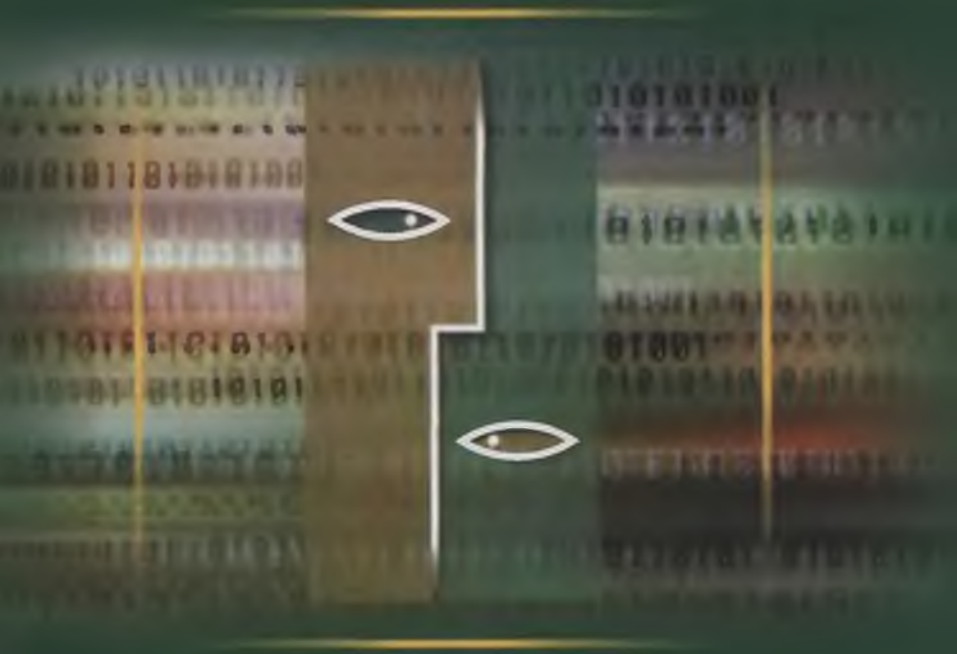
Дункан Крамер

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ В СОЦИАЛЬНЫХ НАУКАХ

Современные методы



ACADEMIA

DUNCAN CRAMER

Advanced Quantitative Data Analysis

Open University Press
Maidenhead · Philadelphia

ДУНКАН КРАМЕР

Математическая обработка данных в социальных науках: современные методы

Перевод с английского

Рекомендовано Советом по психологии
УМО по классическому университетскому образованию
в качестве учебного пособия для студентов высших учебных заведений,
обучающихся по направлению и специальностям психологии

ACADEMIA

Москва
Издательский центр «Академия»
2007

УДК 519.221.25(075.8)
ББК 22.172я73
К777

Рецензенты:
доктор техн. наук *Л. С. Куравский*, зав. кафедрой «Прикладная информатика» Московского городского психолого-педагогического университета;
доктор филос. наук, канд. физ.-мат. наук *А. Н. Кричевец*, вед. науч. сотр. факультета психологии Московского государственного университета им. М. В. Ломоносова

Крамер Д.

К777 Математическая обработка данных в социальных науках : современные методы : учеб. пособие для студ. высших учеб. заведений / Дункан Крамер; пер. с англ. И. В. Тимофеева, Я. И. Киселевой; науч. ред. О. В. Митина. — М. : Издательский центр «Академия», 2007. — 288 с.

ISBN 978-5-7695-2878-1

В пособии американского психолога и математика Дункана Крамера рассматриваются методы статистической обработки данных, применяемые в современных социальных исследованиях. Не останавливаясь на базовых понятиях и критериях, известных из любого начального курса статистики, автор пытается доступным языком, без сложных формул и расчетов, объяснить принципы применения факторного и кластерного анализа, линейной и логистической регрессии, анализа путей, дисперсионного и ковариационного анализа, дискриминантного и, наконец, логлинейного анализа.

Пособие снабжено предисловием и комментариями научного редактора, списками рекомендуемой литературы на русском и английском языках, глоссарием статистических терминов и приложением.

Для студентов высших учебных заведений, изучающих методы математической обработки в социальных и гуманитарных науках, а также для преподавателей и исследователей-практиков.

УДК 519.221.25(075.8)
ББК 22.172я73

ISBN 978-5-7695-2878-1

ПРЕДИСЛОВИЕ НАУЧНОГО РЕДАКТОРА

Издание книги на русском языке, в которой описывается применение современных статистических методов анализа социальных и гуманитарных данных во всем их разнообразии, было и ожидаемо, и необходимо. Ни для кого не секрет, что Россия значительно отстает от стран Западной Европы и США в области приложения количественных методов. В этих странах подобных учебников, адресованных читателям самого различного уровня подготовки (от студентов колледжей, еще не имеющих даже степени бакалавра, до докторантов, избравших прикладную статистику и измерения в гуманитарных науках своей основной специализацией), издается великое множество, однако соответствующих переводов на русский язык не было с тех пор, как прекратилось издание серии «Библиотечка иностранных книг для экономистов и статистиков». За последнюю четверть XX в. за рубежом достигнуты большие успехи в области статистики с точки зрения как развития науки (появились новые методы статистического анализа), так и программного обеспечения (все больше методов реализуется в широко используемых статистических пакетах), а также в области методологии (понимания того, какие методы и в каких случаях использовать). Западные исследователи активно применяют статистические методы разной степени сложности, о которых российские ученые узнают либо из публикуемых статей, либо из выступлений на конференциях.

За последние годы было издано достаточно много монографий и учебных пособий, написанных отечественными авторами. Но они охватывают в основном простейшие и самые распространенные способы статистического анализа: описательную статистику, сопоставление двух выборок, а все выходящее за рамки этого обязательного минимума рассматривается только в ознакомительном формате (А. Д. Наследов, 2004; А. П. Кулаичев, 2006) либо посвящено отдельным методам (А. Н. Гусев, 2000; О. В. Митина, И. Б. Михайловская, 2001). В результате большое количество методов, уже широко известных и активно используемых коллегами за рубежом, позволяющих проверять более интересные и дифференцированные гипотезы, до настоящего времени известны узкому кругу исследователей, применяются крайне редко (например, дискриминантный анализ, анализ ковариаций) или не используются вообще (логлинейный анализ).

Структурное моделирование, оформившееся как методология работы с данными в США и странах Западной Европы в конце 70-х — начале 80-х годов XX в. и по сути органично включающее практически все линейные статистические методы — от определения простейших показателей до многомерного регрессионного и факторного анализа, получившего здесь естественное развитие и объединение (О. В. Митина, 2005), является, по нашему мнению, абсолютно необходимым в психологии и других гуманитарных дисциплинах, но только начинает применяться в России, поэтому учебно-методическое обеспечение особенно актуально.

Хотя статистический анализ имеет многовековую историю, по-настоящему оформление прикладной статистики как методологии работы с числовыми данными можно связывать с именем Ф. Гальтона, который в конце XIX в. впервые применил статистический анализ в биологии и психологии, ввел в психологию тесты и опросники (включая и сам термин «тест»), разработал близнецовый анализ. В 1888 г. ученый выступил с докладом на заседании Королевского общества «Корреляции и их измерение, преимущественно по антропометрическим данным».

Применение статистических методов во многом шло параллельно и взаимосвязанно с развитием метрических дисциплин как в естественных и инженерных (био-, хемометрика, метрология), так и в гуманитарных, социальных (измерения в психологии, социологии, экономике, истории) областях и способствовало их оформлению в науки с точки зрения требований строгости, доказательности, объективности, верифицируемости и пр. Статистика помогает доказывать различные гипотезы о проявлениях психологических свойств личности, целесообразности использования новых средств и методов лечения, причин и следствий тех или иных заболеваний и отклонений, предсказывать результаты политических выборов, проводить разработку месторождений полезных ископаемых, контролировать работу атомных станций и т.д. Как правило, проверяются гипотезы о количественных характеристиках различных параметров (переменных) и связи этих параметров друг с другом попарно или в более сложных конфигурациях. Затем возникает вопрос точности и обоснованности результатов. Насколько надежны результаты исследования? Достаточно ли большая выборка? Существуют ли подвыборки, в которых взаимосвязь между установленными переменными значимо различается? Насколько можно доверять полученным измерениям? На все эти вопросы можно ответить с помощью статистических методов. Широкое распространение компьютеров привело к тому, что возможность провести статистический анализ имеет практически каждый, однако, для того чтобы не сделать ошибочных выводов, необходимы соответствующие знания.

Одним из наиболее ранних примеров такого рода стал так называемый парадокс Симпсона (E. H. Simpson, 1951). Предположим, нам необходимо проверить эффективность определенного вида воздействия. Это могут быть какой-то новый вид психотерапевтического воздействия, медицинский препарат, обучение какому-то предмету в школе или в вузе по новой методике и т.д. На формальном уровне, для того чтобы зафиксировать наличие или отсутствие воздействия, вводится независимая дихотомическая переменная X, принимающая значения «0» при отсутствии воздействия и «1», если воздействие имело место. Необходимо оценить, влияет ли X на значения зависимой переменной Y, соответствующей успешности. Дихотомическая переменная $Y = 1$, если эффект был зафиксирован (успех достигнут), $Y = 0$ — при отсутствии эффекта (успеха). Согласно всем требованиям проведения подобного рода экспериментов, всех испытуемых разделяют на две равные группы: экспериментальную (подвергшуюся воздействию) и контрольную (воздействию не подвергавшуюся). В приведенной ниже таблице общее количественное распределение в контрольную и экспериментальную группы указано в столбцах, озаглавленных «Всего», строки содержат информацию о достигнутом результате (успехе). Ведь можно обучиться какому-либо предмету и по старым учебникам, решить свою психологическую проблему и без вмешательства психотерапевта и т.д.

Согласно данным, представленным в таблице, в эксперименте принимало участие 2000 испытуемых[1]. Они были поровну распределены в экспериментальную и контрольную группы в соответствии со стандартными требованиями экспериментального дизайна. Из 1000 человек, подвергшихся воздействию изучаемого фактора (попавших в экспериментальную группу), успех был зафиксиро-

Таблица. **Оценка успешности воздействия по всей выборке в целом и по каждому полу**

Успех	Количество испытуемых			Воздействие					
				Есть ($X=1$)			Нет ($X=0$)		
	Всего	Мужчины	Женщины	Всего	Мужчины	Женщины	Всего	Мужчины	Женщины
Да ($Y=1$)	1100	375	725	500	300	200	600	75	525
Нет ($Y=0$)	900	625	275	500	450	50	400	175	225
Всего	2000	1000	1000	1000	750	250	1000	250	750

[1] Все данные, содержащиеся в таблице, являются гипотетическими.

ван у 500 человек (у половины — 50 %), у испытуемых, входивших в контрольную группу, успех был зафиксирован у 600 человек (т.е. в 60 % случаях). Предположим теперь, что мы хотим проанализировать эффекты воздействия отдельно на подвыборках мужчин и женщин. И тех и других было поровну, а конкрстные данные проведения эксперимента по каждой из этих подвыборок содержатся в столбцах, озаглавленных «мужчины» и «женщины» соответственно. Успех был зафиксирован у 300 из 750 мужчин, входивших в экспериментальную группу (40 %), и у 75 из 250 мужчин, входивших в контрольную группу (30 %). Аналогично достигли успеха 200 из 250 женщин, входивших в экспериментальную группу (80 %), и 525 из 750 женщин, входивших в контрольную группу (70 %). Таким образом, согласно полученным результатам, изучаемое воздействие оказывает положительный эффект на мужчин и женщин в отдельности, но не на выборку в целом. В итоге имеем парадокс[1].

Чтобы исключить ошибочные построения, научные сообщества вырабатывают стандарты представления результатов статистического анализа. И от рецензентов в научных журналах требуется оценивать предлагаемые публикации не только с точки зрения содержательной ценности представленных результатов, но и с точки зрения статистической грамотности их получения и обоснования.

В нашей стране при высоком уровне математической подготовки студентов и развитии математической науки в целом все, что связано с прикладными областями, исторически считалось чем-то малозначимым. Интересно в связи с этим процитировать письмо великого математика Н. Н. Лузина одному из наиболее талантливых своих учеников А. Н. Колмогорову, написанное в середине 20-х годов XX в.: «...Мое желание, чтобы Вы несколько удалились от работ по теории вероятностей. И вовсе не потому, что Ваш вклад в нее не фундаментален: я прекрасно знаю, что он оценивается всеми, как равноценный вкладу классиков. Но самое-то теория вероятностей не стоит Вас: ее источники сомнительные... и ее действие на работающих в ней не положительное. Вам дан высокий дух, и я хочу, чтобы Вы его силы берегли для вещей, которые под силу очень немногим». В то же время в традиции зарубежных ученых умение измерять и сопоставлять результаты полученных измерений еще в XIX в. рассматривалось как необходимый компонент общей грамотности и культуры. Приведем цитату из Карманной книги солдата того времени: «Подразумевается, что

---

[1] Было бы ошибкой думать, что подобного рода ситуации носят искусственный характер, чтобы их можно было встретить в реальном исследовании. Реальный пример, с которым столкнулись исследователи, выясняя, не отдается ли при отборе кандидатов в аспирантуру преимущество представителям одного пола перед другим, изложен в (P. J. Bickel, E. A. Hammel, J. W. O'Connell, 1985).

вы неплохо знаете арифметику; по крайней мере, две первые книги Евклида; планиметрию; алгебру вплоть до квадратных уравнений и фортификацию. Следует научиться с первого взгляда распознавать обычные разновидности растительного покрова, включая различные виды древесных пород. Для удобства измерения расстояний и т. п. каждому следует знать точную длину своего обычного шага и научиться точно отсчитывать шагами ярды; следует знать точную длину своей ступни, кисти, локтя, сабли, а также руки — расстояние от кончиков пальцев левой кисти до правого уха; следует знать высоту своего колена, талии и линии глаз, а также точное отношение объема своей питьевой кружки к пинте» (G.Wolseley, 1886).

Эти и ряд других обстоятельств привели к тому, что среди отечественных ученых сложилась традиция считать психологию сугубо гуманитарной наукой, а потому в определенном смысле даже несовместимой с количественным анализом, «приводящим к выхолащиванию принципов гуманности».

В последние годы российские психологи все интенсивнее сотрудничают с американскими и западноевропейскими коллегами, получают приглашения опубликовать результаты своих работ в иностранных журналах, но часто сталкиваются с проблемами, связанными с качеством выполнения и описания результатов количественного анализа данных. Для решения этих проблем, на наш взгляд, необходимо обеспечить доступ к определенной информации — удачно составленным лекционным курсам, качественным учебникам и монографиям; возможность присутствовать на докладах и мастер-классах, проводимых ведущими специалистами в этой области. Современная мобильность — возможность участия в зарубежных мероприятиях и Интернет-коммуникации — позволяет решить эти проблемы.

Наиболее образованными в области использования статистических методов оказались ученые-экономисты. И это неудивительно. Исторически сложилось так, что благодаря экономике у многих математиков появилась возможность стать лауреатами Нобелевской премии. Интегрирование математики в эту область науки способствовало интенсивному развитию эконометрики непосредственно (за счет собственных достижений) и косвенным образом (через повышение общего уровня математической грамотности ученых-экономистов). По мнению Нобелевского комитета, в настоящее время эконометрика применяется в качестве стандартного метода микроэкономики, изучающей все, начиная от расходов на ведение домашнего хозяйства и предпринимательских инвестиций и заканчивая организацией производств, рынков труда и эффектами государственной политики.

Россия в данном случае не является исключением. Статистическая и математическая грамотность эконометриков — достой-

ный пример для подражания. Госстандарт, утвержденный Министерством образования РФ для соответствующих специальностей, свидетельствует о возможности хорошей подготовки прикладных статистиков среди гуманитариев и социальных исследователей в нашей стране.

Отметим также, что, несмотря на лидерство статистики в области работы с количественными данными, это не единственный способ анализировать и репрезентировать результаты. Неправомерно сводить эконометрию, биометрию, клиометрию, психометрию и т.д. исключительно к статистике. В настоящее время методология гораздо шире и включает методы нелинейных динамических систем, нейронные сети, симуляционное моделирование и др. Однако даже в этом случае при построении моделей на основе функционального подхода исследователи используют данные предварительного статистического анализа, например аппроксимации или регрессии, прежде чем строить дифференциальную или разностную модель (О.В.Митина, В.Ф.Петренко, 2002). Более того, уместно напомнить, что даже математические физики, традиционно отдающие предпочтение аналитическим методам дифференциальных и интегральных уравнений, функционального анализа и т.п., в настоящее время предлагают учитывать флуктуационные эффекты, полноценно проанализированные с помощью статистических методов (В.И.Кляцкин, 2002). С учетом таких тенденций правомерно прогнозировать развитие количественных методов анализа и математического моделирования в противоположную сторону: от использования статистических методов к построению функциональных моделей. И в этом случае статистика будет играть роль не только иллюстратора результатов, но и своеобразного фундамента для построения дальнейших операциональных моделей, а значит, требования к уровню ее использования существенно возрастают.

Теория и практика анализа данных в той или иной науке, являясь одной из ее отраслей, занимают не какое-то обособленное место, а выполняют важную интегрирующую функцию. Использование сходного математического аппарата при решении исследовательских задач из разных сфер науки позволяет фиксировать их однотипность и тем самым помогает классифицировать исследовательские научные проблемы как интегральные, объединяющие в единый класс частные задачи, возникшие в различных отраслях, но по сути своей являющиеся проявлениями обобщенной латентной проблемы.

Здесь уместно вспомнить идею Л.С.Выготского (1982) о сравнении методологии со скелетом: внешним, наблюдаемым в простейших случаях, когда внутренности остаются мало дифференцированы и слабо детерминированы этим каркасом, и внутренним (являющимся опорой, костью каждого движения), и необхо-

димости различать низшие и высшие типы методологической организации. В настоящее время можно констатировать, что использование математического аппарата соответствует первому типу методологии, т.е. исследователь не имеет точного представления о том, какие методы, в каких случаях целесообразно применять, а исходит в первую очередь из того, что ему знакомо, привычно. Осознанное с этой точки зрения интегрирование математики в такой методологический каркас позволяет осуществить его преобразование из внешнего во внутренний. Можно предположить, что постановка проблемы и тип решаемой задачи существенным образом определяют выбор того или иного метода. Именно поэтому, по всей видимости, неправомерно ставить вопрос о создании учебника по всем методам анализа данных. Возможна лишь выработка основных стратегий, в рамках которой ученый должен проявить свой исследовательский талант, чувство «темы» и «данных». Однако для того чтобы такого рода творчество стало действительно доступным, необходимо в полном масштабе освоить технологию: совокупность приемов, алгоритмов и техник, используемых в той или иной науке на протяжении всей истории ее развития.

Итак, перейдем к анализу предлагаемой читателю книги. Собственный научный интерес автора связан с анализом психологических и педагогических данных. Преподавательскую деятельность он также ведет для студентов этих специальностей, чем и объясняется психологическое содержание рассматриваемых в книге примеров. Поэтому, безусловно, книга в первую очередь привлечет внимание ученых и аспирантов психолого-педагогического профиля; несмотря на то что требования к выполнению статистических процедур достаточно универсальны, тем не менее для каждой дисциплины существует своя специфика: степень строгости, допустимая погрешность, правила интерпретации и т.д. Однако и все остальные исследователи, чьи интересы связаны с анализом данных, найдут в ней много полезного для себя. Точно так же целесообразным является изучение психологами соответствующих книг для экономистов, медиков или биологов.

В каждой главе разбирается хотя и искусственный (сокращенный), но все же достаточно содержательный пример, что делает процедуру интерпретации достаточно осмысленной и стимулирует читателя не только проделать вслед за автором все вычисления, но, быть может, повторить их и на других переменных из этого примера, представляющих содержательный интерес. Все расчеты автор эксплицирует, и это методически очень полезно. Конечно, в настоящее время никто не выполняет вычисления ни на бумаге, ни даже с помощью калькулятора, однако при изучении того или иного метода целесообразно хоть раз проделать все выкладки «вручную», чтобы не относиться к компьютерной программе как к чер-

ному ящику (со страхом и недоверием или, наоборот, возлагая слишком большие надежды).

Необходимо заметить, что применение количественных методов сродни искусству и наивно было бы ожидать, что прочтя одну или даже несколько книг, вы обретете желаемую степень уверенности при работе. Главное — это большая практика. Именно тогда вы обнаруживаете «подводные камни», которые не указаны ни в одной книге — авторы даже порой не рефлексируют их. Поэтому не огорчайтесь, если, просчитав за автором весь пример, а потом выполнив все процедуры на компьютере, вы все еще не обрели желаемой уверенности. Главное — сохранить мотивацию продолжать освоение этой области.

Второй очень правильный методический прием (помимо включения в объяснения всех численных расчетов) — последовательное и подробное описание диалога работы с компьютером. Приведение в книге пошагово всех интерфейсных окон позволяет читателю без особых проблем воспроизвести все шаги самостоятельно и делает этот процесс воспроизводимым даже для читателей самого начального уровня компьютерной грамотности (не секрет, что среди психологов и педагогов таких немало).

Базовой программой, с помощью которой проанализировано большинство примеров, является SPSS. Это действительно наиболее популярная среди западных психологов и педагогов программа, и ее полезно освоить. При этом работать лучше с нерусифицированной версией. Недостатки русификации англоязычных статистических программ связаны с отсутствием устоявшейся статистической терминологии (в силу недостаточной развитости статистики в нашей стране). Поэтому переводчики порой проявляют фантастическую изобретательность, переводя то или иное слово, но при этом увеличивают хаос в сознании пользователя, не имеющего профессиональной математической и статистической подготовки. Отсюда возникает путаница при переводе, например, таких терминов: «part correlation» и «partial correlation». Или человек с удивлением осознает, что «ящик с усами» и «boxplot» — это одно и то же, или обнаруживает, что «индекс пригодности» обозначает надежность альфа-Кронбаха.

Одну «нишу» с SPSS (для исследователей и аспирантов, специализирующихся в гуманитарных и социальных науках) занимает программа STATISTICA. Это более молодая, но быстро развиваемая программа, имеющая множество дополнительных программных блоков, существенно расширяющих ее возможности. Большей частью они не востребованы отечественными учеными (например, алгоритмы нейронных сетей), но в перспективе владение программой STATISTICA может оказаться очень полезным.

Для читателей, имеющих начальный уровень компьютерного и статистического образования, полезно будет в качестве дополни-

тельной точки опоры использование программы STADIA, созданной в России (автор А. П. Кулаичев), а потому во многом учитывающей специфику отечественной аудитории (хорошо продуманный с методической точки зрения, дружественный интерфейс легко осваивается студентами-первокурсниками, встроенный Help достаточно информативен и понятен и содержит статистические рекомендации). Кроме того, автором STADIA недавно выпущен учебник (А. П. Кулаичев, 2006).

Для реализации конфирматорного факторного анализа и путевого анализа, являющихся частными случаями структурного моделирования, в книге используется еще одна программа — LISREL, узконаправленная для анализа структурных моделей. Эта программа не является распространенной в России, однако для примера, рассматриваемого в гл. 2, достаточно возможностей демонстрационной версии, бесплатно загружаемой прямо с сайта. На наш взгляд, кроме LISREL существуют более дружественные программы, с помощью которых выполнять структурное моделирование, в частности конфирматорный факторный анализ, оказывается намного легче. Примером такой программы является EQS (P. M. Bentler, 1996; О. В. Митина, 2005). Также у SPSS есть дополнительная надстройка AMOS, но она не входит в стандартный пакет и ее нужно приобретать дополнительно, в то время как в программе STATISTICA этот блок устанавливается по умолчанию (Л. Я. Дорфман, А. В. Огородников, 2005).

После каждой главы приведен список литературы, рекомендуемый автором, а также редактором перевода. Предлагаемые списки не претендуют на полноту, главное, что указанные в них книги еще не стали библиографической редкостью и их можно найти в книжных магазинах (год издания — не ранее 2000 г.). Правила применения той или иной процедуры можно найти и в справочной информации (Help) используемых программ SPSS, STATISTICA, STADIA, а также на их веб-сайтах.

О. В. Митина

ПРЕДИСЛОВИЕ РЕДАКТОРА СЕРИИ

Серия «Постижение методов социальных исследований» ставит своей задачей помочь студентам познакомиться с тем, как проводятся исследования в области общественных наук, и разобраться в различных проблемах методологии исследований в данной области. Книги этой серии адресованы студентам, аспирантам и начинающим исследователям, в программы обучения которых входят разделы, посвященные методам исследований в науках о человеке и обществе, будущим социологам, исследователям социальной политики, психологам, культурологам, демографам, политологам, криминалистам, а также тем, кто специализируется в области социального взаимодействия, организационных исследований и т. п. Эти книги будут полезны при проведении исследований в рамках курсовых работ, дипломных проектов, диссертаций.

Книги серии помогут читателям «постичь» проблемы и методы исследований в области общественных наук, что означает развитие умения ценить те радости и огорчения, с которыми связано любое исследование в этой области, выработать навыки использования определенных методов обработки данных и знаний основных проблемных моментов в обсуждаемых областях. Относительный акцент на том или ином из этих аспектов меняется от книги к книге, но цель каждой из них — осветить конкретный метод или проблему с позиций практического исследователя, а не представить просто сборник «пошаговых» инструкций. Чтобы достичь этой цели, серия включает освещение основных методов социальных исследований и рассмотрение большого числа проблем и спорных вопросов. Каждая книга серии написана практическим исследователем, имеющим опыт использования рассматриваемых методов или решения обсуждаемых проблем и вопросов. Таким образом, авторы опираются на свои практические знания и личный опыт.

Несмотря на то что существует множество книг, посвященных основам статистического анализа данных, сравнительно небольшое их количество в доступной форме рассматривает более изысканные виды и аспекты анализа. Именно это удалось сделать Дункану Крамеру в предлагаемой вниманию читателя книге. Он опирается на опыт чтения лекций и проведения семинаров, а также

написания книг, освещающих основные методы статистического анализа данных. Его подход состоит в том, чтобы на тщательно разобранных примерах познакомить начинающего студента или исследователя с рассматриваемыми методами. Более того, он связывает изложение статистических методов с объяснением того, какие шаги должны предпринять читатели, чтобы реализовать эти методы, используя компьютерные программы. Насколько это возможно, Д. Крамер уделяет внимание применению SPSS — наиболее широко используемого психологами пакета программ для статистического анализа. Кроме того, проведен разбор большей части результатов, выдаваемых рассматриваемыми в книге компьютерными программами, и указаны моменты, на которые необходимо обращать внимание и как интерпретировать полученные результаты.

Дункан Крамер подробно знакомит читателя с такими широко используемыми в западной экспериментальной психологии методами, как множественная регрессия, логлинейный анализ, логистическая регрессия и дисперсионный анализ. Знание этих методов принципиально важно, если исследователь хочет выйти за рамки простых манипуляций с данными и традиционной описательной статистики. Более того, изучение тех или иных методов анализа требует также знаний того, как их применять на практике, а в наше время это во многом сводится к умению использовать соответствующее программное обеспечение для реализации представленных методов. Именно в этом особая ценность книги Д.Крамера, уделяющего большое внимание изучению компьютерных программ. В то же время исследователь должен знать, как интерпретировать полученные в ходе статистической обработки данных результаты, и в этом отношении книга окажет неоценимую помощь в объяснении того, как правильно читать и понимать файлы, содержащие результаты работы компьютерных программ.

Эта книга очень своевременна, учитывая тот факт, что студентов всячески поощряют приобретать навыки, которые можно передать другим, а высшие учебные заведения, в свою очередь, призваны формировать такие умения у студентов. Знание того, как осуществлять более сложные и тонкие виды статистического анализа данных и как использовать программное обеспечение в связи с таким анализом, является важным компонентом внедрения подобных навыков в практику. В этом качестве данная книга окажет неоценимую помощь студентам, исследователям и преподавателям высших учебных заведений.

Алан Брайман

ПРЕДИСЛОВИЕ

Цель этой книги состоит в том, чтобы служить введением в некоторые из основных статистических методов, которые, однако, нельзя назвать абсолютно общеизвестными и повсеместно используемыми в социальных науках и психологии для анализа количественных данных, и сделать это насколько возможно проще и доступнее, без привлечения технически сложного математического аппарата. Развитие компьютерных программ количественного анализа данных, относительно простых в применении, привело не только к широкому распространению этих методов, но и к стремлению, когда это возможно, использовать данные методы для анализа собственных результатов. Чтобы понимать результаты количественного анализа, представленные в публикуемых статьях, и быть в состоянии критически оценить их, необходимо знать сами методы количественного анализа, ибо авторы статей обычно предполагают такое понимание со стороны читателя само собой разумеющимся, описывая только результаты своей работы. Несмотря на то что использование количественных методов имеет достаточно долгую историю, книг, в которых эти методы излагаются относительно просто и доступно для не очень подготовленного с точки зрения математики читателя, мало. Автор выражает надежду, что эта книга поможет восполнить данный пробел.

