерминации? А как будет вести себя нескорректированный фактор детерминации?

3.8. Какой смысл имеет коэффициент толерантности переменной и для каких целей он используется в регрессионном анализе?

3.9. Как производится анализ статистической значимости регрессионного уравнения в целом? Какая нулевая гипотеза при этом выдвигается?

3.10. Какая статистика используется для оценки статистической значимости регрессионных коэффициентов? Какая гипотеза при этом проверяется?

3.11. Как ставится задача об интервальной оценке регрессионных коэффициентов?

3.12. Какие формы уравнений регрессии могут быть сведены к линейной модели?

3.13. Каким способом можно выявить автокорреляцию во входном наборе данных?

3.14. Как убедиться в гомоскедастичности входного набора данных? В чем состоит смысл теста Левене?

3.15. Какие нелинейные модели однофакторной регрессии можно построить в SPSS, используя закладку Analyze/Regression/Curve Estimation (Анализ/Регрессия/Оценка кривых)?

3.16. Сформулируйте постановку задачи логистической регрессии.

3.17. Какие показатели используются для оценки качества модели логистической регрессии? С помощью какого критерия оценивается значимость регрессионных коэффициентов в модели логистической регрессии?

3.18. Что представляет собой ряд динамики? В чем специфика построения регрессионных моделей для рядов динамики?

3.19. Для каких целей производится сглаживание рядов динамики? Какие процедуры сглаживания рядов динамики имеются в SPSS?

3.20. Какие модели временного ряда обычно используются при анализе рядов динамики в SPSS?

3.21. Какими способами можно устранить автокорреляцию в рядах динамики? В чем суть итерационного процесса Кохрана – Оркатта?

3.22. Какие преобразования временного ряда включает в себя модель ARIMA(p,d,q)?

Задачи и упражнения

3.1. В файле Задача_3.1.xls приведены данные детской смертности (число смертей на 1000 рождений), грамотности взрослого населения в процентах, процента учеников, оканчивающих начальную школу, и ВНП на душу населения (долл. США) ряда стран Азии Африки и Латинской Америки.

а) При уровне значимости 0,05 выясните, с какими из анализируемых факторов связана детская смертность.

б) Постройте линейную регрессионную модель, оцените статистическую значимость модели в целом и ее регрессионных коэффициентов.

в) Сохраните предсказываемые моделью значения результативной переменной и не объясняемые моделью остатки.

г) Убедитесь, что остатки подчиняются распределению, близкому к нормальному, а сумма остатков близка к нулю.

д) Постройте график зависимости не объясняемых регрессионной моделью остатков от предсказываемых моделью значений результативной переменной и на основании этого дайте заключение о наличии или отсутствии гетероскедастичности во входном наборе данных.

3.2. В файле Задача_3.2.xls приведены данные об объеме выпуска продукции (шт.) и производственных затратах (тыс. руб.) небольшого автоагрегатного предприятия за 14 последних месяцев работы. Планируется в следующем месяце довести выпуск агрегатов до 1100 шт. Оцените с 95 % -й доверительной вероятностью размер производственных затрат при таком объеме производства.

а) Постройте линейную регрессионную модель, оцените статистическую значимость модели в целом и ее регрессионных коэффициентов. Используя статистику Дарбина – Уотсона, убедитесь, что авторегрессия во входном наборе данных отсутствует.

б) Сохраните предсказываемые моделью значения результативной переменной и не объясняемые моделью остатки.

в) Убедитесь, что остатки подчиняются распределению, близкому к нормальному, а сумма остатков близка к нулю.

г) Постройте график зависимости не объясняемых регрессионной моделью остатков от предсказываемых моделью значений результативной переменной и на основании этого сделайте заключение о наличии или отсутствии гетероскедастичности во входном наборе данных.

д) Постройте автокорреляционную функцию остатков и убедитесь, что ошибки не коррелируют.

3.3. Агент по продаже недвижимости желает получить формулу, позволяющую предсказывать размер месячной аренды (тыс. руб.) производственных помещений в зависимости от их площади ($м^2$). Для этих целей он располагает небольшой базой данных, основанной на результатах предыдущих сделок (файл Задача_3_3.xls).

а) Определите уравнение взаимосвязи стоимости арендной платы в зависимости от арендуемой площади. Оцените статистическую значимость регрессионного уравнения в целом и регрессионных коэффициентов. Найдите 95 % -й доверительный интервал для арендной платы, если площадь арендуемых помещений равна 1000 $м^2$.

б) Постройте диаграмму рассеяния. Используя возможности редактора диаграмм, постройте 95 % -й доверительный интервал для функции регрессии и для средних значений функции регрессии.

3.4. В файле Задача_3_4.xls содержатся данные об оценочной стоимости садового дома в БТИ, времени (мес.), прошедшего со дня оценки до продажи, типе постройки (деревянный дом – 0, кирпичный – 1) и реальной цене, по которой дом был продан.

а) Постройте линейную регрессионную модель продажной стоимости садовых домиков в зависимости от оценочной стоимости, типа дома и

времени, прошедшего со дня оценки, отобрав значимые объясняющие переменные.

б) Определите толерантность каждой из переменных, включенных в модель, и на основании этого сделайте вывод о наличии или отсутствии мультиколлинеарности во входном наборе данных.

3.5. В центре переподготовки кадров решили выяснить, как число правильно выполненных тестовых заданий зависит от количества дней обучения на курсах переподготовки. Для этой цели контрольным группам, состоящим из нескольких человек, давались задания и контролировалось число правильно выполненных заданий. Данные результатов обследования приведены в файле Задача_3_5.sav. Очевидно, что дисперсия числа правильных ответов скорее всего будет зависеть от числа дней обучения. Поэтому условия гомоскедастичности дисперсии не выполняются, и для получения достоверных результатов необходимо использовать метод взвешенных наименьших квадратов.

а) Постройте диаграмму разброса для входного набора данных и убедитесь, что разброс результативного признака растет с ростом числа дней обучения.

б) Запустите процедуру построения регрессионного уравнения методом взвешенных наименьших квадратов выбором опций Analyze/ Regression/Weight Estimation (Анализ/Регрессия/Оценка весов) и определите параметры регрессионного уравнения, который дает метод взвешенных наименьших квадратов.

в) Получите параметры обычной регрессионной модели и убедитесь, что регрессионная модель, построенная методом взвешенных наименьших квадратов, является более приемлемой.

3.6. В файле Задача_3_6.sav приведены данные обследования 17 предприятий с целью выявления факторов, влияющих на производительность труда:

y1 – выработка на одного работника, тыс. руб.;
y2 – выработка на одного рабочего, тыс. руб.;
x1 – доля рабочих, занятых наблюдением за работой автоматов, %;
x2 – доля рабочих, занятых при машинах и механизмах, %;
x3 – доля рабочих, занятых вручную при машинах и механизмах, %;
x4 – доля рабочих, занятых вручную не при машинах и механизмах, %;
x5 – доля рабочих, занятых вручную по наладке, %;
x6 – процент текучести кадров;
x7 – коэффициент сменности по всем рабочим;
x8 – коэффициент сменности по рабочим основных цехов;
x9 – доля профильной продукции в общем объеме продукции, %;

x10 – количество типов выпускаемой продукции, ед.;
x11 – доля покупных изделий и полуфабрикатов в затратах на производство, %;
x12 – доля оборудования основных цехов в общем количестве оборудования, %;
x13 – доля технологического оборудования в оборудовании основных цехов, %;
x14 – доля основных рабочих в общей численности рабочих, %;
x15 – доля рабочих основных цехов в общей численности рабочих, %;
x16 – доля специалистов и служащих в общей численности работающих, %;
x17 – фондовооруженность на одного рабочего, тыс. руб.;
x18 – фондовооруженность на одного работника, тыс. руб.;
x19 – электровооруженность потенциальная, кВт;
x20 – электровооруженность фактическая на одного рабочего, тыс. кВт.ч;
x21 – электровооруженность фактическая на 1 тыс. отработанных чел-ч, кВт. ч;
x22 – доля полуавтоматов и автоматов в технологическом оборудовании;
x23 – доля полуавтоматов в технологическом оборудовании;
x14 – доля автоматов в технологическом оборудовании;
x25 – доля в технологическом оборудовании автоматических линий.

Переменные y1 и y2 являются результативными, а x1 – x25 – факторными.

а) Определите при уровне значимости 0,05 факторы, влияющие на объем выработки на одного работника и на одного рабочего. Одинаковыми ли получились наборы факторов?

б) Проверьте условия отсутствия мультиколлинеарности для всех переменных, включенных в модель.

3.7. Имеются данные (файл Задача_3_7.sav) о средней продолжительности жизни мужчин и женщин в ряде стран Европы и о тех факторах, которые могут влиять на среднюю продолжительность жизни. В частности, приведены данные о детской смертности на 1000 рождений, числе солнечных часов в году, числе дождливых дней в году, проценте городского населения, средней температуре января и июля.

а) Используя регрессионный анализ, выделите из имеющихся факторов те, которые значимо (при уровне значимости 0,05) влияют на продолжительность жизни мужчин и женщин.

б) Получите прогнозные значения продолжительности жизни мужчин и женщин, а также 95 % -е доверительные границы для индивидуальных значений результативного признака.

в) Убедитесь, что включение всех факторов в модель не приводит к улучшению модели.

3.8. Небольшое предприятие на протяжении 19 месяцев выпускает три вида продукции, которые условно обозначены как A, B, C. В файле Задача_3_8.xls содержатся данные о полных расходах завода (млн руб.) за каждый месяц и объемах произведенной продукции типа A, B, C (штук).

а) Проверите, выполняются ли необходимые условия метода МНК для анализируемого набора данных.

б) Определите, какой из видов продукции в действительности не является объясняющей переменной при уровне значимости 0,05.

в) Постройте регрессионное уравнение, позволяющее предсказывать полные расходы предприятия, значимое при уровне значимости 0,05. Найдите доверительные интервалы для коэффициентов регрессионного уравнения и индивидуальных значений полных расходов.

3.9. В файле Задача_3_9.xls приведены данные о поквартальных продажах автомобилей в США (тыс. шт.) с 1979 по 1986 г. и данные о ВНП (млрд долларов США), уровне безработицы (%) и налоговой ставке (%). Данные, характеризующие ВНП, уровень безработицы и уровень налоговой ставки, приведены с запаздыванием на один квартал. Причины этого понятны, поскольку люди, планирующие расходы на покупку автомобиля, ориентируются на экономические показатели предшествующего временного периода. Очевидно, что продажи автомобилей подвержены сезонным колебаниям. В этом легко убедиться, если построить график продаж от номера квартала. В то же время, хотя в исходном наборе данных имеются сезонные колебания, можно построить линейную регрессионную модель, вводя фиктивные переменные K1, K2, и K3. Значения переменной K1 будет равно 0 для всех кварталов, не совпадающих с первым кварталом года. Для всех первых кварталов значение этой переменной равно единице. Аналогично определяются значения переменной K2 и K3. (Более подробные объяснения применения методики фиктивных переменных находятся в файле Задача_3_9.xls на листе Условие).

а) Постройте регрессионную модель для поквартального объема продаж, используя в качестве объясняющих переменных фиктивные переменные K1, K2, и K3 и данные об уровне ВНП, налоговой ставке и безработице.

б) Выведите в окне редактора данных прогнозные значения и ошибку. Убедитесь, что автокорреляция ошибок не является значимой.

в) Проверьте, можно ли существенно улучшить модель, производя экспоненциальное сглаживание результативной переменной (следует выбрать метод, учитывающий наличие тренда и сезонных составляющих).

г) Постройте на графике исходные, предсказанные моделью значения объема продаж в зависимости от номера квартала.

3.10. Накал страстей в период президентских выборов в США приковывает внимание прессы и населения многих стран мира. Между тем многофакторный регрессионный анализ позволяет с высокой надежностью

предсказать результат этих выборов. В файле Задача_3_10. sav содержатся данные о проценте голосов, набранных правящей партией на президентских выборах с 1916 по 2000 г. (результативный признак) и данные о некоторых социально-экономических показателях США в год выборов (факторные переменные). В качестве факторов используются следующие независимые переменные:

- правящая партия (0 – демократическая, 1 – республиканская партия);
- процент роста ВНП за первые девять месяцев в год выборов;
- темп инфляции за первые девять месяцев в год выборов;
- число кварталов за последние четыре года, когда рост ВНП превышал 3,2 %;
- число сроков подряд, в течение которых правящая партия находится у власти;
- проходят ли выборы в период, когда страна ведет войну (учитывались только глобальные войны);
- выдвигается ли действующий президент на следующий срок.

Большая часть факторных переменных относится к номинативной шкале и кодируется с помощью некоторых числовых значений. Способ кодирования можно посмотреть непосредственно в файле Задача_3_10. sav, если в окне редактора SPSS зайти на закладку Variable View (Отображение переменных).

а) Используя данные за 1916 – 2000 гг., постройте регрессионное уравнение, позволяющее предсказать число голосов, отданных за представителя правящей партии в 2004 г.

б) Проверьте статистическую значимость регрессионного уравнения и регрессионных коэффициентов.

в) Получите интервальную оценку числа голосов, отданных за правящую партию. С какой доверительной вероятностью эта регрессионная модель в состоянии предсказать результаты президентских выборов в США?

3.11. В файле Задача_3_11.sav приведены данные об оценочной стоимости коттеджей (тыс. долларов США), выставляемых на продажу в Подмосковье, площади земельного участка (га), площади коттеджа (кв. м), количестве ванных комнат и мест для автомобилей в гараже.

а) Постройте регрессионную модель, позволяющую оценить стоимость коттеджей.

б) Найдите доверительный интервал для стоимости с 95 % -й доверительной вероятностью.

в) Проверьте, является ли модель действительно линейной по предикторам «Площадь земельного участка» и «Площадь коттеджей». Для этого опре-

делите новые переменные, равные квадратам исходных величин, и постройте регрессионную модель снова. Улучшилось ли качество модели? Как можно интерпретировать новые переменные?

3.12. В файле Задача_3_12.sav приведены данные о 121 автомобиле выпуска 2002 г. Требуется построить регрессионную модель, позволяющую предсказать пробег автомобиля при использовании 1 литра топлива (результативная переменная «Пробег») в зависимости от таких факторов, как тип автомобиля (спортивный или неспортивный), тип привода (полный, передний, задний), качество топлива (высшее, среднее), мощность двигателя (л. с.), длина автомобиля (м), ширина автомобиля (м), вес (кг), индекс грузоподъемности, радиус разворота автомобиля (м).

а) Постройте регрессионное уравнение, определяющее пробег автомобиля на одном литре топлива (км/л) в зависимости от технических характеристик автомобиля (используйте метод Backward исключения переменных).

б) Пробег должен зависеть от аэродинамических характеристик автомобиля, но среди приведенных характеристик такого показателя нет. С увеличением длины автомобиля пробег уменьшается (знак регрессионного коэффициента при переменной «Длина» получился отрицательным). Тем не менее данные содержат скрытую информацию зависимости пробега от аэродинамических качеств автомобиля. Эти рассуждения легко проверить, если ввести новую переменную A1, которая равна квадрату длины, и построить новую регрессионную модель. Коэффициент при новой переменной должен получиться значимым и иметь положительный знак. Это как раз и будет указывать на то, что с ростом длины автомобиля некий фактор приводит к увеличению длины пробега. Создайте новую переменную A1, включите ее в число объясняющих переменных и постройте новую регрессионную модель. Действительно ли в приведенных данных содержатся знания о влиянии аэродинамики на пробег автомобиля? Улучшилось ли при введении новой переменной качество модели?

в) Проверьте, выполняются ли для итоговой модели основные предположения регрессионного анализа (найдите статистику Дарбина – Уотсона, среднее значение остатков, проверьте распределение остатков на нормальность, проверьте факторные переменные на наличие мультиколлинеарности). Сравните свои результаты с файлом Задача_3_12.spo, в котором приведены результаты исследования этого набора данных.

3.13. Начальник отдела кадров крупного промышленного предприятия постоянно сталкивается с проблемой подбора кадров на должность руководителей структурных подразделений. Отбор кандидатов производится с помощью тестирования. Первый тест ориентирован на проверку знаний в области технологии управления. Второй тест позволяет проанализировать

личностные качества кандидатов, такие как целеустремленность, настойчивость, ориентация на достижение конечного результата. Причем, чем выше балл теста, тем больше проявляются выявляемые тестом качества. Поскольку тестовая система подбора кадров функционирует уже не первый год, накопилась некоторая база знаний, используя которую, начальник отдела кадров хотел бы выяснить, какие из факторов существенны при подборе кадров. В файле Задача_3_13.xls содержатся данные о 45 руководителях подразделений, назначенных на свои должности в результате тестового отбора. Результативной переменной является выполнение плана в процентах за последний год. Факторными переменными являются результаты первого и второго теста в баллах, опыт работы на руководящих должностях (лет), а также профильное образование (инженерное или гуманитарное).

а) Постройте регрессионное уравнение, позволяющее прогнозировать процент выполнения планового задания в зависимости от значения факторных переменных.

б) Выясните, какие из перечисленных выше критериев оценки руководителей не оказывают существенного влияния на выполнение производственного плана.

3.14. В файле Задача_3_14.sav содержатся данные о работе 259 управляющих инвестиционных компаний США в 2001 г. Требуется выяснить с 95 %-й доверительной вероятностью, от каких факторов зависят средняя трехлетняя и средняя пятилетняя доходности для управляющих фондов, работающих с крупным, средним и мелким капиталом. Одинаковым ли оказывается результат для инвестиционных фондов, ориентированных на быстрый, медленный и средний рост доходов?

3.15. В файле Задача_3_15.sav приведены данные о минимальных и максимальных дневных температурах (по шкале Цельсия), днях недели и пиковых нагрузках (МВт) на подстанции, обслуживающей промышленный район города. Постройте регрессионную модель, позволяющую предсказать пиковую нагрузку по дням недели в зависимости от температурного режима. Модель должна содержать фиктивные переменные, позволяющие описать неравномерность потребления энергии по дням недели. Для построения модели возьмите выборку из 800 первых случаев, а остальные 199 случаев используйте как проверочную совокупность. Найдите сумму квадратов отклонений предсказанных значений от истинных значений на проверочном множестве и сравните результат с тем, который может быть получен с помощью нейросетевого моделирования. Для проведения нейросетового анализа используйте файл Задача_3_15.xls и программу Pathfinder.exe.

3.16. Имеются поквартальные данные за 1995 – 2002 гг. об итогах деятельности ООО «Вавилон» в г. Чебоксары, сформированные на основе бухгалтерской отчетности этого предприятия, и данные об уровне инфляции, основанные на индексе потребительских цен в Чувашии (Задача_3_16.sav). Требуется проанализировать факторы, от которых зависит себестоимость продукции ООО «Вавилон».

Легко обнаружить, что если просто построить линейную регрессионную модель, то индекс потребительских цен не войдет в число объясняющих переменных и наблюдается сильная автокорреляция первого порядка в остатках. Очевидно, что индекс потребительских цен сказывается на производстве с запаздыванием. Поэтому требуется переопределить переменные. Кроме того, для улучшения качества модели следует произвести экспоненциальное сглаживание для результативной переменной и уже после этого строить линейную регрессионную модель. Сравните свои результаты с данными файлов Задача_3_16.spo, Задача_3_16_Решение.sav.

3.17. Имеются данные о росте численности населения США с 1790 по 1970 г. с 10-летним шагом (файл Задача 3_17.xls). Требуется подобрать вид кривой, наиболее корректно описывающей рост численности населения США на этом временном отрезке.

3.18. Имеются данные о производстве аппаратов факсимильной связи: их суммарном производстве компанией и себестоимости аппарата по годам (файл Задача_3_18.xls). Очевидно, что с ростом выпуска продукции ее себестоимость понижается. Используя подгонку кривых, определите закон, по которому снижалась себестоимость одного аппарата на этом предприятии в зависимости от суммарного объема выпущенной продукции.

3.19. В файле Задача_3_19.sav приведены данные о результатах голосования 145 конгрессменов США по целому ряду законопроектов и их принадлежность к демократической или республиканской партиям. Принадлежность кодируется цифрой «0» для демократической и цифрой «1» – для республиканской партии. Результаты голосования по всем законопроектам кодируются цифрами: 0 – против; 1 – воздержался; 2 – поддержал законопроект. Для 5 конгрессменов имеются результаты голосования, но не известна их принадлежность к той или иной партии. Используя 130 первых случаев как обучающую выборку, постройте модель логистической регрессии для предсказания принадлежности конгрессменов к той или иной партии. Проверьте построенную модель, используя следующие 15 случаев, для которых известна партийная принадлежность (тестовая выборка). Определите партийность конгрессменов, для которых известны результаты голосования, но не известна партийная принадлежность.

3.20. В файле Задача_3_20.xls приведены данные, представляющие собой результаты психологического тестирования учащихся специализированных школ Санкт-Петербурга с физико-математическим и гуманитарным уклоном. Всего предлагалось пять тестов, условное название которых приведено ниже: Тест_1 – дополнение предложений; Тест_3 – нахождение аналогий; Тест_4 – обобщение умозаключений; Тест_5 – способность к устному счету; Тест_7 – образность мышления. Чем выше набранный бал, тем лучше проявляется анализируемое качество. Учащиеся школ с физико-математическим уклоном условно названы – «физиками», а с гуманитарным уклоном – «лириками».

а) Загрузите данные в редактор SPSS, создайте новую переменную Index, которая принимает целые случайные значения в интервале от 1 до 76.

б) Отсортируйте данные так, чтобы переменная Index была упорядочена по возрастанию.

в) Создайте модель логистической регрессии, используя первые 70 случаев. Прогностические способности регрессионной модели проверьте, сравнивая истинные и прогнозные значения для оставшихся 6 случаев.

3.21. Проиллюстрируем процесс добычи знаний из баз данных на примере выборов президента в США.

Имеется таблица данных с результатами 31-й предвыборной ситуации с 1860 по 1980 г. (файл Задача_3_21.sav). Для каждых выборов в таблице содержатся данные по 12 бинарным признакам:

1. Правящая партия была у власти более одного срока?
2. Правящая партия получила более 50% голосов на прошлых выборах?
3. В год выборов была активна третья партия?
4. Была серьезная конкуренция при выдвижении кандидата от правящей партии?
5. Кандидат от правящей партии был президентом в год выборов?
6. Был ли год выборов временем спада или депрессии?
7. Был ли рост среднего национального валового продукта на душу населения более 2,1%?
8. Произвел ли правящий президент существенные изменения в политике?
9. Во время правления были существенные социальные волнения?
10. Администрация правящей партии виновна в серьезной ошибке или скандале?
11. Кандидат от правящей партии – национальный герой?
12. Кандидат от оппозиционной партии – национальный герой?

Также в таблице содержится информация о результатах выборов (победе правящей или оппозиционной партии). Значения бинарных признаков «0» (ответ «нет» для входного признака) и «1» (ответ «да» для входного признака). Значения результативного признака закодированы цифрой «1», если был избран кандидат правящей партии, и цифрой «2», если победил кандидат оппозиционной партии.

а) Постройте модель логистической регрессии, позволяющую по приведенным значениям факторных признаков предсказать результаты президентских выборов в США. В качестве объясняющих переменных выберите факторы под номерами 3, 4, 6, 7, 8, 9 в приведенном выше списке объясняющих переменных. Убедитесь, что модель точно воспроизводит все результаты выборов.

б) Проверьте, правильным ли является предсказание модели для выборов 1992 г. (Д. Буш – Б. Клинтон). Как известно, в этом году победил Клинтон, кандидат от оппозиционной партии.

в) Введите самостоятельно пропущенные значения для объясняющих переменных для выборов 1984, 1988, 1996, 2000 и 2004 гг. и проверьте прогностические возможности модели для этих выборов.

3.22. В файле Задача_3_22.sav содержатся данные о 149 клиентах банка, желавших получить кредит, и решение опытного менеджера о выдаче или отказе в выдаче кредита. На основании этой базы данных можно разработать экспертную систему, позволяющую решать вопрос о выдаче или отказе в кредите.

а) Выбирая случайным образом 70 % случаев, постройте модель логистической регрессии, позволяющую принимать решение о выдаче или отказе в выдаче кредита, 30 % оставшихся случаев используйте для проверки прогностических возможностей модели. Число имеющихся факторных признаков избыточно. Поэтому в качестве объясняющих переменных выберите переменные: Сумма_кредита, Срок_кредита, Площадь_квартиры, Расположение, Должность, Среднемес_доход, Среднемес_расход, Цель_кредитования.

б) Попробуйте улучшить качество модели, изменяя состав и количество объясняющих переменных.

3.23. В файле Задача_3_23.sav представлены данные социологического исследования 46 респондентов. Изучался вопрос, от каких факторов зависит, окажет ли человек другому лицу помощь или нет. Изучались факторы симпатии и агрессии к человеку, нуждающемуся в помощи; польза, которую принесет помощь; оценка сложности проблемы, в которой оказался человек, нуждающийся в помощи, условия, в которых требуется оказать помощь. Кроме того, в число объясняющих факторов была вклю-

чена и эмпатия – способность человека к сопереживанию. Постройте модель, позволяющую отобрать факторы, от которых зависит, будет ли оказана помощь. Используйте методы включения переменных Forward LR и Backward LR. Одинаковыми ли получились результаты? Если результаты получились разными, то попробуйте объяснить, почему это происходит.

3.24. В файле Задача_3_24.sav представлены данные о годовом уровне доходов семей (тыс. долларов США) различных слоев населения Америки по годам, начиная с 1967 по 2005 г. (данные взяты с сайта http://www.census.gov). По уровню доходов население делится на 5 групп. Приведены данные для нижней по уровню доходов группы, второй, третьей и четвертой, а также данные, представляющие верхнюю, 5 %-ю группу семей с максимальными доходами.

а) Создайте линейную авторегрессионную модель для переменной Нижние, взяв в качестве объясняющих переменных Нижние_Лаг1 и Нижние_Лаг2. Позаботьтесь о выводе статистики Дарбина – Уотсона и выводе диагностики коллинеарности объясняющих переменных. Оцените качество модели.

б) Используя опции Analyze/Time Series/Autoregression, постройте авторегрессионную модель для переменной Нижние, взяв в качестве объясняющих переменных Нижние_Лаг1 и Нижние_Лаг2 по методу Кохрана – Оркатта или Прайса – Винстена. Оцените качество полученной регрессионной модели.

в) Постройте модель, основываясь на данных с 1967 по 2000 г., а данные с 2000 по 2005 г. используйте для контроля качества модели. Какая из моделей, построенных выше, дает меньшую суммарную ошибку на контрольной выборке?

г) Повторите те же исследования для других групп населения. Можно ли утверждать на основании этого исследования, что расслоение населения по уровню доходов в США возрастает?

3.25. Файл Задача_3_25.xls содержит данные о валовом внутреннем продукте нескольких европейских стран за 1977 – 2000 г. Все величины ВВП выражены в сопоставимых ценах 1990 года и указаны в миллиардах единиц национальных валют.

а) Используя данные 1977 – 1998 гг., создайте две модели временного ряда как это было рекомендовано сделать в пунктах а и б предыдущей задачи. Объясняющей переменной здесь являются уровни анализируемого временного ряда, сдвинутые на один временной период назад. Для того чтобы создать объясняющую переменную, после загрузки данных в редактор SPSS нужно в окне Variable View объявить новую

переменную, скопировать анализируемые значения временного ряда и вставить их как значения новой переменной со сдвигом на один временной шаг.

б) Используя данные 1999 – 2000 гг., протестируйте созданные модели, найдя суммарную ошибку предсказаний. Проверьте наличие автокорреляции в остатках.

в) Постройте модель ARIMA (0,1,0) для каждого из анализируемых временных рядов. (Рекомендации по созданию такой модели можно найти в примере 3.10). Используя данные 1999 – 2000 гг., протестируйте созданную модель, найдя суммарную ошибку предсказаний. Проверьте наличие автокорреляции в остатках. Какая из трех созданных моделей представляется более правильной? Ответ аргументируйте.

3.26. В файле Задача_3_26.sav содержатся данные об уровне средней часовой производительности труда в США (в % к уровню 1982 г.) и данные о среднечасовой заработной плате (долларов США) в сопоставимых ценах 1982 г. за период с 1960 по 1990 г. Требуется построить модель временного ряда, позволяющую предсказать значение почасовой ставки заработной платы в 1991 г.

а) Постройте диаграмму зависимости почасовой ставки от года. Обратите внимание, что временной ряд неоднороден. До 1973 г. почасовая ставка растет, а после 1973 г. ведет себя неоднозначно. Чтобы учесть эту неоднородность, создана новая переменная I1, принимающая значения «0» на временном интервале от 1960 до 1973 г. и значение «1» позже 1973 г. Причиной изменения тенденции поведения временного ряда может быть какой-либо неучтенный фактор, приведший к резкому повышению производительности труда, например роботизация производства.

б) Выбирая в качестве объясняющих переменных величины Лаг1X, Лаг1Y и I1, постройте регрессионное уравнение для всего временного интервала. Проанализируйте качество модели, оценив статистику Дарбина – Уотсона, мультиколлинеарность переменных и авторегрессию остатков. Постройте графики истинных и прогнозных значений тарифной ставки в зависимости от производительности труда.

в) Попробуйте в качестве объясняющих переменных использовать только лаговые переменные: Лаг1X, Лаг2X, Лаг3X, Лаг4X, Лаг3_2X, Лаг4_2X. Удалось ли построить более качественную модель?

3.27. В файле Задача_3_27.xls приведены данные о фактическом уровне доходов небольшой фирмы и индексе потребительских цен в США в период с 1982 по 2001 г. Используя данные с 1982 по 2000 г., требуется по-

строить модель, позволяющую предсказать уровень реальных доходов фирмы в 2001 г.

а) После загрузки файла в редактор данных SPSS создайте новую переменную Реальные_доходы. Присвойте этой переменной значения, вычислив их по формуле

$$\text{Реальные\_доходы} = \frac{\text{Фактические\_доходы}}{\text{CPIU}} \cdot 100 .$$

б) Создайте новую переменную Лаг1Y и присвойте ей значения переменной Реальные_доходы, сдвинутые на один временной интервал.

в) Постройте линейную регрессионную модель с зависимой переменной Реальные_доходы и объясняющей переменной Код (или Год). Оцените статистическую значимость модели. Изучите поведение частной корреляционной функции остатков.

г) Постройте авторегрессионную (линейную) модель с объясняющей переменной Лаг1Y и модель ARIMA(0,1,0). Какая из трех построенных моделей является статистически более обоснованной и какая из моделей дает наименьшую ошибку прогноза?

3.28. В файле Задача_3_28.xls приведены данные об усредненных по годам значениях индекса Доу – Джонсона за период с 1969 по 2002 г. Постройте линейную авторегрессионную модель $y_t = b_0 + b_1 \cdot y_{t-1} + b_3 \cdot y_{t-3}$ и модели авторегрессии, используя метод Прайса – Винстена и Кохрана – Оркатта с объясняющими переменными y_{t-1} и y_{t-3}. Используя данные за 2001 и 2002 гг. как контрольную выборку, определите, какая из моделей дает меньшую ошибку на этой выборке.

ГЛАВА 4. ДИСКРИМИНАНТНЫЙ, ФАКТОРНЫЙ И КЛАСТЕРНЫЙ АНАЛИЗ. ДЕРЕВО РЕШЕНИЙ

4.1. Дискриминантный анализ

Дискриминантный анализ включает в себя методы классификации наблюдений в ситуации, когда исследователь обладает достаточно большим числом примеров правильной классификации (обучающими выборками). Дискриминантный анализ иногда называют также классификацией с учителем. Типичным примером задачи дискриминантного анализа может быть задача определение клиентов банка с высоким риском невозврата кредита. Ранее такую задачу мы уже решали, используя логистическую регрессию (см. пример 3.6).

Для решения этой задачи банк должен провести специальные исследования, позволяющие определить, какие характеристики клиентов связаны с невозвратом долга, и подготовить базу данных (знаний) достаточно большого числа случаев своевременного возврата и невозврата кредитов (построить обучающую выборку).

Дискриминантный анализ является более универсальной статистической процедурой, нежели уже обсуждавшийся метод логистической регрессии, хотя он преследует ту же самую цель – определить вероятность принадлежности заданного объекта к одной из заранее определенных групп.

Метод может применяться во всех случаях, когда на основании уже имеющейся информации требуется отнести новый случай (наблюдение) к одной из заранее определенных групп, В кадровых центрах, например, этот метод может быть использован для отбора претендентов на престижные должности. В медицине его широко применяют при постановке диагноза в сложных клинических случаях. В экономике и управлении этот метод также незаменим для определения факторов, эффективно влияющих на конечный результат.

В дискриминантном анализе в общем случае число классов, по которым производится классификация объектов, может быть произвольным.

Рассмотрим некоторые принципы дискриминантного анализа в случае, когда объекты нужно распределить по двум классам (A и B, например). На рис. 4.1 схематически представлены объекты, принадлежащие двум различным множествам. Каждый объект характеризуется двумя переменными x_1 и x_2. Если рассматривать проекции

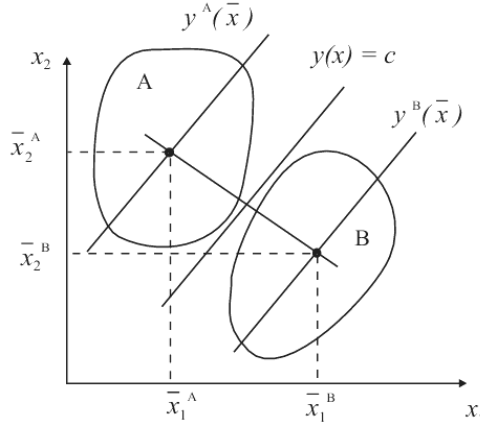


Рис. 4.1. Графическая интерпретация дискриминантного анализа

точек множеств A и B на оси x_1 и x_2, то по каждой из переменных некоторые из объектов множеств A и B имеют сходные свойства. Чтобы наилучшим образом разделить объекты этих двух множеств, нужно построить новую координатную систему. Новые оси (линия $y(x) = c$ и перпендикулярная ей линия, соединяющая центры множеств A и B) должны быть расположены таким образом, чтобы объекты классифицируемых множеств оказались по разные стороны от линии $y(x) = c$.

Дискриминантная функция $y(x)$ чаще всего выбирается линейной и для случая двух переменных определяется выражением

$$y(x) = b_1 \cdot x_1 + b_2 \cdot x_2 , \qquad (4.1)$$

где b_1 и b_2 – коэффициенты дискриминантной функции.

Координаты центров множеств A и B, отмеченные на рис. 4.1 жирными точками (в дискриминантном анализе их принято называть координатами центроидов), вычисляются по формулам

$$\bar{x}_1^{A} = \frac{1}{n^{A}} \sum_{i=1}^{n^{A}} x_{1i}^{A} , \ \bar{x}_2^{A} = \frac{1}{n^{A}} \sum_{i=1}^{n^{A}} x_{2i}^{A} , \ \bar{x}_1^{B} = \frac{1}{n^{B}} \sum_{i=1}^{n^{B}} x_{1i}^{B} , \ \bar{x}_2^{B} = \frac{1}{n^{B}} \sum_{i=1}^{n^{B}} x_{2i}^{B} . \qquad (4.2)$$

Обозначим $y^{A}(\bar{x})$ и $y^{B}(\bar{x})$ значения дискриминантной функции, вычисленной при подстановке координат центра множеств A и B в (4.1):

$$y^{A}(\bar{x}) = b_1 \cdot \bar{x}_1^{A} + b_2 \cdot \bar{x}_2^{A} ; \quad y^{B}(\bar{x}) = b_1 \cdot \bar{x}_1^{B} + b_2 \cdot \bar{x}_2^{B} \qquad (4.3)$$

и построим две линии: $y(x) = y^{\mathrm{A}}(\bar{x})$ и $y(x) = y^{\mathrm{B}}(\bar{x})$. Очевидно, что эти линии будут параллельны линии $y(x) = c$, причем первая линия пройдет через центр множества A, а вторая – через центр множества B. На рис. 4.1 эти линии обозначены просто как $y^{\mathrm{A}}(\bar{x})$ и $y^{\mathrm{B}}(\bar{x})$.