В книге показано конкретное использование определенных методов статистического анализа. Каждая глава книги начинается с общего описания назначения того или иного метода, а затем приводится простой пример-иллюстрация с небольшим набором данных. В целом используется восемь различных наборов данных. Они не велики по объему, включают от 9 до 15 наблюдений и от двух до девяти переменных. Поэтому с ними сравнительно легко работать. Вначале рассматриваются наиболее важные для понимания разбираемого метода аспекты статистики. Хотя в данной книге анализируются более изысканные статистические методы, все статистические термины объясняются. Автор надеется, что это поможет читателям, недостаточно хорошо знакомым со статистикой.

Чтобы не перегружать читателя запоминанием множества символов, вместо них использованы термины. По возможности, про-

водятся необходимые статистические расчеты, чтобы показать, как получены численные величины, а сами вычисления, чтобы отделить их от основного текста, размещены в таблицах. Также, где только возможно, приводится краткое изложение результатов каждого рассматриваемого примера. Разные издательства требуют различного стиля изложения полученных статистических результатов. В данной книге все примеры изложены в одном стиле. Но изменить их в соответствии с требованиями редактора не составляет труда. В конце каждой главы приведена рекомендуемая литература, включающая наименее технически сложные источники, и тем не менее их понимание требует более высокого уровня подготовки, чем тот, на который ориентирована данная книга.

Несмотря на то что более глубокое понимание статистического метода часто достигается путем проведения некоторых сопутствующих расчетов, мы не предполагаем и даже не рекомендуем читателям анализировать данные, непосредственно проводя такие вычисления. Эти вычисления эффективнее и приятнее выполнить, используя статистические компьютерные программы. В настоящее время в наличии имеется несколько различных широко распространенных программных продуктов. Автор использовал для выполнения вычислений в рассматриваемых примерах везде, где только возможно, пакет SPSS, полагая именно его наиболее популярным и доступным программным средством. Последней версией, выпущенной к моменту завершения рукописи, была версия пакета с номером 11. Эта версия подобна трем предыдущим версиям программы. Таким образом, данная часть книги также доступна тем, кто использует более ранние версии пакета. Пакет SPSS не содержит модуля структурного моделирования[1]. Для выполнения этих расчетов был выбран LISREL, поскольку он является одной из наиболее широко используемых программ для подобных вычислений[2]. Применялась последняя версия этой программы, которой на тот момент являлась LISREL 8.51. Рассматриваемый в книге пример может быть выполнен с помощью свободно распространяемой студенческой или ограниченной версий этой программы, которые можно загрузить со следующего вебсайта: http://www.ssicentral.com/other/entry.htm. Инструкции о том, как загрузить данные программы, доступны на этом сайте. Чтобы обозначить тот факт, что используемые термины и величины относятся к пакетам SPSS и LISREL или являются частью их выходных файлов, они выделены в тексте жирным шрифтом. Автор старался

---

[1] Здесь автор не прав, ибо специальный модуль в рамках SPSS существует и называется AMOS (здесь и далее примечания научного редактора).

[2] На наш взгляд, использовать LISREL достаточно сложно, с этой точки зрения заслуживает внимания программа EQS с более дружественным интерфейсом.

по возможности сократить описания программ, с тем чтобы книга в максимальной степени была полезна читателям, которые пользуются другим программным обеспечением.

Данные для разбираемых примеров были подобраны таким образом, чтобы проиллюстрировать важные статистические моменты, и не претендуют на получение результатов, типичных для исследовательской литературы по данной теме. При составлении примеров мы также старались упростить их. Если эти примеры покажутся вам недостаточно интересными или содержательными, то можно по аналогии создать свои собственные. Самостоятельное построение и анализ примеров полезны для проверки общего понимания рассматриваемых методов. Это понимание можно углубить, изучая опубликованные отчеты с примерами подобного анализа в своей области исследований. Анализ количественных данных — умение, которое только выигрывает от должным образом осмысленной практики. Автор надеется, что данная книга поможет читателю развить этот навык.

Хотелось бы выразить свою благодарность Алану Брайману, Тиму Ляо и Аманде Сакер за их отзывы и комментарии к первому проекту рукописи.

Дункан Крамер,
Университет Лохборо

ВВЕДЕНИЕ[1]

Одна из главных задач социальных наук и психологии состоит в объяснении различных аспектов человеческого поведения. Например, нас может интересовать объяснение того, почему одни люди более агрессивны, чем другие. Один из способов определения адекватности или действительности объяснений состоит в том, чтобы собрать данные, характеризующие изучаемые характеристики людей, и найти, до какой степени эти данные совместимы с предлагаемыми объяснениями. Данные, согласующиеся с объяснением, поддерживают его в той степени, в какой они противоречат другим объяснениям. Данные, противоречащие объяснению, являются доказательными лишь настолько, насколько корректно операционализированы признаки, имеющие отношение к рассматриваемому явлению. Операционализация включает управление признаками или их измерение.

Качественные и количественные переменные

Как следует из названия, переменная должна представлять собой изменяющуюся особенность или характеристику изучаемого объекта или явления. Если характеристика не изменяется, ее называют *постоянной*. Для психологии и общественных наук представляет интерес объяснение того, почему изменяются те или иные признаки. Самый простой тип переменной — *бинарный*, в котором данное качество либо присутствует, либо отсутствует. Такая характеристика, как пол, представляет собой бинарную переменную, значения которой в каждом конкретном случае определяются полом объекта — женским или неженским (мужским). Аналогично, бинарной является переменная «быть в разводе». Обе категории, составляющие бинарную переменную, могут быть представлены (закодированы) любой парой чисел, таких, как 1 и 2

[1] Во введении уделяется внимание переменным и измерениям, с помощью которых исследователи получают данные для своего анализа. Специфика того или иного способа сбора данных определяет, какие математические и статистические действия с ними можно выполнять.

или 23 и 71. Например, «быть женщиной» может быть закодировано как 1, а «быть мужчиной» как 2.

Различают два типа переменных. Переменные первого типа называют по-разному: качественными, категориальными, номинальными, или частотными. Примером качественной переменной может служить семейное положение, которое включает пять категорий: (1) холост/не замужем; (2) женат/замужем; (3) проживает отдельно; (4) разведен(а); (5) вдовец/вдова. Эти пять категорий могут быть представлены или закодированы любым набором из пяти чисел: 1, 2, 3, 4 и 5 или 32, 12, 15, 25 и 31. Данные числа просто используются для обозначения различных категорий. Можно подсчитать только число, или частоту, объектов, относящихся к каждой из данных категорий, поэтому такие переменные часто называют *частотными переменными*. Например, из 100 человек 30, возможно, никогда не были женаты, 40 могут состоять в браке, 8 — проживать раздельно, 12 — быть в разводе и у 10 человек супруги скончались. Частота случаев в категории может быть выражена как доля или процент от полной частоты случаев. Так, доля людей, состоящих в браке, равна 0,40 (40/100 = 0,40). Эта же величина, выраженная в процентах, равна 40 (40/100 × 100 % = 40 %). Данные, состоящие из качественных переменных, являются количественными в том смысле, что частота, доля или процент случаев могут быть определены количественно. Категории качественной переменной могут рассматриваться как бинарные переменные. Например, разведенные могут представлять одну категорию бинарной переменной, а оставшиеся четыре группы из никогда не состоявших в браке, состоящих в браке, живущих раздельно и вдовых — другую категорию. К качественным переменным можно отнести также вид потребляемой пищи, страну происхождения, характер заболевания и метод получаемого лечения.

Другой тип переменных называют *количественными переменными*. Числовые значения в этом случае используются, чтобы отобразить и (или) упорядочить уровни увеличения значений этих переменных. Самый простейший пример количественной переменной — бинарная переменная, такая как пол, где одна из категорий интерпретируется как представляющая большее количество данного качества по сравнению с другой. Например, если женщины закодированы как 1, а мужчины как 2, то эта переменная может рассматриваться как отражение маскулинности, и большее значение указывает на бóльшую выраженность данного качества. Следующий простой пример: переменная социального класса, которая включает три уровня значений — высший, средний и низший. Высшее сословие может быть закодировано как 1, средний класс как 2 и низший класс как 3; в данном случае меньшие значения соответствуют более

высокому социальному статусу. Эти числа можно рассматривать как величины, принадлежащие шкале отношений. Значение 1 соответствует в два раза более высокому рангу по сравнению с 2, что дает отношение 1 к 2[1]. Как правило, количественные переменные включают более трех категорий, к их числу относятся возраст, доход или суммарный балл по шкале какого-нибудь опросника. Например, для оценки агрессивности человека может использоваться опросник, в который входят десять вопросов. Для каждого вопроса предусмотрено два варианта ответа «Да» или «Нет». Значение 1 можно присвоить ответам, которые указывают на агрессивность, в то время как значение 0 можно приписать ответам, которые показывают отсутствие агрессивности. Сложив вместе баллы для каждого из десяти вопросов, получим суммарный балл, изменяющийся в пределах от 0 (минимальное значение) до 10 баллов (максимальное значение). Более высокие значения будут означать бóльшую агрессивность. Правомерно считать, что эти числа представляют собой шкалу отношений. Испытуемый, набравший 10 баллов, имеет в два раза больший балл по сравнению с тем, кто набрал 5 баллов по данной шкале (10 : 5 = 2 : 1).

При необходимости всегда можно объединить смежные категории, чтобы сформировать меньшее число групп значений, однако количество вновь образованных категорий не должно быть меньше двух. Например, 11 категорий только что упомянутого полного набора значений агрессивности могут быть повторно сгруппированы в три новые категории, объединяющие значения от 0 до 1, от 2 до 5 и от 6 до 10. Вновь образованные категории не обязаны включать равное количество значений, и именно такая ситуация имеет место в нашем примере, где первая категория состоит из двух значений (0 и 1), вторая включает четыре значения (2, 3, 4 и 5), а третья объединяет пять значений (6, 7, 8, 9 и 10). Эти три новые категории будут теперь иметь новые числовые обозначения: 1 — для первой группы, 2 — для второй группы и 3 — для третьей группы. Однако такая перегруппировка должна быть обоснована. В рассмотренном случае смысл новых категорий менее ясен, чем исходных, а диапазон значений меньше.

В социальных науках и психологии нас обычно интересует вопрос: связана ли переменная с одной или несколькими другими переменными? Чем сильнее зависимость между переменными, тем больше между ними общего. Двумерный анализ исследует отношения между двумя переменными, в то время как много-

[1] На наш взгляд, это не совсем верно, ибо крайне сложно определить, как один статус может быть в ДВА раза выше, чем другой. Система показателей оценки таких характеристик в социальных науках является приближенной.

мерный[1] анализ исследует отношения между тремя или более переменными одновременно. В книге рассматривается только один случай анализа связи между двумя переменными — однофакторный дисперсионный анализ (см. гл. 9). Все другие примеры касаются трех или более переменных одновременно. В однофакторном дисперсионном анализе рассматриваются отношения между качественной переменной «семейное положение» и количественной переменной «уровень выраженности депрессии», измеряемой в баллах по специальной шкале. Однофакторный анализ ковариаций исследует эту зависимость, контролируя уровень значений второй количественной переменной в зависимости от значения первой. Другими словами, он включает три переменные, одна из которых качественная (семейное положение), а две другие — выраженность депрессии и взаимосвязь между семейным положением и депрессией — количественные. Следовательно, этот тип анализа также является многопеременным. Любой аспект поведения человека находится под влиянием нескольких различных факторов. Поэтому анализ, учитывающий их одновременно, будет способствовать лучшему пониманию исследуемых феноменов.

Статистический вывод

Очень часто перед исследователем возникает проблема: как определить, можно ли факт определенной взаимосвязи переменных, обнаруженный на данных, полученных из выборки, распространить на всю генеральную совокупность, из которой осуществлялась выборка. В этом контексте понятие генеральной совокупности относится к значениям переменных, а не к людям или другим организмам. Выборка представляет собой подмножество таких значений. Чтобы установить, можно ли факт, полученный на выборке, распространить на генеральную совокупность, необходимо вычислить вероятность его обнаружения в ситуации, обусловленной случайными событиями. Если вероятность такого обнаружения $\leq 0{,}05$ (1 шанс на 20 или меньше), то можно считать, что данный факт закономерен: присутствует в генеральной сово-

[1] В английском языке используется термин «multivariate», который по сложившейся традиции переводится термином «многомерный», хотя, на наш взгляд, лучше использовать именно термин «многопеременный», оставив термин «мерность» для обозначения размерности структуры анализируемых данных. Плоская таблица является двумерной независимо от того, сколько переменных она содержит, а данные, имеющие структуру куба (в случае, когда каждый респондент заполняет двумерную матрицу, или лонгитюдные наблюдения, требующие от одних и тех же респондентов неоднократного ответа по нескольким пунктам фиксированного опросника), являются трехмерными.

купности, из которой осуществлялась выборка. Другими словами, данная взаимосвязь вряд ли случайна. Такая взаимосвязь называется *статистически значимой*. Однако возможно, что мы как раз столкнулись с редким случаем (вероятность его наступления < 0,05) а именно, связь есть, но она обусловлена случайными событиями. В этой ситуации вывод о существовании закономерности связи будет сделан, в то время как на самом деле ее нет. Такая ошибка называется *ошибкой первого рода*[1]. Хотя за общепринятый уровень статистической значимости принимается величина 0,05 или менее, следует помнить о произвольности подобного соглашения.

Если вероятность установления связи в ситуации, обусловленной случайными причинами, превосходит 0,05, считаем, что на генеральной совокупности, из которой осуществлялась выборка, исследуемой закономерности нет. Такая взаимосвязь называется *статистически незначимой*. Однако возможно, что исследуемая взаимосвязь, вероятность случайного обнаружения которой превышает 0,05, носит все же неслучайный характер и на самом деле закономерна. В этом случае, отвергая имеющуюся в действительности закономерность на основании того, что она с достаточно большой вероятностью (более 0,05) может быть обусловлена случайными событиями, мы также совершаем ошибку. Такая ошибка известна как *ошибка второго рода*[2].

Заметим, что чем больше объем выборки, тем с большей вероятностью можно утверждать, что установленная связь обусловлена неслучайными событиями. Другими словами, вероятность совершить ошибку первого рода (утверждать существование взаимосвязи, когда на самом деле ее нет) увеличивается с ростом выборки. Вероятность обнаружения статистически незначимой взаимосвязи тем больше, чем меньше объем выборки. Другими словами, вероятность совершить ошибку второго рода (утверждать отсутствие взаимосвязи, когда на самом деле она существует) увеличивается с уменьшением выборки. Более сильные взаимосвязи имеют бóльшую вероятность оказаться статистически значимыми. Следовательно, при интерпретации статистической значимости необходимо принять во внимание и величину взаимосвязи, и объем

---

[1] Говорить, что ошибка первого рода — это неправильно сделанный вывод о том, что исследуемая связь закономерна, в то время как она обусловлена случайными событиями, было бы не совсем точно. В ситуации статистического вывода проверяется так называемая гипотеза H_0, которая может утверждать закономерность существования связи между какими-то данными или, наоборот, закономерность отсутствия этой связи и т.д. В дополнение к гипотезе H_0 формулируется альтернативная гипотеза H_1, заключающаяся в отрицании H_0. В терминах гипотез ошибка первого рода — ошибочное отвергание нулевой гипотезы (H_0), когда она на самом деле верна.

[2] В терминах гипотез ошибка второго рода означает принятие нулевой гипотезы в то время, когда она ошибочна.

выборки. Из всех статистических методов, описанных в этой книге, процедура вычисления статистической значимости отсутствует только в двух случаях: кластерном и эксплораторном факторном анализе[1].

Зависимые и независимые переменные

Для многих из описываемых статистических методов необходимо различать зависимые и независимые переменные. В некоторых случаях — множественной регрессии, дисперсионном факторном анализе — зависимую переменную можно назвать откликом, или переменной-откликом, а независимую переменную — предиктором, или переменной-предиктором. Зависимая переменная, или переменная-отклик, — переменная, поведение которой мы стремимся объяснить в терминах независимых переменных, или переменных-предикторов. Зависимая переменная называется так, потому что мы предполагаем, что она находится под влиянием, или зависит от независимых переменных. Предполагается, что независимые переменные не испытывают влияния, т.е. независимы, от других переменных.

В анализе путей (иногда мы будем употреблять синонимический термин «путевой анализ»), обсуждаемом в гл. 6 и 7, будет рассмотрена такая последовательность переменных, когда первая переменная влияет на вторую, вторая — на третью и т.д. Переменную, с которой начинается последовательность, иногда называют *внешней*, или экзогенной[2], переменной, потому что она является внешней по отношению к модели, рассматриваемой в путевом анализе. Переменные следующих за внешними переменными уровней называют *внутренними*, или эндогенными[3], переменными, потому что модель путевого анализа должна их объяснить. В то время как экзогенная переменная является независимой переменной, эндогенные переменные выступают в роли как независимых, так и зависимых переменных. Они зависят от предшествующей и влияют на следующую за ними переменную. Также возможно, что две или более переменных влияют друг на друга. В этом случае их зависимость называют взаимной. Данная тема не освещается в настоящей книге.

Следует ясно понимать, что, используя статистический анализ, можно только установить, связаны ли переменные друг с другом, но нельзя определить, влияет ли одна переменная на

---

[1] Собственно эта ситуация верна не только для выборки примеров, включенных в книгу, но и отражает общую картину статистических методов.

[2] От англ. exogenous — внешний.

[3] От англ. endogenous — внутренний.

другую. Например, можно найти зависимость между уровнем безработицы и преступности, при которой те, кто не имеет оплачиваемой работы, с большей вероятностью вовлечены в преступную деятельность или осуждены за нее. Это соотношение, однако, не означает, что безработица ведет к преступной деятельности. Одинаково правдоподобно и то, что те, кто вовлечен в преступную деятельность, меньше интересуются поиском оплачиваемой работы. Также возможно, что обе переменные влияют друг на друга. Определение причинной природы или характера (направления) зависимости двух переменных в социальных науках и психологии обычно представляет собой сложную проблему, которая предусматривает необходимость веских оснований для хорошо аргументированной точки зрения. Роль статистического анализа для выработки подобной аргументации состоит в том, чтобы предложить показатель величины любой наблюдаемой взаимосвязи и дать оценку вероятности случайного появления такой зависимости.

Выбор адекватного статистического метода

Здесь будет дан краткий обзор статистических методов, рассматриваемых в настоящей книге, чтобы помочь читателям, которые не очень хорошо представляют, какой из методов наиболее адекватен для анализа их данных, и, в первую очередь, заинтересованы в том, чтобы получить информацию только о таком методе. Методы в книге были упорядочены таким образом, чтобы стало ясным, как анализировать потенциально большой набор количественных переменных и выделять статистические идеи, используемые в дальнейшем. Например, множественная регрессия рассматривается перед дисперсионным анализом потому, что в ходе ее выполнения решаются в том числе и задачи анализа вариаций (дисперсий), т.е. дисперсионного анализа. При выборе оптимального метода анализа тех или иных данных необходимо учитывать тип переменных, к которым этот метод может быть применен.

В настоящей книге представлен только один метод, позволяющий одновременно рассматривать три или более качественные переменные. Это — логлинейный анализ, описанный в гл. 13. Например, нас может интересовать соотношение между психическим расстройством (указанным по классификации тревога, депрессия, тревога и депрессия одновременно), религиозной ориентацией (отсутствует, протестант, католик) и детским семейным статусом (жил с обоими родителями, только с матерью, только с отцом). Логлинейный анализ используется, чтобы ответить на два связанных между собой типа вопросов. Первый вопрос: отличается

ли статистически значимо частота интересующих исследователя наблюдений от случайной, в терминах взаимодействия между тремя или более качественными переменными? Другими словами, должны ли мы рассматривать более двух переменных для объяснения распределения интересующих нас наблюдений в зависимости от этих переменных? Второй вопрос: какие из переменных и/или их взаимодействия необходимы для объяснения распределения наблюдений? Этот вопрос отличается от первого тем, что в данном случае наряду с влиянием отдельных переменных и их двусторонних взаимодействий с другими переменными рассматриваются и взаимодействия более высокого порядка.

Если одну из качественных переменных необходимо рассматривать как зависимую (например, классифицированные психические расстройства), а другие качественные переменные как независимые, то логистическая регрессия является наиболее адекватным методом, поскольку она рассматривает только взаимосвязи между зависимой переменной и независимыми переменными. Другими словами, она исключает из рассмотрения взаимосвязи независимых переменных между собой (например, отношения между религиозной ориентацией и детским семейным статусом). Единственным типом логистической регрессии, рассматриваемым в настоящей книге, является бинарная логистическая регрессия, где зависимая переменная состоит из двух категорий; она описана в гл. 8. Множественная логистическая, или логит-регрессия, используется, чтобы определить, какие качественные и количественные переменные и их взаимодействия наиболее сильно связаны с вероятностью осуществления определенной категории зависимой переменной, при этом учитываются их связи с другими независимыми переменными, участвующими в анализе. Качественные переменные должны быть преобразованы в фиктивные двоичные переменные. Эта процедура описана в гл. 9—11 для дисперсионного анализа.

Существует три метода введения предикторов в логистическую регрессию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных могут играть незначительную роль в максимизации вероятности реализации категории. В иерархическом, или последовательном, методе независимые переменные вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад в максимизацию вероятности реализации рассматриваемой категории они вносят. Например, демографические переменные, такие, как возраст, пол, социальный статус, могут вводиться в первую очередь для того, чтобы их влияние можно было зафиксировать на следующем этапе. В статистическом, или пошаговом, методе среди предикторов выбираются те переменные, которые вносят наибольший вклад в максимизацию веро-

ятности реализации категории. Если два предиктора связаны друг с другом и имеют очень близкие вклады в максимизацию вероятности осуществления рассматриваемой категории, то будет выбран предиктор с большей величиной вклада в максимизацию категории, даже если различие в величине коэффициентов максимизации у этих двух предикторов минимально.

Дискриминантный анализ, описанный в гл. 12, может использоваться для определения той количественной переменной, которая лучше всего предсказывает, в какую категорию попадет объект, при условии, что данные отвечают следующим требованиям. Число объектов в категориях зависимой переменной не должно сильно различаться. Независимые переменные должны иметь нормальное распределение, а внутригрупповые дисперсии — быть одинаковыми[1]. Независимые переменные порождают новую составную переменную, называемую дискриминантной функцией. Максимальное число рассматриваемых дискриминантных функций не должно превышать число предикторов, с одной стороны, и быть, как минимум, на единицу меньше числа групп — с другой. Как и в случае с логистической регрессией, существует три метода включения предикторов в дискриминантную функцию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных, возможно, и не позволяют выявить различия между группами. В иерархическом, или последовательном, методе предикторы группами вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад в дискриминацию (различение между группами) они вносят. В статистическом, или пошаговом, методе среди предикторов по одному выбираются те переменные, которые вносят наибольший вклад в дискриминантную функцию. Если два предиктора связаны друг с другом и их вклады в дискриминантную функцию почти равны, то будет выбран предиктор, у которого этот вклад больше, даже если различие минимально.

Другие статистические методы, описанные в этой книге, предназначены для тех случаев, когда либо зависимая переменная, либо все переменные вместе являются количественными. Множественная регрессия используется, чтобы определить, какие количественные и качественные независимые переменные и их взаимодействия наиболее сильно связаны с количественной переменной-откликом. Качественные переменные необходимо рассматривать как фиктивные переменные, как это описано в гл. 9—11 для дисперсионного анализа. Так же, как и в случае с логистической регрессией и дискриминантным анализом, существует три метода

[1] Это предположение называется предположением об однородности дисперсий (или их гомогенности — от англ. homogeneous).

включения предикторов в множественную регрессию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных могут и не быть связаны с переменной-откликом. В статистическом, или пошаговом, методе в качестве предикторов выбираются те независимые переменные, которые вносят наибольший вклад в дисперсию зависимой переменной. В том случае, когда два предиктора связаны (коррелируют) друг с другом и имеют очень близкие коэффициенты связи с переменной-откликом, выбирают предиктор, имеющий больший коэффициент связи с зависимой переменной, даже если различие между соответствующими сравниваемыми коэффициентами минимально. Этот метод описан в гл. 4. В иерархическом, или последовательном, методе независимые переменные вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад они вносят. Этот метод описан в гл. 5. Иерархическая множественная регрессия также используется в основном варианте путевого анализа, в дисперсионном и ковариационном анализе.

Анализ путей применяют, чтобы определить силу связи в гипотетической последовательности, или ряду, количественных эндогенных (экзогенных) переменных и ту степень, в которой выбранные пути обеспечивают удовлетворительное описание или величину критерия согласия между всеми переменными. Качественные переменные могут быть включены в анализ как экзогенные переменные, когда они преобразованы в фиктивные двоичные переменные, как это описано в гл. 9—11. В самом простом случае рассмотрения трех переменных анализ путей может использоваться, чтобы оценить, в какой степени одна из переменных является прямой функцией двух других и косвенной функцией одной из них. Самая простая форма анализа путей рассматривается в гл. 6. Более сложная форма анализа путей, учитывающая надежность переменных, описана в гл. 7.

Дисперсионный анализ используется, чтобы определить, связаны ли значимо одна или несколько качественных независимых переменных и их взаимодействия с количественным откликом или зависимой переменной. Если качественная переменная состоит только из двух категорий (является бинарной), значимая связь означает, что средние значения откликов в двух группах, принадлежность к которым определяется по значению качественной переменной, значимо различаются. Если качественная переменная включает более двух категорий, значимая связь подразумевает, что средние значения откликов двух или более групп (соответствующих различным категориям независимой качественной переменной) значимо различаются. Если имеются веские основания для того, чтобы прогнозировать, какие из этих средних различаются, значимость этих различий может быть опреде-

лена с использованием одностороннего критерия Стьюдента. Если никаких различий не прогнозировалось или не было веских причин предполагать какие-либо различия, то значимость различий необходимо анализировать с помощью апостеорных критериев (post-hoc, т.е. возникающих на основании работы с данными, а не в результате выдвинутых предварительных теоретических соображений, например критерий Шеффе). Дисперсионный анализ с одной качественной переменной описан в гл. 9, а дисперсионный анализ с двумя качественными переменными — в гл. 11. Ковариационный анализ позволяет фиксировать влияние независимых количественных переменных, связанных с количественной переменной-откликом. В гл. 10 рассматривается анализ ковариаций, в котором участвуют две независимые переменные (одна качественная и одна количественная) и одна зависимая количественная переменная.

Наконец, статистические методы, описываемые в гл. 1—3, позволяют определить возможности группировки связанных количественных переменных в меньшее число объемлющих их факторов или кластеров. Например, нас может интересовать, можно ли сгруппировать пункты опросника, измеряющие тревогу и депрессию, соответственно, в два фактора или кластера, представляющие эти два типа вопросов. Разновидность эксплораторного[1] (разведочного) факторного анализа, называемая методом главных компонент, описана в гл. 1. Такой анализ является разведочным в том смысле, что способ возможной группировки переменных не предопределен заранее, как это имеет место в случае конфирматорного[2] (подтверждающего) факторного анализа, описанного в гл. 2. Конфирматорный факторный анализ даст статистическую меру для определения того, насколько удовлетворительно заранее определенная структура группировки наблюдаемых признаков объясняет реально существующие связи (вычисленные по экспериментальным данным) между ними. Другим методом группировки переменных является кластерный анализ. Разновидность кластерного анализа, называемая иерархической агломеративной кластеризацией, описана в гл. 3.

Следует отметить, что одни методы, описанные в этой книге, встречаются в литературе по социальным наукам и психологии чаще, другие — реже. К наименее популярным методам относятся кластерный анализ, логлинейный анализ и дискриминантный анализ. Следовательно, читатели с меньшей вероятностью встретятся с этими методами при чтении публикаций с использованием статистического анализа данных. Поэтому, чтобы лучше познакомиться с более редкими способами примене-

[1] От англ. explore — исследовать.

[2] От англ. confirm — подтверждать.

ния данных методов, имеет смысл попытаться найти примеры использования этих методов, осуществляя специальный поиск в электронных библиографических базах данных, относящихся к области ваших собственных научных интересов[1].

Дополнительная литература, рекомендуемая научным редактором

Кричевец А. Н. Математика для психологов / А. Н. Кричевец, Е. В. Шикин, А. Г. Дьячков. — М.: Флинта: Московский психолого-социальный институт, 2003.

Сидоренко Е. В. Методы математической обработки данных в психологии. — СПб. : Речь, 2003.

Гусев А. Н. Измерение в психологии / А. Н. Гусев, Ч. А. Измайлов, М. Б. Михалевская. — М.: изд-во «Смысл», 2000.

Тюрин Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. — М.: Инфра-М, 2003.

[1] На наш взгляд, кластерный анализ в отечественной литературе встречается несравнимо чаще. Пример дискриминантного анализа см.: О. В. Митина, В. Ф. Петренко, 2002, а вот примеры использования логлинейного анализа нам не удалось найти вообще.

ЧАСТЬ I
ГРУППИРОВКА КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

Глава 1
ЭКСПЛОРАТОРНЫЙ (РАЗВЕДОЧНЫЙ) ФАКТОРНЫЙ АНАЛИЗ

Предисловие научного редактора

Статистические методы, описываемые в ч. I (гл. 1—3), позволяют определить возможности группировки связанных количественных переменных в меньшее число объемлющих их факторов или кластеров.

В гл. 1 описывается метод главных компонент. Этот метод часто рассматривают как одну из наиболее распространенных (установленных в большинстве компьютерных программ по умолчанию) разновидностей эксплораторного (исследовательского, разведочного) факторного анализа. Хотя с математической точки зрения эти два метода существенно отличаются друг от друга: в задачу первого входит объединение исходных признаков (шкал) в минимально возможное число классов с сохранением максимально возможной информации, задаваемой этими переменными, а для второго целью является максимально приближенное воспроизведение матрицы взаимосвязей между переменными, с точки зрения содержательной (для предметной интерпретации) большой роли это не играет — оба метода используются для группировки первичных переменных с установлением весовых коэффициентов, определяющих степень включенности каждой из них в ту или иную группу. Выделить главные компоненты проще (меньше требований к эмпирическим соотношениям, определяемым первичными переменными). Решение, получаемое в результате выполнения именно факторного анализа, надежнее и устойчивее, однако условия, накладываемые на переменные, более жесткие. Поэтому для построения эвристических моделей чаще всего используют главные компоненты, в то время как для построения опросников — инструментов, связанных с практическим применением для получения психодиагностических показателей, предпочтительнее использовать более надежный и математически более корректный факторный анализ.

Факторный анализ представляет собой совокупность методов, призванных определить, насколько связанные (коррелирующие) переменные могут быть сгруппированы так, чтобы

каждую группу можно было рассматривать как одну составную переменную, или фактор, а не как ряд отдельных переменных. Возможно, наиболее распространенное применение факторного анализа в социальных науках и психологии состоит в том, чтобы определить, можно ли объединить совокупность пунктов опросника, оценивающих конкретную характеристику, таким образом, чтобы получить общий индикатор данной характеристики.

Например, нас может интересовать оценка восприятия людьми собственной тревожности. Мы могли бы просто спросить некоторых людей, насколько тревожными они обычно бывают. Однако существуют три главные проблемы при попытке измерить социальную характеристику, индивидное качество, личностную черту и т.д. с помощью единственного вопроса или пункта опросника. Во-первых, потенциальная чувствительность такого показателя будет существенно ограничена. Например, если мы ограничим возможные ответы на задаваемый вопрос только вариантами «Да» или «Нет», то сможем распределить наших респондентов лишь по двум категориям. Чем больше вопросов о тревожности мы зададим, тем больше возможных категорий респондентов в зависимости от данных ими ответов на все вопросы в совокупности можно получить. Так, на основании ответов на два вопроса можно составить четыре категории (вариантов ответов)[1], три вопроса — шесть категорий и т.д.

Вторая проблема с измерением, основанным на ответе на единственный вопрос, состоит в том, что в этой ситуации невозможно определить, насколько надежен показатель. Так, может оказаться, что люди, которых мы спрашиваем, не знают, что значит «испытывать тревожность», и отвечают «Да» или «Нет» в большей степени случайным образом без четкого понимания смысла вопроса. Чтобы определить надежность этого вопроса, можно задать его два или более раз в рамках одного и того же интервью.

Если бы данный вопрос являлся надежным показателем, мы могли бы предполагать, что испытуемые каждый раз дадут на него один и тот же ответ. Поскольку попытка задать один и тот же вопрос несколько раз в течение короткого промежутка времени может быть воспринята как факт недостаточной организованности интервьюеров или их недоверия к интервьюируемым, предпочтительнее задавать вопросы, которые различаются по форме, но сохраняют свою содержательную направленность. Например, можно спросить респондентов: напряжены ли они обычно?

---

[1] Для двух вопросов варианты ответов могут быть следующие: Да-Да, Нет-Да, Да-Нет, Нет-Нет. Читателю может быть полезно выписать все возможные варианты ответов на три вопроса.

Третья проблема, связанная с измерением, построенном на одном пункте опросника, состоит в том, что оно не позволяет исследовать различные аспекты данной характеристики. Например, «тревожиться», «быть в напряжении», «нервничать» или «легко пугаться» может описывать несколько различные аспекты состояния, которое мы называем тревогой. Если это действительно различные проявления тревоги, то можно ожидать, что те, кто описал себя как тревожащихся, также опишут себя как испытывающих напряжение, нервничающих или пугливых. Аналогично, можно предполагать, что те, кто отвечал, что не испытывает тревогу, также скажут, что не испытывают напряжение, мало нервничают и не боятся. Другими словами, мы предполагаем, что ответы на эти четыре вопроса связаны друг с другом и образуют единый фактор. Если дело обстоит именно таким образом, мы могли бы объединить вместе ответы на эти четыре вопроса и создать единый суммарный показатель, а не рассматривать их как четыре различных отдельных показателя, дающих сходную информацию.

Однако интерпретация результатов факторного анализа, опирающаяся только на соображения о том, что какие-то пункты вопросника образуют единый фактор, который можно было бы назвать тревожностью, проблематична. Если, например, обнаружено, что ответы на четыре пункта вопросника, связанных с тревогой, группируются и образуют единый фактор, то без дополнительной информации нельзя узнать, в самом ли деле этот фактор специфичен для тревоги и отражает ее уровень или представляет более общий фактор, который можно было бы назвать склонностью к жалобам. Или же, если эти четыре пункта вопросника объединяются в два отдельных фактора, например, в фактор тревоги (напряженности), с одной стороны, и нервозности (пугливости) — с другой, нельзя узнать, расщепился ли общий фактор тревожности на два более специфичных подфактора. Следовательно, проводя факторный анализ, полезно включать ответы на те пункты, которые, как предполагается, не относятся к исследуемой (проверяемой) характеристике (состоянию). Например, если мы считаем, что депрессия и тревога представляют собой отдельные состояния, то, чтобы лучше интерпретировать результаты, можно включить в анализ ответы на вопросы, связанные с депрессией. Если бы тревога и депрессия группировались в единый фактор, это означало бы, что респонденты, испытывающие тревогу, также испытывают депрессию, и эти характеристики нельзя развести, по крайней мере, на основании самоотчета. Если же пункты, связанные с тревогой, группировались бы в один фактор, а пункты, связанные с депрессией, — в другой, мы могли бы быть более уверены в том, что наши вопросы, связанные с тревогой, являются не просто мерой общей склонности респондентов чувствовать себя несчастными.

Проиллюстрируем интерпретацию и некоторые вычисления, связанные с факторным анализом, пытаясь установить, можно ли выделить тревогу и депрессию, получаемые по результатам самоотчета, в отдельные факторы. Чтобы не усложнять пример, ограничимся тремя короткими вопросами по тревоге (Anxiety, A1 — A3) и тремя — по депрессии (Depression, D1 — D3), соответственно, хотя в большинстве случаев использования факторного анализа оперируют бо́льшим числом переменных:

A1 Я испытываю тревогу
A2 Я становлюсь напряженным
A3 Я спокоен
D1 Я подавлен
D2 Я чувствую себя бесполезным
D3 Я счастлив

Ответы на каждый из этих вопросов даются по 5-балльной шкале, где 1 означает «никогда», 2 — «иногда», 3 — «часто», 4 — «бо́льшую часть времени» и 5 — «всегда».

Рекомендуемый минимальный объем выборки, которую можно использовать для проведения факторного анализа, по-разному определяется разными авторами, но общепринято, что он должен быть больше числа переменных. Например, R. L. Gorsuch (1983) полагает, что для проведения факторного анализа необходимо не менее 100 испытуемых (наблюдений) и, как минимум, 5 наблюдений в расчете на переменную. Наблюдение представляет собой единицу анализа, которой в нашем случае является респондент. Однако это может быть школа, коммерческая организация, город и т. п. Чтобы упростить процедуру ввода данных для анализа, мы использовали вымышленные ответы только девяти испытуемых, приведенные в табл. 1.1[1]. В строке, соответствующей первому испытуемому, видим, что он иногда испытывает тревогу, никогда не напряжен и часто спокоен[2].

Корреляционная матрица

Первым шагом при осуществлении факторного анализа является создание корреляционной матрицы, в которой содержатся коэффициенты корреляции всех переменных друг с другом. Корреляционная матрица для данных табл. 1.1 представлена в табл. 1.2.

---

[1] В соответствии с идеей R. L. Gorsuch (1983) наблюдений должно быть существенно больше.

[2] Данные, содержащиеся в табл. 1.1, называют сырыми, или первичными.

Таблица 1.1. **Ответы девяти испытуемых по шести переменным (вопросам)**

Наблюдения	A1 Тревожный	A2 Напряженный	A3 Спокойный	D1 Подавленный	D2 Бесполезный	D3 Счастливый
1	2	1	3	1	2	5
2	1	2	3	4	3	3
3	3	3	4	2	1	4
4	4	4	3	3	2	3
5	5	5	2	3	4	4
6	4	5	2	4	3	1
7	4	3	2	5	4	1
8	3	3	4	4	4	3
9	3	5	3	3	4	1

Таблица 1.2. **Нижний треугольник корреляционной матрицы для шести переменных**

Переменные	A1 Тревожный	A2 Напряженный	A3 Спокойный	D1 Подавленный	D2 Бесполезный	D3 Счастливый
A1 Тревожный	1,00					
A2 Напряженный	0,74	1,00				
A3 Спокойный	−0,50	−0,40	1,00			
D1 Подавленный	0,22	0,30	−0,37	1,00		
D2 Бесполезный	0,28	0,39	−0,43	0,65	1,00	
D3 Счастливый	−0,25	−0,54	0,41	−0,74	−0,53	1,00

Такая матрица называется *нижнетреугольной* из-за своей формы, при которой корреляции между каждой парой переменных показаны лишь один раз.[1]

[1] Полная матрица получается в результате симметричного отображения чисел нижнего треугольника относительно главной диагонали.

Корреляция отражает направление и величину линейной зависимости между двумя переменными. Коэффициент корреляции принимает значения в диапазоне от –1,00 до +1,00. Значение коэффициента корреляции, равное –1,00, указывает на максимальную обратную зависимость между переменными, при которой наибольшее значение одной переменной (например, 5 — для «Тревожный») связано с наименьшим значением другой переменной (1 — для «Спокойный»), следующее по величине значение первой переменной (4 — для «Тревожный») связано со следующим из наименьших значений второй переменной (2 — для «Спокойный»), и т.д. Коэффициент корреляции, равный +1,00, указывает на максимальную прямую зависимость между двумя переменными, при которой максимальное значение одной переменной (например, 5 — для «Тревожный») связано с максимальным значением другой переменной (например, 5 — для «Напряженный»), следующее по величине наибольшее значение первой переменной (4 — для «Тревожный») связано со следующим по величине наибольшим значением второй переменной (4 — для «Напряженный»), и т.д. Значение коэффициента корреляции, равное 0,00, указывает на отсутствие линейной зависимости между двумя переменными. Обычно максимальные по модулю значения коэффициента корреляции +1,00 или –1,00 очень редко встречаются. Коэффициенты корреляции, стоящие на диагонали матрицы, приведенной в табл. 1.2, представляют собой просто корреляции переменных с самими собой и по своей сути всегда равны 1,00, а потому не представляют никакого интереса и никакой роли не играют.

Чем больше по абсолютному значению коэффициент корреляции, независимо от знака, тем сильнее линейная взаимосвязь между двумя переменными. Наибольшие по абсолютному значению коэффициенты корреляции в табл. 1.2 равны 0,74 (корреляция между «Тревожный» и «Напряженный») и –0,74 (корреляция между «Подавленный» и «Счастливый»). Следующая по абсолютному значению корреляция между «Подавленный» и «Бесполезный» равна 0,65. Наименьший по абсолютному значению коэффициент корреляции между «Тревожный» и «Подавленный» равен 0,22.

Величина совместной дисперсии для двух переменных есть квадрат коэффициента корреляции этих переменных. Так, величина совместной дисперсии, объясняемой тем общим, что присутствует и в ощущении тревожности и в ощущении напряженности, равна $0{,}74^2$, или приближенно 0,55, в то время как величина общей дисперсии между тревожностью и подавленностью равна $0{,}22^2$, или около 0,05. Другими словами, величина совместной дисперсии между тревожностью и напряженностью в 11 раз больше величины совместной дисперсии между тревожностью и подавленностью. Максимальное значение величины совместной дис-

персии равно 1,00 ($\pm 1{,}00^2 = 1{,}00$), а минимальное — 0,00 ($0{,}00^2 = 0{,}00$).

Если посмотреть на значения коэффициентов корреляции в табл. 1.2, можно увидеть, что имеется некоторая тенденция, проявляющаяся в том, что пункты вопросника, связанные с тревогой, сильнее коррелируют друг с другом, чем с пунктами, связанными с депрессией, а последние, в свою очередь, сильнее коррелируют между собой, чем с пунктами, связанными с тревогой. Таким образом, можно предположить существование двух отдельных групп для пунктов, связанных с тревогой и депрессией. Например, коэффициент корреляции между ощущением тревожности и напряженностью равен 0,74, в то время как коэффициент корреляции между ощущением тревожности и подавленностью равен всего лишь 0,22. Однако картина не совсем ясна, поскольку абсолютное значение коэффициента корреляции между «Напряженный» и «Счастливый», входящих в группу депрессивных пунктов, выше (–0,54), чем абсолютное значение коэффициента корреляции между «Напряженный» и «Спокойный» (–0,40). Обычно невозможно сказать, просто глядя на корреляционную матрицу, на сколько групп, или факторов, разобьются переменные; чем больше переменных, тем труднее это сделать. Следовательно, чтобы определить, сколько групп получится, надо использовать более строгий формальный метод. Таковым является эксплораторный факторный анализ.

Метод главных компонент

Существует много различных видов факторного анализа, но, возможно, самым простым и наиболее часто используемым среди них является метод главных компонент. Компоненты — еще один термин для обозначения факторов, и компоненты в методе главных компонент часто называются факторами. В этой книге будем использовать данные термины как взаимозаменяемые[1]. В методе главных компонент величина объясняемой дисперсии равна числу переменных, поскольку дисперсия или общность каждой пере-

[1] Нам кажется, следующий комментарий может облегчить понимание сути эксплораторного факторного анализа. Каждая переменная имеет определенную дисперсию (общность). В методе главных компонент переменные стандартизируются и нормируются, поэтому предполагается, что дисперсия каждой из них равна 1. Исходя из их независимости можно предположить, что суммарная дисперсия всех переменных равна числу переменных. Каждая главная компонента является в определенном смысле переменной величиной, а потому также имеет дисперсию, называемую объясняемой дисперсией. Дисперсия отсутствует только у постоянных величин. Суммарная объясняемая дисперсия всех главных компонент должна быть равна суммарной дисперсии всех переменных, т.е. их общему числу.

менной принимается равной 1,00. Таким образом, в случае шести переменных суммарная (общая) объясняемая дисперсия равна 6,00. Число образованных, или извлекаемых, компонент математически всегда равно числу переменных, участвующих в анализе[1]. Таким образом, в случае шести переменных будет извлечено шесть факторов. Факторы всегда упорядочиваются в соответствии с величиной их дисперсии. Первый фактор всегда объясняет наибольшую долю общей дисперсии, второй фактор — следующую по величине долю общей дисперсии, которая не была объяснена первым фактором, и т.д., последний фактор объясняет наименьшую долю общей дисперсии. Значения коэффициентов корреляции переменной с каждым фактором дают величины нагрузок данной переменной по этим факторам[2]. Так как первый фактор объясняет наибольшую долю общей дисперсии, значения коэффициентов корреляции, или нагрузки, всех переменных по этому фактору будут, в среднем, самыми высокими, следующими по величине будут нагрузки на второй фактор и т.д. Чтобы вычислить долю общей дисперсии, объясняемой каждым фактором, надо суммировать квадраты нагрузок по данному фактору, в результате получаем собственное, или характеристическое, значение этого фактора, которое делим на число переменных.

В табл. 1.3 приведены шесть главных компонент для набора данных из шести переменных, представленных в табл. 1.1. Как можно видеть, все шесть переменных сильнее всего коррелируют с первым фактором, за исключением ощущения тревожности, которое чуть сильнее коррелирует со вторым фактором. В табл. 1.4 приведены доли общей дисперсии, объясненной этими шестью главными компонентами, для чего были вычислены квадраты нагрузок, затем суммированием этих квадратов были получены собственные значения и, наконец, путем деления на число переменных — доли общей дисперсии. Так, нагрузка тревожности по первому фактору, равная 0,66, при возведении в квадрат и округлении до двух десятичных знаков дает 0,43. Суммирование квадратов нагрузок переменных по первому фактору дает собственное значение, рав-

---

[1] Только так будет соблюдено условие о том, что сумма суммарных дисперсий факторов в точности равна сумме дисперсий первичных переменных.

[2] Такая интерпретация коэффициентов корреляции переменных с факторами правомерна, если полагать каждый фактор как некоторую гипотетическую переменную, которую нельзя реально измерить. Эти переменные называются латентными, подробно о них будет рассказано далее, здесь следует упомянуть, что факторы, будучи переменными, хотя и гипотетическими, могут коррелировать с другими реальными переменными, и именно эти гипотетические коэффициенты корреляции и называются факторными нагрузками. Сумма квадратов факторных нагрузок всех переменных по фактору (сумма совместных дисперсий) равна объясняемой этим фактором дисперсии, умноженной на число первичных переменных.

Таблица 1.3. **Начальные главные компоненты**

	1	2	3	4	5	6
A1 Тревожный	0,66	0,67	0,03	0,11	0,29	–0,14
A2 Напряженный	0,76	0,46	0,37	0,02	–0,20	0,18
A3 Спокойный	–0,69	–0,20	0,64	0,25	0,10	–0,05
D1 Подавленный	0,76	–0,53	0,05	–0,04	0,35	0,14
D2 Бесполезный	0,75	–0,34	–0,15	0,52	–0,17	–0,06
D3 Счастливый	–0,80	0,34	–0,27	0,35	0,14	0,17

Таблица 1.4. **Доля общей дисперсии, объясняемая начальными главными компонентами**

	1	2	3	4	5	6	Общности
A1 Тревожный	0,43	0,45	0,00	0,01	0,08	0,02	1,00
A2 Напряженный	0,58	0,21	0,14	0,00	0,04	0,03	1,00
A3 Спокойный	0,48	0,04	0,41	0,06	0,01	0,00	1,00
D1 Подавленный	0,57	0,28	0,00	0,00	0,12	0,02	1,00
D2 Бесполезный	0,56	0,11	0,02	0,27	0,03	0,00	1,00
D3 Счастливый	0,64	0,12	0,07	0,02	0,02	0,03	1,00
Собственные значения	3,26	1,21	0,64	0,47	0,31	0,11	6,00
Доля[1]	0,54	0,20	0,11	0,08	0,05	0,02	

ное 3,26, которое при делении на число переменных, равное шести, дает, с учетом округления, долю общей дисперсии, равную 0,54(3,26/6 = 0,543). Другими словами, первая главная компонента (фактор) объясняет 54 % общей дисперсии шести переменных, второй фактор — еще 20 % общей дисперсии и т.д.[2].

Определение числа главных компонент, оставляемых для дальнейшего анализа

Поскольку число выделяемых компонент (факторов) равно числу переменных, нам необходим некоторый критерий отбора и решения вопроса о том, сколькими меньшими факторами можно

[1] Объясняемой дисперсии.

[2] Читатель может попробовать подсчитать эту величину для того, чтобы убедиться, что основные принципы усвоены правильно.

пренебречь, учитывая незначительную долю объясняемой ими общей дисперсии. Одним из главных критериев, используемых для решения этого вопроса, является критерий Кайзера, или Кайзера — Гутмана, согласно которому не следует рассматривать факторы с собственными значениями, меньшими или равными единице. Обосновать это можно следующим образом. Наибольшая величина дисперсии, объясняемая одной переменной, равна единице, поэтому факторы с собственными значениями, меньшими единицы, могут хорошо объяснить, самое большее, дисперсию одной-единственной переменной. Как следует из табл. 1.4, первые два фактора имеют собственные значения, превосходящие единицу, в то время как собственные значения остальных четырех факторов меньше единицы. Таким образом, если мы примем этот широко используемый критерий, нам следует обращать внимание только на первые два фактора и игнорировать четыре оставшихся меньших фактора.

R. B. Cattell (1966) указал на то, что критерий Кайзера может сохранять слишком много факторов в случае большого числа переменных и слишком мало факторов в случае малого числа переменных. Он предложил альтернативный критерий, критерий «каменистой осыпи»[1], связанный, главным образом, с нахождением отчетливой границы между первыми бо́льшими факторами, объясняющими значительную часть общей дисперсии, и следующими за ними меньшими факторами, объясняющими примерно равные по величине малые доли общей дисперсии. Чтобы определить, где проходит эта граница, собственные значения каждого фактора изображают точками на плоскости, при этом по вертикали откладывают собственные значения, а по горизонтали располагают номера факторов в порядке уменьшения величин соответствующих собственных значений. На рис. 1.1 таким образом изображены собственные значения шести главных компонент (факторов), приведенных в табл. 1.4.

«Каменистая осыпь» представляет собой геологический термин для обозначения обломков горных пород, скапливающихся у подножия крутого склона и скрывающих настоящее основание самого склона. Факторы, образующие сам склон, считаются важными для рассмотрения, а факторы, представляющие «каменистую осыпь», — нет. Последние можно определить путем проведения прямой линии либо через точки, представляющие соответствующие собственные значения, либо в непосредственной близости от них[2]. Это не всегда легко сделать, как на рис. 1.1, где не совсем ясно, начинается ли «осыпь» со второго или с третьего фактора.

---

[1] От англ. scree test. Часто этот критерий называют критерием следа.

[2] В результате число незначимых факторов определяется количеством точек, лежащих вблизи этой прямой линии (каменистой осыпи).

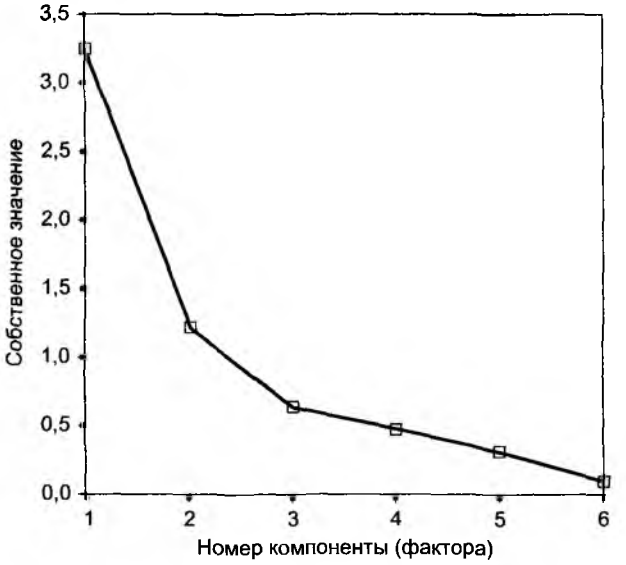


Рис. 1.1. График собственных значений для шести главных компонент

В этом случае может оказаться полезным сравнение переменных, коррелирующих как с первыми двумя, так и с первыми тремя факторами (после соответствующего вращения факторов), с целью определить, какое же из двух решений — двухфакторное или трехфакторное — имеет больший смысл[1]. Причины для осуществления вращения факторов будут обсуждаться в следующем подразделе. Если с помощью проведения прямых линий можно выделить несколько «осыпей», то заслуживающей максимального доверия считается самая верхняя из них.

Вращение факторов

Математическая процедура, позволяющая прояснить содержательный смысл выделенных на предыдущем этапе начальных глав-

[1] Выбор числа значимых факторов, основанный на сравнении решений с точки зрения их осмысленности, имеет существенный недостаток, связанный с отсутствием формальных статистических критериев, т.е. в этом случае исследователь должен выбрать модель исходя из своих субъективных предпочтений. Однако здесь стоит отметить, что даже в тех случаях, когда статистические критерии существуют, субъективность при выборе того или иного решения (гипотезы, модели) также присутствует. Например, выбор границы уровня значимости, т.е. допускаемой вероятности совершить ошибку первого рода, осуществляется исследователем и может быть 0,05; 0,1; 0,01 (а также любое другое число в диапазоне от 0 до 1).

Таблица 1.5. **Первые две главные компоненты после вращения по методу варимакс**

	1	2
A1 Тревожный	0,05	0,94
A2 Напряженный	0,27	0,85
A3 Спокойный	−0,39	−0,60
D1 Подавленный	0,92	0,10
D2 Бесполезный	0,79	0,24
D3 Счастливый	−0,83	−0,27

ных компонент, объясняющих бóльшую часть общей дисперсии переменных, называется *вращением*. Факторные нагрузки преобразуются по специальным формулам[1]. Существует множество методов таких преобразований (вращений). Мы обсудим два из них. Наиболее распространенным методом вращения является *варимакс*[2], при котором факторы остаются независимыми или ортогональными по отношению друг к другу, так что баллы испытуемых по одному из факторов не коррелируют с баллами по другим факторам. Метод варимакс пытается максимизировать дисперсию, объясняемую значимыми (отобранными на предыдущем этапе с помощью критерия Кайзера, следа или другим способом) факторами, путем еще большего увеличения коэффициентов корреляции (факторных нагрузок), высоко коррелирующих с этим фактором переменных, и уменьшения коэффициентов корреляции, низко коррелирующих с этим фактором переменных. В табл. 1.5 представлены результаты вращения двух выделенных главных компонент по методу варимакс.

Сравнение величин коэффициентов корреляции (факторных нагрузок) в табл. 1.5 и 1.3 показывает, что некоторые из коэффициентов корреляции увеличились, в то время как другие, наоборот, уменьшились, а факторная структура компонент после вращения стала более понятной и легче интерпретируемой. Первый из факторов после вращения варимакс можно интерпретировать как фактор депрессии, поскольку все три соответствующих пункта имеют по этому фактору нагрузки ±0,79 и выше, в то время как

[1] Эти преобразования факторов производятся по формулам таким образом, что если сопоставить старые и новые полученные после пересчета факторных нагрузок факторы в многомерном пространстве, координатами которого являются значения факторных нагрузок у этих факторов по всем первичным переменным, то геометрически это будет выглядеть как поворот вокруг начала координат.

[2] Стоит отметить, что вращение варимакс используется, как минимум, в 80 % случаев применения эксплораторного факторного анализа для анализа данных в конкретных эмпирических исследованиях.

Таблица 1.6. **Доли общей дисперсии, объясняемые первыми двумя компонентами после вращения по методу варимакс**

	1	2
A1 Тревожный	0,00	0,88
A2 Напряженный	0,07	0,72
A3 Спокойный	0,15	0,36
D1 Подавленный	0,84	0,01
D2 Бесполезный	0,62	0,06
D3 Счастливый	0,69	0,07
Собственные значения	2,37	2,10
Доля[1]	0,40	0,35

все три вопроса, касающиеся тревожности, имеют нагрузки ±0,39 и ниже. Второй фактор после вращения варимакс можно интерпретировать как фактор тревожности, поскольку пункты, соответствующие тревоге, имеют по этому фактору нагрузки ±0,60 и выше, в то время как пункты, связанные с депрессией, — ±0,27 и ниже. При анализе большого количества переменных полезно расположить все пункты в порядке убывания факторных нагрузок с тем, чтобы видеть, какие из переменных имеют самые высокие нагрузки по рассматриваемым факторам.

Доля общей дисперсии, объясняемая каждым из двух выделенных факторов, после вращения по методу варимакс равна их собственным значениям, или сумме квадратов факторных нагрузок для каждого фактора, деленной на число переменных. Эти данные для двух главных компонент после вращения варимакс представлены в табл. 1.6. Доля общей дисперсии, объясняемой первым фактором, равна 0,40, вторым фактором — 0,35. Эти доли объясняемой дисперсии отличаются от соответствующих величин для начальных, не подвергавшихся вращению, главных компонент, поскольку в результате вращения изменились нагрузки переменных на эти факторы.

Другой метод вращения называется *прямой облимин* (direct oblimin[2]), здесь факторы могут коррелировать, т.е. быть неортогональными по отношению друг к другу. Существует два способа представления результатов косоугольного вращения. Первый — это так называемая матрица факторного отображения[3], которая обычно приводится в выходных файлах, содержащих результаты

---

[1] Объясняемой дисперсии.

[2] От англ. oblique — в геометрии острый или тупой, наклонный, т.е. неортогональный (прямой).

[3] В SPSS — pattern matrix.

Таблица 1.7. **Матрицы факторного отображения и структурная для первых двух компонент после вращения по методу прямой облимин**

	Матрица факторного отображения		Структурная матрица	
	1	2	1	2
A1 Тревожный	–0,16	0,99	0,26	0,93
A2 Напряженный	0,09	0,85	0,45	0,88
A3 Спокойный	–0,28	–0,55	–0,51	–0,67
D1 Подавленный	0,97	–0,11	0,92	0,29
D2 Бесполезный	0,79	0,07	0,82	0,40
D3 Счастливый	–0,83	–0,09	–0,87	–0,44
Доля	0,39	0,34	0,47	0,42

анализа[1]. Матрица факторного отображения показывает, какой вклад каждая переменная вносит в уникальную часть того или иного фактора, и не учитывает вклада, делаемого переменной в совместную с другими факторами часть. Структурная матрица отражает общий вклад каждой переменной в соответствующий фактор[2]. Если факторы не коррелируют между собой, то эти матрицы совпадают, и в этом случае проще и корректнее осуществить вращение по методу варимакс. Если же факторы коррелируют между собой, не имеет смысла приводить доли общей дисперсии, объясняемой каждым фактором, поскольку матрица факторного отображения будет давать заниженную, а структурная матрица — завышенную оценки.