Для того чтобы задачу дискриминантного анализа в рассматриваемом случае двух переменных можно было считать решенной, следует найти коэффициенты дискриминантной функции b_1 и b_2 и константу c в уравнении (4.1). Коэффициенты дискриминантной функции (4.1) определяются из условия максимальности отношения межгрупповой вариации δ^2 к суммарной внутригрупповой вариации $\overline{\sigma^2}$:

$$\frac{\delta^2}{\overline{\sigma^2}} \to \max; \qquad \delta^2 = \left(y^{\mathrm{A}}(\bar{x}) - y^{\mathrm{B}}(\bar{x})\right)^2,$$

$$\overline{\sigma^2} = \sum_{i=1}^{n^{\mathrm{A}}} \left(y(x_i) - y^{\mathrm{A}}(\bar{x})\right)^2 + \sum_{i=1}^{n^{\mathrm{B}}} \left(y(x_i) - y^{\mathrm{B}}(\bar{x})\right)^2. \qquad (4.4)$$

В первой сумме формулы (4.4) суммирование идет по объектам множества A, а во второй – по объектам множества B.

Мы не будем приводить явные формулы для отыскания коэффициентов дискриминантной функции, поскольку в SPSS они вычисляются автоматически при запуске процедуры дискриминантного анализа. Важно лишь понимать, что в основе дискриминантного анализа, по существу, лежит хорошо известный принцип группировки объектов: наиболее правильной является такая группировка, при которой межгрупповая дисперсия максимальна, а внутригрупповая – минимальна.

После того как коэффициенты дискриминантной функции найдены, константа c может быть определена из условия равной удаленности точек линии $y(x) = c$ от центроидов множеств A и B:

$$c = \frac{1}{2}\left(y^{\mathrm{A}}(\bar{x}) + y^{\mathrm{B}}(\bar{x})\right). \qquad (4.5)$$

Уравнения (4.1), (4.5) позволяют построить процедуру дискриминантного анализа для определения принадлежности объекта к одной из двух групп, если число дискриминантных переменных равно двум. В общем случае обучающая выборка позволяет построить дискриминантную функцию, которая напоминает регрессионное уравнение

$$y(x_i) = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \ldots + b_k \cdot x_{ki}. \qquad (4.6)$$

Здесь $x_{1i}, x_{2i}, \ldots, x_{ki}$ $(i = 1, 2, \ldots, n)$ – значения переменных-предикторов, от которых, по мнению исследователя, должна зависеть группировка объектов, $b_0, b_1, \ldots, b_k$ – коэффициенты дискриминантной функции, которые и подлежат определению в дискриминантном анализе, символ i нумерует точки наблюдения (случаи). Как и для любой статистической процедуры, применение дискриминантного анализа возможно, если выполняются определенные условия:

- предикторы не должны сильно коррелировать между собой;
- дисперсия предикторов должна быть одинаковой для разных групп;
- средние значения и дисперсия для каждого из предикторов не коррелируют между собой;
- значения каждого из предикторов нормально распределены.

В табл. 4.1 приведены основные характеристики переменных, которые могут участвовать в дискриминантном анализе.

Таблица 4.1

Основные характеристики переменных, участвующих в дискриминантном анализе

Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Номинативная или порядковая	Любое	Любой

При выполнении дискриминантного анализа в SPSS выводится, как правило, большой объем вспомогательной информации, позволяющей контролировать степень применимости метода для изучаемого набора данных. Разобраться в этом материале для новичка крайне сложно. В этих условиях неоценимую помощь может оказать прекрасная справочная система SPSS (на английском языке), в которой достаточно подробно на конкретном статистическом материале иллюстрируются возможности пакета при использовании той или иной статистической процедуры и обсуждается интерпретация полученных результатов. Чтобы получить пошаговые инструкции по процедуре дискриминантного анализа следует выбрать следующие пункты: Help/Case Studies/base System/Discriminant Analysis (Помощь/Учебные примеры/ Дискриминантный анализ).

Пример 4.1

В файле Пример_4_1.sav собраны кредитные истории 700 клиентов банка и данные еще о 150 клиентах, которые намерены обратиться в банк за кредитом. Информация о клиентах банка содержит следующие данные:

- возраст (числовая переменная);
- образование (номинативная переменная);
- стаж коммерческой деятельности;
- число полных лет постоянного места проживания;
- годовой доход (в тыс. долларов США);
- долги (в процентах к годовому доходу);
- долги по кредитной карте (в тыс. долларов США);
- другие долги (в тыс. долларов США);
- для клиентов, уже бравших кредит, данные о том, являлся ли он должником банка.

Требуется, используя 70%-ю случайную выборку клиентов, уже бравших кредит в банке, создать модель дискриминантного анализа, позволяющую предсказать невозвращение кредита клиентом. Используя оставшиеся 30 % случаев (проверочная совокупность), выяснить степень пригодности построенной модели для предсказания случаев невозвращения кредита клиентами банка. Выяснить, какие переменные могут быть без ущерба исключены из модели.

Решение

Отметим, что эту задачу мы уже рассматривали в примере 3.6, используя метод логистической регрессии. Здесь для достижения той же цели – определения клиентов банка, для которых риск невозврата платежа слишком велик, применим дискриминантный анализ.

После загрузки данных следует отобрать 70 % случаев из 700 для построения базы знаний, на основании которой будет построена модель экспертной системы (если вы затрудняетесь выполнить эту процедуру, вернитесь к примеру 3.6). Оставшиеся 30 % случаев будем использовать как проверочную выборку.

Для запуска процедуры дискриминантного анализа выберем пункты меню Analyze/Classify/Discriminant (Анализ/Классификация/ Дискриминантный анализ). В открывшемся окне разместим переменную Должник в окно Grouping Variable (Группирующая переменная), переменную Выборка в окно Selection Variable (Переменная отбора). Все оставшиеся переменные поместим в окно Independents (Независимые переменные). Размещение переменных в окнах установки параметров дискриминантного анализа показано на рис. 4.2.

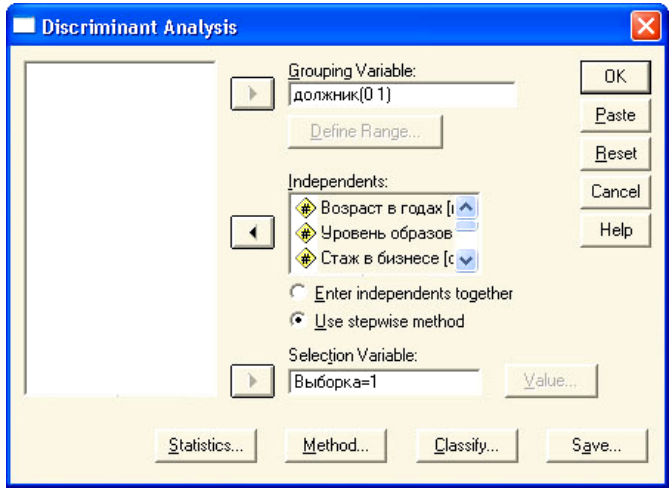


Рис. 4.2. Окно выбора параметров модели дискриминантного анализа

После того как все переменные размещены в нужных окнах, следует задать диапазон изменения переменной в окне Grouping Variable. Для этого нужно выделить имя переменной в окне и нажать кнопку Define Range (Определить область изменения) и в открывшемся окне установить минимальное значение «0» и максимальное «1». Затем нужно определить, какие из случаев будут участвовать в построении модели. Для этого выделим имя переменной в окне Selection Variable и нажмем кнопку Value. В результате появится небольшая панель, в окне которой нужно ввести цифру «1» (в создании модели будут участвовать только случаи, для которых значение переменной Выборка = 1.

Определим другие параметры анализа, последовательно открывая имеющиеся на панели дискриминантного анализа закладки. Нажмем кнопку Statistics (Статистики). В результате откроется окно, изображенное на рис. 4.3, в котором можно заказать получение дополнительной информации в окне вывода результатов.

Установка флажков в группе Discriptives (Описательные) обеспечивает:

- Means (Средние) – вывод средних значений и стандартных отклонений для каждой переменной в каждой группе (эта информация помогает визуально определить переменные, средние значения которых сильно различаются для различных групп);
- Univariate Anova (Однофакторный дисперсионный анализ) – вывод информации о том, значимо ли различаются между собой средние

для групп по каждой переменной (приводится значение статистики Фишера);

– Box′s M (Статистика Бокса) – выводится информация о результатах проверки исходной гипотезы о том, что ковариационные матрицы для различных групп одинаковы. Гипотеза проверяется на основании F -критерия Фишера. Если значимость M -статистики оказалась меньше 0,05, то это свидетельствует о нарушении предположения, что данные, относящиеся к разным группам, имеют близкие многомерные нормальные распределения. При большом объеме данных M -статистика неоправданно завышает степень неодинаковости многомерных нормальных распределений в группах.

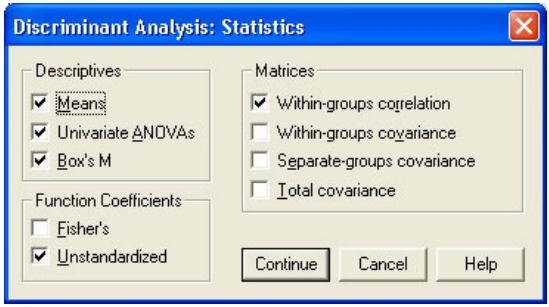


Рис. 4.3. Диалоговое окно выбора параметров Statistics

В группе Function Coefficient (Коэффициенты функции) следует установить флажок в поле Unstandardized (Нестандартизованные), что обеспечит вывод нестандартизованных коэффициентов дискриминантной функции. При желании эти коэффициенты позволяют вычислить значение дискриминантной функции вручную для любого нового наблюдения.

Наконец, в группе Matrices (Матрицы) можно поставить флажок в поле Within-groups correlation (Внутригрупповые корреляции), что позволяет получить информацию о взаимосвязи различных переменных между собой. Далее следует нажать кнопку Continue (Продолжить) и перейти на закладку Method.

На этой закладке все параметры можно оставить по умолчанию. По умолчанию выбран пошаговый метод составления дискриминантной функции, основанный на статистике Λ Уилкса (Wilk′s Lambda). Лямбда Уилкса оценивает отношение внутригрупповой суммы квадратов к общей сумме квадратов:

$$\Lambda \approx \frac{\overline{\sigma^2}}{\delta^2 + \overline{\sigma^2}}.$$

Поэтому эта величина равна единице, если группы не различаются между собой, и уменьшается с ростом доли межгрупповой вариации.

Критерием включения/исключения переменной из модели при построении дискриминантной функции является F -статистика Фишера, которая позволяет выяснить, значимо ли различаются средние значения переменной для разных групп. Если значение F -статистики для переменной больше 3,84, она включается в модель, а если это значение меньше 2,71, она исключается из модели.

Рассмотрим теперь значения параметров, которые можно установить на закладке Classify (Классифицировать), изображенной на рис. 4.4.

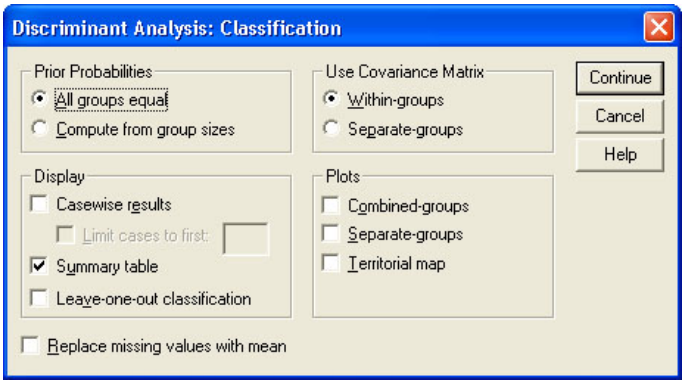


Рис. 4. 4. Выбор параметров дискриминантного анализа на закладке Classification (Классификация)

В группе Prior Probabilities (Априорные вероятности) имеется два переключателя, которые позволяют установить равные априорные вероятности и априорные вероятности, пропорциональные числу объектов в группе. В анализируемом наборе данных в построении модели участвуют 466 случаев. Из них в 342 случаях кредит возвращался вовремя и в 124 случаях имелись задержки возвращения кредита. Как видно, объемы групп сильно различаются, и для более правильного определения вероятности принадлежности объекта к группе следует установить флажок в окне Compute from Group Size (Вычисление исходя из объема группы). Следует сказать, что выбор способа вычисления априорных вероятностей никак не сказывается на вычислении коэффициентов и самих значений дис-

криминантной функции, но оказывает существенное влияние на принадлежность объекта к тому или иному классу.

Группа переключателей Use Covariance Matrix (Использование ковариационной матрицы) позволяет выбрать одну из двух возможностей. Если выполняется предположение о многомерной нормальности распределения для анализируемой совокупности данных, то переключатель следует установить в положение Within - Groups (Внутри групп). Это положение выбрано по умолчанию. Если предположение о многомерной нормальности не выполняется, то переключатель нужно установить в положение Separate Groups (Раздельные группы). Судить о том, выполняются или не выполняются предположения о многомерной нормальности исходного набора данных, можно на основании М-статистики Бокса. Остальные переключатели на этой закладке влияют только на объем выводимой информации. Поэтому мы рекомендуем поставить флажок только в поле Summary Table (Итоговая таблица), что обеспечит вывод итоговой информации о числе правильно и неправильно классифицированных объектов.

Наконец, на закладке Save (Сохранить) следует поставить флажок в поле Predicted group Membership (Предсказываемое членство в группе).

После того как необходимые параметры дискриминантного анализа заданы, можно запустить процедуру на исполнение. Объем выводимой информации достаточно велик, и поэтому мы ограничимся лишь рассмотрением тех результатов, которые существенным образом определяют качество модели дискриминантного анализа. В табл. 4.2 представлены данные, позволяющие определить способность переменных делить объекты на группы.

Таблица 4.2

Критерий равенства групповых средних

Переменная	Лямбда Уилкса	*F*	Ст. св1	Ст. св2	Знач.
Возраст, лет	0,986	6,424	1	464	0,012
Уровень образования	0,980	9,372	1	464	0,002
Стаж в бизнесе, лет	0,907	47,809	1	464	0,000
Не менял адрес, лет	0,984	7,364	1	464	0,007
Годовой доход, тыс. долл.	0,997	1,267	1	464	0,261
Отношение долга к доходу, %	0,849	82,594	1	464	0,000
Долги по кредитной карте, тыс. долл.	0,951	24,096	1	464	0,000
Другие долги, тыс. долл.	0,980	9,528	1	464	0,002

В этой таблице представлены все переменные, характеризующие объекты, значение критерия Фишера при проверке гипотезы о равенстве группо-

вых средних, и соответствующие уровни значимости. Сразу видно, что переменная «Годовой доход, тыс. долл.» должна быть исключена из анализа, поскольку ее средние значения для различных групп различаются незначимо.

Дополнительную информацию о способности переменной дискриминировать объекты дает значение статистики лямбда Уилкса. Чем меньше значение этой величины, тем лучше переменная способна делить объекты на группы.

Следующий блок информации, на который следует обратить внимание – это критерий Бокса равенства ковариационных матриц, представленный в табл. 4.3.

Таблица 4.3

Результаты теста

М Бокса		236,852
F-критерий	Приблизительно	39,094
	Ст. св1	6
	Ст. св2	338440
	Знач.	0,000

Как следует из таблицы, проверка нулевой гипотезы о равенстве ковариационных матриц для групп дала отрицательный результат: предположение о том, что все анализируемые объекты можно описать одним многомерным нормальным распределением, не нашло подтверждения. Это означает, что процедуру дискриминантного анализа нужно повторить, установив на закладке Classify (см. рис. 4.4) в группе переключателей Use Covariance Matrix значение Separate Groups (Раздельные группы).

Дискриминантными переменными являются лишь переменные, включенные в анализ на последнем (третьем) шаге. Эти переменные представлены в табл. 4.4.

Таблица 4.4

Переменные, включенные в анализ на шаге 3

Переменная	Толерантность	*F*- исключения	Лямбда Уилкса
Отношение долга к доходу, %	0,820	26,769	0,763
Стаж в бизнесе	0,715	79,884	0,846
Долги по кредитной карте	0,608	31,576	0,770

Остальные переменные из анализа исключены из-за того, что *F*-статистика включения для этих переменных оказалась меньше порогового зна-

чения, равного 3,84. Эти данные можно найти в таблице с названием «Переменные, не включенные в анализ» (мы не приводим здесь данные этой таблицы, но ее можно увидеть в окне вывода результатов SPSS).

Для практических целей важное значение имеют ненормированные значения дискриминантной функции и ее значения в центроидах групп (результаты приведены в табл. 4.5, 4.6).

Таблица 4.5

Коэффициенты канонической дискриминантной функции

Переменная	Коэффициент
Стаж в бизнесе	–0,136
Отношение долга к доходу, %	0,078
Долги по кредитной карте, тыс. долл.	0,297
Константа	–0,152

Таблица 4.6

Значение дискриминантной функции в центроидах групп

Задерживал возврат кредита ранее	Функция
Нет	–0,374
Да	1,031

Для каждого случая, пользуясь коэффициентами из табл. 4.5, можно рассчитать численное значение дискриминантной функции (эти значения рассчитываются автоматически, если на закладке Save поставить флажок в поле Discriminant scores (Подсчет значений дискриминантной функции). Если значение дискриминантной функции для i-го клиента банка больше полусуммы значений дискриминантной функции в центроидах

$$y_i > c = \frac{1}{2} \cdot (-0{,}374 + 1{,}031),$$

то этот клиент будет отнесен к группе с повышенным риском невозврата кредита, а если

$$y_i < c = \frac{1}{2} \cdot (-0{,}374 + 1{,}031),$$

то этот клиент не входит в группу риска и ему можно смело выдавать кредит.

Общие итоги дискриминантного анализа можно увидеть в итоговой таблице окна вывода информации SPSS с названием «Результаты классификации». Мы не приводим эту таблицу полностью. Для нас важен лишь результат: 76,2% выбранных исходных наблюдений классифицировано правильно; 75,2% невыбранных исходных наблюдений классифицировано правильно. Хотя этот результат и не дает 100%-й гарантии от ошибки при выдаче кредита, но все же снижение числа случаев невозврата кредита даже на 70 % сулит банку существенное увеличение прибыли, покрывающее издержки проведенного исследования. Важно, что база знаний с годами только увеличивается, что, естественно, дает надежду на улучшение прогноза невозврата кредита в будущем.

4.2. Факторный анализ

Факторный анализ – это процедура, с помощью которой большое число переменных, характеризующих имеющиеся наблюдения, сводится к меньшему количеству независимых величин, называемых факторами. К факторному анализу обычно прибегают тогда, когда пытаются определить скрытые (латентные) переменные, которые в действительности определяют изучаемое явление. Очень часто такие факторы или неизвестны заранее, или не поддаются непосредственному измерению. Факторный анализ можно использовать и при построении регрессионного уравнения. Если исходные переменные сильно взаимосвязаны, то в качестве объясняющих переменных можно использовать факторы, полученные в результате факторного анализа. Факторы по построению являются ортогональными, так что проблема мультиколлинеарности при этом будет полностью решена.

В последние годы факторный анализ приобрел значительную популярность в психологических и социальных исследованиях. Во многом этому способствовала разработка Раймондом Кеттелем знаменитого 16-факторного личностного опросника. Именно при помощи факторного анализа ему удалось свести около 4500 наименований личностных особенностей к 187 вопросам, которые в, свою очередь, позволяют измерить 16 различных свойств личности (познакомиться с вопросами и пройти тестирование можно, например, на сайте http://www.bitnet.ru/mirrors/students.ru/kettel-test.html).

Задачей факторного анализа является определение новых переменных f_l, $l = 1, 2, \ldots, m$ (факторов), через которые исходные переменные выражаются с помощью соотношений линейной связи:

$$x_j = a_{j1} \cdot f_1 + a_{j2} \cdot f_2 + \ldots + a_{jm} \cdot f_m \ ; \ j = 1, 2, \ldots, k \ . \qquad (4.7)$$

Здесь k – число исходных переменных, m – число факторов. Коэффициенты a_{jm} называются факторными нагрузками. Факторную нагрузку a_{jm} следует понимать как коэффициент корреляции между переменной x_j и фактором f_m. В общем случае число факторов должно быть меньше числа исходных переменных: $k \le m$. В уменьшении числа факторов и состоит одна из задач факторного анализа. Первоначально при проведении факторного анализа в SPSS число факторов всегда равно числу исходных переменных, и пользователь может по своему усмотрению задать алгоритм отбора факторов. Обычно отбираются те факторы, каждый из которых способен объяснить наибольшую долю дисперсии. Суммарно оставленные факторы должны объяснять 70 – 80 % дисперсии первоначальных. Идеи факторного анализа можно проиллюстрировать графически. Предположим, что n наблюдаемых объектов (автомобилей) оцениваются в двумерном признаковом пространстве с координатными осями x_1 – стоимость автомобиля и x_2 – длительность рабочего ресурса двигателя (рис 4.5).

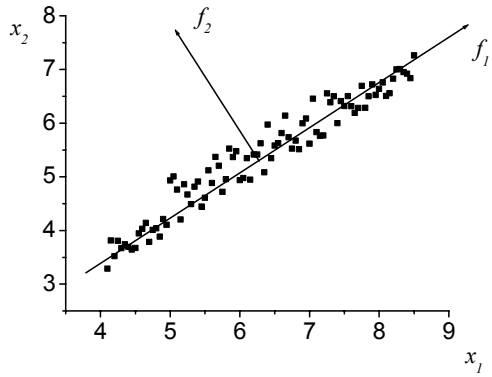


Рис.4.5. Графическая интерпретация идей факторного анализа

При условии коррелированности x_1 и x_2 появляется направленное скопление точек. Те же самые объекты можно описать, выбрав новые координатные оси (факторы f_1 и f_2). Особенностью новых осей является то, что они проходят через плотное скопление точек и лучше описывают дисперсию свойств объектов. Непосредственно из графика

на рис. 4.5 видно, что большую часть дисперсии свойств автомобилей описывает фактор f_1.

После того как факторы найдены, их необходимо интерпретировать. Далеко не всегда эта интерпретация оказывается очевидной. При интерпретации факторов приходится анализировать факторные нагрузки и на основании этих данных пытаться интерпретировать смысл факторов. В рассматриваемом случае с возрастанием фактора f_1 будет расти как стоимость автомобиля, так и длительность ресурса двигателя. Этот фактор можно ассоциировать с надежностью автомобиля.

С ростом второго фактора длительность ресурса двигателя будет также возрастать, а стоимость автомобиля – падать. Поэтому фактор f_2 можно назвать степенью использования проверенных технических решений при конструировании автомобиля. При увеличении доли хорошо зарекомендовавших себя технических решений стоимость может падать с одновременным ростом ресурса двигателя.

Существует большое число различных алгоритмов факторного анализа. Здесь мы кратко рассмотрим лишь сущность метода главных компонент, который является, несомненно, наиболее простым и универсальным методом.

Метод главных компонент позволяет для исходного числа m объясняющих переменных найти m взаимно ортогональных главных компонент. Модель исходит из предположения, что факторы f_l могут быть выражены в виде линейной комбинации исходных переменных x_i:

$$f_l = b_{1l} \cdot x_1 + b_{2l} \cdot x_2 + \ldots + b_{ml} \cdot x_m, \quad l = 1, 2, \ldots, m \qquad (4.8)$$

Уравнение (4.8) определяет линейную взаимосвязь факторов и исходных переменных. Для дальнейших целей удобно перейти к матричной записи. Определим вектор-строку $\mathbf{F}$ как совокупность величин $f_1, f_2, \ldots, f_m$ и $\mathbf{X}$ как совокупность величин $x_1, x_2, \ldots, x_m$ и матрицу $\mathbf{B}$ коэффициентов b_{ij}. В матричных обозначениях уравнение (4.8) имеет очень простой вид: $\mathbf{F} = \mathbf{X} \cdot \mathbf{B}$.

Матрицу коэффициентов $\mathbf{B}$ будем определять из условия, что оценки дисперсии для многомерных случайных величин $\mathbf{X}$ и $\mathbf{F}$ должны быть равны, т. е. $D(\mathbf{X}) = D(\mathbf{X} \cdot \mathbf{B})$. Это условие, по существу, является предположением, что факторы можно определить таким образом, что они полностью объясняют дисперсию исходных переменных.

Поскольку $\mathbf{X}$ – многомерная случайная величина, то ее дисперсия совпадает с ковариационной матрицей $\mathbf{S}$: $D(\mathbf{X}) = \mathbf{S}$.

Поиск главных компонент сводится к последовательному выделению факторов $f_1, f_2, \ldots, f_m$, каждый из которых способен объяснить некоторую долю дисперсии факторных признаков $\mathbf{X}$. Следовательно, задачу можно поставить и таким образом: следует найти последовательным образом факторы $\mathbf{F}_l = \mathbf{X} \cdot \mathbf{B}_l$, для которых

$$D(\mathbf{X}) - D(\mathbf{X} \cdot \mathbf{B}_l) \to \min . \tag{4.9}$$

В формуле (4.9) величина $\mathbf{B}_l$ представляет собой вектор-столбец (один из столбцов матрицы $\mathbf{B}$). Дисперсия произведения многомерного случайного вектора $\mathbf{X}$ на постоянные коэффициенты $\mathbf{B}_l$ вычисляются по формуле $D(\mathbf{X} \cdot \mathbf{B}_l) = \mathbf{B}_l^T \cdot \mathbf{S} \cdot \mathbf{B}_l$.

Будем требовать, чтобы вектора $\mathbf{B}_l$ были нормированы на единицу:

$$\mathbf{B}_l^T \cdot \mathbf{B}_l = b_{l1}^2 + b_{l2}^2 + \ldots + b_{lm}^2 = 1 . \tag{4.10}$$

Вектор $\mathbf{B}_l$ определим из условия (4.9) при выполнении ограничения (4.10). Эта задача является типичной задачей на поиск условного экстремума, и она решается методом неопределенных множителей Лагранжа. Составляем функцию Лагранжа L:

$$L = \mathbf{B}_l^T \cdot \mathbf{S} \cdot \mathbf{B}_l - \mathbf{S} - \lambda \cdot \mathbf{E} \cdot (\mathbf{B}_l^T \cdot \mathbf{B}_l - 1) . \tag{4.11}$$

В этой формуле $\mathbf{E}-$ единичная матрица. Неизвестные величины $\mathbf{B}_l$ и множитель Лагранжа λ находим из условия экстремума функции (4.11)

$$\frac{\partial L}{\partial \mathbf{B}_l} = 2 \cdot \mathbf{S} \cdot \mathbf{B}_l - 2 \cdot \lambda \cdot \mathbf{E} \cdot \mathbf{B}_l = 0; \quad \frac{\partial L}{\partial \lambda} = \mathbf{B}_l^T \cdot \mathbf{B}_l = 1 . \tag{4.12}$$

Из уравнения (4.12) следует, что величины $\mathbf{B}_l$ являются собственными векторами матричного уравнения

$$(\mathbf{S} - \lambda \cdot \mathbf{E}) \cdot \mathbf{B}_l = 0 . \tag{4.13}$$

Уравнение (4.13) имеет однозначное решение, если выполняется второе из уравнений (4.12) и определитель матрицы системы однородных уравнений (4.13) равен нулю:

$$\left| \mathbf{S} - \lambda \cdot \mathbf{E} \right| = 0 . \tag{4.14}$$

Это уравнение называется характеристическим уравнением, а множители Лагранжа являются собственными значениями характеристического уравнения.

Таким образом, задача нахождения уравнения взаимосвязи исходных объясняющих переменных и факторов сводится к поиску собст-

венных векторов уравнения (4.13). Отметим, что по построению система собственных векторов будет ортонормированной.

Если исходные данные предварительно стандартизовать, т. е. вместо величин x_{il} ввести стандартные отклонения z_{il}:

$$z_{il} = \frac{(x_{il} - \bar{x}_i)}{\sigma_i}, \quad i = 1, 2, \ldots, n; \quad l = 1, 2, \ldots, m, \tag{4.15}$$

то для этих переменных ковариационная матрица тождественно совпадает с обычной корреляционной матрицей $\mathbf{R}$. Следовательно, в этом случае величины $\mathbf{B}_l$ можно найти, решая уравнения

$$(\mathbf{R} - \lambda \cdot \mathbf{E}) \cdot \mathbf{B}_l = 0, \qquad \left|\mathbf{R} - \lambda \cdot \mathbf{E}\right| = 0. \tag{4.16}$$

В дальнейшем будем предполагать, что переход к стандартизованным переменным (4.15) выполнен, и вместо ковариационной матрицы будем использовать корреляционную матрицу $\mathbf{R}$.

Размерность матриц в матричных уравнениях (4.16) совпадает с числом объясняющих переменных m. Поэтому второе из уравнений (4.16) имеет m собственных значений λ_i, $i = 1, 2, \ldots, m$. Для каждого значения λ решение первого из уравнений (4.16) дает собственный вектор $\mathbf{B}_l$, компоненты которого $b_{l1}, b_{l2}, \ldots, b_{lm}$, как это следует из уравнения (4.8), являются характеристиками силы связи l-го фактора и объясняющих переменных $x_1, x_2, \ldots, x_m$. Таким образом, мы имеем схему, позволяющую получить все коэффициенты b_{lj}, выражающие взаимосвязь факторов f_l и объясняющих переменных x_j. Естественно, после того как найдены все коэффициенты b_{lj}, можно разрешить систему уравнений (4.8) относительно x_j и выразить независимые переменные через факторы f_l, определив тем самым коэффициенты a_{jl} – так называемые факторные нагрузки.

В действительности на практике представляет интерес использование только тех факторов, которые объясняют наибольшую долю дисперсии объясняющих переменных (обычно считается приемлемым, если факторы объясняют 70 – 80 % дисперсии). В этом случае число используемых факторов будет меньше числа объясняющих переменных и мы получаем возможность найти так называемые скрытые факторы, через которые выражаются все объясняющие переменные x_j.

Дальнейшие детали выполнения факторного анализа мы поясним на простом примере.

Пример 4.2

В таблице представлены данные статистического наблюдения за экологической обстановкой в семи городах Урала с различным уровнем техногенной нагрузки и уровень заболеваемости злокачественными новообразованиями (чел. на 1000 жителей).

Заболеваемость	Пыль	Сера	CO2	NO2	Сброс
3,6	0,14	0,005	1,6	0,02	7578
1,19	0,1	0,004	1,2	0,04	87474
2,87	0,25	0,005	2,4	0,05	38496
5,4	0,27	0,01	1,7	0,04	329000
0,47	0,22	0,07	3	0,08	7200
5,6	0,16	0,012	1,8	0,06	1093102
2,54	0,21	0,03	1,1	0,04	8212

В столбце «Пыль» приведены данные о средней концентрации загрязняющих веществ в атмосферном воздухе (мг/м^3). Столбцы «Сера», «CO2», «NO2» содержат данные о содержании сернистого ангидрида, окиси углерода, двуокиси азота в атмосфере. Переменная «Сброс» содержит данные об объеме промышленных сбросов (м^3).

Считая переменные Пыль, Сера, CO2, NO2, Сброс объясняющими, построить, используя метод главных компонент, матрицу факторных нагрузок. Выделить главные факторы, определяющие экологическую обстановку в городе. Найти численные значения главных компонент для всех объектов наблюдения. Исходные данные содержатся в файлах Пример_4_2.xls и Пример_4_2.sav.

Все вычисления выполнить «вручную», используя лишь электронные таблицы Excel и пакет математических вычислений Maple. Проверить полученные результаты, проведя аналогичные вычисления в SPSS.

Построить регрессионные уравнения, объясняющие заболеваемость в городах действием двух-трех главных факторов.

Решение

В качестве первого шага, пользуясь формулой (4.15), перейдем к нормированным переменным z_{il}. Для каждой переменной нужно сначала вычислить среднее значение и оценку дисперсии по выборке. Для вычисления дисперсии по выборке в электронных таблицах Excel следует использовать функцию ДИСП, а не ДИСПР, а затем провести вычисление по формуле (4.15). При использовании пакета SPSS можно перейти к нормированным переменным, используя пункты меню Analyze/Descriptive Statistics/ Descriptives. Затем в открывшемся окне следует все объясняющие переменные перенести в правую часть окна и поставить флажок в окошке Save standardized values as variables (Сохранить стандартизованные значения как переменные). После запуска процедуры на исполнение в окне редактора возникнут новые переменные, имена которых отличаются от имен нестандартизованных переменных наличием первой буквы Z.

Матрицу коэффициентов корреляции стандартизованных объясняющих переменных можно получить, используя функцию КОРРЕЛ электронных таблиц Excel. Результаты приведены в табл. 4.6.

Таблица 4.6

Корреляционная матрица исходных переменных

	Пыль	Сера	CO2	NO2	Сброс
Пыль	1,000	0,275	0,458	0,295	–0,112
Сера	0,275	1,000	0,597	0,741	–0,222
CO2	0,458	0,597	1,000	0,745	–0,066
NO2	0,295	0,741	0,745	1,000	0,246
Сброс	–0,112	–0,222	–0,066	0,246	1,000

Для проверки результатов, полученных «вручную», можно использовать процедуру факторного анализа.

Опишем сразу всю процедуру установки параметров факторного анализа, а затем вновь вернемся к пошаговым вычислениям «вручную». После загрузки файла Пример_4_2.sav в окно редактора SPSS следует выбрать пункты меню Analyze/Data Reduction/Factor (Анализ/Уменьшение размерности/Факторный анализ). В результате появится окно, вид которого приведен на рис. 4.6.

В этом окне объясняющие переменные следует перенести в правое окошко с названием Variables (Переменные).

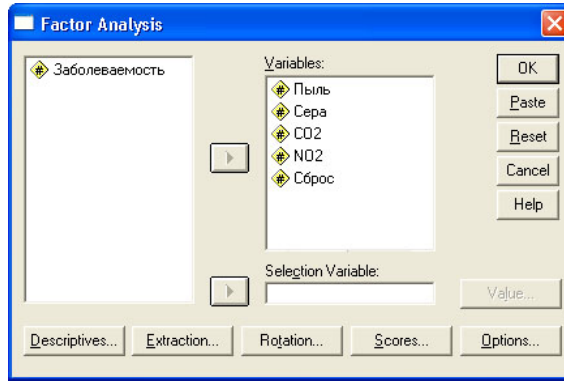


Рис. 4.6. Окно выбора параметров факторного анализа

Затем следует нажать кнопку Descriptives (Описательные). В этом окне, изображенном на рис. 4.7, можно «заказать» отображение в окне вывода некоторых промежуточных результатов вычисления, в частности значений коэффициентов корреляционной матрицы. Мы рекомендуем поставить флажки так, как это изображено на рис. 4.7. Нажав кнопку Continue (Продолжить), вернемся в окно, изображенное на рис. 4.6 и продолжим выбор параметров для проведения факторного анализа.

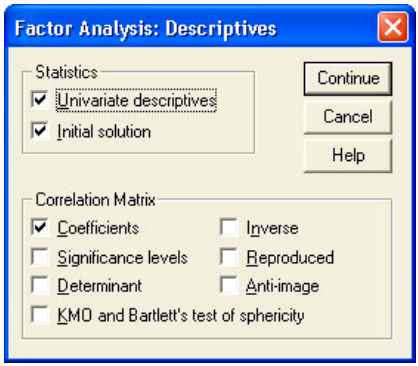


Рис. 4.7. Выбор параметров описательной статистики

Выберем теперь метод извлечения факторов и число факторов, которые следует определить. Для установки этих параметров нажмем кнопку Extraction (Выделение) и в открывшемся окне выберем параметры, отмеченные флажками на рис. 4.8.