В табл. 1.7 приведены матрицы факторного отображения и структурная для двух главных компонент из табл. 1.4 после вращения по методу прямой облимин. Также приведены доли общей дисперсии, объясняемой этими факторами после вращения, с тем чтобы показать их различие, хотя обычно подобные показатели не приводятся.

---

[1] Поскольку в результате косоугольного вращения факторы коррелируют, то можно говорить о том, что дисперсия каждого фактора объясняется частично только этим фактором (уникальная часть), а также совместными корреляциями этого фактора с другими, т.е. каждый фактор можно рассматривать состоящим из двух слагаемых. Во-первых, это уникальная, или специфическая, часть фактора, во-вторых, та часть, которая совместна, т.е. входит в другие факторы тоже. В случае, когда вторая составляющая нулевая, корреляция фактора со всеми остальными факторами равна нулю.

[2] В структурной матрице учитывается не только непосредственная взаимосвязь переменной с фактором (как в матрице факторного отображения), но и опосредованная взаимосвязь факторов друг с другом. В результате получаются значения — коэффициенты корреляции переменных с факторами.

Коэффициент корреляции между двумя выделенными косоугольными факторами равен примерно 0,42. Так как данные факторы коррелируют, величины нагрузок в матрицах факторного отображения и структурной различаются между собой. Нагрузки из матрицы факторного отображения легче интерпретировать, чем нагрузки из структурной матрицы, хотя результаты в целом сходны. Первый фактор, подвергнутый вращению по методу прямого облимина, можно интерпретировать как фактор депрессии, поскольку в матрице факторного отображения все три пункта, связанные с депрессией, имеют по этому фактору нагрузки ±0,79 или выше, в то время как все три элемента, связанные с тревогой, — ±0,28 или ниже. Второй фактор после вращения по методу прямой облимин, видимо, представляет тревогу, так как все три элемента, связанные с тревогой, имеют по этому фактору в матрице факторного отображения нагрузки ±0,55 или выше, в то время как все три пункта, связанные с депрессией, — ±0,11 или меньше. В данном отношении результаты вращения по методу прямого облимина в основном совпадают с результатами вращения по методу варимакс.

Отчет о результатах

Форма отчета о проведении анализа по методу главных компонент в определенной степени зависит от целей его проведения. Один из кратких способов описать результаты примера, использованного в данной главе, состоит в следующем: «Корреляционная матрица шести переменных была подвергнута процедуре анализа по методу главных компонент. Было извлечено два фактора с собственными значениями больше единицы. Эти факторы подвергались вращению по методу варимакс и прямой облимин. Были получены существенно сходные результаты. Первый фактор можно интерпретировать как фактор, отражающий депрессию, так как все три пункта, связанные с депрессией, имеют по нему самые высокие нагрузки. Второй фактор можно интерпретировать как фактор, отражающий тревогу, поскольку все три пункта, соответствующие тревоге, имеют по нему высокие факторные нагрузки. Факторы, полученные в результате вращения по методу варимакс, объясняют 40 и 35 % совокупной (общей) дисперсии соответственно. Коэффициент корреляции между факторами, полученными в результате вращения direct oblimin (прямой облимин), составил 0,42».

Реализация процедуры в программе SPSS для Windows

Приведем алгоритм получения основной информации, связанной с анализом данных из табл. 1.1 по методу главных компонент.

	1 : тревожен	2				
	тревожен	напряжен	спокоен	подавлен	бесполез	весел
1	2	1	3	1	2	5
2	1	2	3	4	3	3
3	3	3	4	2	1	4
4	4	4	3	3	2	3
5	5	5	2	3	4	4
6	4	5	2	4	3	1
7	4	3	2	5	4	1
8	3	3	4	4	4	3
9	3	5	3	3	4	1
10						

Рис. 1.2. Данные в **Редакторе данных (Data Editor)**

Ввести данные из табл. 1.1 с помощью **Редактора данных (Data Editor)**, как показано на рис. 1.2, присвоить названия переменным и сохранить их как файл с каким-либо названием.

В строке **Меню**, расположенной в верхней части главного окна программы, выбрать пункт **Анализ (Analyze)**, а в появившемся ниспадающем подменю — пункт **Сокращение данных (Data reduction)** и затем — **Факторный анализ (Factor)**, открывающий диалоговое окно **Факторный анализ (Factor Analysis)** (рис. 1.3)

Выберите переменные с «тревожен» до «весел» и нажмите первую (верхнюю) ▸ из кнопок, чтобы поместить их в список **Переменные (Variables)**.

Нажмите на кнопку **Описательная статистика (Descriptives)**, открывающую диалоговое окно **Факторный анализ: Описательные статистики (Factor Analysis: Descriptives)**, рис. 1.4. Установите флажок **Коэффициенты (Coefficients)** в группе флажков **Корреляционная матрица (Correlation Matrix)**, чтобы получить корреляционную матрицу, аналогичную приведенной в табл. 1.2. Эта матрица, являющаяся первой таблицей в окне вывода результатов, называ-

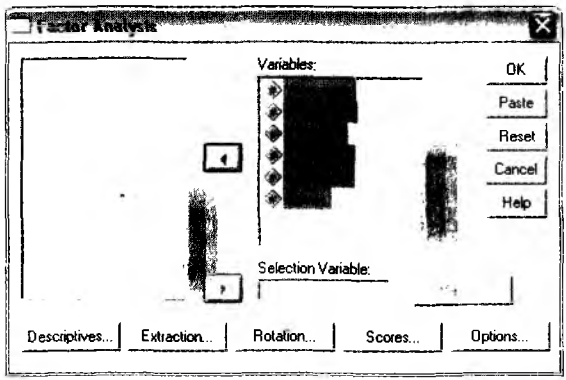
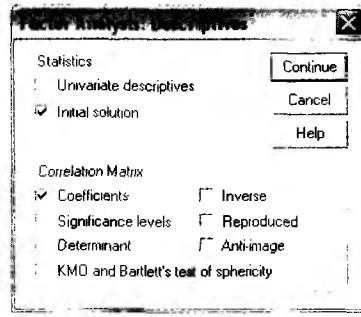


Рис. 1.3. Диалоговое окно **Факторный анализ (Factor Analysis)**

Рис. 1.4. Диалоговое окно **Факторный анализ**: **Описательные статистики** (**Factor Analysis: Descriptives**)



ется **Корреляционная матрица** (**Correlation Matrix**) и имеет квадратную, а не нижнетреугольную форму. Значения коэффициентов корреляции приведены с точностью до трех, а не до двух десятичных знаков. Нажмите на кнопку **Продолжить** (**Continue**), чтобы вернуться в диалоговое окно **Факторный анализ** (**Factor Analysis**), представленное на рис. 1.3.

Нажмите на кнопку **Извлечение факторов** (**Extraction**), чтобы открыть диалоговое окно **Факторный анализ**: **Извлечение факторов** (**Factor Analysis: Extraction**), рис. 1.5. Установите флажок **График собственных значений** (**Scree plot**) в группе флажков **Отображать** (**Display**), чтобы получить график собственных значений (след), представленный на рис. 1.1. По умолчанию в SPSS предполагается выполнять 25 итераций для осуществления процедуры извлечения факторов. В случае шести переменных этого вполне достаточно, однако при анализе других наборов данных может потребоваться более 25 итераций для сходимости алгоритма. В этом случае следует увеличить значение в поле **Максимальное число итераций для сходимости** (**Maximum Iterations for Convergence**) до 50 или 100 и более, в случае необходимости. В качестве критерия для выделения значимого количества факторов в SPSS

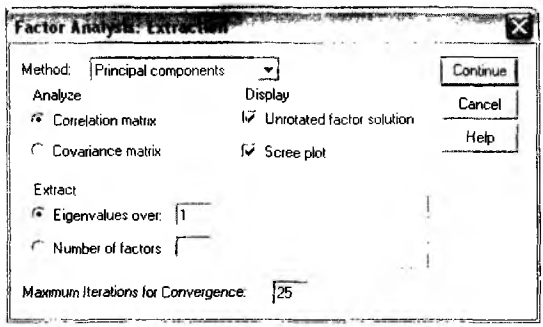


Рис. 1.5. Диалоговое окно **Факторный анализ**: **Извлечение факторов** (**Factor Analysis: Extraction**)

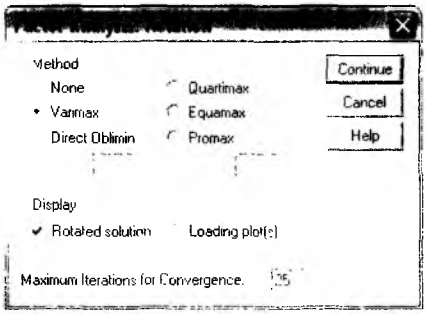
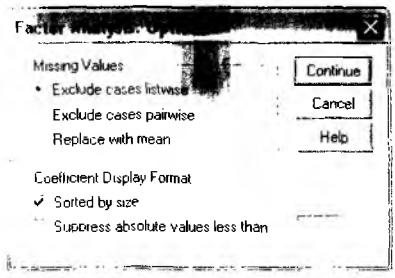


Рис. 1.6. Диалоговое окно **Факторный анализ: Вращение (Factor Analysis: Rotation)**

по умолчанию используется критерий Кайзера, оставляющий факторы с собственными значениями, превосходящими единицу. Однако можно оставить меньше или больше факторов. Для этого нужно указать число в поле рядом с переключателем **Число факторов (Number of factors)** в группе **Извлечение факторов (Extract)**. Процент общей дисперсии, объясняемый оставленными для вращения главными компонентами (факторами), отображается в таблице, озаглавленной **Объясненная общая (совокупная) дисперсия (Total Variance Explained)** в третьем и пятом столбцах с названием **Процент дисперсии (% of Variance)** (под заголовками **Исходные собственные значения (Initial Eigenvalues)** и **Конечные суммы квадратов факторных нагрузок (Extraction Sums of Squared Loadings)** соответственно). Нагрузки по извлеченным главным компонентам (факторные нагрузки), приведенные во втором и третьем столбцах табл. 1.3, отображаются в таблице окна результатов программы, озаглавленной **Матрица факторных нагрузок (Component Matrix)**, в виде десятичных чисел, округленных до третьего знака после запятой. Нажмите кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Факторный анализ (Factor Analysis)**.

Нажмите на кнопку **Вращение (Rotation)** и откройте диалоговое окно **Факторный анализ: Вращение (Factor Analysis: Rotation)**, рис. 1.6. В рамках одной сессии анализа можно выполнить только одно вращение. Установите переключатель **Варимакс (Varimax)**, чтобы получить данные, содержащиеся в табл. 1.5, которые будут представлены в окне результатов программы SPSS под заголовком **Матрица факторных нагрузок после вращения (Rotated Component Matrix)**. Некоторые величины будут округлены до трех десятичных знаков после запятой, в то время как остальные будут представлены в экспоненциальной форме. Например, коэффициент корреляции между тревожностью и первым из преобразованных факторов отображается как 5,237E-02. Обозначение E-02 показывает, что запятая должна быть перенесена на два разряда влево, в обычном формате это число равно 0,05237. Доля общей (совокупной) дисперсии, объясняемая ортогонально вращаемы-

Рис. 1.7. Диалоговое окно **Факторный анализ: Параметры (Factor Analysis: Options)**



ми факторами, отображается в таблице, озаглавленной **Объясненная общая (совокупная) дисперсия (Total Variance Explained)** в девятом столбце с названием **Процент дисперсии (% of Variance)** под рубрикой **Суммы квадратов факторных нагрузок после вращения (Rotation Sums of Squared Loadings)**. Для следующего шага анализа установите переключатель **Прямой облимин** (Direct Oblimin), чтобы получить матрицы факторного отображения и структурную из табл. 1.7. Две таблицы из окна вывода программы называются, соответственно, **Матрица факторного отображения (Pattern Matrix)** и **Структурная матрица (Structure Matrix)**.

Коэффициент корреляции между двумя косоугольными факторами отображается в окне вывода программы в таблице с названием **Корреляционная матрица факторов (компонент) (Component Correlation Matrix)** с округлением до трех десятичных знаков. Нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Факторный анализ (Factor Analysis)**.

Если вы хотите вывести факторные нагрузки в порядке убывания их величин, нажмите кнопку **Параметры (Options)**, чтобы открыть диалоговое окно **Факторный анализ: Параметры (Factor Analysis: Options)**, рис. 1.7. Установите флажок **Сортировать по величине (Sorted by size)** (в группе **Формат вывода коэффициентов (Coefficient Display Format)**). Нажмите кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Факторный анализ (Factor Analysis)**.

Нажмите **OK**, чтобы провести анализ.

Рекомендуемая литература

Child, D. (1990) *The Essentials of Factor Analysis,* 2nd edn. London: Cassell.

Kline, P. (1994) *An Easy Guide to Factor Analysis.* London: Routledge.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: Harper Collins.

Дополнительная литература, рекомендуемая научным редактором

Митина О. В. Факторный анализ для психологов / О. В. Митина, И. Б. Михайловская. — М.: УМК «Психология», 2001.

Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

Кулаичев А. П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

Глава 2

КОНФИРМАТОРНЫЙ (ПОДТВЕРЖДАЮЩИЙ) ФАКТОРНЫЙ АНАЛИЗ

Предисловие научного редактора

Гл. 2 посвящена введению в конфирматорный (подтверждающий) факторный анализ, предназначенный в отличие от исследовательского не для отыскания структур группировки переменных в классы (факторы), а для проверки уже имеющихся априорных гипотез о составе этих классов и степени и характере взаимосвязей между этими классами. Этот метод является частным случаем структурного моделирования. В нашей стране структурное моделирование еще не стало такой общепринятой и общеизвестной методологией, как на Западе, поэтому доступные книги на русском языке пока отсутствуют.

В то время как эксплораторный (разведочный) факторный анализ используется для того, чтобы определить, какова наиболее вероятная факторная структура соотношений между некоторым множеством переменных, с помощью конфирматорного (подтверждающего) факторного анализа проверяют, с какой вероятностью априорно предполагаемая факторная структура подтверждается данными. Если данные не подтверждают модель или не соответствуют постулируемой факторной структуре, то они будут значимо отличаться от того, что могло бы получиться на основе предполагаемой факторной модели. Если данные подтверждают модель, то значимых различий не будет. Проверив несколько моделей, можно установить, какая из них объясняет данные наиболее адекватно. При наличии шести пунктов опросника, относя-

Таблица 2.1. **Однофакторная и двухфакторные модели для шести пунктов**

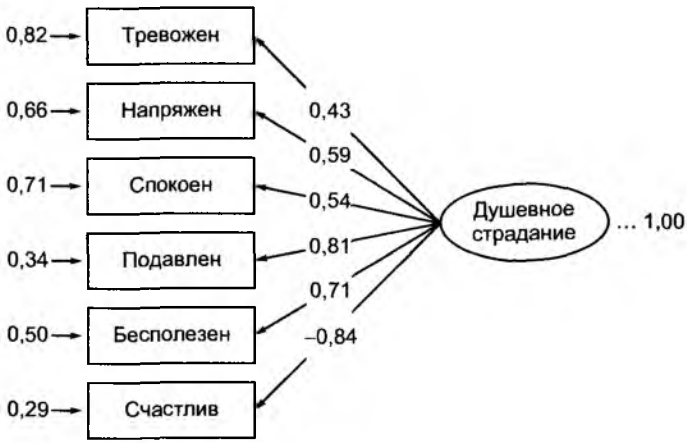
	Один фактор	Два фактора	
	1	1	2
A1 Тревожный	1	1	0
A2 Напряженный	1	1	0
A3 Спокойный	1	1	0
D1 Подавленный	1	0	1
D2 Бесполезный	1	0	1
D3 Счастливый	1	0	1

щихся, по нашему мнению, к измерению тревоги или депрессии и коррелирующих между собой, можно определить, как представить модель — действием одного или двух факторов, и коррелируют ли тогда эти факторы.

В однофакторной модели все шесть пунктов будут иметь нагрузки по единственному фактору, в двухфакторной модели пункты, связанные с депрессией, будут иметь нагрузки только по одному из факторов, а пункты, связанные с тревогой, — только по другому фактору. Факторы могут предполагаться коррелирующими или некоррелирующими. В табл. 2.1 показано, какие из пунктов нагружены (1) и не нагружены (0) по факторам в однофакторной и двухфакторных моделях.

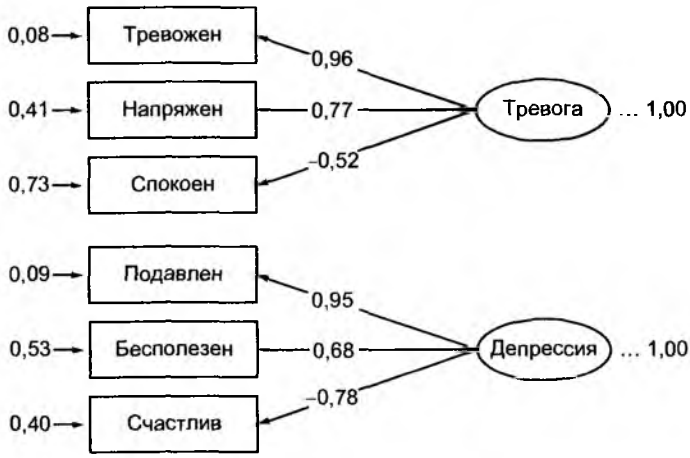
Диаграммы путей

Другой способ представления проверяемой факторной модели — ее графическое изображение с помощью диаграммы путей. Применение конфирматорного факторного анализа проиллюстрируем примером из предыдущей главы с тревогой и депрессией, чтобы можно было сравнить результаты двух методов. Диаграммы путей для однофакторной модели, двухфакторной модели с независимыми (некоррелирующими) факторами и двухфакторной модели с зависимыми (коррелирующими) факторами представлены на рис. 2.1, 2.2 и 2.3 соответственно. Эти диаграммы путей были получены с помощью компьютерной программы LISREL, где аббревиатура LISREL означает линейные структурные связи (за-



Chi-Square = 96,73; *df* = 9; *P*-value = 0,00100; RMSEA = 0,314

Рис. 2.1. Диаграмма путей для однофакторной модели



Chi-Square = 58,78; *df* = 9; *P*-value = 0,00100; RMSEA = 0,236

Рис. 2.2. Диаграмма путей для двухфакторной модели с несвязанными (ортогональными) факторами

висимости) (LInear Structural RELationships). Изображая модели в общем виде, значения числовых параметров обычно не указывают. Это означает, что их надлежит вычислить, или оценить. Информация из строки, расположенной под каждой моделью на рисунках, содержит два показателя степени соответствия модели экспериментальным данным.

Линии со стрелками на конце называются путями. Пункты представлены прямоугольниками, а факторы изображены эллипсами. Взаимосвязь между пунктом (переменной) и фактором обозначена стрелкой, направленной от эллипса к прямоугольнику. Например, на рис. 2.2 стрелка между переменной «Тревожный» и фактором «Тревога» показывает, что эта переменная имеет нагрузку по данному фактору или связана с ним. Между переменной «Тревожный» и фактором «Депрессия» стрелок нет, поскольку предполагается, что данный пункт не связан с данным фактором (не нагружает его). LISREL выводит в результирующий файл только восемь символов из имени переменной, так что последние буквы из названия данного фактора в окне выдачи программы будут отсутствовать. Стрелка направлена от фактора к пункту, а не от пункта к фактору, что означает зависимость пункта от фактора. Факторы иногда называют латентными (скрытыми) переменными, поскольку их нельзя измерить непосредственно, в то время как пункты опросника называют индикаторами, или явными переменными. Величины рядом с этими стрелками можно с некоторой натяжкой считать корреляциями, или нагрузками, пунктов по факторам, поскольку они в общем слу-

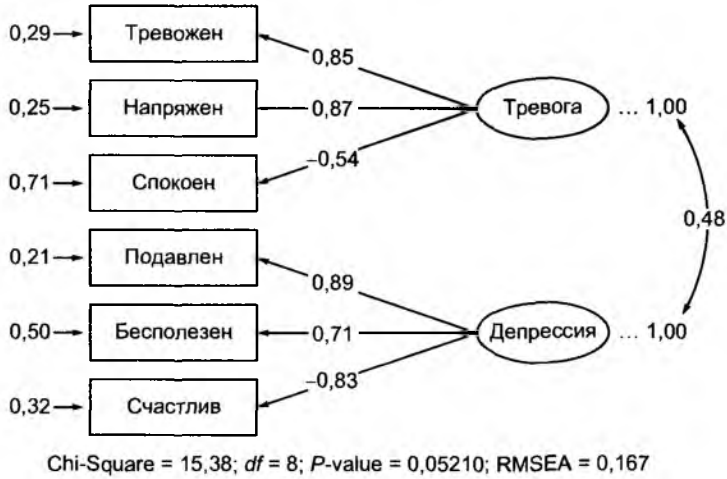


Рис. 2.3. Диаграмма путей для двухфакторной модели с коррелирующими (связанными) факторами

чае принимают значения от –1 до +1[1]. Так, на рис. 2.2 нагрузка пункта «Тревожный» по фактору «Тревога» равна 0,96.

Стрелки, расположенные слева от пунктов и указывающие на них, показывают, что каждый пункт не является совершенной мерой того фактора, который он должен отражать согласно модели. Цифры рядом с каждой стрелкой выражают долю дисперсии, которая соответствует ошибке[2,3]. На рис. 2.2 коэффициент детерминации пункта «Тревожный» его ошибкой[4] равен 0,08. На рис. 2.2 и 2.3 справа от факторов также имеются числа, равные 1,00. Они указывают на то, что дисперсия этих факторов стандартизована и равна 1. На путевых диаграммах эти единицы часто опускают.

Дуга со стрелками на обоих концах, соединяющая факторы «Тревога» и «Депрессия» на рис. 2.3, обозначает, что эти две переменные влияют друг на друга, но направление причинной зависимости не может быть определено. Цифра, стоящая рядом с ду-

---

[1] Но более строго это коэффициенты линейной регрессии независимой переменной (фактора) на зависимые переменные (пункты опросника).

[2] Детерминируется ею как латентной переменной.

[3] Хотя в англоязычной литературе принято для указания дополнительной дисперсии, не связанной с фактором, использовать термин ошибка (Error), нам кажется, более правильным употреблять слова «Специфичность» или «Остаточный член». Это соответствует концепции о том, что никакая модель не может описать всю ситуацию полностью, всегда есть некоторый остаток, не охваченный моделью. Это не ошибка, а та реальность, существование которой всегда следует иметь в виду, даже если не ставить перед собой задачу ее изучения в рамках конкретной модели.

[4] Специфичностью.

гой, есть коэффициент корреляции между этими переменными, который в данном случае равен 0,48.

Показатели согласованности[1]

Хи-квадрат

Нагрузки пунктов в целом выше для двухфакторных моделей, чем для однофакторной. Это наводит на мысль о том, что в данном случае двухфакторные модели точнее соответствуют экспериментальным данным. Было предложено большое количество показателей согласованности, определяющих, насколько хорошо модель соответствует данным, но пока еще нет общепринятого мнения о том, какие из этих показателей являются наиболее адекватными. В данной книге расскажем лишь о некоторых из них. Один из наиболее распространенных показателей приведен первым в строке под каждой из диаграмм путей (см. рис. 2.1—2.3) и для краткости назван хи-квадрат.

Критерий с вычислением этого же показателя используется при сравнении различий между наблюдаемыми и ожидаемыми частотами в таблице сопряженности. Он также используется в гл. 13, посвященной логлинейному анализу. Чем больше разница между наблюдаемой и ожидаемой частотами, тем больше критерий хи-квадрат и тем более вероятно, что эти различия не случайны. Критерий хи-квадрат даже при случайных различиях будет тем больше, чем больше категорий в таблице сопряженности, так что необходимо ввести поправку, учитывающую число категорий. Для этого при расчете критериального значения хи-квадрат в качестве второй переменной используется число степеней свободы[2, 3].

Критерий хи-квадрат в конфирматорном факторном анализе вычисляет различия между корреляционной матрицей исходных данных и корреляционной матрицей, получаемой на основе модели. Чем больше разница между исходной матрицей и матрицей, построенной на основе модели, тем больше будет критерий хи-квадрат и тем больше вероятность того, что эти различия не случайны. Если разница не может считаться случайной, то модель не дает адекватного согласия с данными. Критерий хи-квадрат пропорционален числу параметров, оцениваемых в модели. Следовательно, критерий хи-квадрат должен учитывать число таких пара-

[1] Имеется в виду согласованность модели с экспериментальными данными.

[2] Чем больше степеней свободы, тем больше может быть допустимая величина хи-квадрат, при которой согласие модели с экспериментальными данными остается возможной.

[3] В целом данный абзац никак не связан с конфирматорным факторным анализом, поэтому пусть у читателя не возникает недоумений в связи с этим.

метров. Это делается путем введения числа степеней свободы, которое равно суммарному количеству наблюдаемых дисперсий и ковариаций за вычетом количества оцениваемых параметров.

Количество наблюдаемых дисперсий и ковариаций вычисляется по общей формуле $n(n+1)/2$, где n — число наблюдаемых переменных[1]. Поскольку во всех трех моделях по шесть наблюдаемых переменных, количество наблюдаемых дисперсий и ковариаций равно $6(6+1)/2 = 21$. Это можно проверить, подсчитав число величин в табл. 1.2, включая единичные значения на диагонали. Количество оцениваемых величин или параметров равно числу путей[2]: для моделей, представленных на рис. 2.1 и 2.2, — 12, а для модели, представленной на рис. 2.3, — 13. Таким образом, для первых двух моделей число степеней свободы равно 9 (21 – 12 = 9), а для третьей модели — 8 (21 – 13 = 8).

Вероятность того, что критерий хи-квадрат является значимым по двустороннему критерию, меньше 0,05 для первых двух моделей[3], что указывает на то, что ни одна из них не согласуется с данными должным образом. Критерий хи-квадрат для третьей модели меньше, чем для других, и это позволяет сделать вывод о том, что данная модель лучше всего соответствует экспериментальным данным[4].

Критерий различий хи-квадрат для вложенных моделей

Модель, имеющая больше оцениваемых величин или параметров, будет, как правило, лучше соответствовать данным, чем модель, у которой большинство параметров уже жестко фиксированы и не подлежат оценке, т.е. выбору оптимального значения в ходе анализа. В крайнем случае модель, имеющая максимально возможное число оцениваемых параметров ($n(n+1)/2$), фактически представляет собой сами экспериментальные данные и потому полностью им соответствует. В этом случае величина хи-квадрат будет равна нулю и будет отражать абсолютную согласованность

---

[1] Эта величина определяется исходя из того, что число корреляций равно числу пар, которые можно составить из n элементов $n(n-1)/2$, а число вариаций этих переменных равно n, в итоге получается указанная сумма $n(n+1)/2$.

[2] Т.е. числу изображенных на рисунке стрелок, включая те, которые соответствуют специфической (ошибочной) дисперсии.

[3] Нулевая гипотеза утверждает, что модель соответствует экспериментальным данным (влияние других закономерных событий, не включенных в путевую диаграмму, нулевое). Поскольку уровень значимости меньше 0,05, нулевую гипотезу следует отвергнуть. В результате мы приходим к выводу о том, что существуют другие закономерные события, значимо влияющие на экспериментальные данные, и поэтому должны сделать вывод о том, что модель, представляемая путевой диаграммой, не соответствует экспериментальным данным.

[4] Т.е. в этом случае вероятность того, что нулевая гипотеза верна, больше 0,05, и мы ее принимаем.

модели с данными. Одна из задач в социальных науках и психологии состоит в том, чтобы найти модель, содержащую меньше всего свободно оцениваемых параметров, поскольку она обеспечит самое простое и экономное объяснение данных. Из трех рассматриваемых моделей больше всего свободных параметров имеет двухфакторная модель с коррелирующими факторами. Эта модель, по всей видимости, лучше всего согласуется с данными, поскольку имеет наименьшую величину хи-квадрат[1]. Значимо или нет различаются между собой две модели, также можно определить с помощью критерия хи-квадрат: если одна из сравниваемых моделей получается из другой путем исключения каких-либо параметров из числа оцениваемых[2].