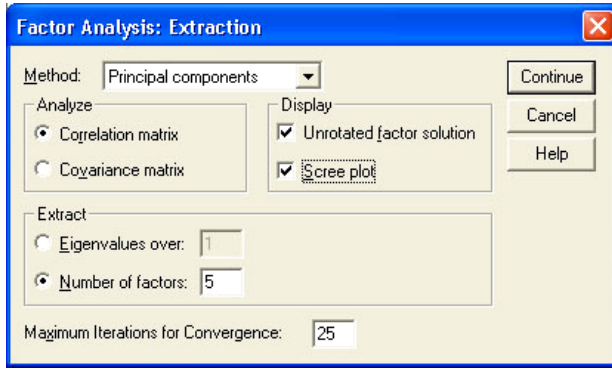


Рис. 4.8. Окно выбора параметров выделения факторов

Для выделения факторов по умолчанию установлен метод главных компонент (Principal components). Эту установку следует оставить без изменения. Во фрейме Extract (Выделение) переключатель следует перевести в положение Number of factors (Число факторов) и поставить в окошке справа значение «5», поскольку мы желаем выделить все факторы, даже если они и окажутся малосущественными. Если оставить значения в этом фрейме по умолчанию, то будут выделены только факторы, для которых собственное значение больше единицы.

Наконец, во фрейме Display (Отображение) следует поставить флажки в имеющихся там двух окошках, что обеспечит нам вывод исходных неповернутых факторов и диаграмму Scree plot (Диаграмма каменистой осыпи), на которой отображаются собственные значения факторов (название связано с тем, что диаграмма чисто визуально напоминает профиль каменистой осыпи).

Следующим шагом после выделения факторов является их вращение. Вращение требуется потому, что изначально полученные факторы, будучи совершенно корректными с математической точки зрения, неудобны для интерпретации. Задачей вращения является такое преобразование факторов, чтобы структура факторных нагрузок была более простой и проще поддавалась интерпретации. В идеале желательно сделать так, чтобы для каждой объясняющей переменной был один фактор, для которого факторная нагрузка была бы большой по абсолютной величине (факторные нагрузки изменяются в пределах от –1 до +1), а все остальные факторные нагрузки были бы близки к нулю. Наиболее популярным методом вращений является метод Varimax. Для установки этого метода вращения нужно нажать кнопу Rotation (Вращение) и в появившемся окне

отметить флажком опцию Varimax. Все остальные параметры в этом окне можно оставить по умолчанию.

Чтобы отобразить численные значения факторов для всех анализируемых объектов, следует нажать кнопку Scores (Подсчет значений) и в появившемся окне поставить флажок в окошке Save as variables (Сохранить как переменные). При запуске процедуры факторного анализа это обеспечит появление в окне редактора данных SPSS пяти новых переменных FACT1, FACT2, FACT3, FACT4, FACT5, которые представляют собой численные значения всех пяти выделенных факторов. Напомним, что обычно выделять все факторы не следует, нужно ограничиться лишь теми факторами, которые объясняют 70 – 80 % дисперсии факторных переменных.

На закладке Option (Опции) можно задать правила работы с пропущенными значениями. Здесь все настройки можно оставить по умолчанию.

После задания всех параметров факторного анализа можно запустить процедуру на исполнение. В результате появится окно выдачи результатов, которые мы собираемся сравнить с результатами, полученными «вручную».

Во-первых, легко убедиться, что корреляционная матрица, вычисленная в SPSS, точно совпадает с матрицей, приведенной в табл. 4.6.

Вычислим теперь собственные значения и собственные вектора корреляционной матрицы. Для этого удобнее воспользоваться пакетом математических вычислений Maple. Ниже мы приведем команды, которые нужно выполнить в среде Maple, и результаты, которые получаются при выполнении этих команд. Maple – это модульный пакет. Поэтому нужно вначале подключить пакет, позволяющий производить матричные вычисления. Для этого нужно выполнить команду

```
> with(linalg);.
```

Затем нужно выполнить команду формирования корреляционной матрицы:

```
> R:=matrix([[1,0.274915899,0.458224311,0.294676383,-0.11183599],
[0.274915899,1,0.597090068,0.740501357,-0.221618665],
[0.458224311,0.597090068,1,0.744524018,-0.065790368],
[0.294676383,0.740501357,0.744524018,1,0.246112759],
[-0.11183599,-0.221618665,-0.065790368,0.246112759,1]]);
```

Собственные значения этой матрицы можно получить, используя команду

```
> L:=eigenvalues(R);
```

В результате будут выведены на экран собственные значения корреляционной матрицы **R** :

L= 0.08998103834, 0.3443877439, 0.8005047714, 1.154359089, 2.610767358.

Эти значения точно совпадают с теми, что выдает программа SPSS. В табл. 4.7 приведены результаты расчета собственных значений в SPSS. Легко убедиться, что собственные значения одинаковы. Следует обратить внимание на то, что сумма собственных значений определяется размерно-

стью матрицы **R** и в данном случае равна пяти. Доля дисперсии, которую объясняет тот или иной фактор f_i, пропорциональна собственному значению и вычисляется по формуле

$$\sigma_i^2 = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}.$$

Таблица 4.7

Полная объясненная дисперсия

Компонента	Начальные собственные значения		
	Всего	% дисперсии	Кумулятивный %
1	2,611	52,215	52,215
2	1,154	23,087	75,303
3	0,801	16,010	91,313
4	0,344	6,888	98,200
5	0,090	1,800	100,000

Для получения пяти собственных векторов, соответствующих собственным значениям, необходимо выполнить команду

> V:=eigenvectors(R);

В результате будет получен массив из пяти собственных векторов. Составим матрицу собственных векторов, каждый из столбцов которой является собственным вектором. Расположим собственные вектора так, чтобы в первом столбце размещался собственный вектор, соответствующий наибольшему собственному значению, во втором столбце – следующему по величине, и т. д.

$$\mathbf{B} = \begin{pmatrix} -0.300 & -0.551 & -0.764 & -0,058 & -0.139 \\ 0.031 & -0.474 & 0.334 & 0.471 & -0.336 \\ -0.247 & -0.129 & -0.001 & 0.335 & 0.900 \\ 0.852 & 0.428 & 0.566 & 0.176 & 0.238 \\ -0.348 & 0.521 & -0.549 & 0.552 & 0.340 \end{pmatrix} \qquad (4.17)$$

Полученная матрица позволяет найти факторы в соответствии с формулой (4.8). Нам же нужно найти факторные нагрузки a_{ij}, позволяющие выразить исходные переменные через факторы в соответствии с уравнением (4.7).

Матрица **B** является ортонормированной и удовлетворяет условию $\mathbf{B}^T \cdot \mathbf{B} = \mathbf{E}$. Матрица факторных нагрузок **A** обычно определяется таким

образом, чтобы $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{\Lambda}$, где $\mathbf{\Lambda}$ – диагональная матрица, на главной диагонали которой расположены собственные значения матрицы коэффициентов корреляции. Очевидно, что матрица $\mathbf{A}$ определяется произведением матрицы $\mathbf{B}$ и матрицы

$$\sqrt{\mathbf{\Lambda}} = \begin{pmatrix} \sqrt{2.610} & & & & \\ & \sqrt{1.115} & & & \\ & & \sqrt{0.800} & & \\ & & & \sqrt{0.344} & \\ & & & & \sqrt{0.090} \end{pmatrix}$$

В результате перемножения получаем матрицу факторных нагрузок, которая позволяет выразить исходные переменные через выделенные факторы. В табл. 4.8 приведены факторные нагрузки a_{ij}, выражающие исходные переменные через выделенные факторы.

Таблица 4.8

Факторные нагрузки

Переменные	Факторы				
	1	2	3	4	5
Пыль	0,563	–0,265	0,763	0,176	–0,009
Сера	0,842	–0,139	–0,383	0,323	0,142
CO2	0,887	–0,001	0,051	–0,448	0,100
NO2	0,892	0,360	–0,157	0,034	–0,222
Сброс	–0,055	0,967	0,213	0,081	0,101

Приведенные в таблице 4.8 данные взяты из программы SPSS. Простой проверкой можно убедиться, что результаты вычисления «вручную» совпадают с приведенными результатами.

Полученная таблица факторных нагрузок совершенно правильная, но она не позволяет интерпретировать факторные нагрузки. Поэтому следует рассмотреть результат вращения факторов. С геометрической точки зрения вращение факторов представляет собой поворот системы координат в факторном пространстве. Результаты вращения факторов, которые дает программа SPSS, представлены в табл. 4.9. Теперь хорошо видно, что фактор 1 связан с наличием в атмосфере серы и двуокиси азота. Второй фактор фактически описывает влияние окиси углерода, третий фактор – влияние сброса промышленных отходов, четвертый характеризует запыленность атмосферы, а пятый фактор сразу можно отбросить, поскольку важность того иди иного фактора для анализа определяется соответствую-

щим собственным значением, которое пропорционально доле дисперсии результативных признаков, которую объясняет этот фактор.

Таблица 4.9

Матрица повернутых факторов

Переменные	Факторы				
	1	2	3	4	5
Пыль	0,111	0,195	–0,054	0,973	0,026
Сера	0,943	0,268	–0,159	0,113	0,006
CO2	0,340	0,905	–0,034	0,248	0,050
NO2	0,664	0,525	0,276	0,126	0,438
Сброс	–0,061	0,000	0,996	–0,054	0,043

Наиболее важным является первый фактор, т. е. содержание серы и двуокиси азота. Далее следует содержание окиси углерода и сброс промышленных отходов. Эти три фактора, как следует из табл. 4.7, объясняют свыше 91 % дисперсии факторов. Фактически это означает, что факторы 4 и 5 можно исключить из рассмотрения.

Для построения регрессионной модели с использованием в качестве объясняющих переменных факторов требуется найти численные значения факторов для всех объектов. Матрица, приведенная в табл. 4. 8, фактически определяет матричное уравнение

$$\mathbf{Z} = \mathbf{A} \cdot \mathbf{F}, \tag{4.18}$$

где $\mathbf{Z}$ – вектор стандартизованных исходных переменных, $\mathbf{A}$ – матрица факторных нагрузок, $\mathbf{F}$ – вектор факторов. Нам нужно разрешить систему (4.18) относительно факторов. Очевидно, что в матричной форме решение имеет вид

$$\mathbf{F} = \mathbf{\Lambda}^{-1} \cdot \mathbf{A}^T \mathbf{Z}. \tag{4.19}$$

Приведем запись уравнения (4.19) в компонентах

$$f_l = \frac{1}{\lambda_l} \cdot \left(a_{l1}^T \cdot z_1 + a_{l2}^T \cdot z_2 + a_{l3}^T \cdot z_3 + a_{l4}^T \cdot z_4 + a_{l5}^T \cdot z_5 \right). \tag{4.20}$$

Непосредственным вычислением легко убедиться, что появившиеся в результате выполнения процедуры факторного анализа в окне редактора данных переменные FACT1, FACT2, FACT3, FACT4, FACT5 как раз вычислены по формуле (4.20). Эти переменные или их часть можно использовать как объясняющие переменные при построении регрессионных уравнений. Как уже отмечалось, достоинством факторов является их ортогональность. Предлагаем читателю самостоятельно сравнить регрессионные модели, используя в качестве объясняющих переменных исходные переменные и выделенные факторы.

В рассмотренном выше примере основное внимание было уделено вычислительным процедурам, а не содержательному анализу данных. Рассмотрим еще один пример применения факторного анализа, взятый из практики социологического исследования общественного мнения.

Пример 4.3

Работникам одного из предприятий были заданы пятнадцать вопросов, касающихся их отношения к временным переселенцам, приезжающим в страну в поисках работы. Ответ на каждый вопрос следовало дать, используя 7-балльную шкалу: от полного несогласия (1) до полного согласия (7). Исходные данные опроса приведены в файле Пример_4_3.sav. Используя факторный анализ исходных данных, требуется выявить основные тенденции в отношении населения к временным переселенцам.

Решение

В рассмотренном выше примере достаточно подробно описаны те настройки, которые нужно использовать при выполнении факторного анализа. При выполнении этого примера нас интересует только матрица факторных нагрузок. Поэтому после загрузки файла в редактор данных SPSS следует открыть окно, изображенное на рис. 4.6, и перенести все переменные в правое окно. Далее нажать кнопку Rotation (Вращение) и выбрать метод вращения Varimax. Все остальные настройки можно оставить по умолчанию. Запустим процедуру факторного анализа на исполнение. В окне вывода результатов в первую очередь обратим внимание на табличку «Полная объясненная дисперсия». Сокращенный вариант этой таблицы приведен ниже (табл. 4.10).

Таблица 4.10

Полная объясненная дисперсия

Фактор	Начальные собственные значения		
	Всего	% дисперсии	Кумулятивный %
1	5,146	34,308	34,308
2	1,945	12,970	47,278
3	1,415	9,433	56,711
4	0,990	6,601	63,312

По умолчанию матрица факторных нагрузок определяется только для факторов с собственным значением больше единицы (собственные значения приведены в колонке «Всего»). В рассматриваемом примере три первых фактора с собственным значением $\lambda > 1$ объясняют 56, 711 % суммар-

ной вариации факторных переменных. Проанализируем матрицу факторных нагрузок только для трех главных факторов. Сразу рассмотрим матрицу повернутых факторных нагрузок, которая приведена в табл. 4.11.

Таблица 4.11

Матрица повернутых факторов

Метки исходных переменных	Факторы		
	1	2	3
Улучшить интеграцию иностранцев	–0,466	0,628	–0,191
Мягче относиться	–0,141	0,657	0,215
Деньги на нужды россиян	0,327	–0,153	0,711
Россия не служба социальной помощи для СНГ	0,533	–0,106	0,394
Необходимо налаживать отношения	–0,362	0,783	0,045
Права беженцев следует ограничить	–0,012	–0,038	0,763
Русские станут меньшинством	0,525	0,036	0,543
Права беженцев нужно охранять	–0,116	0,719	–0,267
Враждебность наносит вред России.	0,026	0,551	–0,088
Сначала нужно создать условия для Россиян	0,252	–0,095	0,685
Россияне также везде являются иностранцами	0,125	0,392	–0,292
Мультикультура означает мультикриминал	0,802	–0,199	0,108
В лодке нет свободных мест	0,685	–0,110	0,465
Иностранцы – вон	0,837	–0,144	–0,025
Интеграция иностранцев – это убийство нации	0,725	–0,048	0,144

Рассмотрим факторные нагрузки для первого фактора. Как следует из приведенной таблицы, наибольшая положительная связь первого фактора имеется с переменными с условными названиями «Иностранцы – вон» (0,837), «Мультикультура означает мультикриминал» (0,802), «Интеграция иностранцев – это убийство нации» (0,725), а наибольшая отрицательная связь – с переменными с условными названиями «Улучшить интеграцию иностранцев» (–0,466), «Необходимо налаживать отношения» (–0,362). Таким образом, первый фактор аккумулирует крайне отрицательную, враждебную позицию к мигрантам.

Второй фактор имеет большую положительную корреляцию с переменными с условными названиями «Улучшить интеграцию иностранцев» (0,628), «Мягче относиться» (0,657), «Необходимо налаживать отношения» (0,783), «Права беженцев нужно охранять» (0, 719). Таким образом,

второй фактор отражает тенденцию к улучшению отношения к мигрантам, их более тесную интеграцию с коренным населением.

Наконец, третий фактор может быть интерпретирован как ощущение угрозы экономическим интересам коренного населения, которая связана с массовой миграцией. Здесь велики положительные значения корреляции с такими переменными, как «Деньги на нужды россиян» (0,711), «Права беженцев следует ограничить» (0,763), «Сначала нужно создать условия для россиян» (0,685).

Естественно, представляет интерес выяснить, какой процент респондентов занимает те позиции, которые характеризуются выделенными факторами. Для такой оценки нужно вывести в окно редактора SPSS численные значения факторов для всех объектов и подсчитать, например, используя опции дескриптивной статистики, для какого количества объектов значение первого фактора больше, скажем, 0,5. Поскольку численные значения факторов приведены в нормализованной форме, то значения первого фактора большее, 0,5 говорят о том, что этот респондент относится крайне отрицательно к мигрантам. Сделанный вывод легко проверить, используя анализируемые данные.

Точно так же следует поступить и с другими факторами. В итоге получается такая оценка: 24 респондента из 90 опрошенных придерживаются крайних шовинистических взглядов на миграцию; 27 респондентов обеспокоены прежде всего экономическими последствиями большого притока дешевой рабочей силы; 30 респондентов настроены на «мягкое» решение проблемы мигрантов, при котором не будут страдать права ни коренного населения, ни мигрантов. Следует обратить внимание на то, что некоторые из респондентов не попадают ни в одну из определившихся трех групп, а взгляды некоторых столь непоследовательны, что они почти с равной вероятностью могут быть отнесены, например, к первой и второй группам.

4.3. Кластерный анализ

Кластерный анализ ставит перед собой задачу классификации объектов. Синонимами термина «кластерный анализ» являются «автоматическая классификация объектов без учителя» и «таксономия».

Если данные понимать как точки в признаковом пространстве, то задача кластерного анализа формулируется как группировка объектов в многомерном признаковом пространстве, разбиение совокупности на однородные подмножества объектов. В этом смысле по своим задачам кластерный анализ похож на дискриминантный анализ, но последний для целей классификации использует обучающие выборки, на

основании которых строится дискриминантная функция, позволяющая классифицировать новые объекты.

Другой возможностью кластерного анализа (более редко используемой) является классификация переменных, т. е. поиск переменных, которые близки по своему смыслу. Классификация переменных в кластерном анализе преследует фактически ту же цель, что и факторный анализ, – сокращение числа переменных. Если переменные оказались близкими (попали в один кластер), то можно оставить для дальнейшего анализа одну из них, удалив из набора данных другие.

Использование факторного анализа предпочтительнее, если переменные относятся к интервальной шкале, Кластерный анализ переменных с целью анализа их близости обычно применяют, если переменные относятся к дихотомической шкале.

SPSS предлагает для использования три различных алгоритма факторного анализа: двухшаговый кластерный анализ, иерархический кластерный анализ и итерационный метод, известный как метод К-средних. Более подробно эти методы будут обсуждаться при решении конкретных практических задач. Здесь лишь отметим, что двухшаговый метод кластерного анализа имеет, пожалуй, лишь одно достоинство: анализируемые переменные могут относиться к интервальной и номинативной шкалам. Иерархический кластерный анализ используется в тех случаях, когда число объектов невелико (несколько сотен) или когда нужно произвести кластерный анализ переменных.

Метод К-средних используется при достаточно большом числе случаев, когда переменные относятся к интервальной шкале.

В кластерном анализе классификация объектов производится на основании понятия расстояния между объектами в многомерном признаковом пространстве, а классификация переменных – на основании понятия схожести переменных.

Поскольку в этом анализе могут участвовать переменные, измеренные в номинативной, порядковой и дихотомической шкалах, возникает неоднозначная проблема определения расстояния между объектами. Для каждого из видов шкал измерения SPSS предлагает несколько различных способов определения расстояния между объектами. Чаще всего можно, не вдаваясь в детали, ограничиться способом определения расстояния, которое предлагается по умолчанию.

Обозначим, как и ранее, символом x_{ij} $(i = 1,2,\ldots,n, \quad j = 1,2,\ldots,k)$ значение j-го признака для i-го объекта. Для переменных, принад-

лежащих к интервальной шкале, SPSS предлагает восемь различных способов определения расстояния между объектами. Например, можно определить евклидово расстояние d_{pq} между объектами p и q, обобщая запись для расстояния между двумя точками на плоскости:

$$d_{pq} = \sqrt{\sum_{i=1}^{k} (x_{pi} - x_{qi})^2} . \qquad (4.21)$$

По умолчанию при проведении кластерного анализа в SPSS в качестве меры расстояния между объектами используется квадрат евклидового расстояния. Благодаря возведению в квадрат меры (4.21) различие между объектами получается существенно больше, что облегчает проведение анализа. Различные способы определения расстояния между объектами приводят и к различной конечной классификации объектов. Поэтому решение задачи классификации объектов в кластерном анализе не является однозначным. Такая же ситуация имеется при обычной группировке, результат которой зависит от способа определения интервала группировки.

Другим способом определяется расстояние между объектами, если значения переменных представляют собой частоты. В этом случае в качестве меры расстояния между объектами SPSS предлагает использовать либо величину χ^2 (2.15), либо связанную с ней меру

$$\varphi^2 = \sqrt{\frac{\chi^2}{N}} . \qquad (4.22)$$

Чтобы пояснить сказанное выше, рассмотрим следующий пример. Пусть, например, имеются данные результатов выборов в Государственную Думу РФ, представляющие собой число голосов избирателей, отданных за кандидатов основных политических партий в различных федеральных избирательных округах Российской Федерации. Требуется определить, имеются ли различия в результатах голосования за представителей различных партий в различных федеральных округах. В этой задаче число голосов, набранных той или иной партией в некотором регионе РФ, как раз являются частотами, и поэтому в качестве меры различия следует выбирать расстояния (2.15) или (4.22).

Наибольшее число мер различия переменных (27) SPSS предлагает для дихотомических (бинарных) переменных при проведении кластерного анализа. Здесь, как правило, значения переменных кодируют факт наступления события (1) или ненаступления события (0). При сопостав-

лении двух переменных возможны четыре случая, которые представлены в табл. 4.12.

Таблица 4.12

Таблица четырех полей, на основании которой определяются меры сходства между переменными

Значения переменной 1	Значения переменной 2	
	Да	Нет
Да	*a*	*b*
Нет	*c*	*d*

В этой формуле число a равно числу случаев, для которых значение переменных 1 и 2 совпадают и равны 1 (событие наступило). Аналогично интерпретируются и другие величины этой таблицы четырех полей. Мы не будем приводить здесь определения всех 27 возможных мер близости между бинарными (дихотомическими) переменными и ограничимся лишь определением меры квадрата бинарного евклидова расстояния, которое предлагается SPSS по умолчанию при кластеризации бинарных переменных:

$$d = b + c . \tag{4.23}$$

Смысл этой меры состоит в том, что величина d представляет собой количество случаев, для которых одно из событий наступило, а другое не наступило.

Завершая краткую характеристику методов кластерного анализа в SPSS, следует подчеркнуть, что кластерный анализ является описательной процедурой. Он не позволяет сделать никаких статистических выводов, но дает возможность провести своеобразную разведку – изучить «структуру совокупности».

Иерархический кластерный анализ

Самым распространенным методом кластерного анализа является иерархический кластерный анализ. Сущность его состоит в том, что на первом шаге каждый объект выборки рассматривается как отдельный кластер. Затем по исходным данным вычисляется матрица расстояний между всеми объектами статистической совокупности. Процесс объединения кластеров происходит последовательно: отыскиваются два объекта, расстояние между которыми является наименьшим,

и они объединяются в один кластер. Затем отыскивается следующий объект, расстояние до которого от объектов, включенных в первый кластер, является наименьшим, и он присоединяется к первому кластеру. Процесс продолжается до тех пор, пока не будет получен один кластер. Описанная процедура геометрически изображается в виде дендрограммы, на которой изображается последовательность шагов объединения исходных объектов в один кластер (рис. 4.9).

Для иллюстрации сказанного рассмотрим небольшой пример. Пусть имеется четыре объекта, для которых рассчитана матрица евклидовых расстояний (табл. 4.13).

Таблица 4.13

Симметричная матрица эвклидовых расстояний для четырех объектов

Объекты	1	2	3	4
1	0	2,06	4,03	2,50
2		0	2,24	4,12
3			0	6,32
4				0

На первом шаге будут объединены объекты 1 и 2, поскольку между ними наименьшее расстояние, равное 2,06.

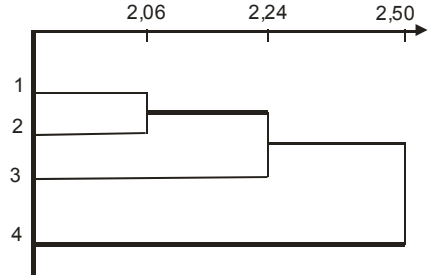


Рис. 4.9. Дендрограмма кластеризации четырех объектов

На втором шаге к первому кластеру будет присоединен объект 3, имеющий наименьшее расстояние, равное 2,24, со вторым объектом первого кластера. На последнем шаге в кластер будет включен четвертый объект, имеющий наименьшее расстояние до первого объекта, включенного в кластер на первом шаге.

Алгоритм иерархического кластерного анализа в SPSS очень похож на описанный выше. Также решающую роль при объединении объектов в кластер играет матрица расстояний между ними. Единственным заметным отличием является то, что шкала расстояний при

построении дендрограммы нормируется на 25, так что максимальное расстояние между кластерами всегда равно 25.

Существенным недостатком иерархического кластерного анализа является то, что этот метод не выявляет число реально существующих кластеров. Это число нужно определять самому исследователю, исходя из анализа расстояний между объектами. К этому вопросу мы еще вернемся при обсуждении конкретных примеров.

Пример 4.4

В файле Пример 4_4.sav приведены данные о результатах голосования избирателей федеральных избирательных округов Свердловской и Тюменской областей за кандидатов различных партий во время выборов в Государственную Думу РФ в 2003 г.

Проведя кластерный анализ объектов, требуется выяснить, существуют ли заметные отличия в результатах голосования избирателей в изучаемых федеральных округах.

Используя кластерный анализ переменных, выяснить, какие из политических партий существенно различаются между собой, с точки зрения избирателей.

Решение

После загрузки данных в SPSS, используя пункты меню Analyze/Classify/Hierarchical Cluster (Анализ / Классификация / Иерархический кластер), откроем окно, изображенное на рис. 4.10.

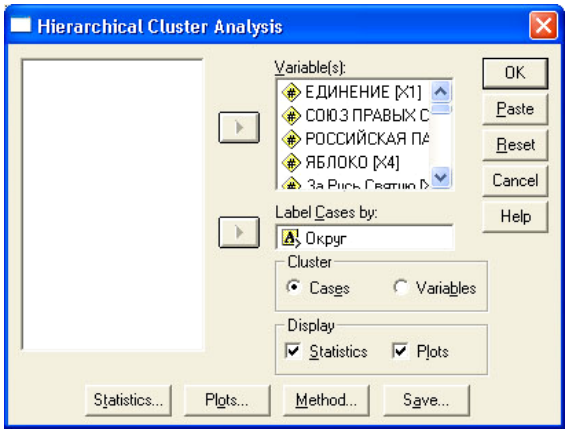


Рис. 4.10. Окно установки параметров иерархического кластерного анализа

Переменную Округ перенесем в окно Label Cases by (Метка объектов), а все остальные переменные, представляющие собой названия партий, кандидаты от которых были внесены в избирательные бюллетени, в окно Variables (Переменные). Поскольку мы хотим выявить структуру результатов голосования по избирательным округам, во фрейме Cluster переключатель следует оставить в положении Cases (Объекты). Следует также оставить настройки по умолчанию в окошке Plots (Графики), чтобы иметь возможность заказать вывод дендрограммы.

Нажав кнопку Statistics (Статистика), в открывшемся окне оставим настройку по умолчанию в окошке Agglomeration Schedule (Дневник агломерации), что позволяет получить в окне вывода записи, отражающие последовательность объединения кластеров. Во фрейме Cluster Membership (Членство в кластере) следует установить переключатель в положение Single Solution (Одно решение) и в маленьком окошке справа ввести цифру 3, поскольку мы желаем получить распределение объектов по трем кластерам.

Затем следует переключиться на закладку Plots (Графики) и поставить флажок в поле Dendrogram (Дендрограмма), а на фрейме Icicle (Сосульчатая диаграмма) поставить галочку в поле None (Не нужно).

Нажав кнопку Method (Метод), выберем метод образования кластеров и способ задания расстояний между объектами (рис. 4.11). По умолчанию выбран метод Within-groups linkage (Внутригрупповая взаимосвязь).

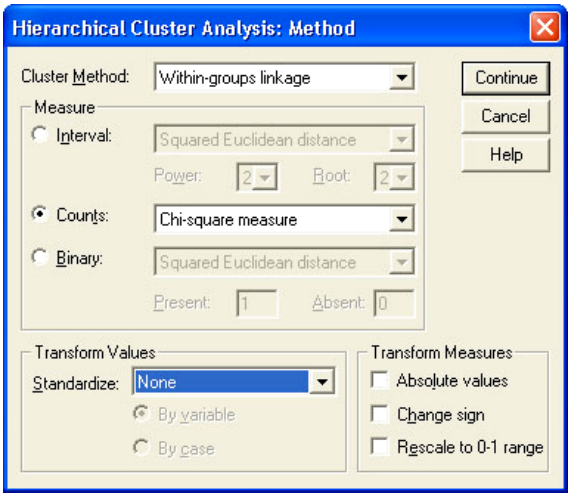


Рис. 4.11. Выбор метода образования кластеров

Этот метод соответствует описанному выше алгоритму образования кластеров и используется чаще других. Во фрейме Measure (Измерение) следует установить переключатель в положение Counts (Подсчеты), поскольку численные данные представляют собой число голосов избирателей, отданных за ту или иную партию. Во фрейме Transform Values (Трансформированные значения) можно оставить без изменения значение, установленное по умолчанию. К методу трансформации исходных переменных следует прибегать тогда, когда значение переменных совершенно несоизмеримы между собой. Переход, например, к стандартизованным значениям переменных приводит к тому, что все переменные будут изменяются в диапазоне от –3 до +3.

Наконец, зайдя на закладку Save (Сохранить), выберем опцию Single Solution (Одно решение) и в небольшом окошке справа введем значение 3. При выполнении процедуры кластерного анализа это обеспечит нам вывод в окно редактора SPSS новой переменной CLU1_1, значения которой для каждого объекта соответствуют номеру кластера, к которому этот объект отнесен.

После того как все установки выполнены, можно запустить процедуру кластерного анализа на исполнение. В окне вывода в первую очередь представляет интерес таблица шагов агломерации (табл. 4.13).

Таблица 4.13

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	1	4	13,325	0	0	7
2	2	5	29,910	0	0	3
3	2	3	60,254	2	0	4
4	2	7	96,620	3	0	6
5	8	9	106,978	0	0	8
6	2	6	109,399	4	0	7
7	1	2	153,477	1	6	8
8	1	8	183,562	7	5	0

На первом шаге объединяются в один кластер первый и четвертый объекты, поскольку между ними оказалось наименьшее расстояние (13,325). Эти кластеры участвуют в объединении впервые, о чем говорят нулевые значения в колонке «Этап первого появления кластера». Следующий раз эти кластеры будут участвовать в объединении на 7-м этапе.

На втором этапе объединяются в один кластер второй и пятый объекты (расстояние 29,910). На третьем и четвертом этапах к этому кластеру добавляются третий и седьмой объекты.

На пятом этапе в результате объединения 8-го и 9-го объектов образуется новый кластер. На шестом шаге к кластеру, образованному на втором шаге, добавляется шестой объект.

На последующих шагах идет процесс слияния кластеров, образованных на первом, втором и пятом шагах, в один кластер. Дендрограмма, полученная в SPSS, приведена на рис. 4.12.

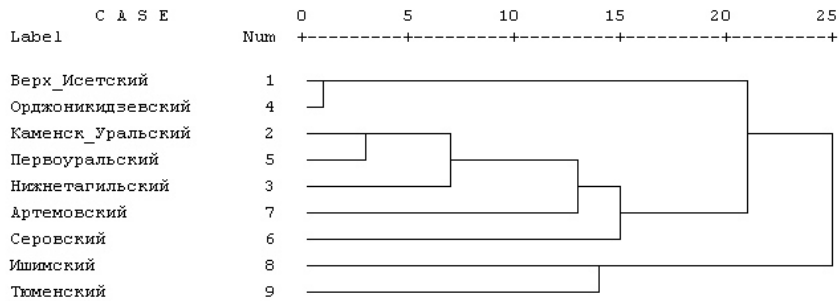


Рис. 4.12. Дендрограмма классификации избирательных округов Свердловской и Тюменской областей по итогам выборов в Госдуму РФ в 2003 г.

Даже после внимательного прочтении описанной процедуры объединения объектов в единый кластер и анализа приведенной выше дендрограммы остается вопрос: сколько же кластеров следует выделить? Есть простой формальный способ подсчета числа кластеров, которые следует выделять для того или иного набора данных. В таблице шагов агломерации следует найти шаг, при котором расстояние возрастает сильнее всего. Для нашего набора данных это шестой шаг. При переходе от шестого к седьмому шагу расстояние увеличивается более чем на 40 единиц. Если из общего числа объектов (9) вычесть шаг, на котором расстояние увеличивается на самую большую величину (6), это и будет вероятное число кластеров, которое следует выделить в этой задаче (3).

Тот же самый вывод можно сделать, анализируя дендрограмму. Число кластеров, которое следует выделить, определяется расстоянием между объектами, которое мы считаем достаточным для разделения кластеров. Если мы примем достаточным приведенное расстояние, несколько большее 15, например 18, и проведем на этом уровне на дендрограмме вертикальную линию вниз, то она пересечет три линии, соответствующие трем кластерам: первому, в который включены объекты 1 и 4; второму, в который включены объекты 2, 5,3 ,7, 6; и третьему, в который включены

объекты 8 и 9. Приведенное расстояние 15 соответствует расстоянию 109,399, при котором объединяются второй и шестой объекты.

Интересно, что произведенная кластеризация позволяет сделать вывод, что результаты голосования избирателей в городских округах Екатеринбурга существенно отличаются от результатов голосования в области, а результаты голосования в округах Тюменской области существенно отличаются от результатов голосования в Свердловской области, поскольку эти округа выделились в отдельный кластер.

Подвергнем теперь кластерному анализу сами переменные с целью отыскать похожие по результатам голосования партии. Кластерный анализ переменных позволит найти партии, которые избиратели девяти избирательных округов практически не различают (эти партии будут принадлежать одному кластеру), и партии, которые, с точки зрения избирателей, отличаются (образуют другой кластер).

Переменную Округ уберем из окна Label cases by (см. рис. 4.10) и во фрейме Cluster переместим переключатель в положение Variables (Переменные). Все остальные установки можно не менять. После запуска процедуры на исполнение в окне вывода результатов проанализируем таблицу шагов агломерации (ради экономии места мы не приводим эту таблицу полностью).

Таблица 4.14

Шаги агломерации при кластеризации партий

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
15	6	10	27,123	0	0	17
16	4	7	30,647	13	12	18
17	1	6	33,816	14	15	18
18	1	4	41,443	17	16	19
19	1	3	55,284	18	0	20
20	1	20	64,336	19	0	21
21	1	12	74,535	20	0	22
22	1	19	88,552	21	0	23
23	1	2	115,307	22	0	0

Анализируя значения расстояний между объектами, приведенное в столбце «Коэффициенты» табл. 4.14, замечаем, что первое существенное увеличение расстояния между объектами произошло на 17-м шаге. Всего объектов 24. Поэтому в этой задаче разумно выделить 7 кластеров. Вернемся к процедуре установки параметров кластерного анализа. На закладке Statistics (см. рис. 4.10) поставим цифру «7» в окошке Single Solution (Одно решение) и запустим процедуру кластерного анализа на

исполнение снова. В итоге получим таблицу распределения партий по кластерам (табл. 4.15).

Таблица 4.15

Принадлежность к кластерам

Название партии	Кластер
Единение	1
За Русь святую	1
Объединенная российская партия «Русь»	1
Народно-республиканская партия России	1
Аграрная партия России	1
Истинные патриоты России	1
Демократическая партия России	1
Партия мира и единства	1
ЛДПР	1
Конституционно-демократическая партия	1
КПРФ	1
Союз правых сил	2
Российская партия пенсионеров и партия социальной справедливости	3
Яблоко	4
Автомобильная Россия	4
Российская экологическая партия «Зеленые»	4
Великая Россия – Евразийский союз	4
Партия СЛОН	4
Родина	4
Развитие предпринимательства	4
Против всех	4
Народная партия	5
Российская партия жизни	6
Единая Россия	7

Как следует из таблицы, в представлении избирателей анализируемых девяти федеральных избирательных округов в первый кластер вошли партии национально-патриотической направленности. Большая группа партий вошла в кластер под номером 4. Анализируя названия партий, вошедших в эту группу, можно сказать, что четвертый кластер объединил

партии умеренно консервативные, связанные с протестным движением или нарождающимся движением некрупных частных предпринимателей.

Все остальные партии принадлежат к различным кластерам, что говорит об узнаваемости этих партий и наличии у них своего собственного электората.

Метод К-средних

Рассмотренный выше иерархический кластерный анализ исходит из анализа матрицы расстояний. При большом числе наблюдений (порядка 1000) кластеризация требует просмотра достаточно большого массива данных и занимает много времени. Поэтому в SPSS имеется процедура быстрого кластерного анализа, которая использует итеративный алгоритм кластеризации, предложенный Дж. Мак-Куином в 1967 г. Этот алгоритм не использует матрицу расстояний, и поэтому в нем число операций линейно растет с ростом числа наблюдений, а не квадратично, как в иерархическом кластерном анализе.

В методе К-средних число кластеров, по которым распределяются анализируемые объекты, должно быть заранее известно. Пусть, например, ставится задача выделить К кластеров. Тогда на первом шаге выбираются случайным образом или задаются пользователем К центров координат будущих кластеров. На втором шаге происходит просмотр всех объектов, и каждый объект присоединяется к тому из кластеров, расстояние до которого оказывается наименьшим. Существует два варианта метода К-средних. В первом варианте после присоединения каждого нового объекта к тому или иному кластеру координаты центра кластера пересчитываются, а во втором варианте пересчет происходит только после просмотра всех объектов.

После того как все объекты просмотрены и пересчитаны новые координаты центров кластеров, процесс распределения объектов по кластерам начинается заново и рассчитываются новые координаты центров кластеров. Итерационный процесс заканчивается тогда, когда после очередного шага координаты центров кластеров остаются практически неизменными. Такая процедура обеспечивает устойчивость распределения объектов по кластерам.

Пример 4.5

В файле Пример_4_5.sav приведены данные о предоставлении услуг 1000 абонентов узла связи. Служба маркетинга компании хотела бы выяснить, в каком направлении должна развиваться служба сервиса. Для этих целей требуется сегментировать рынок потребления услуг, выявить

структуру и численность абонентов, входящих в разные группы потребителей. Очевидно, что целью сегментирования является политика адресного предоставления услуг различным категориям потребителей, позволяющая увеличить объем и качество услуг связи.

Решение

Поставленная задача требует проведения кластерного анализа. Поскольку объем респондентов достаточно велик, будем использовать для кластеризации метод К-средних. После загрузки файла данных в окно редактора SPSS произведем переход к стандартизованным величинам. Необходимость такого преобразования связана с тем, что часть переменных задачи измерена в интервальной, а часть в номинальной шкалах.

Для установки параметров кластерного анализа выберем опции меню Analyze/Classify/ K-Means Cluster (Анализ/Классификация/ Метод К-средних). В результате откроется окно, изображенное на рис. 4.13.

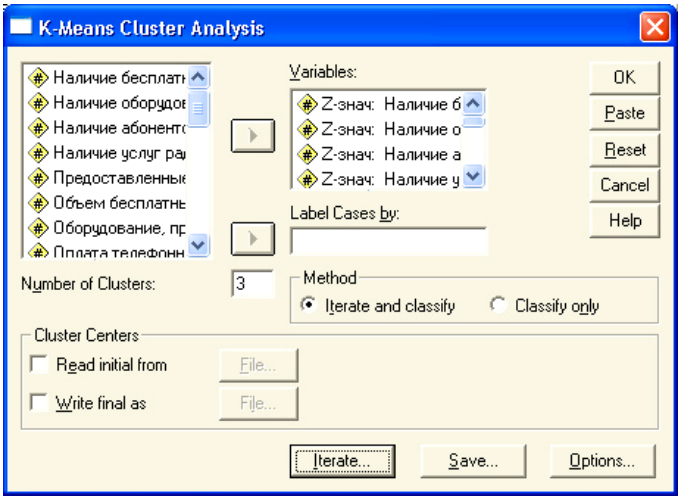


Рис. 4.13. Окно установки параметров кластеризации методом К-средних

Стандартизованные переменные поместим в окно Variables (Переменные). Все другие необходимые установки показаны на рис. 4.13.

Откроем закладку Iterate (Итерации) и установим максимальное число итераций, равное 30. На закладке Save (Сохранение) можно сохранить в новой переменной принадлежность каждого объекта к тому или иному кластеру и расстояние объекта до центра кластера. При большом числе наблюдений эта информация трудно поддается анализу, поэтому можно оставить ус-

тановки по умолчанию и не сохранять новые переменные. На закладке Option (Опции) следует предусмотреть вывод таблицы результатов дисперсионного анализа (ANOVA table), в которой приведены значения F-статистики Фишера для каждой переменной. Чем больше это значение, тем более важную роль играет переменная при разделении объектов на кластеры.

Запустим процедуру кластерного анализа на исполнение и перейдем в окно вывода результатов. Здесь наибольший интерес представляют данные, приведенные в табл. 4.16.

Таблица 4.16

Конечные центры кластеров

Метки переменных, участвующих в анализе	Кластер		
	1	2	3
Наличие бесплатной поддержки	0,615	–0,779	0,802
Наличие оборудования, взятого в аренду	1,013	–0,063	–0,667
Наличие абонентской телефонной карты	0,412	–0,485	0,478
Наличие услуг радиосвязи	1,453	–0,580	–0,158
Предоставленные услуги междугор. связи	–0,043	–0,143	0,266
Объем бесплатных услуг в прошлый месяц	0,531	–0,702	0,741
Оборудование, приобретенное в прошлом месяце	1,275	–0,200	–0,642
Оплата телефонных услуг в прошлом месяце	0,318	–0,358	0,342
Оплата за радио в прошлый месяц	1,469	–0,552	–0,215
Наличие многоканальной связи	0,434	–0,056	–0,238
Наличие голосовой связи	1,202	–0,480	–0,130
Наличие пейджинговой связи	1,265	–0,480	–0,177
Наличие подсоединения к Интернет	0,903	–0,077	–0,559
Наличие автоопределителя	0,663	–0,750	0,719
Наличие сервиса захвата линии	0,610	–0,758	0,772
Наличие службы прохождения сигнала	0,692	–0,744	0,688
Возможность проведения селекторных совещаний	0,647	–0,729	0,697
Возможность получать новостную информацию	0,821	–0,006	–0,613

В табл. 4.16 приведены значения переменных для центров кластеров. Поскольку точки каждого кластера группируются вблизи центра, то объ-

екты, принадлежащие кластеру, будут иметь значения, близкие тем, которые приведены в таблице. Это позволяет интерпретировать смысловое значения разделения объектов на кластеры.

В кластер 1 объединены объекты, которые активно используют все имеющиеся услуги. Они систематически применяют все новые виды связи и неохотно прибегают к услугам междугородней телефонной связи.

Вторую группу образуют абоненты, крайне экономно тратящие свои средства на услуги связи. Видимых предпочтений в выборе средств связи в этой группе нет.

Наконец, третью группу образуют клиенты, которые ориентированы на использование телефонной связи. Они широко используют все сервисы телефонной связи, не уделяя достаточного внимания другим видам связи.

Число объектов, попавших в каждый кластер, приведено в табл. 4.17.

Таблица 4.17

Число наблюдений в каждом кластере

Кластер	1	224,0
	2	481,00
	3	295,0
Валидные		1000,0
Пропущенные значения		0,0

Все переменные, выбранные для анализа, имеют значимую разделяющую силу, поскольку для большинства переменных значение $F > 100$. Ниже приведена лишь часть таблицы дисперсионного анализа (табл. 4.18).

Таблица 4.18

Фрагмент таблицы дисперсионного анализа для трех переменных

Метки переменных	Кластер		Ошибка		F	Знач.
	Средний квадрат	Ст.св.	Средний квадрат	Ст.св.		
Наличие бесплатной поддержки	282,839	2	0,435	997	650,76	0,000
Наличие оборудования, взятого в аренду	181,564	2	0,638	997	284,67	0,000
Наличие абонентской телефонной карты	109,409	2	0,783	997	139,81	0,000

Таким образом, в результате кластерного анализа мы установили, что около 25 % клиентов активно используют все предоставляемые сервисы связи, около 30 % широко используют только сервисы телефонной связи, и более 45 % клиентов образуют группу умеренных пользователей, очень экономно расходующих свои средства на приобретение услуг связи.

4.4. Дерево решений

Основные понятия и определения

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. В главе 1 мы уже обсуждали некоторые принципы построения дерева решений и алгоритмы, лежащие в основе этого метода классификации объектов. Рассмотрим реализацию этого весьма популярного метода в SPSS.

Дерево решений (Decision Tree) – это логический алгоритм классификации, основанный на поиске внутренних закономерностей в данных. *Деревом* называется конечный связный граф с множеством вершин (узлов) n, не содержащий циклов и имеющий единственную вершину (узел), в которую не входит ни одно ребро (ветвь). Эта вершина (узел) называется *корнем дерева*. Вершина (узел), не имеющая выходящих ребер (ветвей), называется *терминальной,* или *листом*. Остальные вершины (узлы) называются *внутренними*.

Дерево называется *бинарным*, если из любой его внутренней вершины выходит ровно два ребра (ветви). Выходящие ветви связывают внутреннюю вершину с левой дочерней вершиной и с правой дочерней вершиной. Дерево решений, изображенное на рис. 1.4, является бинарным. Если переменная, в соответствии со значениями которой производится расщепление узла, имеет более двух градаций, то разумно использовать алгоритм, допускающий множественное расщепление узла. Если узел может быть расщеплен на три и более дочерних узла, то алгоритм допускает *множественное расщепление*.

Формат представления данных для анализа

Дерево решений, по существу, дает решение задачи классификации с учителем (обучающей выборкой). В отличие от дискриминантного анализа, здесь имеется возможность построить интуитивно понятные правила классификации, т. е. непосредственно извлечь знания из данных.

Данные, подлежащие анализу, должны иметь одну зависимую переменную и некоторое число независимых объясняющих переменных.

Алгоритмы построения дерева решений, как правило, ориентированы на работу с переменными, измеренными в номинативной шкале. Результативная (target) и объясняющие переменные могут быть измерены и в интервальной шкале. В этом случае при построении дерева решений в большинстве алгоритмов значения переменных делятся на интервалы, число которых может определить пользователь. Если переменные измерены в ординальной или номинативной шкалах, то это должно быть явно задано в окне Variable View (Обзор переменных) редактора SPSS, поскольку для одних и тех же, но измеренных в разных шкалах числовых данных деревья решений могут получиться различными.

Отбор переменных для расщепления узла

В первой главе мы уже обсуждали проблему выбора переменной для расщепления узла на очередном шаге построения дерева решений. Самый простой способ – это выбрать ту объясняющую переменную, которая имеет самый большой коэффициент корреляции с результативной переменной. Этот метод используется в популярном алгоритме с аббревиатурой CHAID – *Chi-Squared Automatic Interaction Detection* (Автоматическое детектирование взаимодействий методом хи-квадрат). Алгоритм CHAID установлен по умолчанию при построении дерева решений в SPSS. Другой способ выбора переменной для расщепления узла, основанный на понятии энтропии распределения, обсуждался ранее в главе 1. Оба метода имеют один и тот же недостаток: на каждом шаге определяется переменная, которая дает наилучший результат для анализируемого узла, и не принимается во внимание проблема оптимизации построения дерева решения в целом. Такого рода процедуры поиска переменных для расщепления узла называются *рекурсивными*. Рекурсивные процедуры удобны тем, что каждый узел фактически рассматривается независимо от других узлов, что позволяет использовать один алгоритм для расщепления всех узлов.

Правила остановки ветвления дерева решений

Чем больше правил можно сформулировать при построении дерева решений, тем меньшее количество объектов попадает в каждый терминальный узел. Такие деревья состоят из неоправданно большого числа узлов и ветвей, исходное множество разбивается на большое

число подмножеств, состоящих из очень малого числа объектов. Способность к обобщению найденных правил уменьшается, и построенные модели не могут давать верные ответы. В итоге возникает ситуация *переобучения*, когда анализируемые объекты классифицируются верно, но найденные правила слишком подробны, учитывают случайные ошибки, содержащиеся в данных, и поэтому не применимы на других выборках.

Какой размер дерева может считаться оптимальным? Дерево должно использовать ту информацию, которая улучшает качество модели, и игнорировать ту ее часть, связанную со случайными ошибками в данных, которая не может улучшить качество модели.

Существует две возможные стратегии решения проблемы переобучения. Первая состоит в наращивании дерева до определенного размера в соответствии с параметрами, заданными пользователем. К числу таких параметров относится *глубина построения дерева решений*. Под глубиной в данном случае понимается число уровней дерева решений ниже корневого узла. Для алгоритма CHAID в SPSS, например, по умолчанию установлено значение глубины, равное трем.

Другим параметром, позволяющим ограничить «пушистость» дерева решений, является минимальное число узлов в родительском и дочерних узлах (по умолчанию в SPSS установлены значения 100 объектов для родительского узла и 50 для дочернего).

Ряд алгоритмов, например CHAID, оценивают статистическую значимость расщепления узла. Пользователь имеет возможность задать пороговое значение значимости, при котором узел будет расщепляться дальше. В SPSS по умолчанию установлено пороговое значение, равное 0,05. Определение параметров построения дерева решений может основываться на опыте и интуиции аналитика, а также на анализе приемлемости получающихся результатов.

Другая стратегия состоит в том, чтобы подрезать уже построенное дерево решений. После того как дерево построено и выполнены все критерии остановки роста дерева, оно автоматически подрезается на наименьшее из поддеревьев так, чтобы максимальное увеличение риска при подрезании было не больше, чем определил пользователь.

Риск для категориальных переменных – это доля неверно классифицированных случаев, выраженная в процентах. Для переменных, измеренных в интервальной шкале, риск – это средняя из групповых дисперсий в терминальных узлах. Очень часто величина риска являет-

ся формальным признаком качества классификации, произведенной с помощью дерева решений.

Обработка пропущенных значений

При построении регрессионной модели или модели дискриминантного анализа пропущенное значение для любой из объясняющих переменных приводит к тому, что из анализа объект исключается полностью. При построении дерева решений влияние каждой из объясняющих переменных рассматривается независимо, и отсутствие данных для некоторого объекта в одной переменной необязательно должно исключать этот объект из анализа при расщеплении узла по другой переменной.

Трактовка пропущенных значений зависит от метода построения дерева решений и от того, что понимается под пропущенным значением. В SPSS есть так называемые *системные пропущенные значения* (system-missing values), под которыми понимается реально отсутствующие значения в файле данных. В редакторе SPSS системные пропущенные значения помечаются точками. В то же самое время в некоторых случаях данных может не быть, а в файле в закодированном виде указана причина, по которой данные отсутствуют. Например, цифрой «1» кодируется хороший кредитный рейтинг клиента, цифрой «2» – плохой, а цифрой «3» – ситуация отсутствия кредитной истории. При построении профиля кредитоспособного клиента случаи, для которых нет кредитной истории, должны быть исключены из анализа. Таким образом, возникает ситуация, когда необходимо исключить некоторые значения переменной из анализа. В SPSS это так называемые *значения, пропущенные пользователем* (User-missing values).

При выборе параметров построения дерева решений в SPSS можно выбрать различные установки обработки пропущенных пользователем значений: пропущенные пользователем значения трактуются либо как системные пропущенные значения, либо как имеющие смысл новые градации переменной.

В методе CHAID и Exhaustive CHAID (исчерпывающий CHAID) для системных и пользовательских пропущенных данных сначала определяется отдельная градация номинативной переменной, а затем решается вопрос, объединить ли эту градацию с одной из существующих реальных градаций этой переменной либо включить ее в анализ наравне с реально существующими градациями.

Для методов CRT и QUEST[2] случаи с пропущенными значениями исключаются из анализа при построении дерева решений, но есть возможность использовать так называемые суррогатные переменные. *Суррогатные переменные* – это переменные, которые замещают при построении дерева переменные, содержащие пропущенные значения. Строятся они во время работы алгоритма построения дерева решений, исходя из принципа максимальной похожести на замещаемую переменную. Пользователь может задать максимальное число суррогатных переменных, которые можно использовать при анализе (по умолчанию это значение равно $n-1$, где n – число независимых переменных).

Использованные при построении дерева решений алгоритмы обработки пропущенных значений позволяют менее расточительно относиться к данным и не выбрасывать случаи полностью, если хотя бы для одной независимой переменной значение оказалось неизвестным.

Алгоритмы построения дерева решений в SPSS

Алгоритм CHAID

Алгоритм CHAID является основным, установленным по умолчанию в SPSS, алгоритмом построения дерева решений. Алгоритм предложен студентом из Южной Африки Г. В. Кассом (G. V. Kass) в 1975г. и опубликован в 1980 г.[3] Алгоритм ориентирован на работу с переменными, измеренными в номинативной (ординальной) шкале. Переменные, измеренные в шкале отношений, преобразуются в номинативные разбиением значений переменной на интервалы (по умолчанию число таких интервалов равно 10).

При построении дерева решений CHAID использует рекурсивную процедуру расщепления узлов. Вначале находится переменная, имеющая наибольший коэффициент корреляции с результативной переменной. Узел расщепляется на дочерние узлы в соответствии с градациями независимой переменной, отобранной на данном шаге для расщепления узла. Затем попарно на основании критерия хи-квадрат проверяется статистическая значимость различия значений результативной переменной для двух градаций (по умолчанию уровень значи-

[2] CRT – Classification and Regression Trees (дерево классификации и регрессии); QUEST – Quick, Unbiased, Efficient Statistical Tree (быстрый, эффективный, статистически обоснованный алгоритм построения дерева).

[3] *Kass G.V.* An Exploratory Technique for Investigating Large Quantities of Categorical Data// Appl. Stat., 1980. **29**. P. 119 – 127.

мости 0,05). Если различия статистически незначимы, то градации объединяются в одну. Эта процедура продолжается до тех пор, пока все расщепления в данном узле не окажутся статистически значимыми. На этом процедура расщепления узла завершается, и алгоритм переходит к анализу возможности расщепления следующего узла.

Алгоритм завершает работу, если выполняется одно из ограничений, определенных пользователем (достигнуты предельная глубина расщепления, минимальное число объектов в дочерних или родительских узлах), или не остается узлов, расщепление в которых было бы статистически значимым. Следует отметить, что некоторые из независимых переменных могут быть исключены из анализа вообще, поскольку расщепления по градациям этих переменных оказываются статистически незначимыми.

Алгоритм Exhaustive CHAID

Этот алгоритм является развитием алгоритма CHAID. Отличается от него только тем, что дополнительно анализируются для каждой независимой переменной все возможные расщепления узла по градациям этой переменной, и выбирается наиболее значимое расщепление. Алгоритм Exhaustive CHAID приводит к большему числу расщеплений, что далеко не всегда является оправданным.

Алгоритм CRT

Алгоритм CRT (*Classification and Regression Trees*), как видно из названия, решает задачи классификации и регрессии. Он разработан в 1974 – 1984 гг[4]. Этот алгоритм приводит к бинарному расщеплению узлов для факторных и результативных переменных, измеренных как в интервальной, так и ординальной шкалах.

CRT базируется на интуитивной идее уменьшения неопределённости в узле. По существу, это упрощенный вариант идеи минимизации энтропии распределения, обсуждавшейся нами в главе 1, при бинарном расщеплении узла.

Рассмотрим задачу с двумя классами и узлом, имеющим, например, по 50 примеров каждого класса. Такой узел имеет максимальную неопределенность: вероятности вытащить наугад объекты каждого класса равны. Если будет найдено разбиение, при котором родительский узел разбивается на два дочерних узла с содержанием объектов 40 : 5 в одном

[4] *Breiman L.*, *Friedman J.H.*, *Olshen R.A.*, *and Stone C.J.* (1984). Classification and Regression Trees, Belmont, California: Wadsworth, Inc.

и 10 : 45 в другом, то интуитивно неопределенность уменьшится. Она полностью исчезнет, если будет найдено разбиение, которое создаст дочерние узлы с соотношением объектов разных классов 50 : 0 и 0 : 50.

В алгоритме CRT идея неопределенности формализована в индексе Gini (в честь итальянского экономиста Corrado Gini), который оценивает «расстояние» между распределениями классов. Если узел H_1 содержит n объектов, принадлежащих к двум классам, тогда индекс Gini определяется как

$$\text{Gini}\ (H) = 1 - \sum_{i=1}^{2} p_i^2\ , \qquad (4.24)$$

где p_i – вероятность (относительная частота) объектов i-го класса в узле H. Если узел H расщепляется на два узла H_1 и H_2 с числом объектов в каждом n_1 и n_2 соответственно, тогда показатель качества разбиения будет

$$\text{Gini}\,(H_1, H_2) = \frac{n_1}{n} \cdot \text{Gini}\,(H_1) + \frac{n_2}{n} \cdot \text{Gini}\,(H_2)\,. \qquad (4.25)$$

Наилучшим считается то разбиение, для которого значение индекса Gini (4.25) минимально.

Поскольку алгоритм использует только бинарное расщепление, глубина построения дерева решений оказывается существенно выше, нежели в методе CHAID (по умолчанию – 5).

Алгоритм QUEST

Алгоритм QUEST (*Quick, Unbiased, Efficient Statistical Tree*) применим только для тех случаев, когда зависимая переменная измерена в номинативной шкале. Метод может иметь преимущества, когда число градаций независимых переменных велико.

Для простоты рассмотрим лишь случай, когда все независимые переменные измерены в номинативной шкале. В этом случае для отбора переменной, по значениям которой будет производиться расщепление узла, составляются таблицы сопряженности каждой независимой переменной с зависимой переменной (о таблицах сопряженности см. в гл. 2) и анализируется на основании критерия хи-квадрат значимость взаимосвязи результативной и каждой из факторных переменных.

Для расщепления узла выбирается переменная, для которой эта взаимосвязь получается наиболее значимой. Далее запускается процедура дискриминантного анализа, т. е. разделение множества значений

результативной переменной на два класса по значениям найденной дискриминирующей переменной. В результате будет найдено значение дискриминирующей переменной, при котором узел наилучшим образом делится на два дочерних узла: в первом узле будет максимально большое число объектов первого класса, а во втором – второго. Алгоритм расщепления узла является рекурсивным и повторяется по существу без каких либо изменений до тех пор, пока не останется узлов, для которых взаимосвязь результативной и факторных переменных является значимой. Хотя считается, что этот алгоритм позволяет избежать излишней «пушистости» построенного дерева, очень часто алгоритм CHAID приводит к более приемлемым результатом с меньшим значением коэффициента риска.

Завершая рассмотрение раздела, посвященного классификации с помощью дерева решений, рассмотрим простой пример, взятый из справочной системы SPSS. Другие примеры классификации с использованием дерева решений будут рассмотрены на практических занятиях.

Пример 4.6

В файле Пример_4_6. sav содержатся данные о кредитных рейтингах 2464 клиентов банка (результативная переменная), пользовавшихся банковскими услугами в прошлом, а также такие данные о клиентах, как возраст (число лет), уровень дохода (низкий, средний высокий), число используемых кредитных карт (меньше 5, больше 5), уровень образования (высшее, среднее специальное) и число кредитов на покупку автомашины (нет или 1, 2 и более). Построив дерево решений, требуется выяснить, по каким правилам можно определить клиента с плохим кредитным рейтингом (высоким риском невозвращения кредита).

Решение

После загрузки файла в редактор данных активизируем опции Analyze/Classify/Tree (Анализ/Классификация/Дерево). В результате откроется окно установки параметров построения дерева решений, изображенное на рис. 4.14. Результативной переменной является переменная с меткой «Кредитный рейтинг». Поместим эту переменную в окно Depend Variable (Зависимая переменная). Все остальные переменные включим в анализ как независимые переменные, поместив их в окно Independent Variables (Независимые переменные). В качестве метода построения дерева решений оставим метод CHAID, выбранный по умолчанию.

Как уже указывалось, алгоритмы построения дерева решений в SPSS сами выбирают переменную, которую нужно использовать для расщепления узла. Так, в рассматриваемом примере первой переменной для рас-

щепления корневого узла будет выбрана переменная с меткой «Уровень дохода». Если мы хотим, чтобы первой переменной была выбрана переменная, стоящая первой в списке Independent Variables (Возраст – в нашем случае), то нужно поставить флажок в окошке Force First Variables (Принудительно использовать первую переменную). Качество построенного дерева решений будем оценивать с помощью формального критерия «Риск». Чтобы обеспечить вывод этого показателя хотя бы для одной из категорий результативной переменной, нужно нажать кнопку Categories (Категории) и в открывшемся окне поставить флажок в ячейке на пересечении строки Плохой и столбца Target. Это обеспечивает подсчет риска неправильного определения клиентов с плохим кредитным рейтингом.

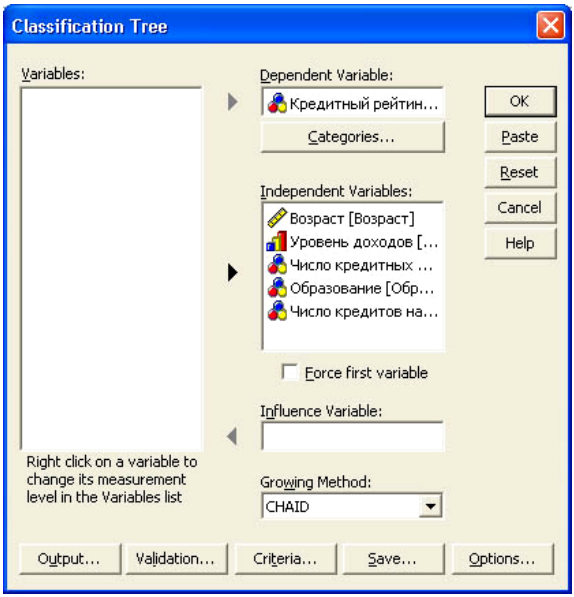


Рис. 4.14. Окно задания параметров построения дерева решений

Управлять выводимой при построении дерева решений информацией можно, нажав кнопку Output (Вывод). Можно не менять большинство установок, выбранных по умолчанию. Но поскольку нас интересуют извлеченные из базы данных правила, по которым можно определить потенциального клиента с плохим кредитным рейтингом, а по умолчанию эта информация не выводится, нам придется произвести дополнительные настройки. Нажав кнопу Output, откроем в появившемся окне закладку Rules (Правила). В результате откроется окно, в котором следует поста-

вить флажок в поле Generate classification rules (Генерировать правила классификации). Нужно также выбрать формат, в котором будут сгенерированные правила. Выберем опцию Simple Text (Простой текст). Нажав кнопку Continue (Продолжить), перейдем к установке других параметров построения дерева решений.

Нажмем кнопку Validation (Валидность, независимая проверка). В открывшемся окне можно выбрать установки для проверки дееспособности сгенерированных правил классификации на независимой выборке. Поскольку в нашем случае выборка достаточно большая, выберем 70 % объема выборки для построения правил классификации, а оставшиеся 30 % случаев используем для проверки построенных правил. Для того чтобы реализовать эту возможность, установим переключатель в положение Split-sample Validation (Валидация расщеплением данных) и в окне Training Sample (%) (Обучающие примеры) установим значение 70,00. В этом случае в окне выдачи результатов будет построено два дерева решений: для обучающей выборки и для контрольной выборки, представляющей собой 30 % случайным образом отобранных случаев исходных данных. Важно отметить, что данные контрольной выборки классифицируются по правилам, найденным на обучающей выборке.

Использование контрольной выборки позволяет обнаружить процесс переобучения дерева решений. Если риск неверной классификации, рассчитанный для контрольной выборки, существенно больше риска, найденного для обучающей выборки, то можно смело утверждать, что дерево решений является переобученным.

Нажав кнопку Criteria (Критерии), можно открыть окно, в котором задаются критерии, определяющие процесс роста дерева решений. С большинством установок здесь можно согласиться. Исправлений требуют лишь два параметра. Во фрейме Minimum Number of Cases (Минимальное число случаев) необходимо установить число 400 в окне Parent Node (Родительский узел) и 200 в окне Child Node (Дочерний узел). При этих установках родительский узел не будет расщепляться на дочерние, если в узле остается менее 400 случаев, а дочерние узлы не будут образовываться, если в них попадает менее 200 случаев. Если оставить в этом фрейме значения по умолчанию, то дерево будет излишне «пушистым», что затруднит извлечение достоверных правил.

Все значения на закладках Save (Сохранение) и Option (Опции) можно оставить по умолчанию. При построении дерева решений наиболее важная информация содержится в самой структуре дерева. Поэтому сохранение данных в окне редактора SPSS не представляет интереса для нашего случая. На закладке Option можно определить правила работы алгоритма с пропущенными значениями, которых в нашем наборе данных нет. Кро-

ме того, здесь можно задать анализ «стоимости» неправильной классификации случаев при построении дерева решений. Эти параметры никак не влияют на процесс построения дерева решений и нужны лишь для оценки результатов. Использование категорий стоимости (*Cost*) и информационного выигрыша (*Gain*) для определения качества произведенной классификации требует определенного навыка и мало информативно для начинающих пользователей. Поэтому анализ этих категорий мы опустим, оценивая качество классификации исключительно с помощью показателя риска.

Завершив настройку параметров построения дерева решений, запустим процедуру на исполнение. Результаты работы программы будут, как обычно, отображаться в окне вывода.

Окно вывода результатов начинается с предупреждения Gain summary Tables are not displayed because profits are undefined (Таблица информационного выигрыша не отображена, поскольку не определена полезность). Это предупреждение – лишь напоминание, что мы отказались использовать категорию информационного выигрыша для анализа качества классификации.

Следующая таблица с названием Model Summary (Сводка по модели) содержит основные параметры модели. Для нас здесь представляет интерес лишь строка, в которой указаны переменные, которые используются в процедуре расщепления узлов. Из пяти переменных (Возраст, Уровень доходов, Число кредитных карт, Образование, Число кредитов на покупку авто) в модель включены только переменные: Уровень доходов, Число кредитных карт. Переменные Образование, Число кредитов на покупку авто и Возраст исключены из модели, поскольку их взаимосвязь с результативной переменной оказалась статистически незначимой.

Построенное программой дерево решений изображено на рис. 4.15. Нулевой узел, содержащий 1697 объектов, из которых 691 имеют плохой кредитный рейтинг и 1006 – хороший, расщепляется по градациям переменной Уровень доходов. Значения критерия хи-квадрат, вычисленное по таблице сопряженности переменных Кредитный рейтинг и Уровень доходов, составляет 467,598, а уровень значимости (вероятность того, что связь в действительности отсутствует) – менее 0,001. В итоге получается три дочерних узла с низким, средним и высоким уровнем доходов. Узел с низким уровнем доходов является терминальным. Клиенты, имеющие низкий доход в 82,2 % случаев, имеют плохой кредитный рейтинг.

Узел, в котором оказались клиенты со средним уровнем доходов, расщепляется по градациям переменной Число кредитных карт на два дочерних узла, в одном из которых аккумулированы случаи клиентов с числом кредитных карт более 5, а в другом – с числом кредитных карт менее 5.

Клиенты со средним достатком и имеющие менее 5 кредитных карт в 84,5 % случаев имеют хороший кредитный рейтинг.

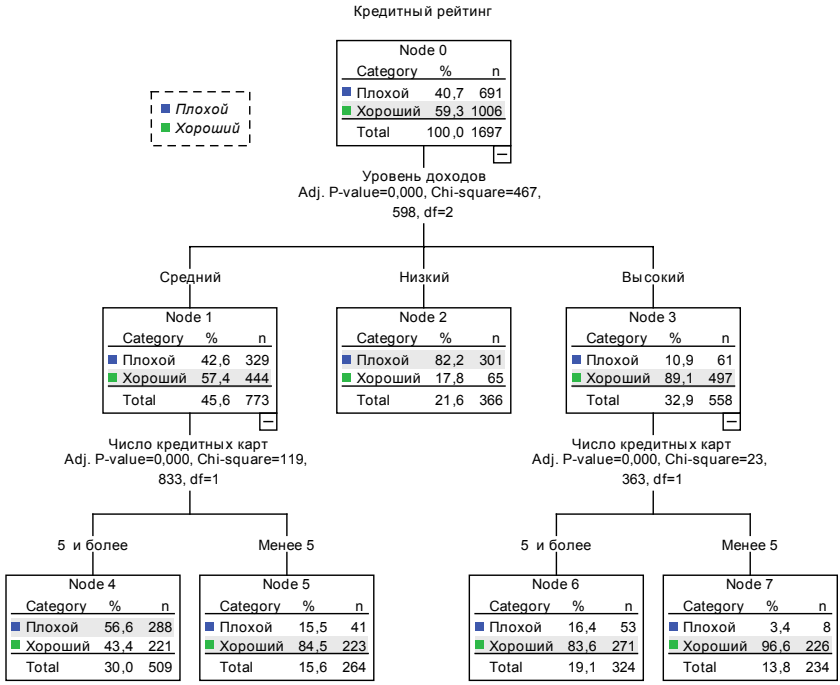


Рис. 4.15. Дерево решений для определения правил построения кредитного рейтинга клиентов банка

Если уровень доходов средний и число кредитных карт больше 5, то имеется примерно равное число клиентов с плохим и хорошим кредитным рейтингом. Этот узел следовало бы расщепить по значению какой-либо другой переменной, но уже достигнут предел на минимальное число объектов в родительском узле.

Наконец, если уровень доходов высокий, то узел 3 расщепляется по градациям переменной Число кредитных карт. В этом случае оказывается, что если число кредитных карт меньше 5, то 96,6 % процентов таких клиентов имеют хороший кредитный рейтинг.

Похожие результаты получаются и для тестовой выборки. Поэтому можно считать, что построенная модель классификации не является переобученной и может использоваться на практике. Суммарная информа-

ция о качестве модели содержится в табл. 4.19, содержащей оценку риска для обучающей и контрольной выборок (вероятности неверной классификации клиента с плохим кредитным рейтингом).

Таблица 4.19

Значение риска для обучающей и контрольной выборок

Выборка	Значение риска	Стандартная ошибка
Обучающая выборка	0,229	0,010
Тестовая выборка	0,231	0,015

Как следует из табл. 4.19, риск неправильной классификации клиентов, имеющих плохой кредитный рейтинг, составляет около 23 % как для обучающей, так и для тестовой выборок. Этот результат можно считать приемлемым для применения в реальной банковской деятельности.

Обратимся, наконец, к правилам определения клиентов, имеющих низкий кредитный рейтинг, извлеченных из анализируемой базы данных. Эти правила достаточно просты, и на основании дерева решений их можно сформулировать самостоятельно. Ниже приведен текст правил, извлеченных программой SPSS, при построении дерева решений. Приводятся правила только для терминальных узлов.

Узел 4:
Если уровень доходов средний и число кредитных карт > 5
Тогда
Рейтинг плохой с вероятностью 0,566;

Узел 5:
Если уровень доходов средний и число кредитных карт < 5
Тогда
Рейтинг хороший с вероятностью 0,845;

Узел 2:
Если уровень доходов низкий
Тогда
Рейтинг плохой с вероятностью 0,822;

Узел 6:
Если уровень доходов высокий и число кредитных карт > 5
Тогда
Рейтинг хороший с вероятностью 0,836;

Узел 7:
Если уровень доходов высокий и число кредитных карт < 5
Тогда
Рейтинг хороший с вероятностью 0,966.