Например, степень соответствия экспериментальным данным двухфакторной модели с коррелирующими факторами можно сравнить с таковой для двухфакторной модели с независимыми факторами, считая, что вторая модель получена из первой в результате удаления корреляции между факторами. В первом случае она оценивается, а во втором — жестко фиксируется равной нулю. Двухфакторная модель с независимыми факторами называется вложенной моделью, поскольку она получается из двухфакторной модели с коррелирующими факторами путем удаления (приравнивания к нулю) корреляции между двумя факторами. Если величина критерия для модели с бóльшим числом свободных параметров значимо меньше величины критерия для вложенной модели, в которой для каких-то из этих параметров были фиксированы значения, то можно считать, что объемлющая модель лучше согласуется с данными. Если величины критериев обеих моделей значимо не различаются, то модель с меньшим числом свободно оцениваемых параметров является более простым и экономным представлением данных. Однофакторную и двухфакторную с коррелирующими факторами модели также можно сравнивать между собой на предмет соответствия данным. В этом случае полагаем, что свободная корреляция между факторами в однофакторной модели заменена фиксированной корреляцией, равной 1. Сравнить однофакторную модель с двухфакторной моделью с независимыми факторами нельзя, так как они содержат одинаковое число параметров[3].

---

[1] Это утверждение не совсем корректно, ибо сравнение значений хи-квадрат должно учитывать соответствующее значение степеней свободы, что, как мы увидим, и делается ниже. Поэтому данное утверждение о том, что третья модель лучше, потому что значения хи-квадрат меньше, мы полагаем, требует уточнения, поскольку в третьей модели число степеней свободы также меньше.

[2] Т.е. жесткого фиксирования этих параметров.

[3] Вернее было бы сказать, что однофакторную модель и двухфакторную модель с независимыми факторами нельзя сравнивать между собой по критерию различий, так как ни одна из них не является вложенной в другую, т.е. из одной модели, жестко фиксировав какой-либо свободный параметр, нельзя получить вторую.

Таблица 2.2. **Критерий различий хи-квадрат для сравнения двух пар моделей**

Сравнения	Разность критериев хи-квадрат	Разность числа степеней свободы
1 *vs* 3	96,73 – 46,15 = 50,58	9 – 8 = 1
2 *vs* 3	58,78 – 46,15 = 12,63	9 – 8 = 1

Чтобы проверить, различаются ли степени соответствия моделей экспериментальным данным, из величины хи-квадрат вложенной модели вычитается величина хи-квадрат модели с бóльшим числом оцениваемых параметров и получается разность хи-квадратов. Поскольку на величину хи-квадрат оказывает влияние число степеней свободы модели, разница хи-квадратов должна учитывать разницу в числе степеней свободы обеих моделей. Из числа степеней свободы вложенной модели вычитается число степеней свободы модели с бóльшим числом параметров и получается разница числа степеней свободы. Статистическая значимость разницы хи-квадратов определяется по таблице критических значений хи-квадрат с числом степеней свободы, равным разнице числа степеней свободы двух моделей.

В табл. 2.2 представлены эти вычисления для сравнения двухфакторной модели, в которой факторы коррелируют, с одной стороны, и двухфакторной модели с независимыми факторами и однофакторной моделью — с другой.

Разность числа степеней свободы в обоих сравнениях равна 1 (9 – 8 = 1). Разность хи-квадратов для первого сравнения равна 50,58, для второго — 12,63. В случае двустороннего критерия хи-квадрат с одной степенью свободы, чтобы быть статистически значимой на уровне 0,05, разность хи-квадратов должна быть больше или равна 3,84. Поскольку обе разности хи-квадратов больше, чем 3,84, двухфакторная модель с коррелирующими факторами статистически значимо лучше согласуется с данными, чем обе другие модели — двухфакторная с независимыми факторами и однофакторная.

Среднеквадратичная ошибка приближения (аппроксимации)

Вторая мера согласованности, приведенная в последней информационной строке каждой диаграммы путей на рис. 2.1—2.3, — среднеквадратичная ошибка приближения (RMSEA). Проблема применения критерия хи-квадрат для конфирматорного факторного анализа состоит в том, что он тем вероятнее будет давать статистическую значимость, соответствующую необходимости отвергнуть нулевую гипотезу, чем больше объем выборки. В общем

случае, большие выборки предпочтительнее меньших по объему, поскольку позволяют оценивать параметры более точно. В нашем иллюстративном примере эта проблема не стоит, поскольку выборка объемом в 100 человек не считается большой. Но в других случаях критерий хи-квадрат, статистически незначимый на выборке такого размера, на выборках большего объема при исследовании одной и той же модели будет значимым, что потребует отвергнуть нулевую гипотезу. Из-за подобной проблемы исследователи разработали другие меры согласия, в меньшей степени зависящие от объема выборки. Одной из них является среднеквадратичная ошибка приближения. Значения ниже 0,100 считаются показателем хорошего согласия с данными (J.C.Loehlin, 1998). Значение этого статистического показателя для всех трех моделей превосходит 0,100, что может означать, что ни одна из них не согласуется с данными в достаточной степени. В нашем случае и хи-квадрат, и среднеквадратичная ошибка приближения позволяют сделать сходные выводы о том, что ни одна из трех моделей не обеспечивает хорошего согласия с экспериментальными данными[1].

Отчет о результатах

Здесь можно привести только краткое описание возможных сведений, которые могут быть включены в отчет о результатах анализа для нашего примера: «Было проведено сравнение статистического соответствия экспериментальным данным однофакторной, двухфакторной с независимыми факторами и двухфакторной с коррелирующими факторами моделей путем проведения конфирматорного факторного анализа по методу оценки максимального правдоподобия с помощью программы LISREL 8.51. В табл. 2.3 приведены результаты двух способов измерения согласованности каждой из трех моделей с экспериментальными данными. Оба измерения показывают, что ни одна из первых двух моделей не обеспечивает удовлетворительного согласия с экспериментальными данными, так как величина хи-квадрат является статистически значимой, а величина среднеквадратичной ошибки приближения превосходит 0,100. Критерий различий хи-квад-

[1] Делая заключение о том, что критерии RMSEA и хи-квадрат не противоречат друг другу во всех трех моделях, автор не совсем последователен. Ибо, согласно принятому их уровню значимости 0,05, критерий хи-квадрат можно считать преодолевшим необходимую границу (хотя и совсем незначительно) для того, чтобы принять нулевую гипотезу, т.е. в этом случае результаты критериев RMSEA и хи-квадрат приводят к разным выводам. Однако исходя из того, что вычисленное значение хи-квадрат в случае третьей модели отличается от граничного 0,05 в очень малой степени, то, делая выбор принимаемой гипотезы, следует руководствоваться критерием RMSEA и принять альтернативную гипотезу.

Таблица 2.3. **Две меры согласия для трех моделей**

Модель	χ^2	*df*	*p*	*RMSEA*
Однофакторная	96,73	9	0,001	0,314
Некоррелированная двухфакторная	58,78	9	0,001	0,236
Коррелированная двухфакторная	46,15	8	0,052	0,167

рат показывает, что двухфакторная модель с коррелирующими факторами гораздо лучше соответствует данным, чем двухфакторная модель с независимыми факторами ($\chi^2 = 50{,}58$, $df = 1$, $p < 0{,}001$) или однофакторная модель ($\chi^2 = 12{,}63$, $df = 1$, $p < 0{,}001$), и эти различия статистически значимы. Стандартизированные оценки максимального правдоподобия для параметров этой модели приведены на рис. 2.3».

Процедура в программе LISREL

Существует несколько различных программ для проведения конфирматорного факторного анализа. Здесь ограничимся демонстрацией процедуры в одной из наиболее популярных программ — LISREL. Для того чтобы провести хотя бы один из показанных в данной главе анализов, необходимо запустить программу. Далее надо создать синтаксический файл, содержащий командные строки на встроенном языке программы. Чтобы сделать это, выбираем пункт **File** из строки меню в верхней части окна программы LISREL, что вызывает ниспадающее меню. В этом меню выбираем пункт **New**, который открывает диалоговое окно **New**. Затем выбираем пункт **Syntax only**, открывающий окно **Syntax**, в котором набираем наши команды. Закончив набор команд, выбираем пункт **File**, а в ниспадающем контекстном меню — **Run LISREL**. Если команды набраны неверно, то программа покажет окно выдачи результатов, в котором будет указано, где допущены ошибки. Если команды были набраны правильно, то программа, прежде всего, отобразит диаграмму путей. Чтобы посмотреть на другие результаты, сопровождающие эту диаграмму, в строке меню в верхней части окна программы выбираем пункт **Window**, а затем — файл с расширением ***.OUT**.

Процедура LISREL для однофакторной модели

Необходимо набрать приведенные ниже жирным шрифтом команды в синтаксический файл (командный файл программы) и

запустить его, в результате получатся результаты для однофакторной модели.

```
CFA: I factor
DAta NInputvar=6 NObserv=100
LAbels
Anxious Tense Calm Depressed Useless Happy
KMatrix
 1,00
 0,74   1,00
-0,50  -0,40   1,00
 0,22   0,30  -0,37   1,00
 0,28   0,39  -0,43   0,65    1,00
-0,25  -0,54   0,41  -0,74   -0,53   1,00
MOdel NXvar=6 NKvar=1 PHi=FLxed TDelta=Diagonal
FRee LX (1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1)
STartval 1 PHi(1,1)
LKvar
Distress
PDiagram
OUtput
```

Опишем кратко, что делают эти команды. Для получения дальнейших подробностей необходимо воспользоваться справочной системой программы, выбрав пункт **Help** в строке меню, или обратиться к руководству по работе с программой LISREL. Полезно дать краткое название списку команд, или программе, как это сделано выше. Первые два символа в названии программы не должны быть следующими друг за другом буквами D и A, поскольку это двухбуквенное сочетание является ключевым словом в тезаурусе LISREL для описания параметров данных (DAta).

Ключевые слова команд могут состоять из символов в верхнем или нижнем регистре, а также из любой комбинации символов в верхнем и нижнем регистрах (прописными или строчными буквами, а также комбинациями прописных и строчных букв). Мы следуем общепринятому соглашению изображать ключевые слова символами в верхнем регистре (прописными буквами). Команды могут состоять из последовательности двухбуквенных имен. Однако поскольку LISREL игнорирует все буквы, которые непосредственно следуют за двумя первыми буквами имени, мы использовали буквы в нижнем регистре (строчные буквы), чтобы дать больше информации о том, что означают имена команд и что эти команды делают.

Строка, начинающаяся с **DA**, показывает, что число (Number) входных (Input variables) переменных, которые программа должна считать, равно 6, а число (Number) наблюдений (Observations), или объектов, равно 100. Чтобы не вводить сырые данные, можно

в качестве данных ввести корреляционную матрицу. В этом случае необходимо дополнительно указать в командном файле число наблюдений, на которых основана данная корреляционная матрица. Хотя число наблюдений в нашем случае было 9, мы изменили это значение до 100, чтобы увеличить вероятность того, что данные значимо отличаются от проверяемой модели. Возможность вводить данные в форме корреляционной матрицы очень полезна при анализе данных, доступ к которым в их исходном виде затруднен или невозможен, подобно данным опубликованных исследований.

Строка, начинающаяся с **LA**, показывает, что мы собираемся присвоить метки (Labels) шести входным переменным. Хотя присваивание имен не является обязательным, полезно это сделать как напоминание о том, что представляют из себя переменные. Сами метки вводятся в следующей строке. В окне вывода программы отображаются только первые восемь символов имени каждой переменной, так что название «Подавленный» (**Depressed**) будет сокращено до **Depresse**.

Строка, начинающаяся с **KM**, указывает на то, что данные должны считываться в виде **к**орреляционной **м**атрицы. Следующие шесть строчек представляют нижний треугольник корреляционной матрицы, приведенной в табл. 1.2.

Строка, начинающаяся с **MO**, описывает спецификации тестируемой **мо**дели. Число (Number) переменных типа **X**, или пунктов (**NX**), равно 6. Число (Number) переменных типа **K**, или факторов (**NK**), равно 1.

Следующую команду трудно описать кратко. Вместе с последующей командой она позволяет стандартизировать дисперсию переменной типа **K**, или фактора, и сделать ее равной 1. Матрица дисперсий-ковариаций переменных типа **K**, или факторов, называется **PHi** и является фиксированной (**FIxed**), так, что ее элементы могут быть заданы непосредственно. Именно это и делает строка, начинающаяся с **ST**, которая устанавливает начальное (**st**arting) значение (value) фактора, равным единице. Хотя в нижнетреугольной Phi-матрице всего один элемент, его необходимо задать явно, для чего используются числа в скобках. Первое число задает номер строки, в которой расположен элемент, второе — номер столбца, в котором расположен данный элемент.

Последняя команда в строке, начинающейся с **MO**, устанавливает, что нижнетреугольная корреляционная матрица ошибок шести пунктов опросника (переменных), называемая Theta **Delta**, является диагональной (т.е. имеет отличные от нуля элементы только на диагонали). Данная модель предполагает, что ошибки шести пунктов не коррелируют друг с другом.

Строка, начинающаяся со спецификации **FR**, определяет, какие из шести пунктов, теперь называемых Lamda переменными

типа **X**, должны быть свободными (**f**ree), чтобы их факторные нагрузки можно было оценить. В данной модели предполагается, что все шесть пунктов коррелируют с единственным фактором, что отображено в виде одного столбца с шестью строками. Положение этих пунктов указывается с помощью двух чисел в круглых скобках. Первое число отвечает номеру строки, а второе — номеру столбца, в котором находится данный элемент.

Строка, начинающаяся с **LK**, не является обязательной и позволяет задать имя переменной типа **K**(**L**abel **K**), или фактору, которое указываем в следующей строке. Мы назвали этот фактор **Distress** (душевное страдание).

Строка, начинающаяся с **PD**, дает нам диаграмму путей (**p**ath **d**iagram), изображенную на рис. 2.1.

Последняя строка, начинающаяся с **OU**, выдает результат, отображаемый в окне вывода (**ou**tput) программы.

Вывод результатов в LISREL для однофакторной модели

Опишем только часть информации, отображенной в окне вывода программы. LISREL начинает с того, что воспроизводит командный файл, который мы запустили, а вслед за ним — матрицу взаимосвязей (называемую ковариационной матрицей) с метками (именами) переменных. Эта часть окна вывода программы здесь не приводится.

Следующая часть окна вывода программы показывает, какие параметры модели должны быть оценены путем перечисления и нумерации их, что отражено в табл. 2.4. В данной модели будут оцениваться 12 параметров. **LAMBDA-X** обозначают нагрузки, или «ламбды», шести пунктов, или переменных типа X, по фактору

Таблица 2.4. **Выводимое LISREL описание параметров однофакторной модели**

```
Parameter Specifications

     LAMBDA-X

              Distress
              --------
  Anxious            1
    Tense            2
     Calm            3
Depresse             4
 Useless             5
    Happy            6

     THETA-DELTA

          Anxious    Tense     Calm Depresse Useless    Happy
          -------  -------  ------ -------- -------  -------
                7        8       9       10       11       12
```

типа K или ksi. **THETA-DELTA** содержит дисперсии ошибок или специфичностей шести пунктов.

Следующая часть окна вывода отображает оценки максимального правдоподобия этих параметров и приведена в табл. 2.5. Так, нагрузка «Тревожность» по единственному фактору составляет 0,43, как показано под заголовком **LAMBDA-X**. Другими словами, при-

Таблица 2.5. **Выводимые LISREL оценки максимального правдоподобия для параметров однофакторной модели**

Number of Irerations = 20

LISREL Estimates (Maximum Likelihood)

LAMBDA-X

	Distress
Anxious	0,43
	(0,10)
	4,16
Tense	0,59
	(0,10)
	6,01
Calm	-0,54
	(0,10)
	-5,43
Depresse	0,81
	(0,09)
	9,17
Useless	0,71
	(0,09)
	7,61
Happy	-0,84
	(0,09)
	-9,64

PHI

Distress
1,00

THETA-DELTA

Anxious	Tense	Calm	Depresse	Useless	Happy
0,82	0,66	0,71	0,34	0,50	0,29
(0,12)	(0,10)	(0,11)	(0,07)	(0,08)	(0,07)
6,79	6,45	6,58	4,79	5,91	4,27

Squared Multiple Correlations for X - Variables

Anxious	Tense	Calm	Depresse	Useless	Happy
0,18	0,34	0,29	0,66	0,50	0,71

мерно $0,43^2$, или 0,18, дисперсии тревожности объясняется данным фактором, оставляя 0,82 в качестве ошибки или необъясненной дисперсии. Значение в круглых скобках, приведенное непосредственно под нагрузкой, представляет собой стандартную ошибку, которая в нашем случае равна 0,10, а ниже — значение критерия Стьюдента (4,16) для определения значимого отличия этой нагрузки от нуля. Для выборки с объемом, равным 100, величина двустороннего критерия Стьюдента должна быть равна +1,98 или выше, чтобы быть значимой на уровне 0,05. Поскольку 4,16 > 1,98, можно сделать вывод о том, что данная нагрузка является статистически значимой с достоверностью не менее 0,95 для двустороннего критерия Стьюдента.

Дисперсия данного фактора была установлена равной 1,00, как показано под заголовком **PHI**.

Дисперсия ошибок для шести пунктов приведена под заголовком **THETA-DELTA**. Дисперсия ошибки «тревожности» равна 0,82.

Когда пункт имеет нагрузку только по одному фактору, как в настоящем примере, квадрат коэффициента множественной корреляции (**Squared Multiple Correlation**), приводимый в конце табл. 2.5, равен квадрату нагрузки этого пункта. Для «тревожности» он равен $0,43^2$, или 0,18.

Последняя бóльшая часть выводимых результатов представляет собой статистики согласия для проверяемой модели, как это показано в табл. 2.6. Величина хи-квадрат, приведенная на рис. 2.1,

Таблица 2.6. **Выводимые LISREL статистики согласия для однофакторной модели**

```
Goodness-of-Fit Statistics

Degrees of Freedom = 9
Minimum Fit Function Chi-Square = 128,43 (P = 0,0)
Normal theory Weighted Least Squares Chi-Square = 96,73 (P = 0,00)
Estimated Non-centrality Parameter (NCP) = 87,73
90 Percent Confidence Interval for NCP = (59,78; 123,13)

Minimum Fit Function Value = 1,30
Population Discrepancy Function Value (F0) = 0,89
90 Percent Confidence Interval for F0 = (0,60; 1,24)
Root Mean Square Error of Approximation (RMSEA) = 0,31
90 Percent Confidence Interval for RMSEA = (0,26; 0,37)
P-Value for Test of Close Fit (RMSEA < 0,05) = 0,00

Expected Cross-Validation Index (ECVI) = 1,22
90 Percent Confidence Interval for ECVI = (0,94; 1,58)
ECVI for Saturated Model = 0,42
ECVI for Independence Model = 3,36

Chi-Square for Independence Model with 15 Degrees of Freedom = 321,02
Independence AIC = 333,02
Model AIC = 120,73
```

```
Saturated AIC = 42,00
Independence CAIC = 354,65
Model CAIC = 163,99
Saturated CAIC = 117,71

Normed Fit Index (NFI) = 0,60
Non-Normed Fit Index (NNFI) = 0,35
Parsimony Normed Fit Index (PNFI) = 0,36
Comparative Fit Index (CFI) = 0,61
Incremental Fit Index (IFI) = 0,62
Relative Fit Index (RFI) = 0,33

Critical N (CN) = 17,70

Root Mean Square Residual (RMR) = 0,14
Standardized RMR = 0,14
Goodness-of-Fit Index (GFI) = 0,75
Adjusted Goodness-of-Fit Index (AGFI) = 0,43
Parsimony Goodness-of-Fit Index (PGFI) = 0,32
```

является второй по порядку в списке выведенных величин и называется **Normal Theory Weighted Least Squares Chi-Square**. Ее значение равно 96,73. Число степеней свободы указано двумя строками выше и равно 9. Среднеквадратичная ошибка приближения (Root Mean Square Error Of Approximation — RMSEA) отображается на девятой по счету строке и равна 0,31. Информацию о других мерах согласия можно найти в книге J. C. Loehlin (1998).

Процедура LISREL для двухфакторной модели с независимыми переменными

Программа на встроенном командном языке пакета LISREL для получения результатов в случае двухфакторной модели с независимыми факторами выглядит следующим образом:

```
CFA: 2 unrelated factors
DAta NInputvar=6 NObserv=100
LAbels
Anxious Tense Calm Depressed Useless Happy
KMatrix
 1,00
 0,74   1,00
-0,50  -0,40   1,00
 0,22   0,30  -0,37   1,00
 0,28   0,39  -0,43   0,65   1,00
-0,25  -0,54   0,41  -0,74  -0,53  1,00
MOdel NXvar=6 NKvar=2 PHi=FIxed TDelta=Diagonal
FRee LX (1,1) LX (2,1) LX (3,1) LX(4,2) LX(5,2) LX(6,2)
STartval 1 PHi(1,1) PHi(2,2)
LKvar
```

Anxiety Depression
PDiagram
OUtput

Прокомментируем только основные различия между данной и предыдущей программой.

Модель содержит две переменные типа **K** (**var**iable), или фактора, поэтому параметру **NKvar** в команде **MOdel** присвоено значение 2.

Первые три **LX**-переменные («Тревожный», «Напряженный» и «Спокойный») имеют нагрузки по первому фактору, тогда как вторая тройка **LX**-переменных («Подавленный», «Бесполезный» и «Веселый») имеет нагрузки по второму фактору. Положение этих элементов в матрице обозначают двумя числами в круглых скобках, первое указывает номер переменной, или строку, а второе — номер фактора, или столбец. Поэтому число **2**, стоящее после запятой, указывает на нагруженность переменной по второму фактору.

Дисперсиям двух факторов — **PHi(1,1)** и **PHi(2,2)** — присвоены значения 1. Поскольку параметр **PHi** является фиксированным (**Fixed**), коэффициенту ковариации между двумя факторами, представленному элементом **PHi(2,1)**, присвоено фиксированное значение, равное нулю.

Двум переменным типа **K**, или факторам, присвоены имена **Anxiety** (Тревога) и **Depression** (Депрессия) соответственно.

Вывод результатов в LISREL для двухфакторной модели с независимыми переменными

Диаграмма путей для этой модели представлена на рис. 2.2. В данном подразделе будут представлены и обсуждены лишь некоторые из выводимых программой результатов для данной модели. В табл. 2.7 приведены оцениваемые параметры для рассматриваемой модели. Их число равно 12, как и для предыдущей модели.

Таблица 2.7. **Выводимое LISREL описание параметров для двухфакторной модели с независимыми перемеиными**

LAMBDA-X

	Anxiety	Depressi
Anxious	1	0
Tense	2	0
Calm	3	0
Depresse	0	4
Useless	0	5
Happy	0	6

THETA-DELTA

Anxious	Tense	Calm	Depresse	Useless	Happy
7	8	9	10	11	12

В табл. 2.8 приведены оценки максимального правдоподобия нагрузок первых трех пунктов по первому фактору и вторых трех пунктов по второму фактору.

Таблица 2.8. **Выводимые LISREL оценки максимального правдоподобия для некоторых из параметров двухфакторной модели с независимыми переменными**

```
Number of Iterations = 7
LISREL Estimates (Maximum Likelihood)

      LAMBDA-X

          Anxiety    Depressi
          -------    --------
Anxious     0,96        - -
           (0,10)
            9,42

  Tense     0,77        - -
           (0,10)
            7,58

   Calm    -0,52        - -
           (0,10)
           -5,15

Depresse     - -        0,95
                       (0,09)
                        10,88

Useless      - -        0,68
                       (0,09)
                         7,25

  Happy      - -       -0,78
                       (0,09)
                        -8,43

        PHI

   Note: This matrix is diagonal.

      Anxiety  Depressi
      -------  --------
         1,00     1,00
```

Процедура LISREL для коррелированной двухфакторной модели с зависимыми переменными

Программа для получения результатов в случае коррелированной двухфакторной модели выглядят следующим образом:

```
CFA: 2 related factors
DAta NInputvar=6 NObserv=100
LAbels
Anxious Tense Calm Depressed Useless Happy
```

```
KMatrix
 1,00
 0,74   1,00
-0,50  -0,40   1,00
 0,22   0,30  -0,37   1,00
 0,28   0,39  -0,43   0,65   1,00
-0,25  -0,54   0,41  -0,74  -0,53  1,00
MOdel NXvar=6 NKvar=2 PHi=FIxed TDelta=DIagonal
FRee LX (1,1) LX (2,1) LX (3,1) LX(4,2) LX(5,2) LX(6,2)
STartval 1 PHi(1,1) PHi(2,2)
FRee PHi (2,1)
LKvar
Anxiety Depression
PDiagram
OUtput
```

Прокомментируем только главное отличие данной программы от предыдущей. Чтобы разрешить ковариацию двух переменных типа **K**, или факторов, необходимо «освободить» соответствующий параметр — **FRee PHi(2,1)**.

Вывод результатов в LISREL для коррелированной двухфакторной модели с зависимыми переменными

Диаграмма путей для этой модели приведена на рис. 2.3. Число оцениваемых параметров в данном случае равно 13, поскольку нам приходится оценивать ковариацию между двумя факторами, как это показано под заголовком **PHI** в табл. 2.9.

Таблица 2.9. **Выводимое LISREL описание параметров для коррелированной двухфакторной модели с зависимыми переменными**

```
Parameter Specifications

      LAMBDA-X

          Anxiety   Depressi
          -------   --------
 Anxious        1          0
   Tense        2          0
    Calm        3          0
Depresse        0          4
 Useless        0          5
   Happy        0          6

        PHI

          Anxiety   Depressi
          -------   --------
 Anxiety        0
Depressi        7          0
```

THETA-DELTA

Anxious	Tense	Calm	Depresse	Useless	Happy
8	9	10	11	12	13

Оценка максимального правдоподобия коэффициента ковариации между двумя факторами приведена в табл. 2.10 и составляет 0,48.

Таблица 2.10. **Выводимые LISREL оценки максимального правдоподобия некоторых параметров для коррелированной двухфакторной модели со связанными переменными**

Number of Iterations = 11
LISREL Estimates (Maximum Likelihood)

LAMBDA-X

	Anxiety	Depressi
Anxious	0,85	- -
	(0,16)	
	5,22	
Tense	0,87	- -
	(0,16)	
	5,37	
Calm	-0,54	- -
	(0,17)	
	-3,13	
Depresse	- -	0,89
		(0,15)
		5,93
Useless	- -	0,71
		(0,16)
		4,40
Happy	- -	-0,83
		(0,15)
		-5,37

PHI

	Anxiety	Depressi
Anxiety	1,00	
Depressi	0,48	1,00
	(0,16)	
	2,92	

Рекомендуемая литература

Jöreskog, K.G. and Sörbom, D. (1989) *LISREL 7: A Guide to the Program and Applications,* 2nd edn. Chicago, IL: SPSS Inc.

Loehlin, J.C. (1998) *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis,* 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1996) *Applied Multivariate Statistics for the Social Sciences,* 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

Дополнительная литература, рекомендуемая научным редактором

Митина О.В. Структурное моделирование: состояние и перспективы // Вестн. Пермского гос. пед. ун-та. Сер. 1. Психология. — № 2, 2005. — С. 3—15.

Дорфман Л.Я. Путевой анализ как метод интегрального исследования индивидуальности / Л.Я.Дорфман, А.В.Огородников // Вестн. Пермского гос. пед. ун-та. Сер. 1. Психология. — № 2, 2005. — С. 15—30.

Глава 3

КЛАСТЕРНЫЙ АНАЛИЗ

Предисловие научного редактора

В гл. 3 описывается разновидность другого известного метода группировки — иерархического агломеративного кластерного анализа. Кластерный анализ дает группировку на основе совокупного интегрального критерия, в то время как при факторизации задается система взаимосвязанных между собой категорий, позволяющих дифференцировать различения классов по набору критериев. Однако при проведении кластерного анализа исследователь имеет возможность учесть **всю** информацию, в то время как при факторизации необходимо искусственно оставить лишь небольшое число факторов, а значит, пренебречь каким-то объемом информации (иногда эти потери информации составляют более половины). Таким образом, можно говорить о том, что кластерный и факторный анализы не конкурируют, а взаимодополняют друг друга, позволяя более выпукло представить модель.

Еще одним методом выявления группировки переменных является кластерный анализ. Однако, по сравнению с факторным анализом, он значительно реже применяется в социальных науках и психологии. Часто его используют для выявления группировки наблюдений, а не переменных.

Необходимо отметить, что как кластерный, так и факторный анализ с равным успехом могут применяться для определения способа группировки не только наблюдений, но и и переменных.

Как и в случае факторного анализа, существует несколько различных методов проведения кластерного анализа, и в настоящее время все еще нет общепринятого соглашения по поводу того, какой из этих методов является наиболее адекватным. Чтобы иметь возможность сравнить результаты кластерного анализа с результатами факторного анализа, проиллюстрируем последовательность этапов кластерного анализа с помощью тех же самых данных, которые использовались при объяснении принципов проведения факторного анализа, а именно, баллов ответов девяти испытуемых по трем пунктам опросника, связанным с депрессией (A1 — «Тревожный»; A2 — «Напряженный»; A3 — «Спокойный»), и трем пунктам, связанным с тревогой (D1 — «Подавленный»; D2 — «Бесполезный»; D3 — «Счастливый»), приведенных в табл. 1.1.