Следует отметить, что изменение параметров построения дерева решений может дать другое дерево решений и другой набор правил. Поэтому при использовании этого метода классификации в практических целях необходимо провести серию экспериментов, а не ограничиться первыми полученными результатами. Например, в рассматриваемом случае полезно изучить, как возраст и уровень образования влияют на кредитный рейтинг клиента.

Пример 4.7

Файл Пример_4_7. sav содержит данные о стоимости первого купленного автомобиля (зависимая переменная), возрасте, поле, уровне дохода, образовании и семейном положении (независимые переменные) клиентов крупной сети магазинов по продаже автомобилей. Используя обучающую выборку из 3110 случаев, построить модель, которая предсказывала бы стоимость приобретаемого автомобиля, если известны значения независимых переменных. Оценить качество модели. Применить построенную модель для предсказания стоимости покупаемого автомобиля клиентами магазина (база данных предполагаемых клиентов находится в файле Пример_4_7a.sav и содержит 3290 записей).

Решение

Поставленная задача по существу является задачей построения регрессионной модели, но решать мы ее будем, используя алгоритм CRT построения дерева решений. Загрузим файл Пример_4_7. sav в редактор данных SPSS, активизируем опции Analyze/Classify/Tree (Анализ/ Классификация/Дерево). В результате откроется окно установки параметров построения дерева решений, изображенное на рис. 4.14. Результативной переменной является переменная с меткой Стоимость автомобиля, тыс. долл. Поместим эту переменную в окно Depend Variable (Зависимая переменная). Все остальные переменные включим в анализ как независимые переменные, поместив их в окно Independent Variables (Независимые переменные). В качестве метода построения дерева решений выберем метод CRT.

Поскольку нас интересуют извлеченные из базы данных правила, по которым можно определить стоимость приобретаемого автомобиля потенциальным клиентом, а по умолчанию эта информация не выводится, нам придется произвести дополнительные настройки. Нажав кнопу Output (Вывод), откроем в появившемся окне закладку Rules (Правила). В результате откроется окно, в котором следует поставить флажок в поле Generate classification rules (Генерировать правила классификации). Нужно еще выбрать формат, в котором будут сгенерированные правила. Выберем опцию SPSS. Для сохранения найденных правил в файле поставим флажок в поле Export

rules to a file и, нажав кнопку Browse (Просмотр), сохраним файл с выбранным именем в нужном месте. Расширение файла должно быть sps.

Нажав кнопку Continue (Продолжить), вернемся в основное окно установки параметров. Все остальные параметры можно оставить по умолчанию. Запустим процедуру на исполнение кнопкой OK. Дерево решений получается достаточно сложным, поэтому анализировать его визуально не представляется возможным. На рис. 4.16 представлены лишь корневой узел и результат его расщепления по переменной Категория дохода. Остальные узлы свернуты.

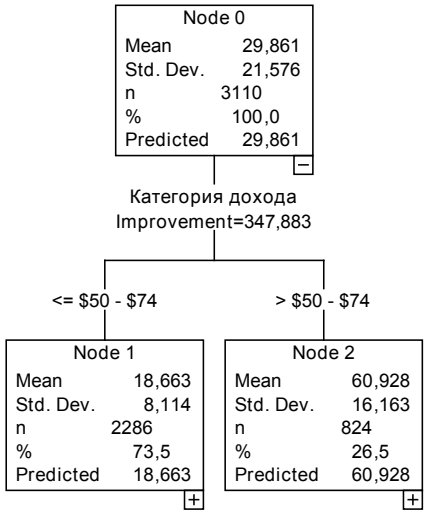


Рис. 4.16. Свернутое дерево решений

О качестве полученной модели будем судить на основания значения параметра Риск (табл. 4.20).

Таблица 4.20

Значение риска для обучающей выборки

Оценка риска	Стандартная ошибка
68,485	2,985

Напомним, что значение риска для переменной, измеренной в интервальной шкале, представляет собой дисперсию $\overline{\sigma_T^2}$ результативной переменной для терминальных узлов. Модель тем лучше, чем меньше эта величина. Но величина дисперсии зависит от выбора единиц измерения. Если

за единицу измерения взять не тысячу долларов, как это сделано в нашем случае, а доллар, то величина риска увеличится в миллион раз. Поэтому нужно определить безразмерный показатель, аналогичный фактору детерминации в дисперсионном или регрессионном анализе (отношение объясненной моделью или межузловой дисперсии к общей дисперсии).

Общую дисперсию найдем как квадрат среднеквадратического отклонения результативной переменной для корневого узла (см. рис. 4.16). Среднеквадратическое отклонение $\sigma = 21{,}576$. Следовательно, общая дисперсия $\sigma^2 = 465{,}523$. Межузловую дисперсию δ^2 (аналог межгрупповой, или объясненной моделью, дисперсии) можно получить, вычитая из общей дисперсии внутриузловую дисперсию $\overline{\sigma_T^2}$: $\delta^2 = \sigma^2 - \overline{\sigma_T^2} = 397{,}039$.

Доля объясненной дисперсии η^2 (аналог фактора детерминации R^2 в регрессионном анализе) тогда определяется выражением

$$\eta^2 = \frac{\delta^2}{\sigma^2} = \frac{397{,}039}{465{,}523} = 0{,}853 .$$

Полученное значение фактора детерминации указывает, что свыше 85 % дисперсии результативного признака объясняется моделью, что говорит о ее хорошем качестве.

Применим теперь полученную модель для определения стоимости автомобиля, который приобретут клиенты. Загрузим файл Пример_4_7а.sav в редактор данных SPSS. В этом файле оставлена переменная Стоимость с тем, чтобы затем можно было сравнить прогнозные и реальные результаты. Важно отметить, что для построения прогнозных значений переменная Стоимость не используется, и при желании ее можно удалить. Выбрав пункты меню File/New/Syntax (Файл /Новый/Синтаксис) откроем окно редактора командных файлов, изображенное на рисунке 4.17, и наберем команду: Insert File='C:\temp\Пример_4_7.sps'.

Отметим, что кавычки, в которые заключена ссылка на файл с извлеченными ранее правилами, и точка в конце команды являются обязательными. Название файла и место его расположения определяет пользователь. Здесь предполагается, что извлеченные правила сохранены в файле Пример_4_7.sps, который размещен в папке C:\temp. После того как команда набрана (курсор должен остаться в той же строке), следует нажать кнопку с изображением треугольника на панели инструментов для запуска команды на исполнение. В результате в окне редактора данных SPSS появятся две новых переменных. Переменная nod_001 будет содержать значение узла, в который попадет тот объект при классификации, а переменная pre_001 – предсказываемые значения результативной переменной для каждого объекта.

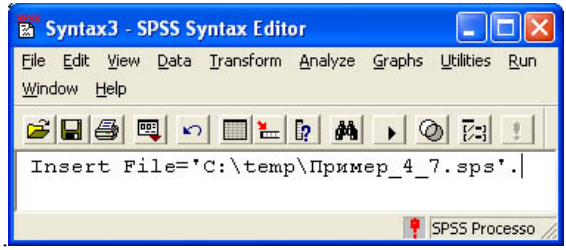


Рис. 4.17. Окно редактора командных файлов

В качестве прогнозирования в этом учебном примере можно убедиться, вычислив коэффициент корреляции истинных и прогнозных значений. Вычисляя парный коэффициент корреляции переменных Стоимость и pre_001, получаем значение, равное 0,919. Этот результат говорит о том, что предсказанные значения очень хорошо коррелируют с истинными.

Контрольные вопросы

4.1. Сформулируйте постановку задачи дискриминантного анализа.

4.2. Дайте графическую интерпретацию задачи дискриминантного анализа для случая двух переменных.

4.3. Какие факторные и результативные переменные могут принимать участие в дискриминантном анализе?

4.4. Сформулируйте математическую постановку задачи дискриминантного анализа.

4.5. Какой критерий используется в SPSS для отбора дискриминирующих переменных?

4.6. Для каких целей может использоваться дискриминантный анализ в задачах государственного и муниципального управления и менеджменте?

4.7. Сформулируйте постановку задачи факторного анализа.

4.8. Дайте графическую интерпретацию факторам для двумерного факторного пространства.

4.9. Сформулируйте математическую постановку задачи факторного анализа.

4.10. Для каких целей производится вращение факторов?

4.11. Для каких целей можно использовать факторный анализ при построении регрессионных моделей?

4.12. Какой смысл имеют коэффициенты матрицы факторных нагрузок?

4.13. Сформулируйте постановку задачи кластерного анализа.

4.14. Каковы принципы кластеризации в методе иерархического кластерного анализа?

4.15. Как происходит образование кластеров в методе К-средних?

4.16. Как определить актуальное число кластеров в методе иерархического кластерного анализа на основании таблицы шагов агломерации?

4.17. Может ли быть процесс кластеризации неоднозначным? Какие факторы влияют на итоговый результат объединения объектов в кластеры?

4.18. Какие меры расстояний чаще всего используются в SPSS для кластеризации объектов?

4.19. Как, имея матрицу расстояний между объектами, можно построить дендрограмму кластеризации?

4.20. Какой из методов кластерного анализа можно использовать в SPSS для определения похожих переменных (кластеризации переменных)?

4.21. Для каких целей используется метод построения деревьев решений?

4.22. Можно ли назвать построение дерева решений методом классификации объектов?

4.23. Дайте определение понятий «корневой узел», «терминальный узел», «дочерний узел».

4.24. Какие существуют методы отбора переменных для расщепления узла на два или несколько дочерних узлов?

4.25. В чем состоит проблема переобучения при построении дерева решений?

4.26. Назовите методы ограничения излишней «пушистости» дерева решений?

4.27. С помощью какого показателя можно судить о качестве классификации объектов при построении дерева решений?

4.28. Являются ли правила классификации объектов, извлеченные при построении дерева решений, однозначными?

4.29. В чем состоят различия обработки пропущенных значений при построении дерева решений и построении регрессионных уравнений?

4.30. В чем сходство и различие алгоритмов построения дерева решений CHAID, Exaustive CHAID, CRT и QUEST? Сформулируйте принципы работы этих алгоритмов.

Задачи и упражнения

4.1. Имеются данные по двум группам промышленных предприятий машиностроительного комплекса: x_1 – фондоотдача основных производственных фондов, руб.; x_2 – затраты на рубль произведенной продукции, коп.; x_3 – затраты сырья и материалов на рубль произведенной продукции, коп.

Группа	Номер предприятия	x_1	x_2	x_3
1-я	1	0,50	94,5	8,50
	2	0,67	75,4	8,79
	3	0,68	85,2	9,10
	4	0,55	98,8	8,47
2-я	5	1,52	81,5	4,95
	6	1,20	93,8	6,95
	7	1,46	86,5	4,70
Неизвестна	8	1,07	93,5	5,30
	9	0,99	84,0	4,85
	10	0,70	76,8	3,50
	11	1,24	88,0	4,95

Требуется произвести классификацию четырех новых предприятий с неизвестной принадлежностью к группе, имеющих значения исходных переменных, указанные в таблице.

Найдите значения коэффициентов дискриминантной функции. Используя найденные коэффициенты, классифицируйте предприятия, принадлежность к группе которых неизвестна, вручную.

4.2. В файле Задача_4_2.sav содержатся данные о кредитных рейтингах 2 464 клиентов банка (результативная переменная), пользовавшихся банковскими услугами в прошлом, а также такие данные о клиентах, как возраст (число лет); уровень дохода (низкий, средний высокий); число используемых кредитных карт (меньше 5, больше 5); уровень образования (высшее, среднее специальное) и число кредитов на покупку автомашины (нет или 1, 2 и более).

а) Сформируйте обучающую выборку, отобрав случайным образом 70 % клиентов банка и 30% -ю тестовую выборку.

б) Поведите дискриминантный анализ для объектов, включенных в обучающую выборку. Найдите значения коэффициентов дискриминантной функции, проверьте их статистическую значимость.

в) Проведите дискриминантный анализ для случая равных априорных вероятностей и априорных вероятностей, пропорциональных объему групп. Проверьте также, выполняется ли предположение о многомерной нормальности данных в группах.

г) Дайте словесное описание «портрета» клиента банка, имеющего плохой кредитный рейтинг, который можно составить на основании анализа коэффициентов дискриминантной функции. Сравните результаты классификации, полученные методом дискриминантного анализа, с результатами классификации на основе дерева решений (см. пример 4.6).

4.3. В файле Задача_4_3.sav[5] приведены данные о результатах голосования 145 конгрессменов в конгрессе США по целому ряду законопроектов и их принадлежность к демократической или республиканской партиям (принадлежность кодируется цифрой 0 для демократической и цифрой 1 – для республиканской партии). Результаты голосования по всем законопроектам кодируются цифрами: 0 – против; 1 – воздержался; 2 – поддержал законопроект. Для 5 конгрессменов имеются результаты голосования, но неизвестна их принадлежность к той или иной партии. Используя 130 первых случаев как обучающую выборку, постройте модель дискриминантного анализа для предсказания принадлежности конгрессменов к той или иной партии. Проверьте построенную модель, используя 15 случаев, для которых известна партийная принадлежность. Определите партийность конгрессменов, для которых известны результаты голосования, но неизвестна партийная принадлежность.

4.4. В файле Задача_4_4.sav приведены данные, представляющие собой результаты психологического тестирования учащихся специализированных школ Санкт-Петербурга с физико-математическим и гуманитарным уклоном. Всего предлагалось пять тестов, условное название которых приведено ниже: Тест_1 – дополнение предложений; Тест_3 – нахождение аналогий; Тест_4 – обобщение умозаключений; Тест_5 – способность к устному счету; Тест_7 – образность мышления. Чем выше набранный бал, тем лучше проявляется анализируемое качество. Учащиеся школ с физико-математическим уклоном условно названы физиками, а с гуманитарным уклоном – лириками.

а) Загрузите данные в редактор SPSS, создайте новую переменную Index, которая принимает целые случайные значения в интервале от 1 до 76.

б) Отсортируйте данные так, чтобы переменная Index была упорядочена по возрастанию.

[5] Задачи 4.3 – 4.5 уже встречались ранее (см. задачи 3.19 – 3.21). Здесь их предлагается решить другим методом.

в) Создайте модель дискриминантного анализа, используя первые 70 случаев. Прогностические способности проверьте, сравнивая истинные и прогнозные значения для оставшихся 6 случаев.

4.5. Имеется таблица данных с результатами 31-й предвыборной ситуации с 1860 по 1980 г. (файл Задача_4_5.sav). Для каждых выборов в таблице содержатся данные по 12 бинарным признакам.

1. Правящая партия была у власти более одного срока.
2. Правящая партия получила более 50 % голосов на прошлых выборах.
3. В год выборов была активна третья партия.
4. Была серьезная конкуренция при выдвижении кандидата от правящей партии.
5. Кандидат от правящей партии был президентом в год выборов.
6. Был ли год выборов временем спада или депрессии.
7. Был ли рост среднего национального валового продукта на душу населения более 2,1%.
8. Произвел ли правящий президент существенные изменения в политике.
9. Во время правления были существенные социальные волнения.
10. Администрация правящей партии виновна в серьезной ошибке или скандале.
11. Кандидат от правящей партии – национальный герой.
12. Кандидат от оппозиционной партии – национальный герой.

Также в таблице содержится информация о результатах выборов (победе правящей или оппозиционной партии). Значения бинарных признаков равны 0 (ответ «нет» для входного признака) и 1 (ответ «да» для входного признака). Значения результативного признака закодированы цифрой «1», если был избран кандидат правящей партии, и цифрой «2», если победил кандидат оппозиционной партии.

а) Постройте модель дискриминантного анализа, позволяющую по приведенным значениям факторных признаков предсказать результаты президентских выборов в США. В качестве объясняющих переменных выберите факторы под номерами 3, 4, 6, 7, 8, 9 в приведенном выше списке объясняющих переменных. Убедитесь, что модель точно воспроизводит все результаты выборов.

б) Проверьте, правильным ли является предсказание модели для выборов 1992 г. (Д. Буш – Б. Клинтон). Как известно, в этом году победил Клинтон – кандидат от оппозиционной партии.

в) Введите самостоятельно пропущенные значения для объясняющих переменных для выборов 1984, 1988, 1996, 2000 и 2004 гг. и проверьте прогностические возможности модели для этих выборов.

4.6. По мнению известного социолога Р. Инглехарта, в более зрелых возрастных группах значимо большее число человек высказываются в пользу материальных ценностей. Склонность к постматериалистическим ценностям зависит, по его мнению, также от уровня образования и профессиональной квалификации. Чем выше образование и профессиональная квалификация, тем выше склонность к постматериалистическим ценностям. Значение имеет и социально-экономический статус отца: чем он выше, тем больше доля постматериалистов.

В файле Задача_4_6.sav содержатся данные об индексе Инглехарта отношения к ценностям, уровне образования, социальном статусе отца, возрасте, уровне специального образования 3058 респондентов. При помощи дискриминантного анализа требуется подтвердить или опровергнуть гипотезу Инглехарта о смене ценностных ориентаций с возрастом.

а) Используя пункты меню Transform/Recode/Into Different Variables (Преобразование/Перекодировка/ В другую переменную), создайте новую переменную номинативного типа с именем Индекс. Эта переменная должна принимать значение «1» (постматериалистический тип), если переменная Индекс_Инглехарта принимает значения «1» или «2»; значение «2» (материалистический тип), если переменная Индекс_Инглехарта принимает значения «3» или «4», и значение «9» во всех остальных случаях. В окне Variable View установите значение «9» для переменной Индекс как значение, пропущенное пользователем.

б) Проведите дискриминантный анализ, взяв в качестве результативной переменной переменную Индекс.

в) Анализируя таблицу групповых статистик, определите, подтверждается ли на изучаемой выборке тезис Инглехарта?

4.7. В файле Задача_4_7.sav содержатся данные клинической диагностики двух групп больных гипертонией, для лечения которых применялось два вида препаратов: «Альфасан» и «Бетасан».

а) Используя дискриминантный анализ, выясните, можно ли считать, что применение этих двух видов лекарственных препаратов дает разную клиническую картину протекания болезни.

б) Какое из лекарств дает более быстрый эффект снижения артериального давления?

4.8. В файле Задача_4_8.sav содержатся данные о числе продаж и перепродаж через четыре года (тыс. шт.) автомобилей разных марок, а также

различные характеристики этих автомобилей (цена, мощность двигателя, база шасси, емкость топливного бака и т. д.).

а) Используя факторный анализ, выделите три главных фактора, определяющих характеристики автомобилей на автомобильном рынке.

б) Используя повернутую матрицу факторных нагрузок, интерпретируйте три первых главных фактора.

в) Определите, от каких факторов зависит объем перепродаж автомобилей.

4.9. Используя данные задачи 3.6, с помощью факторного анализа выясните, какими факторами в действительности следовало бы характеризовать экономическую деятельность изучаемых предприятий. Анализируя матрицу повернутых факторных нагрузок, попытайтесь дать интерпретацию факторам.

4.10. В файле Задача_4_10.sav приведены результаты тестирования интеллекта школьников. Данные содержат результаты 11 различных тестов для 46 респондентов. Смысловая направленность каждого теста определяется меткой, которой характеризуется переменная, содержащая результаты тестирования (значение меток можно увидеть в окне Variable View редактора данных SPSS). Результаты тестирования измерены в ординальной шкале, и большему баллу, набранному за данный тест, соответствует более сильное проявление анализируемого качества.

а) Проведите факторный анализ результатов тестирования, выделив четыре фактора.

б) Интерпретируйте, используя коэффициенты повернутой матрицы нагрузок, выделенные факторы.

4.11. В файле Задача_4_11.sav приведены статистические данные смертности населения РФ от различных причин (тыс. чел.) за период с 1970 по 2004 г. Используя факторный анализ, выясните, можно ли выделить факторы смертности населения РФ по наблюдениям за этот период. Если да, то постарайтесь дать интерпретацию этим факторам.

4.12. В файле Задача_4_11.sav приведены данные о предоставлении услуг 1000 абонентов узла связи. Служба маркетинга компании фиксирует 18 различных показателей, характеризующих диапазон услуг, потребляемых клиентом, и хотела бы выяснить, нельзя ли ограничиться меньшим количеством регистрируемых показателей. Загрузите данные в редактор SPSS и проведите факторный анализ. Выделите три первых главных фактора и, анализируя матрицу повернутых факторных нагрузок, интерпретируйте их. Можно ли интерпретировать факторы, используя исходную матрицу факторных нагрузок?

4.13. В файле Задача_4_13.sav приведены экономические показатели 186 крупнейших компаний РФ за 1997 г. В базу данных включены предприятия, вошедшие в рейтинг журнала «Эксперт» за 1997 год. Экономические показатели предприятий характеризуют восемь параметров. Используя факторный анализ, определите три главных фактора, с помощью которых можно объяснить около 90 % вариации факторных признаков. Используя матрицу повернутых факторных нагрузок, интерпретируйте эти факторы.

4.14. Рассмотрим модель типа «инфляция – производство», основанную на выборке из бухгалтерской отчетности ООО «Вавилон» (г. Чебоксары) поквартальных данных экономической деятельности за 1995 – 2002 гг. Инфляция за текущий год задается переменной I и выражается в процентах, а параметры экономической деятельности предприятия – переменными x1 – x7 (смысловое содержание этих переменных можно увидеть в окне Variable View редактора данных SPSS).

а) С помощью факторного анализа выясните, можно ли экономическую деятельность предприятия ООО «Вавилон» характеризовать меньшим числом параметров.

б) Какой из факторов экономической деятельности в действительности связан с уровнем инфляции?

в) Постройте регрессионную модель взаимосвязи инфляции (переменная I) с параметрами экономической деятельности ООО «Вавилон», которые характеризуются переменными x1 – x7.

г) Постройте аналогичную модель линейной регрессии с выделенными факторами в качестве объясняющих переменных. Объясните, почему регрессионная модель, где объясняющими переменными являются факторы, оказалась хуже.

4.15. База данных риелторской фирмы содержит 1721 запись оценок стоимости жилья в одном из городов РФ. Для создания правил оценки стоимости жилья, кроме стоимости, фиксировалось еще 11 различных параметров квартиры (база данных содержится в файле Задача_4_15.sav).

а) Используя факторный анализ, выделите 4 фактора, которые характеризуют примерно 70 % вариации факторных признаков.

б) Используя матрицу повернутых факторных нагрузок, дайте словесную интерпретацию трем первым факторам.

в) Постройте регрессионную модель стоимости жилья, в которой в качестве объясняющих переменных выступают три первых фактора.

4.16. В файле Задача_4_16.sav приведен поименный состав сборной РФ по баскетболу в 2005 г., физические характеристики баскетболистов (возраст, рост, вес) и их амплуа в команде (защитник, нападающий, центровой).

а) Используя метод К-средних с тремя кластерами, убедитесь, что физические характеристики игроков почти однозначно определяют их функции в команде.

б) Используя иерархический кластерный анализ, убедитесь, что процедура кластеризации не является однозначной и зависит как от избранного метода, так и от способа измерения расстояний.

4.17. Классической задачей кластерного анализа является задача классификации ирисов по длине и ширине чашелистиков и длине и ширине листков, рассмотренная впервые Р. А. Фишером в 1936 г. В файле Задача_4_17.sav содержатся данные о длине и ширине чашелистиков, длине и ширине листков 150 ирисов трех типов – *Iris Setosa*; *Iris Versicolor*; *Iris Virginica*.

а) Используя метод К-средних для трех кластеров, решите задачу кластеризации ирисов.

б) Определите, какой процент случаев классифицирован верно.

4.18. В файле Задача_4_18.sav приведены данные голосования избирателей за кандидатов различных партий, баллотировавшихся в Государственную Думу в 2003 г. по различным федеральным избирательным округам Москвы и Московской области, Екатеринбурга и Свердловской области, Тюмени и Тюменской области.

а) Используя иерархический кластерный анализ, выделите группы избирательных округов со схожими итогами голосования. Правильно выберите меру расстояния между объектами. Определите на основании таблицы шагов агломерации разумное количество кластеров, которое следует анализировать в данной задаче. Постройте дендрограмму. Дайте интерпретацию полученным результатам.

б) Используя иерархический кластерный анализ переменных, определите, какие группы партий имеют схожие результаты по итогам голосования избирателей. Какие группы партий наиболее слабо и наиболее сильно различаются между собой? Число кластеров, которые здесь следует анализировать, определите на основании таблицы шагов агломерации. Дайте словесную интерпретацию полученным результатам.

4.19. В файле Задача_4_19.sav приведены некоторые данные, характеризующие производственную деятельность основных нефтяных компаний РФ в 1996 г. Используя иерархический кластерный анализ, определите группы схожих по производственным показателям компаний. Поскольку числовые характеристики показателей сильно различаются по величине, используйте преобразование переменных к стандартизованным значениям. Определите на основании таблицы шагов агломерации разумное количество кластеров, которое следует анализировать в данной задаче. Постройте дендрограмму, а от построения сосульчатой диаграммы откажитесь.

4.20. База данных риелторской фирмы содержит 1721 запись оценок стоимости проданного жилья в одном из городов РФ. Для создания правил предварительной оценки стоимости жилья, кроме продажной стоимости, фиксировалось еще 11 различных параметров квартиры. База данных содержится в файле Задача_4_15.sav.

а) Методом К-средних проведите кластерный анализ квартир с целью выделить четыре типа квартир, имеющихся на рынке. Поскольку данные измерены как в номинативной, так и интервальной шкалах, предусмотрите переход к стандартизованным значениям переменных при проведении вычислений. Используя таблицу конечных центров кластеров, интерпретируйте полученное разбиение квартир на классы.

б) Методом иерархического анализа проведите кластеризацию переменных, выделив группы переменных, характеризующих одну сущность. Пользуясь таблицей принадлежности к кластерам, интерпретируйте полученное разделение переменных на классы.

4.21. В файле Задача_4_21.sav приведены статистические данные миграционного оттока граждан РФ в различные страны дальнего и ближнего зарубежья.

а) Используя иерархический кластерный анализ, выделите группы стран с похожими миграционными потоками. Число кластеров, которые следует выделить в этой задаче, выберите на основании дендрограммы. Для анализа используйте стандартизованные значения переменных.

б) Выполнив иерархический кластерный анализ переменных, выясните, произошли ли какие-либо изменения в процессах миграции за период 1997 – 2005 гг.

в) Повторите проведенные выше исследования, используя данные миграционного притока граждан различных стран на территорию РФ. Данные миграционного притока по годам за период с 1997 по 2005 г. содержатся в файле Задача_4_21а.sav. Сравните миграционные процессы притока и оттока населения в РФ и дайте словесное описание происходящих процессов.

4.22. В гл. 1 мы рассматривали проблему классификации изображений на два класса (см. рис. 1.3). В файле Задача_4_22.sav содержатся значения 16 дихотомических переменных x1 – x16, характеризующих изображения лиц, а переменная x17 содержит данные о том классе, к которому следует отнести это изображение.

а) Используя метод CRT построения дерева решений, сформулируйте правила, которыми руководствовался художник при отнесении изображения к тому или иному классу. Поскольку объектов очень мало и нужно получить исчерпывающие правила классификации, то следует установить минимальное число объектов: для родительского узла – 2, а для дочернего узла – 1.

б) Получите правила классификации, используя другие методы построения дерева решений (CHAID, QUEST). Возможно, что для построения дерева решений потребуется увеличить значения уровня значимости при решении вопроса о расщеплении узла до значения, равного 0,1. Объясните, почему результаты классификации в этом случае оказались менее удовлетворительными.

в) Попробуйте в качестве первой переменной, по градациям которой производится расщепление нулевого узла, выбрать переменную по своему усмотрению. Проанализируйте, как при этом изменяются правила классификации.

4.23. Имеется 100 объектов, которые характеризуются двумя переменными x_1 и x_2, принимающими числовые значения от 0 до 9 (рис. 1.5). Результативная переменная может принимать всего лишь два значения: «крестик» или «нолик» в соответствии с правилами:

если $x_1 > 4$ и $x_2 < 5$, тогда класс 1 – крестики;

если $x_1 < 5$ и $x_2 > 4$, тогда класс 1 – крестики;

если $x_1 < 5$ и $x_2 < 5$, тогда класс 2 – нолики,

если $x_1 > 4$ и $x_2 > 4$, тогда класс 2 – нолики.

В соответствии с этими правилами подготовлен файл Задача_4_23.sav.

а) Используя алгоритм CHAID построения дерева решений, извлеките из данных этого файла правила, приведенные выше. При запуске алгоритма на исполнение следует установить число интервалов разбиения независимых переменных, равное 2, число объектов в родительском узле – 10, в дочернем – 5. Правила извлекайте в форме простого текста.

б) Попробуйте извлечь те же правила, используя другие алгоритмы построения дерева решений.

4.24. Используя базу данных риелторской фирмы (файл Задача_4_15), с помощью дерева решений создайте правила, по которым назначается низкая (меньше 50 тыс. долл.), средняя (от 50 тыс. до 100 тыс. долл.) и высокая (более 100 тыс. долл.) продажная стоимость квартиры.

а) Перекодируйте целевую переменную или создайте новую номинативную переменную, имеющую три упомянутых выше градации.

б) Для предотвращения излишней «пушистости» дерева решений установите минимальное значение числа объектов в родительском и дочерних узлах, равное 200 и 100 соответственно.

в) Запустите алгоритм CHAID построения дерева решений. Дайте словесную интерпретацию извлеченным правилам.

4.25. В файле Задача_4_25.sav содержатся данные о длине и ширине чашелистиков, длине и ширине листков 150 ирисов трех типов: *Iris Setosa* (кодируется цифрой 1); *Iris Versicolor* (кодируется цифрой 2); *Iris Virginic*а (кодируется цифрой 3). Требуется построить правила классификации ирисов по типам, используя 70 % данных, приведенных в файле, как обучающую выборку и оставшиеся 30 % – как тестовую выборку.

а) После загрузки файла в окно редактора данных SPSS создайте новую переменную I, значениям которой с помощью функции RV.UNIFORM(0,1) присвойте равномерно распределенные в интервале 0 – 1 случайные величины. Затем отсортируйте данные по возрастанию значений переменной I. Эти преобразования желательно выполнить, потому что исходные данные упорядочены по типу ирисов. Строго говоря, эта дополнительная рандомизация не является обязательной, поскольку тестовая выборка в любом случае формируется случайным выбором объектов.

б) Постройте дерево решений и оцените его качество по значению показателя «Риск». При запуске алгоритма на исполнение следует установить число интервалов разбиения независимых переменных, равное 3, число объектов в родительском узле – 2, в дочернем –1. Правила извлекайте в форме простого текста. Интерпретируйте словесно правила классификации ирисов, извлеченные из базы данных.

4.26. На 43 опытных участках выращивался рис с различной агротехникой возделывания культуры. Фиксировалось пять параметров агротехники: x_1 – предшественник (в баллах); x_2 – количество внесенных удобрений (ц на 1 га); x_3 – прополка (раз); x_4 – число дней от залива чеков до сброса воды; x_5 – число дней от косовицы до обмолота. Кроме того, фиксировалась урожайность культуры, которая считалась низкой, если урожайность была меньше 40 ц/га, и высокой, если она превышала значение 40 ц/га. Результаты исследования приведены в файле Пример_4_26.sav. Используя различные варианты построения дерева решений, дайте обоснованные рекомендации по агротехнике возделывания риса, которые можно сделать на основании проведенного эксперимента.

4.27. В объявлениях о приеме на работу новых сотрудников очень часто встречаются ограничения по возрасту. Насколько обоснованны эти ограничения? Были проведены специальные исследования на группе из 60 станочников, выпускающих одинаковую продукцию в одинаковых условиях. Эта группа, в свою очередь делилась на подгруппы по возрасту работников и по стажу работы. Производительность измерялась количеством произведенной продукции за единицу времени в натуральном выражении. Результаты исследования приведены в файле Задача_4_27.sav.

а) После загрузки файла данных в редактор SPSS перекодируйте переменную Выработка, задав градации: 18 –25 единиц – Низкая; 25 – 35 единиц – Средняя; 35 – 45 единиц – Высокая. Не забудьте явно задать номинативный тип вновь созданной переменной.

б) Используя дерево решений, выясните, каким критериям удовлетворяют работники с низкой и высокой производительностью труда.

4.28. Решите задачу, сформулированную в примере 4.7, используя алгоритмы построения дерева решений CHAID, Exhaustive CHAID. Сравните полученные результаты для объясненной доли дисперсии и коэффициента корреляции истинных и предсказанных значений с результатами, полученными при использовании метода CRT в примере 4.7.

4.29. База данных риелторской фирмы (файл Задача_4_29.sav) содержит 1300 записей о параметрах квартиры и ее продажной стоимости.

а) Используя метод CRT, создайте правила определения стоимости квартиры и сохраните эти правила в файле. Оцените качество модели, вычислив долю объясненной моделью дисперсии η^2.

б) Используя созданные правила, предскажите продажную стоимость квартир, параметры которых приведены в файле Задача_4_29a.sav. Поскольку в файле оставлены данные и об истинной продажной стоимости квартир, найдите коэффициент корреляции истинных и предсказанных значений.

в) Повторите проделанные выше вычисления, используя метод CHAID.

ГЛАВА 5. НЕЙРОННЫЕ СЕТИ КАК СРЕДСТВО ДОБЫЧИ ЗНАНИЙ

5.1. Принципы организации нейронных сетей

Под *нейронными сетями* понимаются вычислительные структуры, которые моделируют процессы хранения и обработки информации в биологических системах. Архитектура этих вычислительных систем принципиально отличается от архитектуры традиционных ЭВМ. Они представляют собой распределенные системы, способные к параллельным вычислениям и обучению на примерах, путем накапливания информации о «положительных» и «отрицательных» воздействиях.

Элементарным преобразователем данных в нейронных сетях является *нейрон*, названный так по аналогии с биологическим прототипом, который, как предполагается, выполняет в нейронной сети примерно те же функции, что и биологический нейрон в коре головного мозга человека. В действительности аналогия является достаточно поверхностной, но квазибиологическая терминология довольно прочно закрепилась в этой области информационных технологий, и мы будем ее придерживаться.

Нейронная сеть представляет собой совокупность нейронов, соединенных между собой в соответствии с некоторой выбранной архитектурой. Как правило, нейронная сеть имеет входы, на которые поступает обрабатываемая информация, и выходы, на которые направляется информация о результатах работы сети.

Искусственные нейроны и нейронные сети могут представлять собой как реальные аналоговые устройства, реализованные в виде определенных радиотехнических устройств, так и эмулироваться с помощью цифровых вычислительных машин. При обсуждении принципов работы нейронных сетей их конкретная реализация не играет большой роли, но в дальнейшем мы всегда будем иметь в виду нейронные сети, работа которых эмулируется с помощью программных средств на обычных ЭВМ.

Искусственный нейрон

Несмотря на большое разнообразие вариантов нейронных сетей, все они имеют общие черты. Так, все они, так же как и мозг человека, состоят из большого числа связанных между собой однотипных эле-

ментов – нейронов, которые имитируют нейроны головного мозга. На рис. 5.1 показана схема искусственного (формального) нейрона.

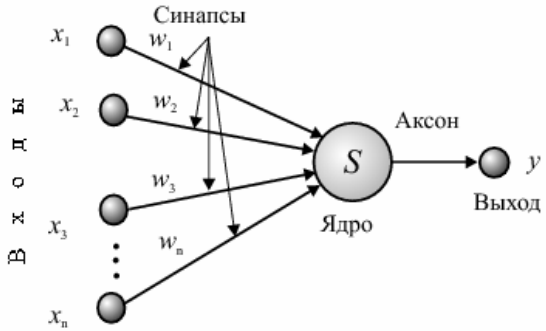


Рис. 5.1. Структура искусственного нейрона

На рис. 5.1 нейрон имеет n входов $x_1, x_2, \ldots, x_n$, на которые могут поступать сигналы либо из внешней среды, либо от других нейронов. Формальный нейрон, так же как и биологический, состоит из синапсов, связывающих входы нейрона с ядром, ядра нейрона, которое осуществляет обработку входных сигналов, и аксона, который связывает нейрон с нейронами следующего слоя. Каждый синапс имеет вес w, который определяет, насколько соответствующий вход нейрона влияет на его состояние. В итоге на выходе из ядра возникает итоговый сигнал

$$S = \sum_{i=1}^{n} w_i \cdot x_i + b, \tag{5.1}$$

где w_i – вес синапса, b – значение смещения. Аксон связывает ядро нейрона с его выходом и производит дополнительное в общем случае нелинейное преобразование сигнала:

$$y = f(S). \tag{5.2}$$

Функция $f(S)$ называется функцией активации нейрона. Чаще других используется функция активации, которая известна как логистическая функция, или сигмоид

$$f(S) = \frac{1}{1 + e^{-a \cdot S}}. \tag{5.3}$$

Величина a является параметром этой функции. При очень малых значениях этого параметра функция (5.3) вырождается в прямую линию, а при $a \to \infty$ $f(S)$ превращается в ступенчатую функцию

$$f(S) = \begin{cases} 0 & \text{при } S \le 0; \\ 1 & \text{при } S > 0. \end{cases}$$

Широкая распространенность сигмоида в качестве активационной функции объясняется тем, что она дифференцируема на всей действительной оси, имеет простую производную и выходные значения этой функции лежат в интервале (0, 1).

Таким образом, типичный формальный нейрон производит простейшую операцию – взвешивает значения своих входов со своими же локально хранимыми весами и производит над их суммой нелинейное преобразование. Нелинейность выходной функции активации принципиальна. Если бы нейроны были линейными элементами, то любая последовательность нейронов также производила бы линейное преобразование и вся нейросеть была бы эквивалентна одному нейрону (или одному слою нейронов в случае нескольких выходов). Нелинейность разрушает линейную суперпозицию и приводит к тому, что возможности нейросети существенно выше возможностей отдельных нейронов.

5.2. Классификация нейронных сетей

Нейронная сеть представляет собой совокупность формальных нейронов, определенным образом соединенных друг с другом и с внешней средой. В зависимости от функций, выполняемых нейронами в сети, можно выделить три их типа:

1) *входные нейроны*, на которые подается сигнал из внешней среды; в них обычно не осуществляется вычислительных процедур и информация передается с входа на выход с использованием функции активации типа (5.2), (5.3);
2) *нейроны промежуточных слоев – основа нейронной сети; пре*образование сигнала на этих нейронах производится в соответствии с формулами (5.1), (5.2);
3) *выходные нейроны* – это нейроны, выходы которых представляют собой выходы сети; преобразование сигнала на этих нейронах производится также в соответствии с формулами (5.2), (5.3).

С точки зрения топологии нейронных сетей можно выделить *полносвязные* сети, в которых все нейроны сети связаны между собой. Такая архитектура используется лишь в специализированных нейронных сетях, и здесь мы не будем на них останавливаться.

В *многослойных, или слоистых*, нейронных сетях нейроны объединяются в слои. Слой содержит совокупность нейронов, с одинаковыми входными сигналами. Число нейронов в слое может быть любым и не зависеть от числа нейронов в других слоях. Слои нейронов упорядочены слева направо, и выходной сигнал предыдущего слоя является входным для всех нейронов следующего слоя. Внешние сигналы подаются на входы нейронов первого слоя, а выходами сети являются выходные сигналы последнего слоя. Схема такой нейронной сети представлена на рис. 5.2. Если в слоистой нейронной сети оставить лишь некоторую часть связей между нейронами, то такие сети называются *слабосвязанными,* или *сетями с локальными связями*.

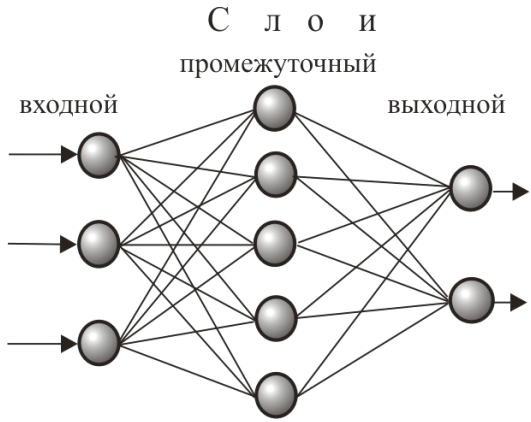


Рис. 5.2. Слоистая сеть с одним промежуточным слоем

Представленная на рис. 5.2 нейронная сеть является сетью прямого распространения без обратных связей. В сетях с *обратными связями* информация с последующих слоев может снова подаваться на вход предыдущего слоя.

В учебном пособии мы познакомим читателя лишь с принципами использования нейросетевого подхода для решения задач экономики и управления. Чаще других для этих целей используются слоистые сети без обратных связей. По этой причине в дальнейшем мы будем рас-

сматривать исключительно только сети с такой архитектурой. Впервые такого рода нейросети были предложены Ф. Розенблатом и получили название *персептрона* (от англ. *perception* – восприятие). Персептрон был впервые смоделирован на универсальной ЭВМ IBM-704 в 1958 г. Аппаратный вариант Mark I Perceptron был построен в 1960 г. и предназначался для распознавания зрительных образов. Его рецепторное поле состояло из 400 пикселей (матрица фотоприемников 20 х 20), и он успешно справлялся с решением ряда задач – мог различать некоторые буквы.

В последующие годы это направление информационных технологий находилось в забвении по причине ограниченных возможностей вычислительной техники того времени, с одной стороны, и отсутствия строгих результатов, позволяющих обосновать возможность применения нейросетевого подхода для решения практически важных задач, с другой стороны.

Настоящий бум практического применений нейросетей начался после публикации Румельхартом с соавторами в 1986 г. метода обучения многослойного персептрона, названного ими методом обратного распространения ошибки. В это же время (1987 г.) была опубликована работа Р. Хехт-Нильсена, которая в неконструктивной форме доказывала возможность представления произвольной функции многих переменных на нейронной сети и позволяла оценить минимальное число нейронов сети, необходимых для решения поставленной задачи.

Таким образом, было доказано, что многослойный персептрон с достаточным количеством внутренних нейронов и подходящей матрицей связей потенциально способен осуществлять любое преобразование сигнала «вход – выход», аппроксимировать любую функцию с наперед заданной точностью. Для этого необходимо:

– выбрать архитектуру персептрона, а именно число слоев и число нейронов в слое, форму активационной функции;

– настроить веса связей (синапсов) w нейронов с помощью некоторого алгоритма обучения.

После этого нейронная сеть будет готова к работе, и каждому входному сигналу будет сопоставлен выходной сигнал, который нейронная сеть найдет, используя полученные в результате обучения веса связей.

5.3. Обучение нейронной сети. Алгоритм обратного распространения ошибки

Процедура обучения нейронной сети зависит от ее архитектуры и класса решаемых задач. Как уже указывалось, мы будем рассматривать только многослойные сети без обратных связей, которые являются основной «рабочей лошадкой» современного нейрокомпьютинга. Подавляющее большинство приложений связано именно с применением таких многослойных персептронов.

В немалой степени популярность персептронов обусловлена широким кругом доступных им задач. В общем случае, они решают задачу аппроксимации многомерных функций, т. е. построения многомерного отображения, обобщающего заданный набор примеров.

В зависимости от типа выходных переменных (тип входных не имеет решающего значения), аппроксимация функций может сводиться:

а) к задаче классификации – аналогу задачи дискриминантного или кластерного анализа (дискретный набор выходных значений переменных);

б) к задаче построения прогноза для функции многих переменных – аналогу регрессионной задачи (непрерывные выходные значения переменных).

Причина популярности персептронов кроется в том, что для своего круга задач они являются, во-первых, универсальными, а во-вторых, эффективными с точки зрения вычислительной сложности устройствами.

Для решения каждого из классов задач требуется свой алгоритм обучения нейронной сети.

Начнем изучение процессов обучения нейронной сети с рассмотрения базового алгоритма, который известен как *алгоритм обратного распространения ошибок*. Ключ к обучению многослойных персептронов был настолько простым и очевидным, что, как выяснилось, он неоднократно переоткрывался различными авторами. Тем не менее рождение алгоритма обратного распространения ошибок связано с именами Румельхарта, Хинтона и Уильямса, опубликовавшими в 1986 г. статью, в которой была изложена теория обучения многослойного персептрона.

Алгоритм обратного распространения ошибок – это итеративный градиентный алгоритм обучения, который используется с целью минимизации среднеквадратического отклонения текущих и требуемых выходов многослойной нейронной сети с последовательными связя-

ми. Минимизируемой целевой функцией ошибки нейронной сети является величина суммарной ошибки

$$E(w) = \sum_{j,k} (y_{j,k}^{Q} - d_{j,k})^2 , \qquad (5.4)$$

где $y_{j,k}^{Q}$ – реальное выходное состояние нейрона j выходного слоя нейронной сети при подаче на ее входы значений факторных переменных k-го объекта; $d_{j,k}$ – требуемое выходное состояние этого нейрона. Суммирование в формуле (5.4) ведется по всем нейронам выходного слоя и всем обрабатываемым сетью объектам.

В классическом алгоритме обратного распространения ошибок в качестве целевой функции используется величина ошибки сети для одного из обучающих сигналов:

$$E(w) = \sum_{j} (y_{j}^{Q} - d_{j})^2 , \qquad (5.4^*)$$

и задача оптимизации суммарной ошибки (5.4) в классическом методе обратного распространения рассматривается как набор частных задач оптимизации (5.4*): на каждой итерации происходят изменения значений параметров сети, улучшающие работу лишь с одним примером обучающей выборки, а суммарная ошибка может использоваться как критерий окончания процесса обучения.

На начальном этапе обучения всем весам присваиваются некоторые случайные значения. Затем, в результате многократно повторяющейся процедуры предъявления данных обучающей выборки, веса подстраиваются, с тем чтобы минимизировать ошибку (5.4*). Когда функционал ошибки задан и задача сводится к его минимизации, можно предложить, например, следующую итерационную процедуру подбора весов:

$$w_{ij}^{q}(n) = w_{ij}^{q}(n-1) + \Delta w_{ij}^{q} . \qquad (5.5)$$

В этой формуле $w_{ij}^{q}(n)$ – весовой коэффициент синаптической связи, соединяющей i-й нейрон слоя $q-1$ с j-м нейроном слоя q на шаге итерации с номером n,

$$\Delta w_{ij}^{q} = -\eta \cdot \frac{\partial E}{\partial w_{ij}} . \qquad (5.6)$$

В формуле (5.6) η – коэффициент, определяющий скорость обучения.

Для того чтобы алгоритм позволял действительно минимизировать функцию (5.4), необходимо найти поправки Δw_{ij}^{q} к синаптическим весам всех нейронов. Выведем для этих целей рекуррентное соотношение, связывающее, по существу, поправки к весам связей нейронов, находящихся в слое q и слое $q-1$. Учитывая, что в соответствии с формулами (5.1) – (5.3)

$$y_j^q = f(S_j),\ S_j = \sum_i w_{ij} y_i^{q-1} \tag{5.7}$$

и пользуясь правилом дифференцирования сложной функции, производную в правой части формулы (5.6) запишем в виде

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial S_j} \cdot \frac{\partial S_j}{\partial w_{ij}}. \tag{5.8}$$

Для простоты мы считаем, что смещения $b_j = 0$, и поэтому в формуле (5.7) смещения опущены. Поскольку активационной функцией является сигмоид (5.3) или гиперболический тангенс, то производная $\partial y / \partial S$ легко может быть найдена. Для сигмоидной функции (5.3) получаем

$$\frac{\partial y_j}{\partial S_j} = a \cdot y_j(S_j) \cdot (1 - y_j(S_j)). \tag{5.9}$$

Последняя производная в формуле (5.8) определяет выход нейрона j предыдущего слоя. Это становится очевидным, если учесть второе из равенств в формуле (5.7). Таким образом,

$$\frac{\partial S_j}{\partial w_{ij}} = y_i^{q-1}. \tag{5.10}$$

Для того чтобы получить рекуррентное соотношение, связывающее между собой слои q и $q+1$, выразим первую производную в формуле (5.8) через характеристики нейронов следующего слоя. Сделать это всегда возможно, поскольку выходные сигналы нейронов слоя q являются входными для последующего слоя – см. формулу (5.7). В результате получаем

$$\frac{\partial E}{\partial y_j^q} = \sum_r \frac{\partial E}{\partial y_r^{q+1}} \cdot \frac{\partial y_r^{q+1}}{\partial S_r^{q+1}} \cdot \frac{\partial S_r^{q+1}}{\partial y_j^q} = \sum_r \frac{\partial E}{\partial y_r^{q+1}} \cdot \frac{\partial y_r^{q+1}}{\partial S_r^{q+1}} \cdot w_{jr}^{q+1}. \tag{5.11}$$

В формуле (5.11) суммирование по r выполняется по всем нейронам в слое $q+1$.

Соотношение (5.11) позволяет получить искомое рекуррентное соотношение. Введем новую величину

$$\delta_j^q = \frac{\partial E}{\partial y_j^q} \cdot \frac{\partial y_j^q}{\partial S_j^q}. \qquad (5.12)$$

Умножая левую и правую части формулы (5.11) на величину $\partial y_j^q \Big/ \partial S_j^q$, получаем

$$\delta_j^q = \left(\sum_r \delta_r^{q+1} \cdot w_{jr}^{q+1} \right) \cdot \frac{\partial y_j^q}{\partial S_j^q}. \qquad (5.13)$$

Формула (5.13) позволяет вычислять поправки к весам синаптических связей слоя q, если известны веса для слоя $q+1$. Действительно, учитывая формулы (5.6), (5.8), (5.10) и (5.13), получаем

$$\Delta w_{ij}^q = -\eta \cdot \delta_j^q \cdot y_i^{q-1} = -\eta \cdot \left(\sum_r \delta_r^{q+1} \cdot w_{jr}^{q+1} \right) \cdot \frac{\partial y_j^q}{\partial S_j^q} \cdot y_i^{q-1}. \qquad (5.14)$$

Теперь, если заданы веса w_{jr}^Q и функции δ_j^Q для выходного слоя, по формулам (5.14) можно рассчитать веса синаптических связей для нейронов всех слоев, двигаясь от выходного слоя к входному в направлении, обратном естественному распространению сигнала в сети.

Синаптические веса в выходном слое найти нетрудно, поскольку известны значения входных сигналов y_j^{Q-1} для выходного слоя и известны выходные сигналы d_j. Величины δ_j^Q можно найти прямым дифференцированием формулы (5.4*):

$$\delta_j^Q = 2 \cdot (y_j^Q - d_j) \cdot \frac{\partial y_j^Q}{\partial S_j^q}. \qquad (5.15)$$

Таким образом, полный алгоритм обучения нейронной сети может быть построен следующим образом.

Шаг 1. Задать случайные веса для всех синаптических связей из некоторого набора возможных значений.

Шаг 2. Подать на вход сети один из возможных образов (переменные, характеризующие некоторый объект в обучающей выборке) и в режиме обычного функционирования сети, когда сигналы распространяются от входа к выходу, рассчитать значения величин

$$y_j^q = f(S_j^q) \text{ и } S_j^q = \sum_i w_{ij}^q \cdot y_j^{q-1}.$$

Шаг 3. Рассчитать по формуле (5.15) значения величин δ_j^O для выходного слоя (5.14) поправки к синаптическим весам, двигаясь от последнего слоя к первому.

Шаг 4. Скорректировать по формуле (5.5) веса синаптических связей и запомнить новые веса. Если суммарная ошибка, вычисляемая по формуле (5.4), меньше некоторого наперед заданного значения, то завершить процедуру подбора весов, иначе вернуться к шагу 1.

Возможен и другой критерий окончания обучения. Обучение сети прекращается, если ошибка для каждого примера на данной итерации не превысила заданного значения (т. е. сеть успешно обработала все поданные на ее вход примеры).

Рассмотренный алгоритм имеет линейную скорость сходимости: при общем числе весов в сети N число операций, необходимых для подбора весов, растет пропорционально N. Алгоритмы прямого перебора имеют по меньшей мере квадратичную сходимость, т. е. объем вычислений здесь растет пропорционально второй степени числа узлов сети.

Для ускорения процесса обучения были предложены многочисленные модификации алгоритма обратного распространения. В частности, упоминания заслуживает модификация алгоритма обратного распространения ошибки, которая была предложена М. Ридмиллером и Г. Брауном. Этот алгоритм использует только знаки частных производных для подстройки весовых коэффициентов. Алгоритм использует так называемое «обучение по эпохам», когда коррекция весов происходит после предъявления сети всех примеров из обучающей выборки.

Эпохой при таком подходе называется цикл обучения, при котором предъявляются все объекты обучающей выборки. В современных нейросетях чаще других используется подход Ридмиллера и Брауна.

Нейронная сеть, обученная в соответствии с алгоритмом обратного распространения ошибок, пригодна для решения задач, традиционно относящихся к регрессионному и дискриминантному анализу. Для решения задач кластеризации требуется совершенно иной алгоритм обучения, который мы рассмотрим в одном из следующих пунктов.

Оценка числа нейронов в скрытых слоях

Адекватный выбор количества нейронов и слоев – серьезная и нерешенная проблема для нейронных сетей. Если емкость сети недостаточна, то она не в состоянии научиться правильно распознавать все

объекты, содержащиеся в обучающей выборке, и найти взаимосвязи, имеющиеся в данных. Если число слоев и нейронов слишком велико, то сеть может переобучиться, т. е. она научится точно распознавать все детали обучающей выборки с учетом и случайных ошибок, неизбежно содержащихся в данных, но совершенно не в состоянии будет осуществлять прогноз для новых данных, не входящих в выборку.

Для оценки числа нейронов в скрытых слоях однородных сетей (имеющих одну и ту же передаточную функцию) можно пользоваться формулой для необходимого числа синаптических весов L_w в многослойной сети с сигмоидальными передаточными функциями

$$\frac{m \cdot N}{1+\log_2 N} \le L_w \le m \cdot (\frac{N}{m}+1) \cdot (n+m+1) + m\,, \qquad (5.16)$$

где n – размерность входного сигнала (число объясняющих переменных при построении аналога регрессионной модели), m – размерность выходного сигнала ($m=1$ для аналога однофакторной регрессии), N – число элементов в обучающей выборке[6].

На практике формула (5.16) может использоваться лишь для весьма грубых оценок. Основным способом выбора остается прямой перебор различного количества слоев, числа нейронов в слое и определение лучшего варианта по конечному результату. Критерием здесь, как правило, является величина среднеквадратической ошибки сети. Для этого требуется каждый раз заново создавать сеть. Информация, накопленная в предыдущих сеансах обучения, теряется полностью. Начинать перебор количества нейронов можно как с заведомо избыточного, так и с заведомо недостаточного их числа. При этом новая сеть с другим количеством нейронов требует полного переобучения.

Проблема обобщения и контроля качества обучения нейронной сети

Обобщение – это способность сети давать близкий к правильному результат для входных объектов, которых не было в обучающем множестве. Если бы нейросети не обладали такой способностью, они были бы лишь механизмом запоминания, а не обработки информации.

---

[6] *Widrow B.*, *Lehr M. A*. Proceedings of the IEEE. 1990. Vol. 78, nr. 9, Sept./ P. 1415-1442.

Поэтому важнейшим качеством нейросети является способность дать хороший результат для объектов, с которыми сеть раньше не встречалась. Можно сформулировать некоторые предпосылки для успешного обобщения.

• Неизвестные объекты должны не слишком отличаться от объектов обучающего множества. По крайней мере, взаимосвязь факторных и результативных признаков для этих объектов должна быть такой же, как и для объектов обучающей выборки. На языке математической статистики это условие означает, что обучающие примеры и новые объекты принадлежат одной генеральной совокупности и их статистические характеристики одинаковы.

• Основной закон, по которому сетью должно быть проведено обобщение, не должен быть скрыт несущественными закономерностями в обучающем множестве. Поэтому входы и выходы сети должны быть подготовлены так, чтобы максимально выявить закон, по которому они должны быть обобщены. В частности, грубые ошибки должны быть предварительно устранены из входного набора данных.

Рассмотрим теперь проблему переобучения сети, которая тесно связана с проблемой обобщения.

Переобучение – это чрезмерно точная подгонка ответов сети по данным обучающей выборки, которая имеет место в том случае, если алгоритм обучения работает достаточно долго, а сеть слишком сложна для поставленной задачи или имеющегося объема данных.

Суть этой проблемы лучше всего объяснить на конкретном примере. Пусть обучающие примеры порождаются некоторой функцией одной переменной, которую нам и хотелось бы воспроизвести. В теории обучения такую функцию называют учителем. При конечном числе обучающих примеров N всегда возможно построить нейросеть с нулевой ошибкой обучения, т. е. ошибкой, определенной на множестве обучающих примеров. Для этого нужно взять сеть с числом весов L_w большим, чем число примеров. Действительно, чтобы воспроизвести каждый пример, у нас имеется L_w уравнений для N неизвестных. Поскольку число неизвестных меньше числа уравнений, то такая система допускает бесконечное множество решений.

В этом-то и состоит основная проблема: у нас нет информации, чтобы выбрать правильное решение – веса, моделирующие функцию-учителя. В итоге выбранная случайным образом функция даст плохие предсказания на новых примерах, отсутствовавших в обучающей выборке, хотя последнюю сеть воспроизвела без ошибок. Вместо того

чтобы обобщить известные примеры, сеть запомнила их. Этот эффект и называется переобучением.

На самом деле задачей обучения является не минимизация ошибки обучения на обучающей выборке, а минимизация ошибки обобщения, определенной для всех возможных в будущем примеров. Именно такая сеть будет обладать максимальной предсказательной способностью. И трудность здесь состоит в том, что реально наблюдаемой является только ошибка обучения. Ошибку обобщения можно лишь оценить, опираясь на те или иные соображения.

Таким образом, стоит задача правильного выбора сложности сети. Почти всегда более сложная сеть дает меньшую ошибку на обучающей выборке, но это может свидетельствовать не о хорошем качестве построенной модели, а о переобучении сети.

Поскольку ошибка обобщения определена для данных, которые не входят в обучающее множество, очевидным решением проблемы служит разделение всех имеющихся в нашем распоряжении данных на два множества: обучающее, на котором подбираются конкретные значения весов, и кросс-проверочное (валидационное), на котором оцениваются предсказательные способности сети и выбирается оптимальная сложность модели. На самом деле должно быть еще и третье – тестовое множество, которое вообще не влияет на обучение и используется лишь для оценки предсказательных возможностей уже обученной сети.

В итоге можно себе представить следующую процедуру обучения сети. В начале обучения ошибки сети на обучающей выборке и на кросс-проверочном множестве будут одинаковыми. По мере обучения ошибка сети будет убывать как для обучающего множества, так и для кросс-проверочного (еще раз следует подчеркнуть, что веса корректируются только с учетом данных обучающего множества). Если начиная с некоторого момента обучения ошибка на кросс-проверочном множестве перестала убывать или даже стала расти, то это говорит о переобучении сети. В этом случае сеть нуждается в упрощении.

Если имеет место другая ситуация – ошибка уменьшается в процессе обучения как на обучающем, так и на проверочном множестве, но остается все-таки недопустимо большой, то выбранная сеть является слишком простой для выбранной задачи, и следует увеличить либо число слоев, либо число нейронов в отдельных слоях.

Необходимость многократных экспериментов приводит к тому, что кросс-проверочное множество начинает играть ключевую роль в выборе модели и само становится частью процесса обучения. Для того,

чтобы окончательно убедиться в прогностических возможностях обученной сети, как уже указывалось выше, заранее из обучающей выборки, если позволяет ее объем, выделяют тестовое множество. Тестовое множество используется только один раз. Если ошибка на тестовом множестве оказалась приемлемой, обученную сеть можно использовать для решения прикладных задач.

Обучение без учителя

Рассмотрим новый тип обучения нейросети – обучение без «учителя», когда сеть самостоятельно формирует свои выходы, адаптируясь к поступающим на ее входы сигналам.

При обучении «классической» многослойной нейросети на вход подаются данные или индикаторы, а выход нейросети сравнивается с эталонным значением (с так называемым «учителем»). Разность этих значений называется ошибкой нейронной сети, которая и минимизируется в процессе обучения. Таким образом, обычные нейронные сети выявляют закономерности между входными данными и прогнозируемой величиной. Если такие закономерности есть, то нейросеть их выделит и прогноз будет успешным.

При решении задачи кластеризации какие-либо предварительные сведения об анализируемых данных, как правило, отсутствуют, и нейросеть должна выявить закономерности во входных данных, самообучаясь. «Учителем» сети могут служить лишь сами данные, т. е. имеющаяся в них информация, закономерности, отличающие входные данные от случайного шума. Лишь такая избыточность позволяет находить более компактное описание данных, что, согласно общему принципу, изложенному в предыдущей главе, и является обобщением эмпирических данных. Сжатие данных, уменьшение степени их избыточности, использующее существующие в них закономерности, может существенно облегчить последующую работу с данными, выделяя действительно независимые признаки. С точки зрения математики задача кластеризации по существу является задачей построения отображения многокомпонентного вектора в пространство признаков, позволяющее различать кластеры. С этой точки зрения задача оценки знаний студента преподавателем на экзамене также является задачей кластеризации, которую реализует преподаватель. Многокомпонентную информацию о знаниях студента преподаватель трансформирует в экзаменационную оценку, с помощью которой все студенты делятся на кластеры знающих предмет отлично, хорошо и т. д.

В математике разработано достаточно много алгоритмов кластеризации, о которых уже говорилось в предыдущей главе. Здесь мы только напомним читателю, что не существует абсолютно правильного алгоритма кластеризации. Разные алгоритмы приводят к разным конечным результатам. Более того, не существует методов, позволяющих с абсолютной достоверностью оценить ошибку кластеризации (поэтому студентам придется смириться с определенной субъективностью экзаменационной оценки).

Нейросетевые алгоритмы кластеризации не являются исключением и содержат те же недостатки, что и алгоритмы численных методов.

Задача сжатия имеющейся информации очень важна сама по себе и является непременной составляющей любого процесса осмысления информации. Поэтому важно понять, как кластеризация может осуществляться с помощью нейронной сети.

Было предложено несколько различных алгоритмов обучения нейросетей без учителя. Здесь мы остановимся только на алгоритме, сформулированном финским ученым Тойво Кохоненом в 1982 г. На самом деле этот алгоритм является дальнейшим развитием соревновательного алгоритма обучения нейронной сети.

Рассмотрим вначале базовый алгоритм соревновательного обучений нейронной сети. Для задач классификации нелинейность передаточной функции активации не является принципиальной. Поэтому можно упростить рассмотрение, ограничившись линейной функцией активации. Выход такого нейрона является линейной комбинацией его входов:

$$y_i = \sum_{j=1}^{m} x_j \cdot w_{ij} .$$

В этой формуле m – число входов такой сети. На все нейроны соревновательного слоя обучения подается одинаковый сигнал с входов нейронной сети.

На рис. 5.3 изображен пример сети, содержащей три входа (x_1, x_2, x_3) и слой нейронов в виде плоской квадратной решетки. Выходы соревновательного слоя на рисунке не изображены. Предполагается, что все нейроны плоской решетки имеют свой индекс (номер).

Пусть на входы такой сети предъявляется образец (объект) с номером τ из имеющегося набора данных, подлежащих классификации.

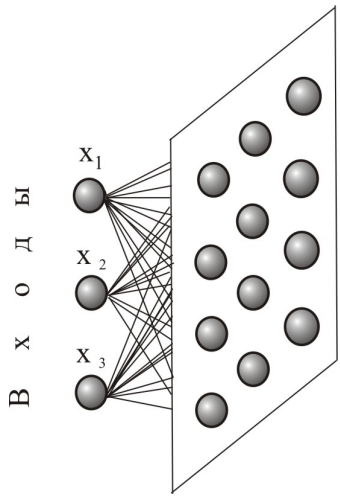


Рис. 5.3. Нейронная сеть, позволяющая решать задачи кластерного анализа (нейроны слоя Кохонена образуют квадратную решетку)

На первом этапе производится обучение такой сети в соответствии с итеративным алгоритмом подстройки весов

$$\mathbf{w}_i^{\tau}(t) = \mathbf{w}_i^{\tau}(t-1) + \eta \cdot y_i^{\tau}(t-1) \cdot \left(\mathbf{x}^{\tau} - \sum_{k=1}^{m} y_k^{\tau}(t-1) \cdot \mathbf{w}_k^{\tau}(t-1) \right). \quad (5.17)$$

В этой формуле мы используем векторные обозначения для весов: $\mathbf{w}_i^{\tau}(t-1)$ – это весовая характеристика i-го нейрона при предъявлении на вход объекта $\mathbf{x}^{\tau}$ на шаге обучения с номером $t-1$. Параметр η характеризует скорость обучения. Обучение по формуле (5.17) в идеале стремится к тому, чтобы для входа $\mathbf{x}^{\tau}$ подобрать веса для некоторого нейрона, точно отображающие входной сигнал, а остальные веса обратить в нуль. Если амплитуда выходного сигнала для этого нейрона будет равна единице, то скобка в выражении (5.17) обратится в нуль и обучение прекратится. Проверить, к чему приводит алгоритм обучения по формуле (5.17), можно в Excel, взяв, например, два нейрона и два входа. На практике, конечно, алгоритм обучения прекращается не потому, что скобка в выражении (5.17) обращается в нуль, а потому, что, например, веса перестали достаточно заметно изменяться при переходе к следующей итерации.

После завершения первого этапа обучения определяется нейрон-победитель, т. е такой нейрон, веса которого наиболее точно соответствуют входному сигналу. Для этого достаточно найти скалярное произведение P_j весов w_{ij}^{τ} j-го нейрона и компонент входного вектора x_i^{τ}:

$$P_j^{\tau} = x_i^{\tau} \cdot w_{ij}^{\tau}. \qquad (5.18)$$

Нейрон, для которого величина P_j^{τ} окажется максимальной, объявляется победителем, веса остальных нейронов полагаются равными нулю. Номер этого нейрона запоминается, и сеть переходит к обучению другому примеру.

Обратите внимание на то, что обучение является локальным. В итоге нейрон-победитель, свой для каждого входного вектора, будет служить прототипом этого вектора. В принципе, один и тот же нейрон может оказаться победителем для различных входных объектов. Очевидно, что это и будет означать возможность отнести их к одному классу.

После того как все объекты будут предъявлены сети, можно построить отображение многомерной входной информации в пространство номеров нейронов-победителей. Так достигается сжатие информации при этом способе обучения нейросети.

В табл. 5.1 приведен фрагмент выходной информации, которую выдает программа NeuroSolutions 5 при решении известной задачи Фишера классификации ирисов (*Iris sentosa*, *Iris versicolor*, *Iris virginica*) в зависимости от характеристик цветка: x1 (длина чашелистика), x2 (ширина чашелистика), x3 (длина лепестка), x4 (ширина лепестка).

Как следует из приведенных данных, имеется три объекта, для которых нейрон-победитель имеет номер «0», и четыре объекта, для которых победителем является нейрон с номером «4». Очевидно, что эти объекты относятся к двум разным классам.

Приведенный алгоритм имеет несколько недостатков. Нейронов в сети, как правило, больше, чем классов, а рассматриваемый выше алгоритм «навязывает» число кластеров, равное числу нейронов в сети. Кроме того, в этом алгоритме никак не используется информация о весах нейронов и их расположении. Нейроны-победители расположены в узлах решетки случайно. Рядом расположенные нейроны-победители не соответствуют близким по свойствам объектам. Иначе говоря, нейроны выходного слоя не упорядочены: положение нейрона-победителя в соревновательном слое не имеет ничего общего с координатами его весов во входном пространстве.

Таблица 5.1

Фрагмент выходной информации программы NeuroSolutions 5 в задаче классификации ирисов Фишера

Нейрон-победитель	x1	x2	x3	x4
20	5,80	2,60	4,00	1,20
0	4,80	3,40	1,90	0,20
24	6,40	3,20	4,50	1,50
4	7,10	3,00	5,90	2,10
9	6,70	2,50	5,80	1,80
0	5,40	3,40	1,70	0,20
4	6,40	2,80	5,60	2,10
4	6,50	3,00	5,20	2,00
4	7,20	3,00	5,80	1,60
0	4,60	3,10	1,50	0,20
20	4,90	2,50	4,50	1,70

Оказывается, что небольшой модификацией соревновательного обучения можно добиться того, что положение нейрона в выходном слое будет коррелировать с положением прототипов в многомерном пространстве входов сети: близким нейронам будут соответствовать близкие значения входов. Тем самым появляется возможность строить топографические карты, чрезвычайно полезные для визуализации многомерной информации.

Кохонен предложил ввести в базовое правило соревновательного обучения информацию о расположении нейронов в выходном слое. Для этого нейроны выходного слоя упорядочиваются, образуя одно- или двумерные решетки. Теперь положение нейронов в такой решетке маркируется векторным индексом $\mathbf{i}$. Такое упорядочение естественным образом вводит расстояние между нейронами в слое. Модифицированное Кохоненым правило соревновательного обучения учитывает расстояние нейронов от нейрона-победителя $\mathbf{i}^*$. После определения нейрона-победителя начинается итерационное обучение нейрона-победителя и его ближайшего окружения по следующему алгоритму:

$$\mathbf{w}_i^{\tau}(t) = \mathbf{w}_i^{\tau}(t-1) + \eta \cdot \Lambda\left(\left|\mathbf{i} - \mathbf{i}^*\right|\right) \cdot (\mathbf{x}^{\tau} - \mathbf{w}_i^{\tau}(t-1)) . \qquad (5.19)$$

Входящая в формулу (5.19) функция соседства Λ равна единице для нейрона-победителя с индексом $\mathbf{i}^*$ и постепенно спадает с расстоянием, например, по закону

$$\Lambda(\left| \mathbf{i} - \mathbf{i}^* \right|) = \exp(-\frac{(\mathbf{i} - \mathbf{i}^*)^2}{\sigma^2(t)}) . \qquad (5.20)$$

Величина σ представляет собой радиус взаимодействия нейронов, который на каждом шаге итерации уменьшается. На рис. 5.4 изображены три области нейронов, ближайшие к нейрону-победителю.

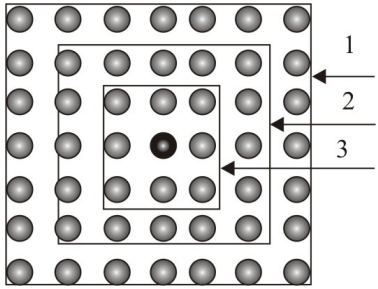


Рис. 5.4. Области в окрестности нейрона-победителя подлежащие обучению на шагах итерации 1, 2 и 3 (нейрон-победитель в центре рисунка)

В результате такого обучения веса нейронов в окрестности нейрона-победителя также подстраиваются, и теперь близкие по характеристикам входные объекты будут отображаться нейронами, расположенными недалеко друг от друга. Таким образом, мы получаем не только квантование входов, когда каждый входной сигнал отображается одним нейроном-победителем, но и упорядочение входной информации в виде одно- или двухмерной карты. Каждый входной многомерный вектор имеет свою координату на этой сетке, причем, чем ближе координаты двух нейронов на карте, тем ближе они и в исходном пространстве. Такая топографическая карта дает наглядное представление о структуре данных в многомерном входном пространстве, геометрию которого мы не в состоянии представить себе иным способом.