Меры и матрица сходства

Первым этапом в проведении кластерного анализа является решение вопроса об измерении степени сходства между переменными. Одной из возможных мер сходства является коэффициент корреляции. Чем более сходны значения двух переменных, тем выше величина положительного коэффициента корреляции между этими переменными. Чем более рассогласованы значения переменных, тем больше абсолютное значение отрицательного коэффициента корреляции между ними. Наверное, самой часто используемой мерой сходства является квадрат Евклидова расстояния (т.е. квадрат длины прямолинейного отрезка) между двумя точками, соответствующими векторам переменных. Это просто сумма квадратов разностей значений переменной по всем наблюдениям данной выборки[1]. Квадраты разностей рассматриваются отчасти из-за того, чтобы устранить влияние знака, или направления, этих разностей. Другими словами, разность значений, равная –2, и разность значений, равная +2, считаются одинаковыми. В табл. 3.1 показано вычисление квадрата Евклидова расстояния между значениями баллов пунктов «Тревожный» и «Напряженный», результат которого равен 8.

Таблица 3.1. **Квадрат Евклидова расстояния между баллами пунктов «Тревожный» и «Напряженный»**

Наблюдения[2]	Тревожный	Напряженный	Разность	Квадрат разности
1	2	1	1	1
2	1	2	–1	1
3	3	3	0	0
4	4	4	0	0
5	5	5	0	0
6	4	5	–1	1
7	4	3	1	1
8	3	3	0	0
9	3	5	–2	4
				8

[1] Фактически строится геометрическое пространство, размерность которого равна числу наблюдений. В этом пространстве переменные являются точками с координатами, соответствующими измерению данной переменной по тому или иному наблюдению.

[2] Испытуемые.

Таблица 3.2. **Матрица сходства, содержащая квадраты Евклидовых расстояний между пунктами «Тревожный» и «Подавленный»**

	Тревожный	Напряженный	Спокойный	Подавленный	Бесполезный	Счастливый
Тревожный						
Напряженный	8					
Спокойный	25	31				
Подавленный	18	20	23			
Бесполезный	16	18	21	8		
Счастливый	38	56	15	52	42	

Эти квадраты Евклидовых расстояний обычно вводятся в виде нижнетреугольной матрицы, как показано в табл. 3.2. Видим, что наиболее близкими (сходными) являются пары «Тревожный» — «Напряженный» и «Подавленный» — «Бесполезный», а квадрат Евклидова расстояния между переменными, составляющими эти пары, в обоих случаях равен 8. Наиболее рассогласованными (удаленными) являются переменные «Напряженный» и «Счастливый», квадрат Евклидова расстояния между которыми равен 56.

Метод иерархической агломеративной кластеризации

Одним из наиболее часто используемых методов для формирования групп является метод иерархической агломеративной кластеризации. В методе иерархической агломеративной кластеризации первоначальное число кластеров равно числу переменных. Формирование групп осуществляется последовательно, или иерархически. На первом шаге в один кластер объединяются две переменные, расположенные на кратчайшем расстоянии друг от друга. Например, переменная «Подавленный» может быть объединена с переменной «Бесполезный». На втором шаге возможны два варианта: 1) либо к первому кластеру, содержащему две переменные, добавляется, или агломерируется, третья переменная; 2) либо две другие переменные объединяются и образуют новый кластер. Например, переменная «Тревожный» может быть объединена либо с парой «Подавленный» — «Бесполезный», либо с переменной «Напряженный». На третьем шаге могут быть объединены две переменные, третья переменная может добавляться к существующей группе переменных, или две группы переменных могут сливаться в один кластер. Таким образом, на каждом этапе формируется только один новый кластер. На последнем шаге все переменные группируются в один-единственный кластер.

Метод межгруппового связывания (связывания средних внутри групп)

Имеются различные методы для формирования кластеров на каждом этапе. Одним из наиболее часто используемых является метод межгруппового связывания, или связывания средних внутри групп. На первом шаге в кластер объединяются две переменные, расстояние между которыми минимально. В нашем примере кратчайшее расстояние равно 8, и существует две пары переменных с таким расстоянием. Поэтому первый кластер будет состоять либо из «Тревожный» и «Напряженный», либо из «Подавленный» и «Бесполезный»[1]. Будем считать, что первый кластер образован парой «Подавленный» — «Бесполезный». На втором шаге в качестве критерия для формирования нового кластера используется минимальное среднее расстояние между кластерами. На данном этапе у нас есть только одна группа. Среднее расстояние между этой группой и, например, переменной «Спокойный» представляет собой среднее арифметическое расстояний между следующими парами переменных: «Подавленный» — «Спокойный» и «Бесполезный» — «Спокойный». Это среднее значение равно 22[(23 + 21)/2 = 22]. Таким образом, среднее расстояние между группами, или кластерами, есть среднее арифметическое расстояний между всевозможными парами переменных, одна из которых принадлежит первому, а другая — второму из рассматриваемых кластеров. Поскольку у нас пока только одна переменная в первом кластере («Спокойный») и две переменные во втором кластере («Подавленный» и «Бесполезный»), среднее вычисляется на основе всего лишь двух пар переменных («Спокойный» — «Подавленный» и «Спокойный» — «Бесполезный»). Аналогичным образом вычисляются средние расстояния между данной группой из двух переменных («Подавленный» — «Бесполезный») и каждой из трех остальных переменных («Тревожный», «Напряженный» и «Счастливый»). В табл. 3.3 в виде нижнетреугольной матрицы сходства приведены расстояния между кластерами после осуществления первого шага алгоритма. Так как минимальное расстояние между двумя кластерами по-прежнему равно 8, следующий кластер будет образован переменными «Тревожный» и «Напряженный». Таким образом процесс продолжается до тех пор, пока все переменные не будут сгруппированы в один кластер.

---

[1] Хотя в таблице для обозначения переменных для краткости используется только одно слово, например «Бесполезный», при описаниях в тексте, чтобы более точно выразить смысл анализируемых переменных, мы будем иногда добавлять уточняющие слова.

Таблица 3.3. **Матрица сходства расстояний между кластерами после выполнении первого шага алгоритма кластеризации**

Кластеры	Тревожный	Напряженный	Спокойный	Подавленный — Бесполезный	Счастливый
Тревожный					
Напряженный	8				
Спокойный	25	31			
Подавленный — Бесполезный	17	19	22		
Счастливый	38	56	15	47	

Результаты кластерного анализа пунктов опросника, связанных с депрессией и тревогой, приведены в табл. 3.4. В целях экономии места каждая переменная обозначается двумя символами. Первый символ представляет собой первую букву названия переменной, а второй — ее порядковый номер в списке, состоящем из шести переменных. Таким образом, переменная «Тревожный» обозначена как Т1, «Напряженный» как Н2 и т.д. Первый столбец содержит номера шагов (этапов) алгоритма. Первоначально у нас имеется столько же кластеров, сколько и переменных. На каждом шаге формируется один новый кластер. Во втором столбце показано, какие переменные объединяются в общие кластеры на каждом шаге алгоритма. Круглые скобки обозначают кластер. Переменные, выделенные жирным шрифтом, показывают, какой кластер был образован на данном этапе. Первый кластер образован переменными «Подавленный» и «Бесполезный», второй — «Тревожный» и «Напряженный» и т.д. В третьем столбце приведены расстояния между двумя кластерами, объединенными в одну группу на данном шаге алгоритма. Если кластер состоит более чем из

Таблица 3.4. **Иерархический агломеративный кластерный анализ**

Шаг	Кластеры	Расстояние между кластерами	Число кластеров
0	(Т1) (Н2) (Сп3) (П4) (Б5) (Сч6)		6
1	(Т1) (Н2) (Сп3) **(П4—Б5)** (Сч6)	8	5
2	**(Т1—Н2)** (Сп3) (П4—Б5) (Сч6)	8	4
3	(Т1—Н2) (П4—Б5) **(Сп3—Сч6)**	15	3
4	**(Т1—Н2—П4—Б5)** (Сп3—Сч6)	18	2
5	**(Т1—Н2—П4—Б5—Сп3—Сч6)**	36	1

одной переменной, то указано среднее расстояние между группами. В последнем столбце приведено количество кластеров на каждом шаге. На последнем шаге остается один-единственный кластер, содержащий все шесть переменных.

Последовательность слияния (агломерации)

Результаты кластерного анализа могут быть отображены в виде последовательности слияния (табл. 3.5), полученной с помощью SPSS.

В первом столбце указаны номера этапов анализа, во втором и третьем — номера кластеров, объединяемых на каждом этапе. Вновь образованному кластеру присваивается номер первой по порядку из объединяемых переменных. Так, первый образованный кластер состоит из переменных «Подавленный» и «Бесполезный». «Подавленный» является четвертой по порядку в списке переменных, а «Бесполезный» — пятой. Четвертый создаваемый кластер должен объединить кластер, содержащий переменные «Тревожный» и «Напряженный», с кластером, состоящим из переменных «Подавленный» и «Бесполезный». Переменная «Тревожный» идет первой в списке переменных, а «Подавленный» — четвертой.

В четвертом столбце даны расстояния (средние расстояния) между кластерами. Так, расстояние между кластером, состоящим из единственной переменной «Подавленный», и кластером, состоящим из единственной переменной «Бесполезный», которые объединяются в пару на первом шаге, равно 8. В пятом и шестом столбцах приведены номера шагов, на которых были образованы первый и второй из объединяемых кластеров соответственно. Первые три из образованных кластеров формировались путем объединения отдельных переменных, которые существовали до начала

Таблица 3.5. **Выводимые SPSS результаты последовательности слияния**

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	5	8,000	0	0	4
2	1	2	8,000	0	0	4
3	3	6	15,000	0	0	5
4	1	4	18,000	2	1	5
5	1	3	36,000	4	3	0

кластеризации (на нулевом этапе), поэтому в соответствующих клетках стоят нули. Кластер, созданный на четвертом этапе, объединяет два кластера, один из которых образован на втором этапе («Тревожный» — «Напряженный»), а другой («Подавленный» — «Бесполезный») — на первом этапе. В последнем столбце указан номер следующего этапа, на котором этот кластер объединяется с другим кластером. Так, кластер «Подавленный» — «Бесполезный» объединяется с другим кластером на четвертом этапе.

Дендрограмма

Одним из способов графического представления результатов кластерного анализа является дендрограмма (рис. 3.1). Эта дендрограмма была получена в результате работы SPSS (Dendron — древнегреческое слово, обозначающее дерево, а дендрограмма некоторым образом напоминает структуру ветвей дерева, способность ветвей образовывать новые побеги).

Заголовок имена/метки (Labels) относится к переменным, а не к объектам (case), а номера (Numbers) — к порядку переменных. Первыми объединяются в пару переменные «Подавленный» и «Бесполезный», потом «Тревожный» и «Напряженный», а за ними — «Спокойный» и «Счастливый». Штриховая горизонтальная линия с отмеченными на ней числами от 0 до 25 показывает расстояние (среднее расстояние) между кластерами, причем 1 соответствует наименьшему расстоянию, а 25 — наибольшему. В нашем случае 1 соответствует расстоянию, равному 8, а 25 — расстоянию, равному 36. Таким образом, расстоянию, равному 15, соответствует отметка 7 на нашей относительной шкале [так как $(15 - 8)/(36 - 8)\,24 + 1 = 7{,}00$], в то время как расстоянию 18 будет соответствовать отметка 9,57[$(18 - 8)/(36 - 8)\,24 + 1 = 9{,}57$].

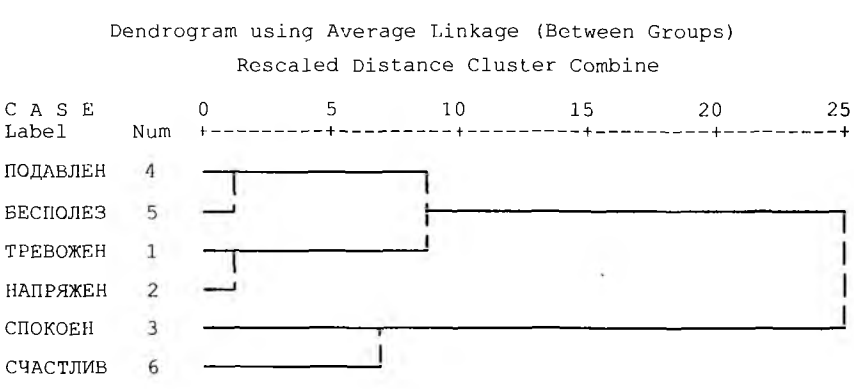


Рис. 3.1. Выводимая SPSS дендрограмма

Диаграмма накопления[1]

Другим способом графического представления результатов кластерного анализа является вертикальная диаграмма накопления, полученная в результате работы SPSS (рис. 3.2). Кому-то крестики напоминают сосульки, свисающие с горизонтальной поверхности. В первом столбце отображено количество кластеров, начиная с конечного решения в виде единого кластера, объединяющего все переменные, и заканчивая начальным этапом кластеризации, на котором были объединены переменные «Подавленный» и «Бесполезный», а остальные переменные представляли свои собственные индивидуальные кластеры. Переменные указаны крестиками во всех строках тех столбцов, которые обозначены их именами. Наличие кластера обозначается крестиком в столбце или столбцах, разделяющих переменные. Таким образом, в нижней строке диаграммы стоит крестик в столбце, отделяющим столбец, соответствующий переменной «Подавленный», от столбца, соответствующего переменной «Бесполезный». По мере того как мы поднимаемся вверх по строкам диаграммы, в каждой последующей строке появляется новый кластер.

Vertical Icicle

Number of clusters	Case										
	СЧАСТЛИВ		СПОКОЕН		БЕСПОЛЕЗ		ПОДАВЛЕН		НАПРЯЖЕН		ТРЕВОЖЕН
1	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X		X	X	X	X	X	X	X
3	X	X	X		X	X	X		X	X	X
4	X		X		X	X	X		X	X	X
5	X		X		X	X	X		X		X

Рис. 3.2. Выводимая SPSS диаграмма накопления

[1] Существующий в русском языке перевод этого термина как «вертикальная древовидная диаграмма» нам кажется неудовлетворительным, так как может обозначать вертикально расположенную дендрограмму. На английском языке эта форма графического представления называется «icicle plot» — «изображение сосульки», так как чем-то напоминает сосульки, увеличивающиеся в диаметре сверху вниз (т.е. накапливающиеся). Именно поэтому мы выбрали именно такой перевод.

Выбор числа кластеров

Так как в данном примере используется всего лишь шесть переменных и число кластеров, получаемых в результате агломерации, мало́, то сделать выбор окончательного числа кластеров, представляющих адекватную информацию о соотношениях между переменными, достаточно легко. По всей видимости, следует остановиться на решениях, содержащих два или три кластера. Можно привести доводы в пользу каждого их этих решений. К сожалению, не предусмотрены статистические критерии, позволяющие осуществить данный выбор. Решение из трех кластеров содержит кластеры, представляющие депрессию, тревогу и положительные чувства, соответственно, в то время как решение из двух кластеров включает кластеры, соответствующие положительным и отрицательным ощущениям. Одним из критериев может быть выбор числа кластеров в точке, где происходит значительный разрыв в величине среднего расстояния между кластерами. Как можно видеть из рис. 3.1, этот разрыв происходит после образования четвертого кластера и проявляется в том, что третий и четвертый кластеры относительно близки по отношению друг к другу и каждый из них примерно одинаково удален от конечного кластера.

Отчет о результатах

Форма отчета о проведении кластерного анализа в определенной степени зависит от целей его проведения. Краткий отчет мог бы выглядеть следующим образом: «Матрица сходства квадратов Евклидовых расстояний, построенная по результатам ответов на шесть пунктов опросника, была подвергнута процедуре иерархического агломеративного кластерного анализа с использованием метода межгруппового связывания для объединения кластеров. Дендрограмма анализа представлена на рис. 3.1. В решении, содержащем два кластера, первый кластер состоит из пунктов опросника, соответствующих негативному воздействию, а второй — из пунктов, соответствующих положительному воздействию».

Реализация процедуры кластерного анализа в программе SPSS для Windows

Для проведения кластерного анализа, описанного в данной главе, следует использовать следующий алгоритм.

Введите данные, представленные в табл. 1.1, в окно **Редактора данных (Data Editor)**, как показано на рис. 1.2. Если эти данные

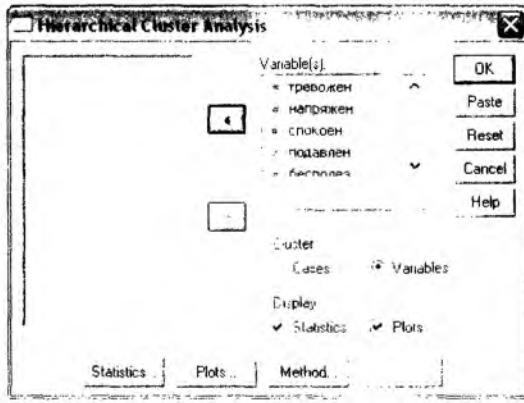


Рис. 3.3. Диалоговое окно: **Иерархический кластерный анализ (Hierarchical Cluster Analysis)**

были сохранены в файле, откройте его, используя команды **Файл (File)**, **Открыть (Open)**, **Данные (Data)** и выбрав в диалоговом окне **Открытие файла (Open File)** имя файла и команду **Открыть (Open)**.

В строке меню в верхней части окна программы выберите пункт **Анализ (Analyze)**, в открывающемся ниспадающем меню — пункт **Классификация (Classify)**, а затем — **Иерархическая кластеризация (Hierarchical Cluster)**, что приведет к открытию диалогового окна **Иерархический кластерный анализ (Hierarchical Cluster Analysis)**, рис. 3.3.

Выберите переменные, начиная с «тревожен» и заканчивая «счастлив», и нажмите на первую из кнопок ▸, чтобы переместить их в список **Переменные (Variable(s))**:

Установите переключатель **Переменные (Variables)** в группе **Кластеризация (Cluster)**, чтобы провести кластерный анализ переменных, а не объектов.

Нажатие кнопки **Статистики (Statistics)** приводит к открытию диалогового окна **Иерархический кластерный анализ: Статистики (Hierarchical Cluster Analysis: Statistics)**, изображенного на рис. 3.4.

Установите флажок **Матрица сходства (Proximity Matrix)**, чтобы получить матрицу сходства, подобную приведенной в табл. 3.2, с тем отличием, что это будет квадратная матрица, в которой элементы представлены десятичными числами, округленными до третьего знака после запятой. Флажок **Последовательность слияния (Agglomeration Schedule)** установлен по умолчанию. Это позволяет вывести последовательность слияния, представленную в табл. 3.5.

Нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Иерархический кластерный анализ (Hierarchical Cluster Analysis)**, показанное на рис. 3.3.

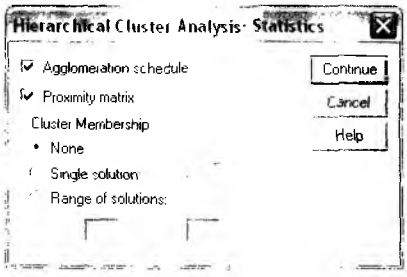


Рис. 3.4. Диалоговое окно **Иерархический кластерный анализ: Статистики (Hierarchical Cluster Analysis: Statistics)**

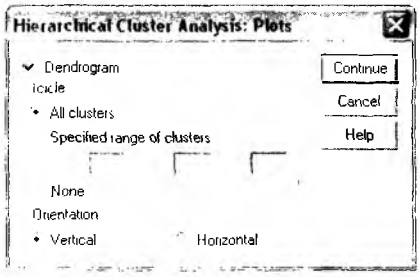


Рис. 3.5. Диалоговое окно **Иерархический кластерный анализ: Диаграммы (Hierarchical Cluster Analysis: Plots)**

Нажатие кнопки **Диаграммы (Plots)** открывает диалоговое окно **Иерархический кластерный анализ**: **Диаграммы (Hierarchical Cluster Analysis**: **Plots)**, показанное на рис. 3.5.

Установка флажка **Дендрограмма (Dendrogram)** позволяет включить в результаты дендрограмму, изображенную на рис. 3.1. В группе **Диаграмма накопления (Icicle)** уже установлен соответствующий переключатель, позволяющий вывести диаграмму накопления, изображенную на рис. 3.2.

Для возврата в диалоговое окно **Иерархический кластерный анализ (Hierarchical Cluster Analysis)** нажмите на кнопку **Продолжить (Continue)**.

Нажатие кнопки **Метод (Method)** открывает диалоговое окно **Иерархический кластерный анализ**: **Метод (Hierarchical Cluster Analysis: Method)**, показанное на рис. 3.6. Это диалоговое окно по-

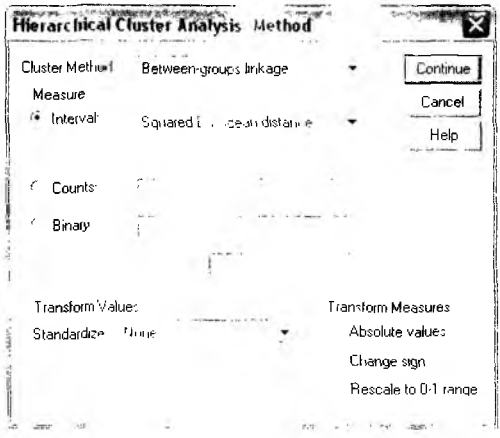


Рис. 3.6. Диалоговое окно **Иерархический кластерный анализ: Метод (Hierarchical Cluster Analysis: Method)**

казывает, какие процедуры будут выполнены, если оставить значения параметров без изменения. Это так называемые значения параметров по умолчанию, и именно они нам нужны. В списке **Метод кластеризации (Cluster Method)** пунктом по умолчанию является **Межгрупповое связывание (Between-groups linkage)**, соответствующее методу связывания средних внутри групп. **Мерой сходства (Measure)** по умолчанию является **квадрат Евклидова расстояния (Squared Euclidean Distance)**.

Нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Иерархический кластерный анализ (Hierarchical Cluster Analysis)**.

Нажмите **OK**, чтобы провести требуемый анализ.

Рекомендуемая литература

Diekhoff, G. (1992) *Statistics for the Social and Behavioral Sciences.* Dubuque, IA: Wm. C. Brown.

Hair, J.F., Jr., Anderson, R.E., Tatham, R.L. and Black, W.C. (1998) Multivariate *Data Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Дополнительная литература, рекомендуемая научным редактором

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

Кулаичев А.П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

ЧАСТЬ II
ОБЪЯСНЕНИЕ ДИСПЕРСИИ КОЛИЧЕСТВЕННОЙ ПЕРЕМЕННОЙ

Глава 4
ПОШАГОВАЯ МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Предисловие научного редактора

В ч. II рассматриваются варианты множественной регрессии, позволяющие определить, какие количественные независимые переменные (предикторы) и их взаимодействия наиболее сильно связаны с количественной переменной-откликом.

В гл. 4 в качестве предикторов последовательно выбираются независимые переменные, согласно тому, какой вклад в дисперсию зависимой переменной они вносят (начиная с максимального). Выбор осуществляется автоматически по ходу выполнения программы без учета содержательных интерпретаций.

Множественная регрессия представляет собой метод определения того, какая доля дисперсии непрерывной, предпочтительно нормально распределенной, переменной может быть объяснена двумя или более переменными с учетом связей между этими переменными. Предположим, необходимо выяснить, какие переменные теснее всего связаны с академической успеваемостью детей. Вероятно, что с академической успеваемостью связан целый ряд факторов, включая интеллект ребенка, его интерес к учебе, заинтересованность родителей в академических успехах своего ребенка, заинтересованность учителя в академической успеваемости ребенка и т.д. Ожидается, что эти факторы связаны друг с другом таким образом, что более умный ребенок больше заинтересован в учебе и может иметь более заинтересованных в его школьных успехах родителей и учителей. Множественная регрессия позволяет выявить, какая доля дисперсии академической успеваемости детей объясняется этими факторами, с учетом их взаимозависимости, а также выяснить, отличаются ли статистически значимо объясняемые этими факторами доли дисперсии от случайных.

Существует два основных варианта использования множественной регрессии. Первый из них состоит в том, чтобы определить,

какие переменные объясняют самые большие и статистически значимые доли дисперсии зависимой переменной и чему равны эти доли. Чаще всего это осуществляется с помощью метода *пошаговой множественной регрессии*, являющегося предметом обсуждения в данной главе. Другой способ состоит в том, чтобы последовательно определить, объясняет ли отдельная переменная или группа переменных значительную часть дисперсии зависимой переменной, а также установить величины этих долей дисперсии. Если это не первый этап процедуры, то необходимо учитывать переменные, выделенные на предыдущих этапах. Этот анализ осуществляется с помощью *иерархической множественной регрессии* (см. гл. 5).

Предикторы (независимые переменные) могут представлять собой либо количественные переменные, аналогичные рассматриваемым выше, либо качественные переменные (место рождения, национальность и религиозная принадлежность). Чтобы оценить влияние таких качественных переменных, необходимо преобразовать их в фиктивные бинарные переменные. Эта процедура описана в гл. 9—11, посвященных дисперсионному анализу. Проиллюстрируем применение пошаговой множественной регрессии на количественных переменных. В табл. 4.1 приведен небольшой массив данных, который будем использовать в настоящем примере. Каждая переменная принимает значения от 1 до 5, причем большее значение соответствует более сильной выраженности рассматриваемого признака. Например, больший балл по академической успеваемости означает, что ребенок лучше учится в школе.

Таблица 4.1. **Значения пяти переменных в девяти наблюдениях (испытуемых)**

Наблюдения	Успеваемость ребенка	Способности ребенка	Интерес ребенка	Заинтересованность родителей	Заинтересованность учителей
1	1	1	1	2	1
2	2	3	3	1	2
3	2	2	3	3	4
4	3	3	2	2	4
5	3	4	4	3	2
6	4	2	3	2	3
7	4	3	5	3	4
8	5	4	2	3	2
9	5	3	4	2	3

Девять наблюдений — слишком маленькая выборка для проведения множественной регрессии. Необходимый размер (объем) используемой выборки зависит от многих факторов, таких, как величина ожидаемых корреляций между зависимой и независимыми переменными. Вероятность того, что корреляции окажутся значимыми, растет с увеличением размера выборки. Чтобы приблизить ситуацию к реально существующей, искусственно увеличим объем выборки, повторив данные тридцать раз, в результате общий объем выборки возрастет с девяти наблюдений до 270. Такой способ увеличения объема выборки не повлияет на величины и знаки корреляций между переменными, поскольку структура данных не изменилась, но позволит сделать эти корреляции значимыми.

Корреляционная матрица и первый предиктор

Полезным шагом в понимании того, какие преобразования совершаются во множественной регрессии, является вычисление корреляционной матрицы анализируемых переменных, приведенной в табл. 4.2.

Объясняемая переменная называется *зависимой переменной*, или *откликом* (*критерием*). Переменные, которые мы используем для объяснения или предсказания зависимой переменной, называются *независимыми переменными*, или *предикторами*. Переменная-предиктор, имеющая самый высокий коэффициент корреляции с переменной-откликом, всегда первой включается в регрессион-

Таблица 4.2. **Нижний треугольник корреляционной матрицы для пяти переменных**

	Успеваемость ребенка	Способности ребенка	Интерес ребенка	Заинтересованность родителей	Заинтересованность учителей
Успеваемость ребенка	1,00				
Способности ребенка	0,59	1,00			
Интерес ребенка	0,44	0,42	1,00		
Заинтересованность родителей	0,30	0,30	0,29	1,00	
Заинтересованность учителей	0,28	0,07	0,47	0,27	1,00

ный анализ при условии того, что эта корреляция статистически значима. Независимой переменной, имеющей наибольшую корреляцию с академической успеваемостью, оказывается «Умственные способности (интеллект) ребенка». Коэффициент корреляции в данном случае равен 0,59, а значимость на уровне ниже 0,05 по двустороннему критерию[1]. Следовательно, интеллект ребенка войдет в качестве первой независимой переменной в уравнение множественной регрессии. Чтобы вычислить долю дисперсии академической успеваемости ребенка, объясняемой его способностями, возводим в квадрат коэффициент корреляции и получаем *приблизительно* 0,34 ($0,59^2 \approx 0,34$). Другими словами, около 34 % дисперсии академической успеваемости ребенка объясняется его способностями. Так как знак корреляции между академической успеваемостью и способностями положительный, это означает, что более умные дети лучше учатся. Если бы знак данного коэффициента корреляции был отрицательным, это означало бы, что менее умные дети успешнее учатся.

Если две или более независимых переменных (предикторы) имеют практически равные коэффициенты корреляции с зависимой переменной (критерием, откликом), то в качестве первого предиктора все равно выбирается переменная, имеющая наибольший коэффициент корреляции, даже если различия соответствующих коэффициентов очень малы и каждая переменная объясняет примерно одинаковый процент дисперсии отклика. Например, если бы корреляция между академической успеваемостью ребенка и его интересом к учебе была равна 0,60 вместо 0,44, то интерес ребенка к учебе первым вошел бы в уравнение регрессии, несмотря на то что обе переменные имели бы очень близкие корреляции с академической успеваемостью и объясняли примерно равные доли ее дисперсии. В этом случае интерес ребенка к учебе объяснял бы 0,36 ($0,60^2 \approx 0,36$) дисперсии академической успеваемости. Другими словами, выбор переменной на очередном шаге пошаговой множественной регрессии осуществляется согласно чисто статистическим показателям.