Способы визуализации карт Кохонена

Нейросеть, обученная с использованием алгоритма Кохонена, в состоянии дать визуальное отображение многомерных входных данных. Информативными здесь являются не только выходы нейронов (как в случае обычной нейросети), но также веса нейронов и распределение

примеров по нейронам. Рассмотрим, как может быть визуализирована информация о результатах решения задачи кластеризации с использованием алгоритма Кохонена.

Задаться вопросом о том, как должна быть представлена результирующая информация, полезно еще на стадии проектирования нейросети. Если нужно, чтобы построенная нейросеть хорошо аппроксимировала данные (находила статистически значимые закономерности), то общее число ячеек сети должно быть как минимум на порядок меньше числа обучающих примеров.

Кластерную структуру данных, напротив, лучше исследовать, когда число ячеек сравнимо с числом примеров. При этом появятся пустые ячейки, разграничивающие области данных и появляется то, что, собственно, и можно назвать топографической картой.

В качестве примера рассмотрим данные рейтинга 185 крупнейших компаний России по объему реализации продукции (по данным журнала «Эксперт» за 1997 год; http://www.expert.ru/). Для кластеризации предприятий выберем следующие индикаторы:

а) добыча/обработка (профиль деятельности предприятий);

б) темпы роста;

в) объем прибыли на единицу объема продукции;

г) объем производства в расчете на одну штатную единицу.

Карта Кохонена для этого набора данных, построенная с помощью программы Excel Neuro Package (надстройки Excel, позволяющей решать задачи нейросетового моделирования), представлена на рис. 5.5. Каждый нейрон-победитель на этом рисунке отмечен квадратиком. Величина стороны квадратика пропорциональна числу объектов, для которых этот нейрон оказался победителем. Многомерная структура исходных характеристик промышленных предприятий точно отображается расположением на плоскости нейронов-победителей. Если предприятия близки по своим характеристикам, то на карте они окажутся рядом.

Дополнительную информацию карта Кохонена дает благодаря раскраске. Дело в том, что веса нейрона-победителя для каждого объекта пропорциональны соответствующим компонентам его входного вектора. Поэтому можно взять, например, входную характеристику «объем производства в расчете на одну штатную единицу» и раскрасить нейроны-победители в различные цвета (или различные градации серого цвета) в соответствии с величиной этого признака у различных объектов. При этом более светлые тона соответствуют большим значениям

признака, а более темные – меньшим. На рис. 5.5 как раз приведен пример такой раскраски по удельному объему производства.

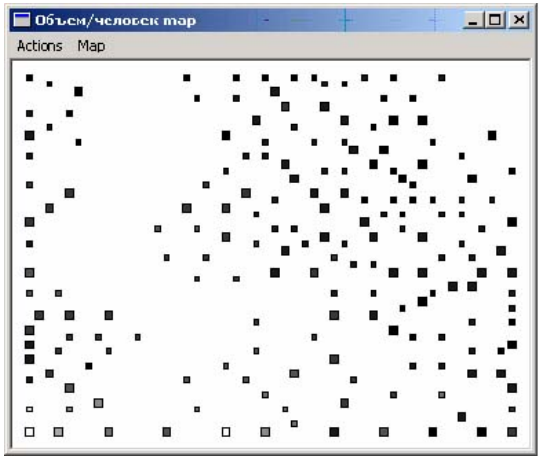


Рис. 5.5. Карта Кохонена крупнейших предприятий РФ в 1997 г. (раскраска узлов выполнена по фактору удельного объема производства)

Совершенно аналогично можно построить карту с раскраской узлов по величине других входных признаков. При этом очевидно, что структура узлов останется прежней, а раскраска узлов поменяется.

В программе Excel Neuro Package имеется еще одна весьма полезная возможность. Если щелкнуть левой кнопкой мыши в тот момент, когда курсор находится в поле того или иного квадратика, то открывается дополнительное окно, в котором отображены все объекты и их характеристики, для которых выбранный нейрон оказался победителем.

Таким образом, вся совокупность нейронов в выходном слое точно моделирует структуру распределения обучающих примеров в многомерном пространстве. Уникальность технологии самоорганизующихся карт состоит в преобразовании многомерного пространства признаков в двух- или одномерное. При этом используется информация о весах обученной нейронной сети (интенсивность цвета раскраски пропорциональна весам), информация о взаимном расположении нейронов-победителей и информация о числе объектов, для которых один и тот же нейрон является победителем.

Количество нейронов в соревновательном слое определяет максимальное разнообразие выходов и выбирается в соответствии с требуемой степенью детализации входной информации.

5.4. Примеры решения задач с использованием нейросетевого моделирования

В этом разделе мы рассмотрим несколько примеров применения нейросетевого подхода для решения задач регрессионного анализа и задач кластеризации.

Хотя известно достаточно большое число программных пакетов, реализующих использование нейронных сетей для анализа данных[7], мы остановили свой выбор на продукте Deductor компании BaseGroup Labs. Причиной такого выбора является то, что это свободно распространяемый, современный, многофункциональный пакет обработки данных, поддерживающий основные технологии Data Mining.

Пакет русскоязычный, разработан в России, имеет достаточно простой, интуитивно понятный интерфейс. Недостатком свободно распространяемой версии является ограничение на число объектов в анализируемой базе данных – всего 150. Естественно, что для практического использования в коммерческих или иных целях это ограничение является весьма существенным, но для учебных целей предельный объем выборки размером 150 объектов не представляется критическим.

Другим достоинством пакета является возможность использовать данные, сохраненные в различных форматах. В частности, можно использовать базы данных в формате Excel.

К недостаткам пакета можно отнести почти полное отсутствие в справке информации об использованных алгоритмах обработки данных и способах оценки ошибок. Впрочем, это характерно для большинства программных продуктов, ориентированных на автоматизированную обработку данных.

Программа Deductor содержит в своем составе все основные алгоритмы обработки данных, характерные для технологии Data Mining:

- линейный регрессионный анализ;
- нейросетевое моделирование;

[7] Краткий обзор программных продуктов для нейросетевого анализа данных можно найти в книге Круглова В. В., Борисова В. В. Искусственные нейронные сети. /М., 2002.

- построение дерева решений;
- создание карт Кохонена;
- построение ассоциативных правил;
- построение произвольной модели взаимосвязи вручную.

В программе Deductor процедура анализа данных разбивается на следующие этапы:

1) импорт исходных данных для обработки;
2) построение одного или нескольких сценариев обработки данных;
3) визуализация результата (выбор способа представления полученных результатов);
4) экспорт результатов.

Все действия аналитика по обработке данных оформляются в программе в виде сценариев обработки данных, к которым на любом этапе можно обратиться, внести какие-либо коррективы и пересчитать результаты, пользуясь новым сценарием.

Подготовка сценариев

Весь процесс анализа данных осуществляется с помощью четырех мастеров – импорта, обработки, визуализации и экспорта. Для построения сценария достаточно использование только этих мастеров и ничего более. Сценарий отображается на панели сценариев (рис. 5.6). Показать или скрыть эту панель можно, выбрав пункт Сценарии меню Вид или нажав на кнопку  на панели инструментов. Сверху на панели сценариев расположены кнопки для вызова мастеров.

Построение сценария начинается с вызова мастера импорта. Мастер импорта предназначен для автоматизации получения данных из любого источника, предусмотренного в системе. Чтобы вызвать это действие, достаточно воспользоваться кнопкой  Мастер импорта в верхней части панели или нажать функциональную клавишу F6. На первом шаге мастера импорта открывается список всех предусмотренных в системе типов источников данных. Среди них следует выбрать нужный тип источника и для перехода на следующий шаг щелкнуть по кнопке Далее. Число шагов мастера импорта, а также набор настраиваемых параметров отличаются для разных типов источников. Например, если исходные данные хранятся в файле Excel, то нужно выбрать строку MS Excel. Более подробно процедура импорта данных из файла Excel будет описана в примере 5.1.

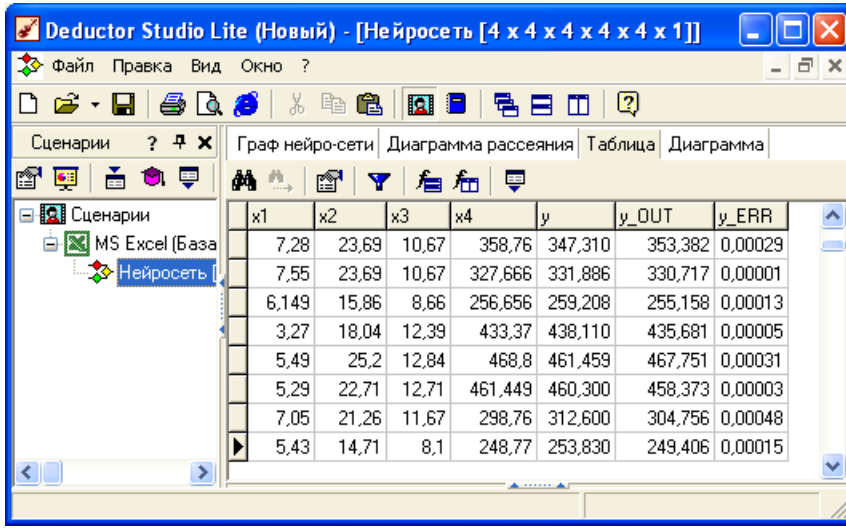


Рис. 5.6. Окно программы Deductor с отображенным окном сценариев и табличными результатами нейросетевого анализа

При импорте данных важно сразу указать правильно тип переменных. В программе Deductor переменные могут иметь следующие типы:

- *Логический* – данные в поле могут принимать только два значения – 0 или 1 (ложь или истина);
- *Дата/время* – поле содержит данные типа дата/время;
- *Вещественный* – значения поля – числа с плавающей точкой;
- *Целый* – данные в поле представляют собой целые числа;
- *Строковый* – данные в столбце представляют собой строки символов.

Затем указывается вид данных: *непрерывный* – значения в столбце могут принимать любое значение в рамках своего типа.

Как правило, непрерывными являются только числовые данные.

Дискретный – данные в столбце могут принимать ограниченное число значений. Как правило, дискретный характер носят строковые данные.

К выбору типа и вида данных нужно относиться серьезно, так как это влияет на возможность дальнейшего использования этого поля. Неправильное указание типа данных может привести к потере информации.

Мастер обработки предназначен для настройки всех параметров выбранного алгоритма. Для вызова мастера обработки достаточно воспользоваться кнопкой  Мастер обработки в верхней части панели или нажать функциональную клавишу F7. В окне первого шага мастера обработки представлены все доступные в системе методы обработки данных. Как правило, на следующем шаге мастера обработки производится настройка назначения переменных. Поле таблицы данных (переменная) может быть объявлено как:

Входное – будет являться входным полем при анализе данных;

ВЫХОДНОЕ – будет являться выходным полем обработчика (например, целевым полем для обучения нейронной сети);

Информационное – поле содержит вспомогательную информацию, которую часто полезно отображать, но не следует использовать при обработке;

Измерение – поле будет использоваться в качестве измерения в многомерной модели данных;

Факт – значения поля будут использованы в качестве фактов в многомерной модели данных;

Свойство – поле содержит описание свойств или параметров некоторого объекта;

Транзакция – поле, содержащее идентификатор событий, происходящих совместно (одновременно): например, идентификатор – номер чека, по которому приобретены товары, покупка товара – это событие, а их совместное приобретение по одному чеку – транзакция;

Элемент – поле, содержащее элемент транзакции (событие);

Непригодное – данные поля не пригодны для данного способа обработки (программа автоматически назначает полю это значение): например, для преобразования даты поле должно иметь тип Дата/время. Если оно будет иметь, например, строковый тип, то программа автоматически укажет для него назначение *Непригодное;*

Неиспользуемое – запрещает использование поля в обработке данных и исключает его из выходного набора. В отличие от непригодного поля, такие поля в принципе могут использоваться, просто в этом нет необходимости;

Первичный ключ – поле будет использоваться в качестве первичного ключа.

После того как алгоритм обработки данных завершит работу, с помощью мастера отображений в пошаговом режиме необходимо выбрать и настроить наиболее удобный способ представления данных. К

мастеру визуализации можно вернуться на любом шаге, изменив способ отображения данных. Для вызова мастера отображений можно воспользоваться кнопкой  Мастер визуализации на панели сценариев или нажать функциональную кнопку F5, предварительно выделив нужную ветвь в сценарии обработки.

В зависимости от метода обработки, в результате которого была получена ветвь сценария обработки, список доступных для нее видов отображений будет различным. Если одновременно выбрано несколько способов отображения результатов, то все они будут доступны простым переключением закладок в окне вывода результатов. На рис. 5.6, например, изображена ситуация, когда доступны для отображения граф нейронной сети, диаграмма рассеяния, таблица и диаграмма.

Экспорт результатов анализа зависит от формы их представления. Мастер экспорта позволяет в пошаговом режиме выполнить экспорт данных в файлы наиболее распространенных форматов, различные базы данных. Для вызова мастера экспорта можно воспользоваться кнопкой  Мастер экспорта на панели сценариев. На первом шаге мастера экспорта представлен список всех форматов, в которые может быть выполнен экспорт данных. Среди них следует выбрать нужный и далее следовать шагам мастера. В результате набор данных будет выгружен в выбранный приемник в виде обычной таблицы.

Выбранная методика построения сценариев является достаточно удобной, поскольку обладает возможностями, схожими с проводником Windows. В частности, можно:

- вставить узел перед текущим узлом и вызывать для него Мастер обработки;
- вырезать текущий узел из дерева (все его «потомки» при этом перемещаются на один уровень вверх и начинают подчиняться «родителю» удаленного узла;
- копировать ветвь сценария, начиная с текущего узла (родителем новой ветви станет родительский узел оригинальной ветви);
- удалить ветвь сценария, начиная с текущего узла;
- сохранить ветвь сценария, начиная с текущего узла в файл ветви или файл проекта для последующего использования;
- загрузить ветвь из файла проекта или файла ветви сценария обработки.

Перечисленные операции позволяют легко вносить изменения в дерево сценариев, а значит, изменять порядок и свойства обработки.

Визуализация данных

На любом этапе обработки можно визуализировать данные. Система самостоятельно определяет, каким способом она может это сделать: например, если будет обучена нейронная сеть, то помимо таблиц и диаграмм можно визуализировать граф нейросети. Пользователю необходимо выбрать нужный вариант из списка и настроить несколько параметров.

Возможны следующие способы визуализации данных.

• *OLAP*. Многомерное представление данных. Любые данные, используемые в программе, можно посмотреть в виде кросс-таблицы и кросс-диаграммы. Пользователю доступен весь набор механизмов манипуляции многомерными данными: группировка, фильтрация, произвольное размещение измерений, детализация, выбор любого способа агрегации, отображение в абсолютных числах и в процентах.

• *Таблица*. Стандартное табличное представление с возможностью фильтрации данных и быстрого расчета статистики (он-лайн-статистика).

• *Диаграмма*. График изменения любого показателя.

• *Гистограмма*. Гистограмма предназначена для визуальной оценки распределения данных.

• *Статистика*. Статистические показатели выборки.

• *Диаграмма рассеяния*. График отклонения значений, прогнозируемых при помощи модели. Используется для визуальной оценки качества построенной модели.

• *Таблица сопряженности*. Таблица сопряженности отображает результаты сравнения категориальных значений исходного выходного столбца и категориальных значений рассчитанного выходного столбца. Используется для оценки качества классификации.

• *«Что – если»*. Позволяет «прогонять» через построенную модель любые интересующие пользователя данные и оценивать влияние того или иного фактора на результат. Активно используется для решения задач оптимизации.

• *Обучающая выборка.* Выборка, используемая для построения модели. Выделяются цветом данные, попавшие в обучающее, тестовое и валидационное множества с возможностью фильтрации. Эта ин-

формация необходима для понимания того, какие записи и каким образом использовались при построении модели.

• *Диаграмма прогноза.* Применяется после использования метода обработки Прогнозирование. Прогнозные значения выделяются цветом.

• *Граф нейросети*. Визуальное отображение обученной нейросети. Отображается структура нейронной сети и значения весов.

• *Дерево решений*. Отображение дерева решений, полученного при помощи соответствующего алгоритма. Имеется возможность посмотреть детальную информацию по любому узлу и фильтровать попавшие в него данные.

• *Дерево правил*. Отображение в иерархическом виде (в виде дерева) ассоциативных правил. Содержит всегда два уровня. На первом – условие, на втором – следствие правила (или наоборот).

• *Правила*. Отображает в текстовом виде правила, полученные при помощи алгоритма построения дерева решений или поиска ассоциаций. Такого рода информация легко интерпретируется человеком.

• *Карта Кохонена*. Отображение карт, построенных при помощи соответствующего алгоритма. Широкие возможности настройки: выбор количества кластеров, фильтрация по узлу/кластеру, выбор отображаемых полей. Мощный и гибкий механизм отображения кластеризованных данных.

• *Популярные наборы*. Отображение наиболее часто встречающихся в ассоциативных правилах множеств в виде списка.

• *Описание*. Текстовое описание параметров импорта/ обработки/экспорта в дереве сценариев обработки.

После краткого обзора основных подходов к анализу данных в программе Deductor рассмотрим несколько конкретных примеров.

Пример 5.1

S&P 500 – это список 500 избранных акционерных компаний США, имеющих наибольшую капитализацию. Список принадлежит компании Standard & Poor's и ею же составляется. Акции всех компаний из списка S&P 500 торгуются на крупнейших американских фондовых биржах, таких как Нью-Йоркская фондовая биржа и NASDAQ. Среднее взвешенное значение цен акций этих компаний известно также как индекс S&P 500.

Индекс S&P 500 конкурирует по популярности с промышленным индексом Доу-Джонса и заслуженно называется барометром американской

экономики. В файле Пример_5_1.xls содержатся данные о средних за неделю значениях индекса S&P 500 (переменная x4), а также некоторые другие характеристики фондового рынка: x1 – доходность по годовым казначейским векселям; x2 – средняя доходность в расчете на акцию активов из списка Standard & Poor, x3 – средние дивиденды в расчете на акцию. Кроме этих данных, файл содержит и значения индекса S&P 500 на следующую неделю (переменная y), которые будут использоваться для обучения нейронной сети. Требуется, оценив необходимое число нейронов, создать нейронную сеть, обучить ее для предсказания значений индекса S&P 500 на следующую неделю. Используя обученную сеть, предсказать значения индекса S&P 500 для другого входного набора данных.

Решение

После запуска программы Deductor откроется окно, изображенное на рис. 5.6, которое разделено на три части: верхняя часть представляет собой типичную для всех приложений Windows панель инструментов. Нижняя левая часть окна отведена для размещения сценариев – процедур анализа данных. Нижняя правая, и большая, часть окна приложения отведена для представления данных и результатов их анализа. Для организации импорта данных следует нажать на клавиатуре функциональную клавишу F6 или кнопку с пиктограммой  на панели инструментов (во фрейме Сценарии). В результате этих действий откроется мастер импорта данных, который имеет интуитивно понятный интерфейс.

Процедура импорта разбита на 6 шагов. На первом шаге нужно просто выделить мышкой строку с надписью MS Excel и нажать кнопку Далее. На втором шаге в окне База данных нужно указать путь к файлу Excel с импортируемыми данными, а в окне Таблица в базе данных выбрать лист, в котором содержатся нужные данные. Третий шаг не нуждается в комментариях. На четвертом шаге нужно определить свойства импортируемых переменных. Как уже указывалось выше, при импорте данных нужно указать их вид и назначение. В данном примере для всех переменных следует определить вид Непрерывный, x1–x4 по назначению отнести к входным переменным, y – к выходным. На пятом шаге нужно задать способ отображения данных. Программа по умолчанию предлагает в этом случае отображать данные в виде таблицы. Эту настройку следует оставить без изменения. На шестом шаге импорта данных предлагается определить заголовок для окна. Эту настройку также можно оставить без изменения. В результате выполненных шагов импорта данные будут загружены в программу и подготовлены для дальнейшего анализа.

Нажав функциональную клавишу F7 или кнопку с пиктограммой  , запустим Мастер обработки, который позволяет в ходе процедуры из 9

шагов выбрать необходимый алгоритм анализа данных, задать нужные параметры обработки данных и визуализации результатов. На первом шаге выбираем из предлагаемого списка процедур интересующую нас процедуру с названием Нейросеть.

На втором шаге следует задать способ нормализации данных. Можно согласиться со способом, предлагаемым по умолчанию. При этом значения всех факторных переменных будут приведены к диапазону –1 … +1, а значения результативной величины к диапазону 0 … +1.

На третьем шаге построения нейронной сети задаются параметры обучающего, валидационного и тестового множеств. По умолчанию 95 % случаев будет отнесено к обучающей выборке и 5 % – к тестовой. Валидационная выборка не предусматривается. Эти настройки также оставим без изменения.

На четвертом шаге следует определить структуру нейросети, задав число скрытых слоев и число нейронов в каждом скрытом слое, а также тип передаточной функции и степень ее крутизны. Окно настройки параметров нейронной сети изображено на рис. 5.7.

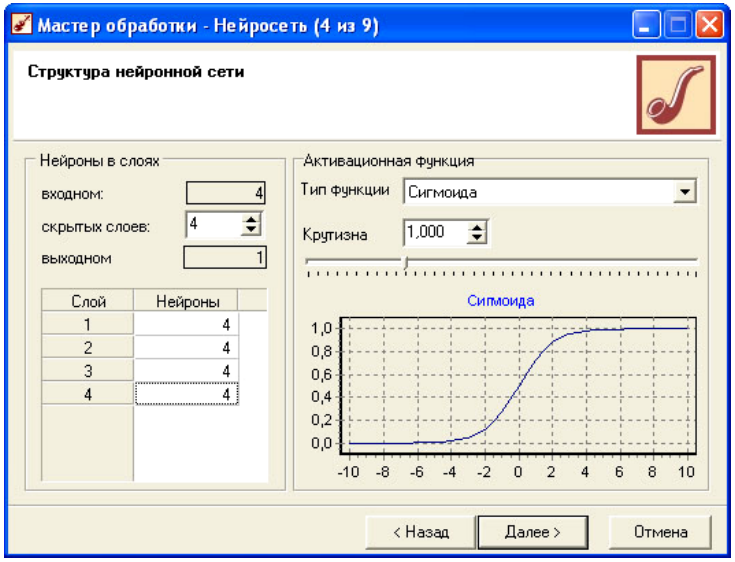


Рис. 5.7. Окно настройки параметров нейронной сети

Очевидно, что сеть должна иметь 4 входных нейрона (по числу входных переменных) и один выходной. Число нейронов в скрытых слоях можно приближенно оценить по формуле (5.16) (напомним, что эта формула дает оценку числа синаптических весов, а не нейронов). Известны и

другие формулы для таких оценок. Например, в книге В. В. Круглова и В. В. Борисова (см. список литературы) приводится простая оценочная формула, которой мы и будем следовать:

$$\frac{N}{10} - n - m \le L \le \frac{N}{2} - n - m, \qquad (5.21)$$

где L – это число нейронов в скрытых слоях. Остальные обозначения такие же, как и в формуле (5.16). В нашем случае $N \approx 150$, $n = 4$, $m = 1$. Поэтому минимальное число нейронов в сети – 10, а максимальное – 70. Выберем сеть, состоящую из 4 скрытых слоев, в каждом из которых будет находиться по 4 нейрона. Передаточную функцию и крутизну ее характеристики изменять не будем.

На пятом шаге следует выбрать алгоритм обучения нейронной сети. Здесь на выбор предлагается два алгоритма: алгоритм обратного распространения ошибок с корректировкой весов после каждого обучающего примера и алгоритм обратного распространения ошибок с корректировкой весов по эпохам (т. е. после предъявления всех примеров). Первый алгоритм является более устойчивым, а второй – значительно более быстрым. По умолчанию установлен алгоритм обучения с корректировкой весов по эпохам. Согласимся с установками по умолчанию на этом шаге.

На шестом шаге следует задать критерии остановки алгоритма обучения. По умолчанию обучение закончится после 10 000 эпох (т. е. все обучающие примеры будут предъявлены 10 000 раз для корректировки весов синапсов). Здесь также следует согласиться с установками по умолчанию.

На седьмом шаге осуществляется процесс обучения нейронной сети. Окно мастера построения нейронной сети, позволяющее следить за процессом обучения, представлено на рис. 5.8.

Запустим с помощью кнопки Пуск процесс обучения нейронной сети. Обучение можно контролировать, наблюдая за максимальной и средней ошибками в обучающей и тестовой выборках. Текущие данные о средних и максимальных ошибках в обучающей и тестовой выборках по эпохам отображаются в виде разноцветных кривых на графике. Здесь по оси абсцисс отложен номер эпохи, а по оси ординат – величина соответствующей ошибки (см. рис. 5.8). Следует сказать несколько слов о способе вычисления ошибки. В качестве средней ошибки по эпохе берется ошибка, определенная формулой (5.4), деленная на число наблюдений (для нормированных значений переменных), а в качестве максимальной ошибки – максимальная из ошибок (5.4*) после однократного предъявления всех объектов. Эта ошибка также вычисляется для нормированных значений переменных.

Процесс обучения можно остановить, нажав кнопку Стоп. Остановить процесс обучения следует тогда, когда средняя ошибка на тестовом множестве перестанет заметно уменьшаться или даже станет увеличиваться.

На рис. 5.8 показана ситуация, когда процесс обучения был остановлен на эпохе175. Видно, что обучение сети еще не закончено, поскольку все контролируемые ошибки имеют устойчивую тенденцию к снижению.

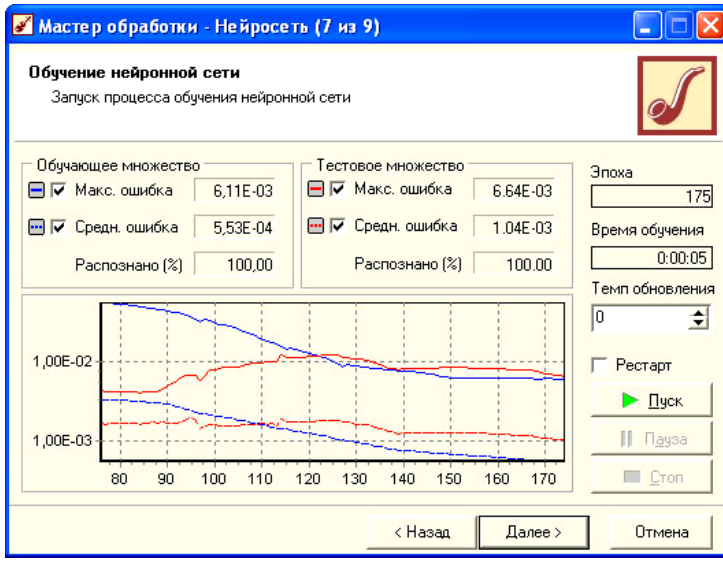


Рис. 5.8. Окно, позволяющее контролировать процесс обучения нейронной сети в программе Deductor

После остановки процесс обучения можно снова продолжить, нажав кнопку Пуск.

После того как сеть обучена, на восьмом шаге мастера построения нейросети следует выбрать способы отображения результатов. Выберем следующие названия из списка доступных возможностей отображения результатов: Граф нейросети, Диаграмма рассеяния и Таблица.

На последнем шаге следует просто нажать кнопку Готово, и программа перейдет в режим отображения результатов.

Наибольший интерес для нас представляют табличные результаты, часть которых представлена на рисунке 5.6. Здесь переменная y_OUT представляет собой выход нейросети (прогнозное значение для переменной y), а переменная y_ERR представляет собой ошибку нейросети. На первый взгляд совершенно непонятно, каким образом при истинном значении переменной y = 347, 310 и прогнозном значении y_OUT = 353,382 ошибка оказалась столь малой: y_ERR = 0,00029. В действительности ничего необычного в этом нет, если вспомнить, что нейросеть использует

при вычислениях нормированные величины, а ошибка определяется формулой (5.4*). Чтобы убедиться в этом, вычислим ошибку нейросети для первого значения входного набора данных. Для этого нужно произвести нормализацию переменной y. Найдем максимальное значение $y_{макс}$ во входном наборе, равное 572,68, и минимальное значение $y_{мин}$, равное 215,97. При проведении вычислений данные приводятся к интервалу 0 … 1. Очевидно, что нормализованные значения $y_{норм}$ могут быть получены из исходных данных с использованием простой формулы

$$y_{норм} = \frac{y - y_{мин}}{y_{макс} - y_{мин}}. \qquad (5.22)$$

Точно такому же преобразованию должны подвергнуться и прогнозные значения

$$y\_OUT_{норм} = \frac{y\_OUT - y_{мин}}{y_{макс} - y_{мин}}. \qquad (5.22*)$$

Найденные по формулам (5.22) и (5.22*) значения нормированных величин равны: $y_{норм} = 0{,}3682$ и $y\_OUT_{норм} = 0{,}3852$. Тогда ошибка, вычисленная по формуле (5,4*), равна

$$y\_ERR = (y_{норм} - y\_OUT_{норм})^2 = (0{,}3682 - 0{,}3852)^2 = 0{,}00029. \qquad (5.23)$$

Приведенный выше результат совпадает со значением ошибки для первой строки таблицы на рис. 5.6. Для всех остальных случаев ошибка рассчитывается аналогично. Средняя ошибка, о которой шла речь выше, равна просто среднему значению величин в столбце y_ERR.

В большинстве программных пакетов по нейромоделированию в качестве ошибки фигурирует величина, имеющая аббревиатуру RMSE (*Root Mean Squared Error)*, равная корню квадратному из средней квадратической ошибки. Для задач однофакторной регрессии ошибка RMSE вычисляется по формуле

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y\_OUT_i)^2}, \qquad (5.24)$$

где n – число наблюдений, y_i и y_OUT_i – истинный и прогнозируемый результаты для i -го наблюдения (используются нормированные данные).

Таким образом, мы обучили нейронную сеть на обучающей выборке. Выясним, способна ли эта нейросеть предсказывать значение индекса S&P 500 на следующую неделю с приемлемой точностью на другом наборе входных данных.

Сохраним обученную нейросеть в виде сценария. Для этого нужно выделить пиктограмму нейросети в окне сценариев, нажать правую кнопку мыши и в выпадающем меню выбрать диалог сохранения сценария, а также папку, в которой будет сохранен файл сценария (расширение файла сценариев deb).

Очистим окно сценариев и откроем диалог импорта данных из формата Excel (импортируем файл Пример_5_1а.xls). Поскольку значения переменной у нам теперь не нужны (мы будем использовать уже обученную сеть), то можно установить назначение этой переменной как Информационное.

После импорта файла данных выделим пиктограмму импорта данных в окне сценариев и, щелкнув правой кнопкой мыши, откроем диалог загрузки ветви сохраненного ранее сценария, который полностью напоминает диалог открытия файла. После загрузки ветви сценария для анализа нового набора данных достаточно дважды щелкнуть левой кнопкой мыши на пиктограмме сценария Нейросеть.

Важно отметить: для того чтобы обработка данных прошла успешно, новый файл должен иметь ту же самую структуру данных. Выходные данные (переменную у) в файле Пример_5_1а.xls мы оставили с целью дополнительного контроля построенной нейросети, но при получении результата эти данные не принимаются во внимание. В этом легко убедиться, если в файле Пример_5_1а.xls предварительно все значения в столбце у заменить нулями (удалять столбец нельзя, поскольку программа в этом случае заметит несоответствие формата данных в обучающей выборке и тестовом примере).

Чтобы убедиться, что на новом наборе данных нейросеть дала удовлетворительный прогноз, найдем среднюю относительную ошибку абсолютных отклонений, определяемую формулой

$$\Delta = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{abs}(y_i - y\_\text{out}_i) , \qquad (5.25)$$

где y_i и y_out_i – истинные и прогнозные значения результативного признака в исходной (ненормированной) форме.

Для того чтобы произвести подсчеты по формуле (5.25), экспортируем таблицу, в которой отражены результаты работы нейросети, в Excel. С этой целью на панели отображения данных анализа нажмем кнопку с пиктограммой  , в результате чего откроется диалог экспорта данных. Выберем экспорт данных в файл формата Excel. Все дальнейшие шаги экспорта данных имеют интуитивно понятный интерфейс и представляют собой обычный для приложений Windows диалог по сохранению файла.

По умолчанию файл будет иметь имя export.xls, но в процессе диалога по сохранению файла это имя можно поменять.

После выполнения экспорта данных провести вычисления по формуле (5.25) не представляет труда. Для сети, которая сохранена под именем Пример_5_1.deb, эта ошибка составила 1,97 %. Возможно, что, подбирая другие параметры нейросети, можно добиться и лучших результатов.

Рассмотрим пример использования нейросети для задачи классификации с учителем. В главе 4 такого типа задачи решались с использованием дискриминантного анализа или логистической регрессии.

Пример 5.2

В файле Пример_5_2.xls на листе Данные приведены результаты голосования 130 конгрессменов США по 16 различным проектам (градации результатов голосования: Да, Нет, Воздержался) и принадлежность их к одной из партий (республиканская или демократическая). Используя эту выборку как обучающую, построить нейронную сеть, которая была бы в состоянии по результатам голосования по этим же проектам определить принадлежность 20 конгрессменов к одной из партий. Данные содержатся в том же файле на листе Контрольная выборка.

Решение

Импорт исходных данных осуществляется с использованием мастера импорта, как это было описано в предыдущем примере. Выходные данные, по которым будут настраиваться веса синапсов, содержатся в переменной Класс. Все остальные переменные должны быть определены как входные.

После импорта данных запустим мастер создания нейронной сети. Здесь важно отметить, что переменные содержат только текстовые значения, но программа Deductor правильно определила количество градаций входных и выходной переменной и произвела их нормализацию.

Создадим нейронную сеть с одним внутренним слоем, в котором будут располагаться два нейрона. В качестве алгоритма обучения выберем алгоритм обратного распространения ошибки (Back Propogation). В этом примере сеть обучается очень быстро, поэтому можно выбрать медленный, но более устойчивый алгоритм обучения.

На шестом шаге мастера выберем в качестве критерия останова значение средней ошибки на обучающем множестве $< 0{,}00005$.

Запустив сеть на обучение, дождемся остановки процесса обучения. Может встретиться ситуации, когда сеть обучается крайне медленно. В этом случае не следует торопиться усложнять сеть. Разумно сделать не-

сколько независимых попыток обучить сеть, а затем, если попытки не принесли результата, усложнить ее.

На восьмом шаге следует выбрать способы отображения результатов. Из списка имеющихся возможностей выберем следующие: Граф нейросети, Таблица сопряженности и Таблица. Для контроля качества классификации наибольший интерес представляет таблица сопряженности (рис. 5.9).

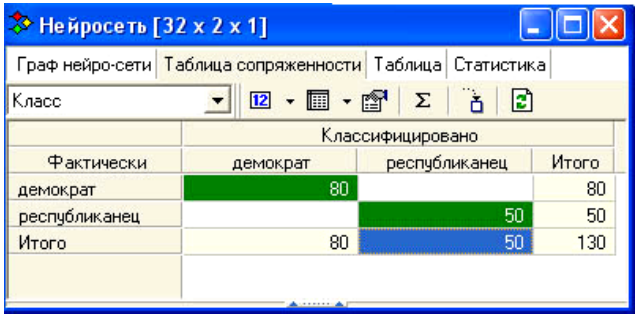


Рис. 5.9. Результаты классификации сенаторов по партийной принадлежности

Из этой таблицы видно, что нейронная сеть правильно классифицировала все 130 случаев.

Используем обученную нейросеть для классификации новых примеров. В предыдущем примере мы сохраняли сценарий на диске, а затем его вновь загружали для анализа новых данных. Точно так же можно поступить и теперь, но есть и более простой способ получить результат классификации для контрольной выборки. Выделим в окне сценариев сценарий импорта данных (строка MS Excel) и воспользуемся кнопкой Настроить узел. Диалог настройки узла можно вызвать и комбинацией клавиш Alt + Enter.