Последующие предикторы

Следующей переменной для включения в регрессионный анализ будет та, которая имеет наибольшую частную корреляцию (partial correlation) с зависимой переменной. Ранее включенная

[1] Нам кажется, более осмысленным и понятным использовать терминологию уровня доверия. Уровень доверия вычисляется, как (1 — уровень значимости). Тогда последнее предложение будет звучать следующим образом: значимость коэффициентов корреляции на уровне доверия выше 0,95 по двустороннему критерию.

переменная при этом «*поддерживается постоянной*» или «*частично исключена*»[1]. Если эта частная корреляция является значимой, то вторая переменная включается в регрессионное уравнение. Если переменная, введенная первой, по-прежнему объясняет значительную долю дисперсии зависимой переменной при частичном исключении второй переменной, то ее сохраняют в регрессионном анализе. Если же первая переменная при частичном исключении второй переменной больше не объясняет значимой доли дисперсии зависимой переменной, ее исключают из регрессионного анализа. Этот процесс продолжается пошагово до тех пор, пока становится невозможным объяснить сколько-нибудь значительное увеличение доли дисперсии зависимой переменной за счет новых независимых переменных.

По корреляционной матрице (см. табл. 4.2) невозможно сказать, какая из независимых переменных имеет наибольшую частную корреляцию с академической успеваемостью при учете влияния способностей. Например, интерес ребенка к учебе имеет следующий по величине коэффициент корреляции с академической успеваемостью (0,44), однако этот интерес также высоко коррелирует со способностями (0,42). Возможно, бóльшая часть корреляции между интересом ребенка к учебе и его академической успеваемостью объясняется его умственными способностями (интеллектом) и поэтому интерес ребенка к учебе может сам по себе и не объяснять еще сколько-нибудь значительной доли дисперсии академической успеваемости.

Частная корреляция

Для следующего этапа необходимо вычислить частные корреляции между академической успеваемостью и каждой из остальных независимых переменных при частичном исключении уровня способностей ребенка. Это можно сделать по следующей формуле:

$$r_{12,3} = \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - r_{13}^2) \times (1 - r_{23}^2)}},$$

где r — корреляция, индекс 1 относится к переменной, обозначающей академическую успеваемость; индекс 2 — к независимой

[1] В англоязычной литературе применяется термин *controlling* (*сдерживание*). Мы решили использовать иные термины вслед за авторами хорошо зарекомендовавшего себя среди отечественных психологов перевода (см.: Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стэнли. — М.: Прогресс, 1976; пер. с англ. Л. И. Хайрусовой; под ред. Ю. П. Аллера).

переменной, отличной от способностей ребенка; индекс 3 — к переменной, обозначающей способности ребенка.

Если подставить соответствующие корреляции из табл. 4.2 в эту формулу и вычислить частную корреляцию между академической успеваемостью и интересом ребенка к учебе при частичном исключении его умственных способностей[1], то окажется, что она приближенно равна 0,26:

$$\frac{0,44-(0,59\times 0,42)}{\sqrt{(1-0,59^2)\times(1-0,42^2)}}=\frac{0,44-0,25}{\sqrt{(1-0,35)\times(1-0,18)}}=$$

$$=\frac{0,19}{\sqrt{0,65\times 0,82}}=\frac{0,19}{\sqrt{0,53}}=\frac{0,19}{0,73}=0,26.$$

Частная корреляция между академической успеваемостью и заинтересованностью родителей составляет 0,15, а частная корреляция между академической успеваемостью и заинтересованностью учителей равна 0,30. Так как наибольшая величина частной корреляции из рассматриваемых — между академической успеваемостью и заинтересованностью учителей, и эта корреляция является значимой, то второй переменной, включаемой в уравнение регрессии, будет заинтересованность учителей. Положительный знак этой частной корреляции означает, что чем больше интереса к работе ребенка проявляют учителя, тем лучше ребенок учится. Отрицательное значение данной частной корреляции означало бы, что чем меньше интереса проявляют учителя к учебе ребенка, тем выше его успехи в школе. Поскольку доля дисперсии академической успеваемости, объясняемой способностями ребенка, даже когда в рассмотрение включена заинтересованность учителей, значительна, эта переменная (способности ребенка) остается в уравнении регрессии вместе с заинтересованностью учителей. Частная корреляция между академической успеваемостью и способностями ребенка при частичном ис-

---

[1] Данный показатель можно интерпретировать следующим образом. Можно предположить, что существует положительная корреляция между академической успеваемостью (1) и интересом ребенка к учебе (2). Однако вывод о том, что некоторые дети успевают лучше в школе вследствие своего интереса к учебе, делать рано. С одной стороны, очевидно, что способные дети лучше учатся, а с другой — быть может, более способные дети более заинтересованы в учебе, т.е. вполне может быть так, что и показатели академической успеваемости, и показатели заинтересованности в учебе более высоки у способных детей. Чтобы ответить на вопрос, как взаимосвязаны академическая успеваемость и интерес ребенка к учебе при частично исключенном уровне способностей, и вычисляется частная корреляция первых двух переменных при третьей, сохраняющей постоянное значение. Заметим, что если первые две переменные никак не связаны с третьей, т.е. $r_{13} = r_{23} = 0$, то $r_{12.3} = r_{12}$.

ключении заинтересованности учителей приближенно равна 0,60 и является статистически значимой[1].

На третьем шаге в регрессионный анализ включается та из независимых переменных, которая имеет наибольшую частную корреляцию с академической успеваемостью при частичном исключении умственных способностей ребенка и заинтересованности учителей. Это уже частная корреляция второго порядка, формула для ее вычисления сложнее, чем формула коэффициента частной корреляции первого порядка (см. с. 88), и поэтому не будет приведена здесь. Ее можно найти в других источниках (D. Cramer, 1998, P. 159). Частные корреляции второго порядка для интереса ребенка к учебе и заинтересованности родителей равны 0,13 и 0,08 соответственно. Поскольку частная корреляция интереса ребенка к учебе выше, чем частная корреляция родительской заинтересованности, и является статистически значимой, то в качестве третьей переменной в уравнение регрессии будет включен именно интерес ребенка к учебе. Положительный знак этой частной корреляции означает, что более заинтересованные дети лучше учатся. Если бы эта частная корреляция была отрицательна, то можно было бы предположить, что менее заинтересованные в учебе дети лучше успевают в школе. И способности ребенка, и заинтересованность учителей остаются в составе уравнения регрессии, поскольку они по-прежнему объясняют значимую долю дисперсии академической успеваемости, даже когда в уравнение включается интерес ребенка к учебе. Коэффициент частной корреляции второго порядка между академической успеваемостью и заинтересованностью учителей при частичном исключении способностей ребенка и его интереса к учебе равен 0,21 и является статистически значимым. Коэффициент частной корреляции второго порядка между академической успеваемостью и способностями ребенка при частичном исключении интереса ребенка к учебе и заинтересованности учителей равен 0,53 и является также статистически значимым.

Последний предиктор, соответствующий заинтересованности родителей, не включается в регрессионный анализ, поскольку коэффициент частной корреляции третьего порядка между акаде-

---

[1] Попробуйте выписать формулу для расчета коэффициента частной корреляции самостоятельно и сопоставить с тем, что приведено у нас. В данном случае в общую формулу нужно подставлять коэффициенты корреляции r_{12}, r_{23}, r_{13} исходя из того, что индекс 1 относится к переменной, обозначающей академическую успеваемость, индекс 2 — к независимой переменной, обозначающей способности ребенка, а индекс 3 — к переменной, соответствующей заинтересованности учителей. Тогда вычисления будут иметь вид:

$$\frac{0,59-(0,28\times 0,07)}{\sqrt{(1-0,28^2)\times(1-0,07^2)}}=\frac{0,59-0,02}{\sqrt{(1-0,08)\times(1-0,01)}}=\frac{0,57}{\sqrt{0,92\times 0,99}}=\frac{0,57}{\sqrt{0,91}}=\frac{0,57}{0,95}=0,60.$$

мической успеваемостью и родительской заинтересованностью при частичном исключении влияния способностей ребенка, заинтересованности учителей и интереса ребенка к учебе равен 0,07 и статистически не значим. Следовательно, *три независимые переменные*, объясняющие значительную долю дисперсии академической успеваемости, соответствуют способностям ребенка, заинтересованности учителей, а также интересу ребенка к учебе.

Важно помнить, что эти результаты основаны не на реальных, а на специально подобранных данных исключительно в целях объяснения метода пошаговой множественной регрессия. Поэтому полученные при объяснениях содержательные результаты не являются достоверными.

Доля объясненной дисперсии

Доля дисперсии зависимой переменной (критерия, отклика), объясняемой первым предиктором, просто равна квадрату корреляции между этими переменными. Так как корреляция между академической успеваемостью и способностями ребенка равна 0,59, доля дисперсии академической успеваемости, объясняемая способностями ребенка, приближенно равна 0,34 ($0{,}59^2 \approx 0{,}34$). Дополнительные доли дисперсии академической успеваемости, объясняемые другими предикторами, входящими в уравнение, вычисляются путем возведения в квадрат показателей, называемых «часть корреляции»[1].

Получастная корреляция (часть корреляции)

Формула для расчета получастной корреляции первого порядка выглядит следующим образом:

$$r_{12.3} = \frac{r_{12} - (r_{13} \times r_{23})}{\sqrt{(1 - r_{23}^2)}}.$$

Частная и получастная корреляции отличаются знаменателями соответствующих формул. В знаменателе формулы получастной корреляции стоит необъясненная дисперсия первого и второго пре-

---

[1] Этот термин взят также из перевода (см.: Гласс Дж. Статистические методы в педагогике...); по-английски «part correlation». Другие англоязычные авторы используют термин semipartial correlation (получастная корреляция) (B.G.Tabachnik, L.S.Fidell). Нам кажется, что использование последнего термина вносит меньше путаницы, поэтому в данном переводе будем применять именно его. Существуют и другие переводы, например А.П.Кулаичев называет частную корреляцию частичной, а получастную — частной.

диктора[1], а в знаменателе частной корреляции добавляется еще множитель — необъясненная дисперсия зависимой переменной (критерия) и второго предиктора[2,3].

Чтобы вычислить получастную корреляцию первого порядка между академической успеваемостью и заинтересованностью учителей при частично исключенном влиянии способностей ребенка, подставим соответствующие корреляции из табл. 4.2 в нашу формулу и получим примерно 0,24:

$$\frac{0,28-(0,59\times 0,07)}{\sqrt{(1-0,07^2)}}=\frac{0,28-0,04}{\sqrt{(1-0,00)}}=\frac{0,24}{\sqrt{1,00}}=\frac{0,24}{\sqrt{1,00}}=\frac{0,24}{1,00}=0,24.$$

Чтобы вычислить долю дисперсии академической успеваемости, объясняемую заинтересованностью учителей помимо и сверх той доли, которая объяснена способностями ребенка, возведем в квадрат эту получастную корреляцию и получим около 0,06 ($0,24^2 = 0,06$).

Формула для получастной корреляции второго порядка более сложна и здесь ее опустим. Однако если вычислить получастную корреляцию второго порядка между академической успеваемостью и интересом ребенка к учебе при частичном исключении влияния способностей ребенка и заинтересованности учителей, то обнаружим, что она приближенно равна 0,10, что при возведении в квадрат дает около 0,01 ($0,10^2 = 0,01$). Иными словами, около 0,01 дисперсии академической успеваемости объясняется интересом ребенка помимо и сверх того, что объясняется умственными способностями ребенка и заинтересованностью учителей.

---

[1] Переменные 2 и 3.

[2] Переменные 1 и 3.

[3] Чтобы не путать частную и получастную корреляции, будем обозначать вторую как $r_{1.23}$. В данной ситуации вычисляется часть корреляции переменной 1 с переменной 2 после того, как часть переменной 2, которую можно было бы предсказать по переменной 3, на основании корреляции переменных 2 и 3 была удалена из переменной 2. Однако подобное неоднозначное словесное определение может лишь служить иллюстрацией к приведенной вычислительной формуле. Если сравнивать формулы $r_{1.23}$ и $r_{12.3}$, то хотя они и похожи, можно отметить, во-первых, что частная корреляция переменной 1 с переменной 2 симметрична, т.е. совпадает с корреляцией переменной 2 с переменной 1. Получастная корреляция таким свойством не обладает. Во-вторых, частная корреляция не может быть меньше получастной корреляции, ибо последняя вычисляется с «усеченной» второй переменной. На уровне формул $r_{12.3}=\frac{r_{1.23}}{\sqrt{(1-r_{13}^2)}}$, поскольку значение знаменателя лежит в диапазоне от 0 до 1 (включительно), то $r_{12.3} \geq r_{1.23}$. Знак равенства достигается при условии $r_{13} = 0$, т.е. если переменные 1 и 3 независимы.

Статистическая значимость объясненной дисперсии

Статистическая значимость доли дисперсии зависимой переменной (критерия), объясняемой независимой переменной (предиктором), определяется F-отношением, которое вычисляется по следующей формуле:

$$F = \frac{\dfrac{[\text{Изменение } R^2]}{[\text{Число добавленных предикторов}]}}{\dfrac{[1 - R^2]}{[N - \text{число предикторов} - 1]}}.$$

Величина F имеет две степени свободы: одна равна знаменателю числителя формулы, а другая — знаменателю знаменателя. При пошаговой множественной регрессии степень свободы числителя всегда равна единице[1]. Число степеней свободы знаменателя равно общему количеству переменных (N) минус число предикторов и минус единица. Число предикторов включает переменные, уже введенные на предыдущих этапах, а также переменную, вводимую на данном этапе. Таким образом, на первом этапе один предиктор, на втором — два и т.д. Уровень значимости величины F можно посмотреть в соответствующей таблице, но статистические пакеты типа SPSS выдают эти значения в выходных файлах. Чем больше величина F-отношения, тем вероятнее, что она окажется статистически значимой.

Когда в регрессионном анализе рассматривается только один предиктор, величина изменения R^2 и само R^2 равны и представляют собой квадрат корреляции между откликом и предиктором. Так, величина F для доли дисперсии академической успеваемости, объясняемой детскими умственными способностями, приближенно равна 145,83, поскольку квадрат корреляции между этими двумя переменными приближенно равен 0,35 ($0{,}59^2 = 0{,}35$):

$$F = \frac{\dfrac{0{,}35}{1}}{\dfrac{(1 - 0{,}35)}{(270 - 1 - 1)}} = \frac{0{,}35}{\dfrac{0{,}65}{268}} = \frac{0{,}35}{0{,}0024} = 145{,}83.$$

Две степени свободы для данной F-величины равны 1 и 268 соответственно. Это F-отношение статистически значимо с уровнем значимости менее 0,001.[2]

Когда в анализе участвует более одного предиктора, изменение R^2 равно квадрату получастной корреляции данного предик-

---

[1] Так как на каждом шаге добавляется по одному новому предиктору.

[2] Т.е. при уровне доверия более 0,999.

тора на текущем шаге, а R^2 равно сумме изменений R^2 на всех предыдущих этапах, включая и текущий.

Квадрат получастной корреляции между академической успеваемостью и заинтересованностью учителей с учетом влияния умственных способностей ребенка приближенно равен 0,06 ($0{,}24^2 = 0{,}06$).

Следовательно, величина F-отношения приближенно равна 27,27:

$$F = \frac{\frac{0{,}06}{1}}{\left[\frac{(1-(0{,}35+0{,}06)}{(270-2-1)}\right]} = \frac{\frac{0{,}06}{0{,}59}}{\frac{267}{}} = \frac{0{,}06}{0{,}0022} = 27{,}27.$$

Две степени свободы для данной F-величины равны 1 и 267 соответственно. Это F-отношение значимо с уровнем значимости менее 0,001.

Квадрат частной корреляции между академической успеваемостью и интересом ребенка к учебе с учетом влияния умственных способностей ребенка и заинтересованности учителей приближенно равен 0,01 ($0{,}10^2 = 0{,}01$). Следовательно, величина F-отношения приближенно равна 4,55:

$$F = \frac{\frac{0{,}01}{1}}{\left[\frac{(1-(0{,}35+0{,}06+0{,}01)}{(270-3-1)}\right]} = \frac{\frac{0{,}01}{0{,}58}}{\frac{266}{}} = \frac{0{,}01}{0{,}0022} = 4{,}55.$$

Две степени свободы для данной F-величины равны, соответственно, 1 и 266. Данное F-отношение является значимым с уровнем значимости менее 0,05. Основные результаты этой пошаговой множественной регрессии отражены в табл. 4.3.

Другие статистики являются важной составной частью множественной регрессии, однако они не так существенны для понимания пошаговой множественной регрессии. Некоторые из них будут описаны в гл. 5.

Таблица 4.3. **Основные результаты пошаговой множественной регрессии**

Этапы	Предикторы	R^2	Изменение R^2	F	df_1	df_2	p
1	Способности ребенка	0,35	0,35	145,83	1	268	0,001
2	Заинтересованность учителей	0,41	0,06	27,27	1	267	0,001
3	Интерес ребенка	0,42	0,01	4,55	1	266	0,05

Отчет о результатах

Существует много различных способов написания отчета о результатах множественной пошаговой регрессии, выполнение которой было рассмотрено в данной главе. Очень краткий отчет можно было бы сформулировать следующим образом: «В процессе пошаговой множественной регрессии умственные способности ребенка оказались первой наиболее важной независимой переменной в уравнении и объясняли примерно 35 % дисперсии академической успеваемости ребенка ($F_{1,268} = 145{,}83$; $p < 0{,}001$). Заинтересованность учителей была второй переменной и объясняла дополнительные 6 % ($F_{1,267} = 27{,}27$; $p < 0{,}001$). Интерес ребенка к учебе рассматривался в качестве третьей переменной и объяснял еще 1 % дисперсии ($F_{1,266} = 4{,}55$; $p < 0{,}05$). Родительская заинтересованность не давала значительной прибавки в доле объясняемой дисперсии. Таким образом, более высокая академическая успеваемость связана с более высокими умственными способностями ребенка, заинтересованностью учителей и интересом ребенка к учебе».

Реализация процедуры в программе SPSS для Windows

Приведем алгоритм процедуры пошаговой множественной регрессии, описанной в данной главе, с помощью SPSS.

Введите данные, представленные в табл. 4.1, в **Редактор данных (Data Editor)**, как показано на рис. 4.1. Шестой столбец в матрице данных, называемый частотой, был создан с целью показать, что данные для каждой строки, или объекта, повторяются 30 раз и размер выборки увеличивается, таким образом, с 9 до 270. Названия пяти переменных являются сокращениями детской успеваемости, детских способностей, детского интереса, родительского интереса и учительского интереса соответствен-

	детуспев	детспос	детинтер	родинтер	учинтер	частота
1	1	1	1	2	1	30
2	2	3	3	1	2	30
3	2	2	3	3	4	30
4	3	3	2	2	4	30
5	3	4	4	3	2	30
6	4	2	3	2	3	30
7	4	3	5	3	4	30
8	5	4	4	3	3	30
9	5	3	4	2	3	30

Рис. 4.1. Значения пяти переменных для девяти объектов в **Редакторе данных (Data Editor)**

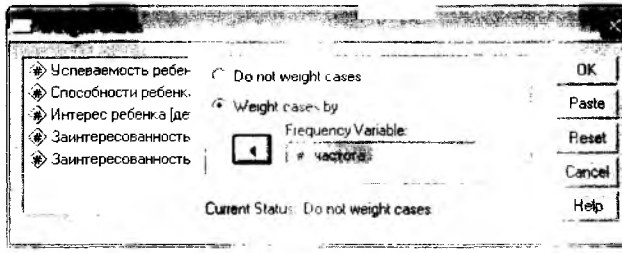


Рис. 4.2. Диалоговое окно **Весовые коэффициенты (Weight Cases)**

но[1]. Сохраните введенные данные в отдельном файле для дальнейшего использования.

Чтобы приписать веса или увеличить число объектов, выберите пункт **Данные (Data)** из строки меню в верхней части окна программы, а в появившемся ниспадающем меню выберите пункт **Весовые коэффициенты для объектов (Weight cases)**, чтобы открыть диалоговое окно **Весовые коэффициенты для объектов (Weight cases)**, изображенное на рис. 4.2. Установите переключатель **Задать весовые коэффициенты для объектов (Weight cases by)**, щелкнув на нем указателем мыши. Выберите частоту в списке переменных и переместите ее с помощью кнопки ▸ в поле **Частотная переменная (Frequency variable)**, а затем нажмите **OK**, чтобы закрыть данное диалоговое окно.

В правом нижнем углу окна **Редактора данных (Data Editor)** появилась надпись **Весовые коэффициенты заданы (Weights On)**, напоминающая вам о том, что данные рассматриваются с учетом весовых коэффициентов.

Выберите пункт **Анализ (Analyze)** в строке меню в верхней части окна программы, а в ниспадающем меню — пункт **Регрессия (Regression)** и в нем — подпункт **Линейная регрессия (Linear)**, открывающий диалоговое окно **Линейная регрессия (Linear Regression)**, изображенное на рис. 4.3.

Выберите **Успеваемость ребенка** и нажмите верхнюю кнопку ▸, чтобы переместить эту переменную в поле **Зависимая переменная (Dependent)**.

Выберите переменные, начиная со **Способность ребенка** и заканчивая **Заинтересованность учителей**, и с помощью второй кнопки ▸ переместите их в список **Независимые переменные (Independent(s))**.

Выберите пункт **Ввод (Enter)** в раскрывающемся списке **Метод (Method)** и после появления ниспадающего меню — пункт **По шагам (Stepwise)** для пошаговой множественной регрессии.

---

[1] Названия переменных SPSS не должны состоять более чем из восьми символов. Поэтому полная расшифровка переменных для того, чтобы было ясно, что они обозначают, приводится в графе метки (labels). Здесь число символов не ограничено.

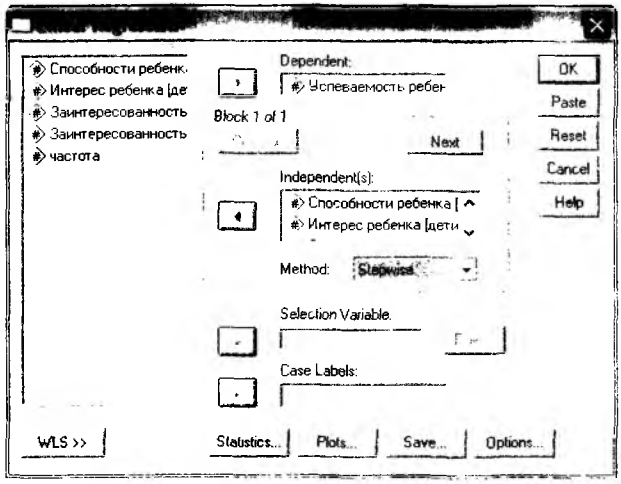


Рис. 4.3. Диалоговое окно **Линейная регрессия (Linear Regression)**

Нажмите на кнопку **Статистики (Statistics)** и откройте диалоговое окно **Линейная регрессия: Статистики (Linear, Regression: Statistics)**, изображенное на рис. 4.4.

Установите флажок **Изменение квадрата R (R squared change)**, чтобы получить статистические результаты в столбцах под общим заголовком **Статистика изменений (Change statistics)** окна вывода программы, приведенного в табл. 4.4.

Установите флажок **Описательные статистики (Descriptives)** для вывода средних, стандартных отклонений и числа объектов для пяти переменных, а также их полной корреляционной матрицы с указанием уровня статистической значимости этих корреляций и числа объектов, на основании которого они вычислялись.

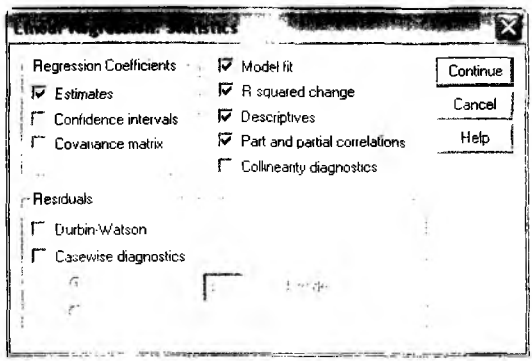


Рис. 4.4. Диалоговое окно **Линейная регрессия: Статистики (Linear regression: Statistics)**

Установите флажок **Частные и части корреляции (Part and partial correlations)** для вывода данных статистических показателей, как показано в табл. 4.5.

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть данное диалоговое окно и вернуться в предыдущее диалоговое окно.

Нажмите **OK**, чтобы выполнить данный анализ.

Выводимая SPSS информация по результатам работы

Программа SPSS выдает огромное количество статистических результатов помимо упоминаемых в нашем описании пошаговой множественной регрессии. Ниже воспроизведем и прокомментируем только отдельные выборочные аспекты выводимых результатов.

В табл. 4.4 воспроизведена **Общая информация о моделях (Model Summary)**, которая была использована в отчете о результатах. *Моделью* называется каждый этап проводимого анализа. В нашем случае имеется три этапа, или модели. Предикторы, вводимые в уравнение регрессии на каждом из трех шагов, отображаются непосредственно под таблицей. Так, способности ребенка вводятся на первом шаге, способности ребенка и заинтересованность учителей — на втором и т.д.

Доля дисперсии, с точностью до трех десятичных знаков, объясняемая на каждом этапе, приведена в шестом столбце под заголовком **Изменение квадрата R (R Square Change)**. Так, 0,533 дисперсии академической успеваемости объясняется умственными способностями ребенка в первой модели. Дополнительные 0,060 дисперсии академической успеваемости объясняются заинтересованностью учителей, а еще 0,011 — интересом ребенка к учебе.

Величина *F*-отношения для каждого этапа анализа отображена в седьмом столбце под заголовком **Изменение F (F-change)**. *F*-отношение на первом шаге равно 146,451, что немного выше, чем значение 145,83, которое было вычислено ранее, поскольку мы проводили округление до двух знаков после запятой, а SPSS проводит округление до трех знаков. Числа степеней свободы числителя и знаменателя данной *F*-величины приводятся в восьмом и девятом столбцах под заголовками df1 и df2 соответственно. Они равны 1 и 268. Статистическая значимость или вероятность *F*-величины быть такой высокой из-за случайных отклонений приведена в десятом столбце под заголовком **Значимость изменения F (Sig. F Change)**. Значение вероятности дано с точностью до трех знаков после запятой и в нашем случае меньше 0,0005, поскольку оно никогда не может быть равно нулю.

Таблица 4.4. **Выводимая SPSS общая информация о моделях**

Model Summary

Model	R	R Square	Change Statistics						
			Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F.Change
1	0,594[a]	0,353	0,351	1,061	0,353	146,451	1	268	0,000
2	0,643[b]	0,413	0,409	1,013	0,060	27,118	1	267	0,000
3	0,651[c]	0,424	0,417	1,006	0,011	4,917	1	266	0,027

a. Predictors: (Constant), Способности ребенка.

b. Predictors: (Constant), Способности ребенка, Заинтересованность учителей.

c. Predictors: (Constant), Способности ребенка, Заинтересованность учителей, Интерес ребенка.

Таблица 4.5. **Выводимая SPSS таблица коэффициентов**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	0,853	0,206		4,137	0,000			
	Способности ребенка	0,853	0,070	0,594	12,102	0,000	0,594	0,594	0,594
2	(Constant)	0,049	0,250		0,198	0,843			
	Способности ребенка	0,830	0,067	0,578	12,310	0,000	0,594	0,602	0,577
	Заинтересованность учителей	0,312	0,060	0,245	5,208	0,000	0,283	0,304	0,244
3	(Constant)	0,011	0,249		0,046	0,963			
	Способности ребенка	0,757	0,075	0,528	10,147	0,000	0,594	0,528	0,472
	Заинтересованность учителей	0,239	0,068	0,187	3,510	0,001	0,283	0,210	0,163
	Интерес ребенка	0,148	0,067	0,130	2,218	0,027	0,439	0,135	0,103

a. Dependent Variable: Успеваемость ребенка.

Квадрат множественной корреляции, или квадрат R (**R Square**), на каждом этапе равен сумме изменений квадрата R (**R Square Change**) для всех предшествующих и текущего этапов. Так, на втором шаге квадрат R (**R Square**) равен 0,413 (0,353 + 0,060 = 0,413). Множественная корреляция, или R, есть корень квадратный из квадрата изменений R (**R Square Change**). Так, для второго этапа корень квадратный из 0,413 равен 0,643.