Диалог настройки узла позволяет просто заменить обучающую выборку контрольной. Как и в предыдущем примере, в контрольной выборке мы оставили истинную принадлежность конгрессменов к той или иной партии. При импорте файла это поле следует объявить информационным по назначению, что полностью исключает значения этого поля из анализа, но позволяет сравнить прогнозные и истинные значения на контрольной выборке. Качество классификации опять легко определить, используя таблицу сопряженности. Все 20 случаев оказались классифицированы правильно: 12 демократов и 8 республиканцев.

Завершая этот раздел, рассмотрим примеры использования нейросети для задачи классификации.

Пример 5.3

Как уже указывалось в предыдущей главе, классической задачей кластерного анализа является задача классификации ирисов по длине и ширине чашелистиков и длине и ширине листков, рассмотренная впервые Р. Фишером в 1936 г. В файле Пример_5_3.xls содержатся исходные данные для 150 ирисов трех типов: *Iris Setosa*; *Iris Versicolor*; *Iris Virginica*.

Требуется произвести кластеризацию ирисов по трем видам и выяснить, какие из характеристик цветков наиболее тесно связаны с их типом.

Решение

Импортируем данные из файла Пример_5_3.xls в программу Deductor и запустим мастер построения Карты Кохонена. На втором шаге мастера обработки определим назначение переменной Класс_цветка как Выходное, а назначение всех остальных переменных как Входное. Переменная с назначением Выходное никак не участвуют в построении карты Кохонена. Кластеризация строится полностью исходя из значений входных переменных, но мы сохранили во входном наборе данных переменную Класс_цветка, поскольку это позволяет проконтролировать, как согласуются результаты кластеризации с реальными данными.

На втором, третьем и четвертом шагах все предустановленные параметры построения карты Кохонена можно оставить без изменения.

На пятом шаге мастера обработки следует снять флажок в поле Автоматически определить количество кластеров и в поле Фиксированное число кластеров установить значение, равное трем.

На шестом шаге мастера обработки следует запустить процедуру обучения сети Кохонена и дождаться ее завершения по прошествии 500 эпох (как это установлено по умолчанию) или остановить обучение раньше, если распознаны 100 % случаев и средняя ошибка перестала уменьшаться.

На седьмом шаге следует определить способы отображения результатов анализа. Предустановленным является отображение итогов в виде карты Кохонена. Добавим к этому еще возможность представления данных в виде таблицы.

На восьмом шаге мастера обработки следует выбрать режимы отображения карты Кохонена, показанные на рис. 5.10. Как следует из рисунка, флажком отмечены все входные данные. Это означает, что на экран будут выведены четыре карты Кохонена с раскраской нейронов-победителей, зависящей от значения признака Длина_чашелистика, Ширина_чашелистика, Длина_лепестка, Ширина_лепестка. Способ раскраски можно выбрать: это либо цветная палитра, либо градации серого.

Чтобы была визуализирована карта Кохонена с отмеченными на ней кластерами, следует поставить флажок в поле Кластеры, а для появления

разделительной линии между кластерами поставим флажок в поле Границы кластеров. Галочка в поле Класс цветка позволяет визуализировать положение объектов, принадлежащих к реальным классам. Напомним, что мы оставили переменную Класс_цветка в наборе данных. Эта переменная не участвовала в создании карты, но теперь с помощью ее мы можем проверить, правильно ли распределяются объекты по кластерам.

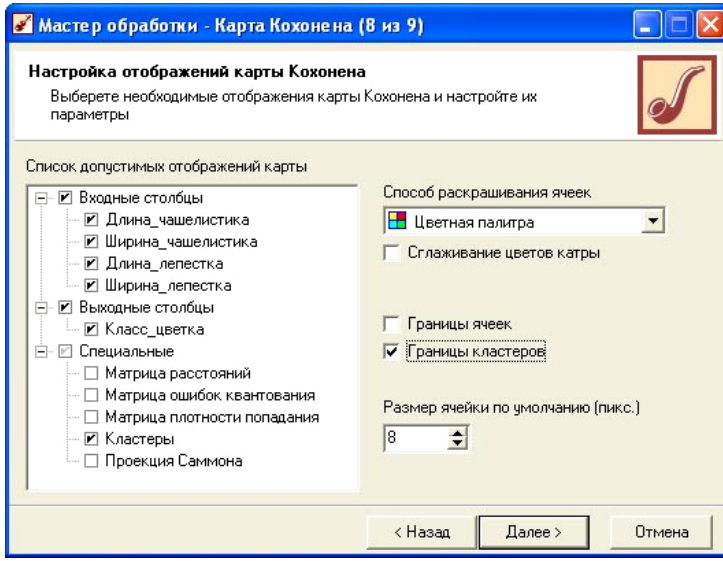


Рис. 5.10. Диалог выбора параметров отображения карты Кохонена в программе Deductor

Если карта Кохонена правильно представляет классификацию объектов, то цветки, принадлежащие к разным типам, попадут в разные кластеры и мы сможем визуально проконтролировать качество классификации.

Полученные в результате сделанного выбора представления карты Кохонена изображены на рис. 5.11.

Обратимся вначале к карте, названной Кластеры. Здесь разными цветами (различными оттенками серого цвета) показаны три кластера и граница между ними. Та же самая граница просматривается и на других картах. Степень правильности классификации мы можем проконтролировать по карте Класс_цветка на рис. 5.11. Пятна разного цвета (различных градаций серого цвета) изображают по существу нейроны-победители, соответствующие реальным объектам. На рисунке изображен случай, когда в создании карты участвует 400 нейронов (поле 20×20). Поскольку объектов все-

го 150, то нейроны-победители не сливаются в одно сплошное пятно, а выглядят изолированными. Из рисунка видно, что все объекты, принадлежащие к типу цветков *Setosa*, классифицированы правильно.

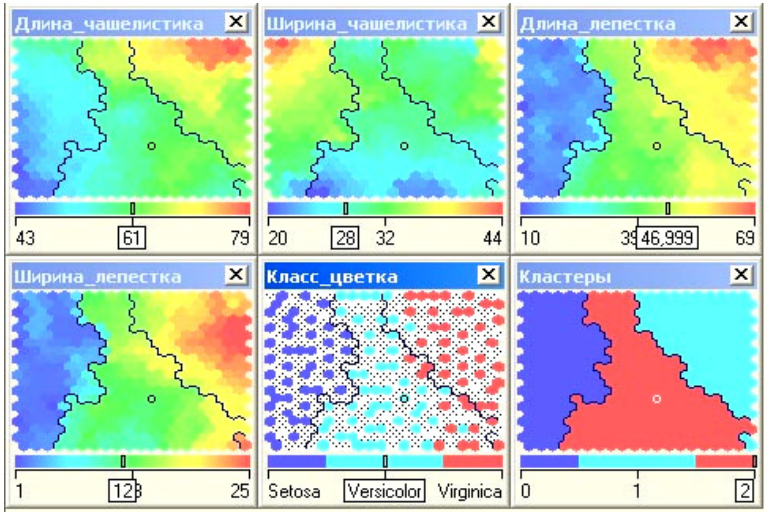


Рис. 5.11. Различные представления карт Кохонена

Различие цветков *Versicolor* и *Virginica* не столь сильное, и имеются несколько случаев неправильной классификации объектов этого вида.

Как говорилось выше, при анализе карт Кохонена проводится оценка не только выходов нейронов, но также и весов нейронов. Для каждого входа нейрона рисуется своя карта, которая раскрашивается в соответствии со значением соответствующего веса нейрона. У нейронной сети, обучаемой с «учителем», веса нейронов не имеет физического смысла и не используются в анализе. При обучении же без «учителя» веса нейронов подстраиваются под точные значения входных переменных и отражают их внутреннюю структуру. Для идеально обученной нейронной сети вес нейрона равен соответствующей компоненте входного примера.

Четыре первых карты на рис. 5.11 представляют собой раскраску нейронов-победителей по значениям входных переменных. Более светлую раскраску имеют нейроны-победители, которые представляют объекты с большим значением переменной, по градациям значений которой ведется раскраска. Рассмотрим, например, раскраску нейронов по значениям переменной Длина_лепестка. Более короткие лепестки имеют ирисы, относящиеся к классу Setosa, и более длинные – относящиеся к классу Virginica. Длина лепестка довольно точно делит цветки на классы. Поэто-

му раскраска нейронов по этому свойству практически точно соответствует разделению объектов на классы. То же самое можно сказать и о раскраске нейронов-победителей по свойству ширина лепестка. Раскраска по свойствам длины и ширины чашелистика в значительно меньшей степени соответствует разделению объектов на классы.

Таким образом, изучая раскраску нейронов-победителей по градациям входных переменных, можно визуально определить, какие из переменных имеют большую разделяющую силу.

Обратимся снова к рис 5.11. Дополнительную информацию о кластеризации объектов можно получить, если вывести карты матрицы расстояний, матрицы ошибок квантования, матрицы плотности попадания и карту проекции Саммона. На рис. 5. 12 приведены упомянутые выше карты для задачи классификации ирисов.

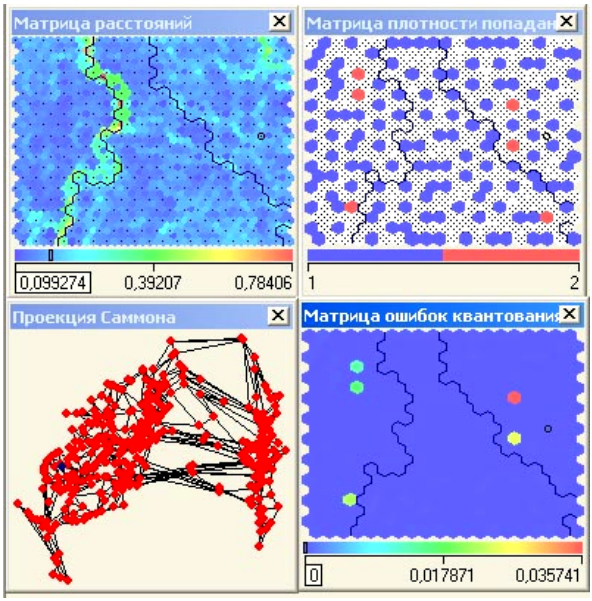


Рис. 5.12. Карты расстояний, плотности попаданий, проекции Саммона и карта ошибок квантования для задачи классификации ирисов

Матрица расстояний применяется для визуализации структуры кластеров, полученных в результате обучения карты. Элементы матрицы определяют расстояние между весовыми коэффициентами нейрона и его ближайшими соседями. Большое значение говорит о том, что данный нейрон сильно отличается от окружающих и относится к другому классу. На рис. 5.12 преобладают темные оттенки, что говорит о небольшом различии весовых коэффи-

циентов рядом расположенных нейронов. Лишь по границе кластера 0 и кластера 1 имеются нейроны, имеющие соседей с сильно отличающимися весами (эти нейроны раскрашены в более светлые тона).

Матрица плотности попаданий характеризует частоту выбора одного и того же нейрона-победителя. На карте видно, что практически все нейроны представляют один объект (вышли победителями всего один раз при предъявлении всех объектов нейросети). Лишь несколько нейронов, имеющих более светлую окраску, вышли победителями дважды (представляют два очень похожих объекта).

Проекция Саммона – это способ отображения положения объекта в многомерном признаковом пространстве на плоскости. Близкие в признаковом пространстве объекты должны располагаться недалеко друг от друга и в проекции Саммона. Если внимательно посмотреть на карту проекции Саммона на рис. 5.12, то можно выделить три группы объектов, соответствующие трем кластерам. Построение проекции Саммона очень часто не дает убедительной картины разделения объектов на кластеры и требует достаточно много процессорного времени.

Матрица ошибок квантования отображает расстояние от расположения объектов до центра ячейки. Объект располагается в многомерном признаковом пространстве, где количество измерений равно числу входных полей. Центр ячейки – это точка пространства с координатами, равными весам нейрона-победителя. Расстояние рассчитывается по формулам евклидовй метрики. Матрица ошибок квантования показывает, насколько хорошо обучена нейросеть. Чем меньше среднее расстояние от объекта до центра ячейки (нейрона-победителя), тем ближе к ней расположены объекты и тем лучше построена модель. Из рис. 5.12 видно, что для большинства нейронов-победителей ошибка квантования очень мала, и всего лишь в одном случае ошибка оказалась несколько больше 0,035.

Пример 5.4

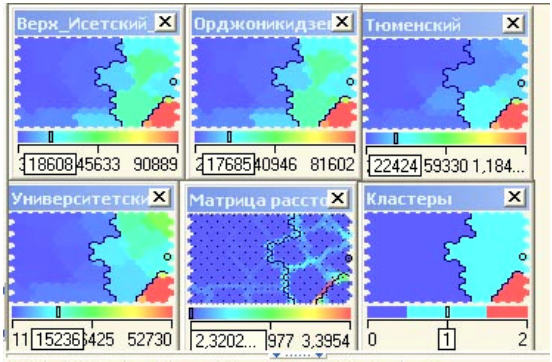
Файл Пример_5_4.xls содержит данные итогов голосования избирателей по 21 федеральному избирательному округу при выборе депутатов в Государственную Думу РФ в 2003 г. Данные приведены по 23 партиям, зарегистрированным для участия в выборах. Кроме того, файл содержит итоги голосования в графе «Против всех». Требуется произвести кластеризацию партий по итогам голосования в этих федеральных округах, выделив три наиболее типичные группы партий.

Решение

Последовательно выполним шаги импорта файла, создания и обучения сети Кохонена. Как и раньше, все установки параметров создания и обучения сети можно оставить без изменения. Исключение составляет шаг 2,

на котором нужно отказаться от тестовой выборки. Поскольку число объектов мало (всего 24), то нет никакого смысла создавать тестовую выборку из одного объекта. На пятом шаге следует отказаться от автоматического определения числа кластеров, указав, что нужна классификация объектов по трем группам.

Результаты классификации партий по итогам избирательной компании в 21 федеральном избирательном округе в 2003 г. представлены на рис. 5.13.



Партия	Верх_Исетски	Каменск_Ура	Нижнетагильски
СОЮЗ ПРАВЫХ СИЛ	38342	10088	11558
ЯБЛОКО	24566	9905	12938
Родина	24692	13083	13351
ЛДПР	34143	35004	32906
КПРФ	18608	19899	17418
Против всех	15638	10042	11072

Рис. 5.13. Кластеризация политических партий РФ
по итогам выборов в Госдуму в 2003 г.
(в таблице отражены партии, попавшие в первый кластер)

Очень часто полезно анализировать графические карты и таблицу результатов кластеризации одновременно. Для этого необходимо на панели инструментов, позволяющих управлять отображением карт Кохонена, нажать кнопку с пиктограммой . В результате в нижней части окна отображения данных откроется таблица результатов, фрагмент которой изображен на рис. 5.13. Используя фильтр (кнопка с пиктограммой), можно отсортировать данные, например, по значениям поля Номер_кластера (это поле генерируется программой при построении итоговой таблицы). На рис. 5.13 с помощью фильтрации отображен список партий, попавших в первый кластер.

Если выделить в таблице одну из строчек и нажать на панели управления отображением таблицы кнопку с пиктограммой , то на всех картах будет поставлена точка, отмечающая нейрон-победитель, соответствующий этому объекту.

Интересно, что кластер 2 (на рис. 5.13 – это кластер в правом нижнем углу карты Кластеры) состоит всего лишь из одной парии – «Единая Россия». Все остальные партии попали в наиболее многочисленную группу слабо различаемых избирателями партий, образующих кластер с номером «0». Отмеченная выше структура кластеров отчетливо проявляется при анализе итогов голосования в Екатеринбурге. Этот результат подтверждает мнение политологов, что результаты голосования в городе Екатеринбурге очень тесно коррелируют с итогами голосования в РФ в целом.

Мы выделили три кластера при построении карт Кохонена. Остается вопрос о том, какие все-таки партии избиратели различают? Чтобы ответить на него, достаточно посмотреть на карту расстояний между кластерами. Видно, что партии, попавшие в нулевой кластер, избиратели практически не различают, поскольку здесь расстояние между соседними нейронами-победителями мало. Партии, попавшие в первый кластер, также различаются достаточно слабо, но здесь все-таки просматривается некоторая структура. Таким образом, для определения разумного числа кластеров, которое следует выделять при построении карты Кохонена, следует проанализировать карту матрицы расстояний между кластерами. Ее структура отражает внутреннюю структуру данных.

Контрольные вопросы

5.1. Что представляет собой искусственный нейрон? Какое преобразование информации он выполняет?

5.2. Какие задачи могут решать нейронные сети?

5.3. Что понимается под архитектурой нейронной сети? Какие виды нейронных сетей могут использоваться для анализа данных?

5.4. Чем отличаются функции нейронов во входном слое, промежуточных слоях и выходном слое?

5.5. Объясните, почему функция активации нейрона обычно является нелинейной.

5.6. Что понимается под процессом обучений нейронной сети?

5.7. В чем состоит сущность алгоритма обратного распространения ошибок при обучении нейронной сети?

5.8. Как можно оценить необходимое число нейронов в сети, позволяющее анализировать данные заданного объема?

5.9. В чем состоит смысл понятия «эпоха» в задаче обучения нейронной сети?

5.10. В чем состоит суть проблем обобщения и переобучения для нейронной сети?

5.11. Зачем необходимо наряду с обучающей использовать тестовую и валидационную выборки?

5.12. В каком виде обученная нейронная сеть хранит информацию о свойствах обучающей выборки?

5.13. Чем различаются процедуры обучения нейронной сети с «учителем» и без «учителя»?

5.14. В чем состоит смысл соревновательного обучения нейронов сети при обучении без «учителя»?

5.15. Как можно определить нейрон-победитель на очередном шаге соревновательного алгоритма обучения?

5.16. В чем состоит модификация соревновательного алгоритма обучения Кохоненом?

5.17. Какую информацию можно извлечь из сети, обученную в соответствии с соревновательным алгоритмом Кохонена?

5.18. Какие существуют способы визуализации карт Кохонена?

5.19. Назовите основные достоинства и недостатки программы Deductor. Для решения каких задач предназначена эта программа?

5.20. Интерфейс программы Deductor. Сценарии импорта, экспорта и обработки данных.

5.21. Имеется ли в программе Deductor возможность сохранения обученной нейронной сети для дальнейшего ее использования в качестве экспертной системы?

5.22. Какие способы визуализации результатов имеются в программе Deductor при построении нейронных сетей и карт Кохонена?

5.23. Как оценивается среднеквадратическая ошибка работы нейронной сети?

5.24. Какой смысл имеют карта кластеров, карты раскраски нейронов-победителей по значениям входных переменных, карты матрицы расстояний, матрицы плотности попаданий, карта проекции Саммона?

Задачи и упражнения

5.1. В файле Задача_5_1.xls содержатся данные об оценочной стоимости садовых домиков в БТИ, времени (мес.), прошедшего со дня оценки до продажи, типе постройки (деревянный дом – 0, кирпичный – 1) и реальной цене, по которой дом был продан.

а) Используя нейронную сеть с двумя скрытыми слоями и двумя нейронами в слое, постройте нейросетевую модель, позволяющую предсказывать продажную стоимость домиков.

б) Экспортируйте таблицу результатов в файл Excel и найдите среднюю относительную ошибку прогноза по обучающей выборке.

5.2. В файле Задача_5_2.xls приведены данные о поквартальных продажах автомобилей в США (тыс. шт.) с 1979 по 1986 г. и данные о ВНП (млрд долларов США), уровне безработицы (%) и налоговой ставке (%). Данные, характеризующие ВНП, уровень безработицы и уровень налоговой ставки, приведены с запаздыванием на один квартал. Причины этого понятны, поскольку люди, планирующие расходы на покупку автомобиля, ориентируются на экономические показатели предшествующего временного периода.

а) Оцените минимальное число нейронов или число синаптических узлов, которые нужно использовать при построении нейронной сети.

б) Постройте нейросетевую модель, позволяющую предсказывать поквартальный объем продаж автомобилей в США.

б) Экспортируйте таблицу результатов в файл Excel и постройте график зависимости истинных и прогнозных значений объемов продаж в зависимости от порядкового номера квартала.

5.3. В файле Задача_5_3.xls приведены данные о 121 автомобиле выпуска 2002 г. Требуется построить регрессионную модель, позволяющую предсказать пробег автомобиля при использовании 1 литра топлива (результативная переменная Пробег) в зависимости от таких факторов, как тип автомобиля (спортивный или неспортивный), тип привода (полный, передний, задний), качество топлива (высшее, среднее), мощность двигателя (л. с.), длина автомобиля (м), ширина автомобиля (м), вес (кг), индекс грузоподъемности и радиус разворота автомобиля (м).

а) Используя формулу (5.21), оцените минимальное и максимальное число нейронов в нейронной сети для анализа этих данных.

б) Постройте нейронную сеть, используя 80 % данных как обучающую выборку и 20 % данных – как тестовую. Оцените качество прогноза, используя формулу (5.24) для ошибки RMSE .

5.4. В файле Задача_5_4. xls содержатся данные о проценте голосов, набранных правящей партией на президентских выборах с 1916 по 2000 г. (результативный признак), и данные о некоторых социально-экономических показателях США в год выборов (факторные переменные). В качестве факторов используются следующие независимые переменные:

- правящая партия (демократическая, республиканская);
- процент роста ВНП за первые девять месяцев в год выборов (%);
- темп инфляции за первые девять месяцев в год выборов (%);
- число кварталов за последние четыре года, когда рост ВНП превышал 3,2 %;
- число сроков подряд, в течение которых правящая партия находится у власти;
- проходят ли выборы в период, когда страна ведет войну (учитывались только глобальные войны);
- выдвигается ли действующий президент на следующий срок.

а) Используя все 100 % случаев как обучающую выборку, постройте простейшую нейронную сеть с одним скрытым слоем и двумя нейронами в нем. Обучите нейронную сеть, используя алгоритм обратного распространения ошибок.

б) Выберите в качестве способов отображения результатов пункты: Граф нейросети; Что-если; Диаграмма рассеяния; Таблица.

в) Используя возможность в диалоге Что-если подставить в качестве входных параметров новые значения переменных, получите прогноз нейросети на результаты выборов президента США в 2004 г. (финальная пара Буш – Керри), если факторные переменные в этом году имели следующие значения: Республиканцы; 4,7; 1,3; 6; 2; Нет; Да. Значения факторных данных приведены в том же порядке, что и их описание в условии задачи.

5.5. В файле Задача_5_5.xls приведены данные, представляющие собой результаты психологического тестирования учащихся специализированных школ Санкт-Петербурга с физико-математическим и гуманитарным уклоном. Всего предлагалось пять тестов, условное название которых приведено ниже: Тест_1 – дополнение предложений; Тест_3 – нахождение аналогий; Тест_4 – обобщение умозаключений; Тест_5 – способность к устному счету; Тест_7 – образность мышления. Чем выше набранный балл, тем лучше проявляется анализируемое качество. Учащиеся школ с физико-математическим уклоном условно названы физиками, а с гуманитарным уклоном – лириками. Используя 75 % случаев в качестве обучающей и 25 % – в качестве тестовой выборок, постройте и обучите нейронную сеть таким образом, чтобы доля правильно распознанных объек-

тов в тестовой выборке превышала 80 %. Для отображения результатов работы нейросети используйте таблицу сопряженности.

5.6. В файле Задача_5_6.xls содержатся данные о 149 клиентах банка, желавших получить кредит, и решение опытного менеджера о выдаче или отказе в выдаче кредита.

а) Используя 75% случаев как обучающую выборку и 25 % случаев как тестовую, убедитесь в том, что сложная нейронная сеть легко переобучается, т. е. легко распознает объекты в обучающей выборке и очень плохо – в тестовой.

б) Оставьте в качестве объясняющих переменных только переменные Сумма_кредита, Срок_кредита, Площадь_квартиры, Расположение, Время_работы_предприятия, Должность, Среднемес_доход, Среднемес_расход, Срок_проживания_в_регионе. Остальные объясняющие переменные переведите в разряд информационных. Улучшилось ли при этом качество модели? О результатах обучения следует судить по доле правильно распознанных случаев в обучающей выборке. Для отображения результатов работы нейросети используйте таблицу сопряженности.

5.7. База данных риелторской фирмы содержит 1721 запись оценок стоимости проданного жилья в одном из городов РФ. Кроме продажной стоимости, фиксировалось еще 11 различных параметров квартиры. База данных содержится в файле Задача_5_7.xls.

а) Выберите из этой базы 150 записей случайным образом, скопируйте их на новый лист рабочей книги и импортируйте эти данные в программу Deductor.

б) Определите число нейронов в сети, необходимое для распознавания правил, содержащихся в данных. Обучите нейронную сеть правильно прогнозировать продажную стоимость квартир, не допуская ее переобучения. Сохраните обученную нейросеть в виде файла сценария. Экспортируйте данные в файл Excel и вычислите относительную ошибку, используя формулу (5.25).

в) Сформируйте другую выборку из исходной базы данных риелторской фирмы, сохраните ее на новом листе рабочей книги и импортируйте в программу Deductor.

г) Загрузите сохраненную ранее ветвь сценария для анализа данных тестовой выборки. Найдите ошибку предсказания нейронной сети на тестовой выборке, используя формулу (5.25). Сравните ошибку на обучающей и тестовой выборках.

5.8. В файле Задача_5_8.xls содержатся данные о кредитных рейтингах 2464 клиентов банка (результативная переменная), пользовавшихся бан-

ковскими услугами в прошлом, а также такие данные о клиентах, как возраст (лет), уровень дохода (низкий, средний высокий), число используемых кредитных карт (меньше 5, больше 5), уровень образования (высшее, среднее специальное) и число кредитов на покупку автомашины (нет или 1, 2 и более).

а) Выберите из этой базы 150 записей случайным образом, скопируйте их на новый лист рабочей книги и импортируйте эти данные в программу Deductor.

б) Определите число нейронов в сети, необходимое для распознавания правил, содержащихся в данных. Обучите нейронную сеть правильно прогнозировать кредитный ранг клиентов банка, не допуская ее переобучения. Сохраните обученную нейросеть в виде файла сценария. Для оценки качества обучения используйте данные таблицы сопряженности.

в) Сформируйте другую выборку из исходной базы данных, сохраните ее на новом листе рабочей книги и импортируйте в программу Deductor.

г) Загрузите сохраненную ранее ветвь сценария для анализа данных тестовой выборки. Долю правильно интерпретируемых случаев оцените с помощью таблицы сопряженности.

5.9. В файле 5_9_xls приведены данные о рейтингах стран по уровню развития демократии, уровню свободы прессы, уровню коррумпированности аппарата чиновников, опубликованные организацией Transparency International на сайте http://www.worldaudit.org/democracy.htm. Хотя эти данные тенденциозны и представляют точку зрения только западных аналитиков, произведите анализ этих данных с помощью карт Кохонена.

а) Импортируйте данные в программу Deductor, постройте нейронную сеть Кохонена для кластеризации стран по уровню развития демократических институтов, выделив 3 кластера.

б) Постройте карты раскраски нейронов-победителей по уровню развития демократии, уровню свободы прессы, уровню коррупции, а также карту разбиения стран на кластеры. Раскраска по каким входным переменным более тесно коррелирует с выделенными кластерами?

в) Откройте таблицу результатов работы нейросети и постройте списки стран, попавших в первый, второй и третий кластеры, используя возможность фильтрации данных по номеру кластера.

5.10. В файле Задача_5_10.xls приведены некоторые данные, характеризующие производственную деятельность основных нефтяных компаний РФ в 1996 г.

а) Используя построение карт Кохонена, произведите деление предприятий на три группы. Какие из показателей деятельности компаний более важны при классификации компаний? Постройте проекцию Саммона.

б) С помощью фильтрации данных по номеру кластера постройте списки компаний, попавших в разные кластеры. Проанализируйте показатели компаний, попавших в различные кластеры, и дайте интерпретацию этим группам.

5.11. Используя данные задачи 5.8, произведите кластеризацию клиентов банка с целью выявить группы клиентов, которым, возможно, следует предлагать разные условия кредитования, выделив сначала два, а затем три кластера.

а) Проанализируйте карты раскраски нейронов по значениям входных переменных, а также карты раскраски по значениям выходной переменной Кредитный_ранг. Как влияют на кредитный рейтинг возраст, число кредитных карт, образование, число кредитов на автомобили? Какая из входных переменных обладает наибольшей разделяющей силой?

б) Используя обученную сеть Кохонена, повторите исследования на тестовой выборке. Можно ли утверждать, что обученная сеть Кохонена может быть использована как экспертная система?

в) Используя способ представления результатов Что-если, смоделируйте работу экспертной системы, подставив новые значения параметров, характеризующих клиента банка, взяв их из базы данных клиентов банка (лист База файла Задача_5_8.xls).

5.12. Используя условие задачи 5.5, постройте кластеризацию учеников школ по итогам тестирования с помощью карт Кохонена, выделив два кластера. Отобразите карту кластеров, матрицу расстояний, проекцию Саммона, карты раскраски нейронов-победителей по градациям входных переменных. Какая из входных переменных позволяет лучше понять структуру объектов, попавших в разные кластеры? Насколько хорошо разделение на кластеры коррелирует с делением учеников на «физиков» и «лириков»?

5.13. Используя условие задачи 5.7, дайте классификацию квартир с помощью карт Кохонена, выделив три кластера.

а) Отобразите карту кластеров, матрицу расстояний, проекцию Саммона, карты раскраски нейронов-победителей по градациям таких входных переменных, как стоимость квартиры, общая площадь, индекс района.

б) Изучая карты Кохонена, раскрашенные по значениям входных переменных, сформулируйте некоторые правила, по которым произошло разбиение квартир на классы.

в) Сохранив обученную сеть, проверьте, сохраняются ли сделанные выше выводы на другой выборке объектов из базы данных. При сравнении карт Кохонена, полученных при каждом новом этапе обучения, следует помнить, что нумерация и расположение кластеров могут измениться. Полезная информация получается при сопоставлении раскраски нейронов по значениям входных переменных и кластеров.

5.14. В файле Задача_5_14.xls приведены данные о результатах хозяйственной деятельности 142 крупнейших предприятий РФ в 1997 г. по объему реализации продукции. Используя индикаторы Доб_обр (добыча/обработка; добывающая отрасль кодируется цифрой 1, обрабатывающая – цифрой 0), Темпы_роста (темп роста по отношению к 1996 г.), Прибыль_на_объем (отношения прибыли к объему выпускаемой продукции в 1997 г.), Объем_чел (удельный объем выпускаемой продукции на одного занятого в производстве, млн руб./чел.), постройте классификацию предприятий, используя карты Кохонена.

а) Используя первые 142 объекта и перечисленные выше индикаторы как входные переменные, обучите нейронную сеть, выделив 4 – 5 кластеров.

б) Отобразите карту кластеров, матрицу расстояний, проекцию Саммона, карты раскраски нейронов-победителей по градациям входных переменных. Используя фильтр, постройте списки предприятий, вошедших в различные кластеры. Попробуйте сформулировать некоторые из правил, по которым объекты отнесены к различным кластерам.

в) Сохраните обученную нейросеть.

г) Экспортируйте результаты работы нейросети (таблицу) в Excel.

д) Загрузите результаты работы нейросети в редактор данных SPSS (перед импортом данных в SPSS необходимо устранить ненужные для анализа столбцы) и, используя дерево решений, найдите правила, по которым предприятия отнесены к различным кластерам. Согласуются ли эти правила с теми, которые вы смогли сформулировать, анализируя списки объектов, попавших в различные кластеры?

ж) Используя обученную нейронную сеть, повторите пункты исследования б – д для данных, находящихся в файле Задача_5_14.xls на листе Контрольная выборка.

з) Повторите пункты исследования а – ж, используя в качестве входных переменных исходные данные работы компаний: Объем_1997, Объем_1996, Баланс_приб_1997, Кол-во_раб, Доб_обр в качестве входных переменных.

5.15. В главе I мы рассматривали проблему классификации изображений на два класса (см. рис. 1.3). В файле Задача_5_15.xls содержатся значения 16 дихотомических переменных x1 – x16, характеризующих изображения лиц, а переменная x17 содержит данные о том классе, к которому следует отнести это изображение. Произведите разбиение данных этого примера на два класса, используя построение карт Кохонена. Ранее предлагалось решить эту задачу с помощью построения дерева решений (задача 4.22).

а) Используя все 100 случаев как обучающую выборку, обучите сеть Кохонена, ограничив число распознаваемых кластеров значением, равным двум. Переменные x1 – x16 следует определить как входные, а переменную x17 – как выходную (выходная переменная не участвует в обучении сети Кохонена, но позволяет проверить, насколько хорошо данные классификации совпадают с известными результатами).

б) Визуализируйте таблицу результатов, карту кластеров, матрицу расстояний, карту раскраски нейронов-победителей по градациям выходной переменной.

в) Анализируя таблицу, убедитесь, что нейросеть распознала все объекты правильно (значения переменных x17 и x17_OUT полностью совпадают), но кластеризация оказалась неудовлетворительной.

г) Выполните кластерный анализ этого набора данных с помощью метода К-средних и сравните результаты с данными, полученными с помощью карт Кохонена. Как вы можете объяснить, что оба метода дают плохие результаты для этого примера?

СПИСОК ЛИТЕРАТУРЫ

Басегян А. А. и др. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004.

Бююль А., *Цёфель П.* SPSS: искусство обработки информации: Анализ стат. данных и восстановление скрытых закономерностей: Пер. с нем. СПб.: ООО «ДиаСофтЮП», 2001.

Дубнов П. Ю. Обработка статистической информации с помощью SPSS. М.: АСТ: НТ Пресс, 2004.

Дюк В.,. Самойленко А. Data Mining: учеб. курс. СПб.: Питер, 2001.

Ежов А.А., Шумский С.А. Нейрокомпьютинг и его применение в экономике и бизнесе [Электрон. ресурс]. Режим доступа: http://neuroschool.narod.ru/books.html

Круглов В. В., Борисов В. В. Искусственные нейронные сети: Теория и практика. 2-е изд., стереотип. М.: Горячая линия–Телеком, 2002.

Левин Д. М. и др. Статистика для менеджеров с использованием Microsoft Excel. 4-е изд.: Пер. с англ. М.: Изд. дом «Вильямс», 2004.

Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005.

Таганов Д. SPSS: Статистический анализ в маркетинговых исследованиях. СПб.: Питер, 2005.

Толстова Ю.Н. Анализ социологических данных. М.: Науч. мир, 2000.

Томас Р. Количественные методы анализа хозяйственной деятельности: Пср. с англ. М.: Дело и Сервис,1999.

Тюрин Ю. Н. Макаров А. А. Анализ данных на компьютере / Под ред. В. Э. Фигурнова. 3-е изд. перераб и доп. М.: ИНФРА-М, 2003.

Чубукова И. А. Data Mining: учеб. курс. М.: БИНОМ, 2006.

Larose D. T. Discovering knowledge in data. An Introduction to Data Mining. A John Wiley & sons, inc., 2005.

З

И

К

Л

М

Н

О

П

Р

С

Т

Ф

Ш

Э

Учебное издание

Биккин Халид Мирхасанович
Полтавец Андрей Васильевич
Шашкин Сергей Юрьевич

Компьютерный анализ данных для менеджеров

Учебное пособие по курсу «Математические методы и компьютерные технологии в управлении»

Для студентов специальности 080500.068 «Менеджмент»

Компьютерный набор и верстка О. А. Тлашадзе

Подписано в печать 27.04.2007.
Формат 60x84/16. Гарнитура *Times New Roman.* Уч-изд. л. 16 12
Отпечатано на ризографе. Тираж 150 экз. Заказ 13/08.
Уральская академия государственной службы.
620148 Екатеринбург ул. 8 Марта 66.