Величина квадрата множественной корреляции оказывается несколько завышенной и тем больше, чем больше число предикторов и объектов в выборке. Менее смещенная оценка дается следующей формулой:

$$[\text{Уточненная величина } R^2] = R^2 - \frac{[(1 - R^2) \times \text{число предикторов}]}{[N - \text{число предикторов} - 1]}.$$

Чтобы проиллюстрировать применение этой формулы, вычислим исправленный квадрат множественной корреляции для второй модели, который оказывается приближенно равен 0,409:

$$0,413 - \frac{(1 - 0,413) \times 2}{270 - 2 - 1} = 0,413 - \frac{0,58 \times 2}{270 - 2 - 1} = 0,413 - 0,0044 = 0,4086.$$

Частная и получастная корреляции приведены с точностью до трех знаков после запятой в двух последних столбцах таблицы коэффициентов (см. табл. 4.5). Частная корреляция между академической успеваемостью и заинтересованностью учителей при учете влияния умственных способностей ребенка на втором этапе равна 0,304. Получастная корреляция между академической успеваемостью и заинтересованностью учителей при учете влияния умственных способностей ребенка на втором этапе равна 0,244.[1]

Рекомендуемая литература

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows.* London: Routledge.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction,* 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[1] Кроме того, в табл. 4.5 приведены стандартизированные коэффициенты. От нестандартизированных они отличаются тем, что вычисляются по стандартизованным z-значениям переменных.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

Дополнительная литература, рекомендуемая научным редактором

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

Кулаичев А.П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

Глава 5

ИЕРАРХИЧЕСКАЯ МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Предисловие научного редактора

В гл. 5 рассматривается иерархический метод, позволяющий исследователю на каждом этапе самому принимать решение о вводе переменных.

Иерархическая множественная регрессия используется для определения доли дисперсии индивидуальной переменной, объясняемой другими переменными в том случае, когда они входят в уравнение регрессионного анализа в определенном порядке, а также для выяснения вопроса о том, являются ли эти доли объясняемой дисперсии статистически более значимыми, чем это было бы, если предположить, что их взаимосвязь носит случайный характер. Например, нас может интересовать, какая доля дисперсии академической успеваемости изначально объясняется заинтересованностью учителей и родителей в школьных успехах ребенка, далее собственным интересом ребенка к учебе, а затем его умственными способностями. Другими словами, нас интересует, какая дополнительная доля дисперсии академической успеваемости объясняется интересом ребенка к учебе и его умственными способностями при исключении других, внешних, факторов — влияния родителей и учителей. Вполне ожидаемо, что доля дисперсии академической успеваемости, объясняемой умственными способностями ребенка, окажется гораздо меньше, если учитывать другие названные факторы. Это могло бы означать, что связь между академической успеваемостью и умственными способностями ребенка в какой-то степени обусловлена тем, что они оценивают одну и ту же реальность. Точная формулировка вопроса зависит от тех идей и гипотез, которые мы хотим проверить.

Примеры, используемые в тексте, были выбраны таким образом, чтобы показать разницу между иерархической и пошаговой множественной регрессией, и не должны рассматриваться с точки зрения соответствия содержательным теориям.

В случае иерархической множественной регрессии исследователь сам принимает решение о том, какие переменные включать в регрессионный анализ на каждом этапе, исходя из содержательных соображений, в то время как в случае пошаговой множественной регрессии это решение основывается исключительно на статистических критериях. В случае иерархической множественной

регрессии в анализ на каждом этапе можно включать более одной переменной, в то время как в случае пошаговой множественной регрессии на каждом шаге анализа добавляется только одна переменная. Во всех остальных отношениях процедуры вычисления соответствующих статистик при пошаговой и иерархической множественной регрессии совпадают.

Опосредующие эффекты

Ряд исследователей приводит аргументы в пользу того, что иерархическая множественная регрессия является более адекватным методом для выявления опосредующего влияния некоторой количественной переменной на взаимосвязь двух других количественных переменных (R.M.Baron и D.A.Kenny, 1986). Например, возможно, что взаимосвязь между депрессией и социальной поддержкой сильнее в том случае, когда люди испытывают стресс. В состоянии стресса люди, у которых есть возможность рассказать о своем состоянии другим, могут испытывать меньшую депрессию, чем те, кто не имеет такой возможности. Для людей, не испытывающих стресс, данное соотношение может быть слабее или не существовать вообще. Если дело обстоит именно таким образом, говорят, что стресс оказывает опосредующий эффект на взаимосвязь между депрессией и социальной поддержкой. По всей вероятности, величина стресса, испытываемого людьми, варьируется в пределах от полного отсутствия до очень сильного. Один из способов удостовериться в том, что стресс оказывает опосредующий эффект на указанную взаимозависимость, состоит в том, чтобы разбить всю выборку на две группы — тех, кто испытывает высокий стресс, и тех, кто испытывает незначительный стресс, вычислить коэффициенты корреляции между депрессией и социальной поддержкой в обеих группах, сравнить их между собой и выяснить, достигает ли различие статистической значимости. Если значимых различий нет, то нет оснований предполагать опосредующий эффект. Одним из недостатков этого подхода является то, что разбиение испытуемых на две группы создает две меньшие по объему выборки, а это, в свою очередь, уменьшает вероятность различий между коэффициентами корреляции, особенно если исходная выборка невелика.

Альтернативный подход состоит в использовании иерархической множественной регрессии, в которой эффект опосредования стрессом учитывается за счет ввода новой переменной — «Взаимодействия между стрессом и социальной поддержкой», вычисляемой перемножением значений показателей стресса и социальной поддержки. Однако при этом помимо совместного эффекта учитывается влияние каждой из двух переменных по

отдельности[1]. Чтобы определить, какой процент дисперсии депрессии объясняется исключительно взаимодействием, необходимо устранить влияние этих двух переменных по отдельности. Для этого нужно включить в регрессионный анализ сначала переменные стресса и социальной включенности, а затем переменную, отражающую их взаимодействие. Если слагаемое, соответствующее взаимодействию, объясняет значительное увеличение дисперсии депрессии, то имеет место опосредующий эффект. Чтобы определить, в чем состоит суть этого опосредования (усиление или ослабление), следует разбить выборку на две группы с высоким и низким уровнем стресса и проанализировать взаимосвязь между депрессией и социальной включенностью внутри каждой выборки.

Более подробно рассмотрим применение иерархической множественной регрессии на тех же данных, которые были использованы в случае пошаговой множественной регрессии[2] (см. табл. 4.1). На первом этапе будут введены переменные родительской и учительской заинтересованности, на втором — переменная интереса ребенка к учебе, на третьем — переменная, отражающая его умственные способности. Вычислять долю дисперсии зависимой переменной, привносимую на каждом этапе, будем аналогично случаю пошаговой множественной регрессии, когда на каждом шаге вводилась только одна переменная. Так, доля дисперсии, объясняемая первым предиктором, равна квадрату коэффициента корреляции между этим предиктором и зависимой переменной; доля дисперсии, объясняемая последующими предикторами, — квадрату коэффициента части корреляции между этим предиктором и зависимой переменной при исключении всех предшествующих предикторов.

Ввод двух или более предикторов на одном этапе

Когда на одном этапе множественной регрессии в анализ одновременно включаются два предиктора или более, невозможно вычислить долю дисперсии зависимой переменной, объясняемой этими предикторами, путем простого суммирования квадратов получастных корреляций между каждым из предикторов и переменной-откликом. Подобное суммирование принимает в расчет только специфический вклад в дисперсию каждого из предикто-

[1] В терминах регрессионного уравнения эта модель выглядит следующим образом:

Депрессия = β_1(*Стресс*) + β_2(*Социальная поддержка*) + β_3(*Стресс* × *Социальная поддержка*) + *Остаточный член.*

[2] Чтобы предостеречь читателя от недопонимания, отметим, что в рассматриваемом ниже примере речи об опосредованном взаимодействии идти не будет.

ров и не учитывает доли дисперсии, объясняемые взаимосвязью предикторов между собой и с зависимой переменной. Например, если необходимо определить долю дисперсии академической успеваемости, объясняемой родительской и учительской заинтересованностью в успехах ребенка, нельзя просто сложить квадраты получастных корреляций с родительской и учительской заинтересованностями. Квадрат получастной корреляции между академической успеваемостью и заинтересованностью родителей в успехах ребенка не включает те доли дисперсии как в академической успеваемости, так и в заинтересованности родителей, которые объясняются заинтересованностью учителей. Тот же общий вклад в дисперсию также не учитывается квадратом получастной корреляции между академической успеваемостью и заинтересованностью учителей. Следовательно, оба эти квадрата получастных корреляций не учитывают данный вклад в дисперсию и просто отражают специфическую долю дисперсии, общую для академической успеваемости и каждой из соответствующих переменных.

Квадрат множественной корреляции

Доля дисперсии зависимой переменной, объясняемой предикторами на любом этапе многомерного регрессионного анализа, равна квадрату множественной корреляции. Когда на одном этапе вводятся два предиктора или более, эта величина может быть вычислена как сумма произведений стандартизированного частного коэффициента регрессии каждого предиктора и соответствующей корреляции между этим предиктором и зависимой переменной, взятая по всем предикторам:

[Квадрат множественной корреляции] = Сумма по всем предикторам [стандартизированный частный коэффициент регрессии × коэффициент корреляции].

Прежде чем мы опишем, что из себя представляет стандартизированный частный коэффициент регрессии, вычислим квадрат множественной корреляции для родительской и учительской заинтересованности. Корреляции между академической успеваемостью и заинтересованностью родителей и учителей равны 0,30 и 0,28 соответственно. Стандартизированные частные коэффициенты регрессии для родительской и учительской заинтересованности, которые будут вычислены позже, приближенно равны 0,24 и 0,22 соответственно. Следовательно, квадрат множественной корреляции для родительской и учительской заинтересованности приближенно равен 0,13:

$$(0{,}24 \times 0{,}30) + (0{,}22 \times 0{,}28) = 0{,}07 + 0{,}06 = 0{,}13.$$

Иными словами, примерно 0,13 дисперсии академической успеваемости объясняется совместной заинтересованностью родителей и учителей.

Стандартизированный частный коэффициент регрессии

Стандартизированный коэффициент регрессии, или бета (β), может рассматриваться как весовой коэффициент для двух или более предикторов в формуле модели многомерной регрессии, учитывающей их взаимосвязь с другими переменными этой модели. Поскольку эти коэффициенты стандартизированы, они принимают значения от –1,00 до +1,00. Чем больше значение коэффициента по модулю, тем сильнее взаимосвязь между предиктором и зависимой переменной. Отрицательный коэффициент означает, что более высоким значениям предиктора соответствуют более низкие значения отклика. В нашем примере стандартизированные регрессионные коэффициенты приближенно равны 0,23 и имеют положительный знак. Эти весовые коэффициенты используются для определения того, какая доля взаимосвязи между откликом и предиктором объясняется данным предиктором. Так как и корреляции, и стандартизированные регрессионные коэффициенты для заинтересованности родителей и учителей близки по величине, доли дисперсии академической успеваемости, объясняемые этими переменными, оказываются примерно равны.

Формула для стандартизированного частного коэффициента регрессии предиктора с зависимой переменной, при частичном исключении еще одного предиктора, выглядит следующим образом:

$$\beta_2 = \frac{r_{12} - (r_{13} \times r_{23})}{1 - r_{23}^2}.$$

Эта формула очень похожа на формулу получастной корреляции первого порядка, за исключением того, что в знаменателе нет корня квадратного. Можно вычислить стандартизированный регрессионный коэффициент для родительской заинтересованности, если считать, что индекс 1 относится к академической успеваемости, индекс 2 — к родительской заинтересованности, а индекс 3 — к заинтересованности учителей. Подставив соответствующие корреляции из табл. 4.2, получим, что стандартизированный коэффициент регрессии приближенно равен 0,24:

$$\frac{0,30 - (0,28 \times 0,27)}{1 - 0,27^2} = \frac{0,30 - 0,08}{1 - 0,07} = \frac{0,22}{0,93} = 0,24.$$

Аналогично можно вычислить стандартизированный частный коэффициент регрессии для заинтересованности учителей, если индекс 1 относится к академической успеваемости, индекс 2 — к заинтересованности учителей, а индекс 3 — к родительской заинтересованности. Подставив соответствующие значения корреляций из табл. 4.2, получим, что данный стандартизированный регрессионный коэффициент приближенно равен 0,22:

$$\frac{0,28-(0,30\times 0,27)}{1-0,27^2}=\frac{0,28-0,08}{1-0,07}=\frac{0,20}{0,93}=0,22.$$

Последующие предикторы

Ту же процедуру можно использовать для вычисления доли дисперсии, объясняемой предикторами, вводимыми в многомерный регрессионный анализ на втором и третьем этапах. Интерес ребенка к учебе вводится на втором этапе. Квадрат множественной корреляции, или доля дисперсии академической успеваемости, объясняемая тремя предикторами — родительской заинтересованностью, заинтересованностью учителей и интересом ребенка к учебе, равен сумме произведений стандартизированных частных регрессионных коэффициентов на соответствующие корреляции для этих трех предикторов.

Стандартизированные частные коэффициенты регрессии для родительской заинтересованности, заинтересованности учителей и интереса ребенка к учебе приближенно равны 0,17; 0,07 и 0,36 соответственно. Формулы вычисления стандартизированных регрессионных коэффициентов второго и третьего порядков здесь не приводим, так как по мере увеличения порядка коэффициента они существенно усложняются. Обратите внимание на то, что стандартизированные частные регрессионные коэффициенты для родительской заинтересованности и заинтересованности учителей, вычисляемые на втором этапе (с учетом интереса ребенка), отличаются от стандартизованных коэффициентов, вычисленных на первом этапе, так как учитывается еще одна переменная (интерес ребенка). Корреляции между академической успеваемостью, с одной стороны, и родительской заинтересованностью, заинтересованностью учителей и интересом ребенка к учебе — с другой, равны 0,30; 0,28 и 0,44 соответственно. Следовательно, квадрат множественной корреляции для этих трех предикторов оказывается приближенно равным 0,23:

$$(0,17\times 0,30)+(0,07\times 0,28)+(0,36\times 0,44)=0,05+0,02+0,16=0,23.$$

Чтобы определить долю дисперсии академической успеваемости, объясняемой интересом ребенка к учебе помимо и сверх той

доли, которая объясняется родительской и учительской заинтересованностью, вычтем из полученного квадрата множественной корреляции квадрат множественной корреляции для родительской и учительской заинтересованности, вычисленный на первом этапе. В результате приближенно имеем 0,10 (0,23 – 0,13 = 0,10).

Способности ребенка являются четвертой и последней переменной, включаемой в иерархический многомерный регрессионный анализ. Стандартизированные коэффициенты регрессии для родительской заинтересованности, заинтересованности учителей, интереса ребенка к учебе и умственных способностей ребенка приближенно равны 0,06; 0,18; 0,13 и 0,53 соответственно. Корреляции между академической успеваемостью, с одной стороны, и родительской заинтересованностью, заинтересованностью учителей, интересом ребенка к учебе и умственными способностями ребенка — с другой, равны 0,30; 0,28; 0,44 и 0,59 соответственно. Следовательно, квадрат множественной корреляции для этих четырех предикторов приближенно равен 0,44:

$$(0{,}06 \times 0{,}30) + (0{,}18 \times 0{,}28) + (0{,}13 \times 0{,}44) + (0{,}53 \times 0{,}59) = \\ = 0{,}02 + 0{,}05 + 0{,}06 + 0{,}31 = 0{,}44.$$

Доля дисперсии академической успеваемости, объясняемой способностями ребенка помимо и сверх той доли, которая объясняется родительской и учительской заинтересованностью и интересом ребенка к учебе, может быть получена как разность квадратов множественной корреляции, вычисленных на третьем (после ввода переменной «Способности») и втором этапах, учитывавшем только переменные родительской заинтересованности, заинтересованности учителей и интереса ребенка к учебе. Она приближенно равна 0,21 (0,44 – 0,23 = 0,21).

Статистическая значимость объясненной дисперсии

Статистическая значимость изменения доли дисперсии зависимой переменной, объясняемой предиктором, вводимым на текущем шаге, вычисляется с помощью F-отношения точно так же, как и в случае пошаговой множественной регрессии. Формула для F-отношения выглядит следующим образом:

$$F = \frac{\dfrac{|\text{Изменение } R^2|}{|\text{Число добавленных предикторов}|}}{\dfrac{|1 - R^2|}{|N - \text{число предикторов} - 1|}}.$$

Величина F имеет две степени свободы, одна равна знаменателю числителя формулы, другая — знаменателю знаменателя. Чис-

ло степеней свободы числителя равно количеству добавленных на данном этапе предикторов. В нашем примере на первом этапе оно равно двум, а на втором и третьем этапах — единице. Число степеней свободы знаменателя равно количеству наблюдений минус количество предикторов на предшествующих этапах и текущем этапе включительно минус единица.

На первом этапе регрессионного анализа изменение R^2 и само R^2 совпадают и приближенно равны 0,13. Величина F-отношения для доли дисперсии академической успеваемости, объясняемой заинтересованностью родителей и учителей, приближенно равна 19,70:

$$\frac{\frac{0,13}{2}}{\frac{(1-0,13)}{(270-2-1)}} = \frac{0,065}{\frac{0,87}{267}} = \frac{0,065}{0,0033} = 19,70.$$

Две степени свободы F-отношения равны 2 и 267 соответственно. Уровень значимости менее 0,001 свидетельствует о неслучайности взаимосвязи между анализируемыми переменными.

На втором этапе величина изменения R^2 равна 0,10, а само R^2 равно 0,23. F-отношение для доли дисперсии академической успеваемости, объясняемой интересом ребенка к учебе, приближенно равно 34,38:

$$\frac{\frac{0,10}{1}}{\frac{(1-0,23)}{(270-3-1)}} = \frac{0,10}{\frac{0,77}{266}} = \frac{0,10}{0,0029} = 34,48.$$

Данное F-отношение обладает двумя степенями свободы, соответственно равными 1 и 266. Величина F более чем на 99,9 % свидетельствует о том, что взаимосвязь между анализируемыми переменными не может быть обусловлена только случайными факторами (уровень значимости менее 0,001).

На третьем этапе величина изменения R^2 равна 0,21, а само R^2 равно 0,44. Величина F-отношения для доли дисперсии академической успеваемости, объясняемой умственными способностями ребенка, приближенно равна 100,00:

$$\frac{\frac{0,21}{1}}{\frac{(1-0,44)}{(270-4-1)}} = \frac{0,21}{\frac{0,56}{265}} = \frac{0,21}{0,0021} = 100,00.$$

Это F-отношение имеет две степени свободы, равные 1 и 265 соответственно, и значимо на уровне менее 0,001.

Таблица 5.1. **Основные результаты иерархической множественной регрессии**

Шаги	Переменная	R^2	Изменение R^2	F	df_1	df_2	p
1	Заинтересованность родителей Заинтересованность учителей	0,13	0,13	19,70	2	267	0,001
2	Интерес ребенка	0,23	0,10	34,48	1	266	0,001
3	Способности ребенка	0,44	0,21	100,00	1	265	0,001

Результаты рассмотренного иерархического многомерного регрессионного анализа отражены в табл. 5.1.

Статистическая значимость частных коэффициентов регрессии

Когда на каком-то этапе в анализ включаются две или более независимые переменные, как в нашем случае, величина F-отношения позволяет лишь определить, будет ли статистически значимым увеличение доли дисперсии зависимой переменной, объясняемое включаемыми на данном шаге предикторами. В таком случае полезно выяснить, являются ли частные регрессионные коэффициенты статистически значимыми. Статистическая значимость частного регрессионного коэффициента выявляется с помощью критерия Стьюдента (t-критерия) по формуле:

$$t = \frac{[\text{Нестандартизированный частный коэффициент регрессии } (B)]}{[\text{Стандартная ошибка}]}.$$

Чем больше стандартная ошибка, тем меньше уверенность в том, что величина соответствующего нестандартизированного частного регрессионного коэффициента отражает его истинное значение в генеральной совокупности. Иными словами, тем больше вероятность того, что величина этого коэффициента будет отличаться от его истинного значения. Бóльшие значения t имеют бóльшую вероятность быть статистически значимыми[1]. Число степеней свободы для данного t-критерия равно количеству наблюдений минус 2.

Формулы для вычисления нестандартизированных частных коэффициентов регрессии и их стандартных ошибок можно найти в

[1] Т.е. значимо отличаться от нуля.

других источниках (E.J. Pedhazur, 1982; D. Cramer, 1998). Нестандартизированные частные регрессионные коэффициенты для родительской и учительской заинтересованности приближенно равны 0,47 и 0,28 соответственно, а их стандартные ошибки — 0,12 и 0,08. Следовательно, соответствующие t-величины приближенно равны 3,92 (0,47/0,12 = 3,92) и 3,50 (0,28/0,08 = 3,50). Обладая 268 степенями свободы, они являются статистически значимыми с уровнем значимости менее 0,001.

Отчет о результатах

Один из возможных способов кратко описать результаты данного иерархического многомерного регрессионного анализа состоит в следующем: «При проведении иерархического многомерного регрессионного анализа на первом этапе были одновременно введены переменные, отражающие родительскую и учительскую заинтересованность в школьных успехах ребенка, которые объяснили примерно 13 % дисперсии академической успеваемости ребенка ($F_{2,267} = 19{,}70$; $p < 0{,}001$), причем доли объясненной дисперсии для каждой из переменных были примерно одинаковы. Частные регрессионные коэффициенты были статистически значимыми как для заинтересованности родителей ($B = 0{,}47$; $t_{268} = 3{,}92$; $p < 0{,}001$), так и для заинтересованности учителей ($B = 0{,}28$; $t_{268} = 3{,}50$; $p < 0{,}001$). Интерес ребенка к учебе включался в анализ на втором этапе и объяснял дополнительные 10 % дисперсии ($F_{1,266} = 34{,}48$; $p < 0{,}001$). Способности ребенка включались в анализ на третьем этапе и объясняли еще 21 % дисперсии зависимой переменной ($F_{1,265} = 100{,}00$; $p < 0{,}001$). Более высокая академическая успеваемость оказалась связана с более высокой родительской заинтересованностью, заинтересованностью учителей, интересом ребенка к учебе и способностями ребенка».

Реализация процедуры в программе SPSS для Windows

Приведем алгоритм процедуры иерархического многомерного регрессионного анализа.

Если данные из табл. 4.1 были сохранены в файле, откройте этот файл в **Редакторе данных (Data Editor)**, выбирая последовательно пункты меню **Файл (File)**, **Открыть (Open)** и **Данные (Data)**, указывая имя файла в появившемся диалоговом окне **Открытие файла (Open File)** и нажимая в этом окне на кнопку **Открыть (Open)**. Если же данные не были сохранены, введите их так, как показано на рис. 4.1, и припишите наблюдениям соответствующие весовые коэффициенты с помощью процедуры **Задание ве-**

совых коэффициентов (Weight Cases), описанной в предыдущей главе.

Выберите пункт **Анализ (Analyze)** в строке меню в верхней части окна программы, в появившемся ниспадающем меню — пункт **Регрессия (Regression)**, а в нем подпункт **Линейная регрессия (Linear)**, открывающий диалоговое окно **Линейная регрессия (Linear Regression)**, показанное на рис. 4.3.

Выберите переменную **Успеваемость ребенка («детуспев»)** и с помощью верхней кнопки ▸ переместите ее в поле **Зависимая переменная (Dependent)**.

Выберите переменные **Заинтересованность родителей («родинтер»)** и **Заинтересованность учителей («учинтер»)** и с помощью второй сверху кнопки ▸ переместите их в список **Независимые переменные (Independent(s))**.

Нажмите на кнопку **Следующий (Next)**, расположенную непосредственно выше списка **Независимые переменные (Independent(s))** и справа от метки **Блок 1 из 1 (Block 1 of 1)** и задайте второй блок переменных, который в нашем случае состоит из единственной переменной.

Выберите переменную **Интерес ребенка к учебе («детинтер»)** и с помощью второй сверху кнопки ▸ переместите ее в список **Независимые переменные (Independent(s))**.

Нажмите на кнопку **Следующий (Next)**, расположенную непосредственно выше списка **Независимые переменные (Independent(s))** и справа от метки **Блок 2 из 2 (Block 2 of 2)** и задайте третий (и последний) блок переменных.

Выберите переменную **Способности ребенка («детспос»)** и, нажав вторую сверху кнопку ▸, переместите ее в список **Независимые переменные (Independent(s))**.

Нажав на кнопку **Статистики (Statistics)**, откройте диалоговое окно **Линейная регрессия: Статистики (Linear Regression: Statistics)**, представленное на рис. 4.4.

Установите флажок **Изменение квадрата R (R squared change)**, чтобы вывести статистическую информацию, находящуюся в столбцах под общим заглавием **Статистика изменений (Change Statistics)** в табл. 5.1.

Установите флажок **Описательные статистики (Descriptives)**, чтобы отобразить средние значения, стандартные отклонения и число наблюдений для пяти рассматриваемых переменных, а также их полную корреляционную матрицу, величины статистической значимости этих корреляций и число наблюдений, на которых они основаны.

Установите флажок **Часть (получастные)** и **Частные корреляции (Part and partial correlations)**, чтобы вывести данные статистические величины, как это показано в двух последних столбцах табл. 5.2.

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это диалоговое окно и вернуться в основное диалоговое окно.

Нажмите **OK**, чтобы провести данный анализ.

Выводимая SPSS информация по результатам работы

В данной книге приведем только две таблицы из окна вывода SPSS — Model Summary таблицу основных характеристик моделей (см. табл. 5.2) и Coefficients таблицу коэффициентов (табл. 5.3). Все расхождения между значениями величин, приводимых в данных таблицах, и результатами наших расчетов происходят из-за ошибок округления, при этом величины в таблицах, выводимых SPSS, точнее. В таблице коэффициентов (Coefficients) приведены значения стандартизированных регрессионных коэффициентов (Бета) (Standardized Coefficients (или Betas)) и нестандартизированных регрессионных коэффициентов B (Unstandardized Coefficients (или Bs)), их стандартных ошибок (Std. Error), величин *t* и их статистических значимостей (Sig.). Можно проверить величину доли дисперсии, объясняемой на каждом шаге, складывая произведения стандартизированных регрессионных коэффициентов и коэффициентов корреляции. Например, на первом этапе доля объясняемой дисперсии приближенно равна 0,132, что совпадает со значением в таблице основных характеристик моделей (Model Summary):

$$(0{,}237 \times 0{,}296) + (0{,}219 \times 0{,}283) = 0{,}070 + 0{,}062 = 0{,}132.$$

Значения *t*-величин можно проверить путем деления нестандартизированных коэффициентов регрессии на их стандартные ошибки. Так, на первом этапе величина *t* для нестандартизированного регрессионного коэффициента заинтересованности родителей приближенно равна 3,99 (0,467/0,117 = 3,99), что практически не отличается от значения 4,000, приведенного в табл. 5.3.

Рекомендуемая литература

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows.* London: Routledge.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction,* 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Таблица 5.2. **Выводимая SPSS таблица основных характеристик моделей**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	0,363[a]	0,132	0,125	1,232	0,132	20,274	2	267	0,000
2	0,477[b]	0,228	0,219	1,164	0,096	33,000	1	266	0,000
3	0,653[c]	0,426	0,418	1,005	0,199	91,791	1	265	0,000

a. Predictors: (Constant), Заинтересованность учителей, Заинтересованность родителей.

b. Predictors: (Constant), Заинтересованность учителей, Заинтересованность родителей, Интерес ребенка.

c. Predictors: (Constant), Заинтересованность учителей, Заинтересованность родителей, Интерес ребенка, Способности ребенка.

Таблица 5.3. **Выводимая SPSS таблица коэффициентов**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	1,357	0,305		4,449	0,000			
	Заинтересованность родителей	0,467	0,117	0,237	4,000	0,000	0,296	0,238	0,228
	Заинтересованность учителей	0,279	0,076	0,219	3,693	0,000	0,283	0,220	0,211
2	(Constant)	0,958	0,297		3,229	0,001			
	Заинтересованность родителей	0,344	0,112	0,174	3,057	0,002	0,296	0,184	0,165
	Заинтересованность учителей	0,088	0,079	0,069	1,118	0,264	0,283	0,068	0,060
	Интерес ребенка	0,406	0,071	0,357	5,745	0,000	0,439	0,332	0,310
3	(Constant)	-0,133	0,280		-0,476	0,635			
	Заинтересованность родителей	0,112	0,100	0,057	1,121	0,263	0,296	0,069	0,052
	Заинтересованность учителей	0,223	0,069	0,175	3,216	0,001	0,283	0,194	0,150
	Интерес ребенка	0,143	0,067	0,125	2,135	0,034	0,439	0,130	0,099
	Способности ребенка	0,736	0,077	0,513	9,581	0,000	0,594	0,507	0,446

a. Dependent Variable: Успеваемость ребенка.

Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

Дополнительная литература, рекомендуемая научным редактором

Кулаичев А.П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

ЧАСТЬ III
УСТАНОВЛЕНИЕ ПОСЛЕДОВАТЕЛЬНЫХ СООТНОШЕНИЙ МЕЖДУ ТРЕМЯ И БОЛЕЕ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

Глава 6
АНАЛИЗ ПУТЕЙ ПРИ ДОПУЩЕНИИ ОБ ОТСУТСТВИИ ОШИБКИ ИЗМЕРЕНИЯ

Предисловие научного редактора

В ч. III рассматривается путевой анализ, являющийся частным случаем структурного моделирования и, как уже отмечалось, не имеющий достаточной методической базы на русском языке.

В гл. 6 представлен анализ путей в предположении отсутствия ошибки измерения, а в гл. 7 ошибки измерения учитываются.

Анализ путей включает задание моделей, показывающих, каким образом три или более переменные соотносятся друг с другом. Часто бывает полезно изобразить такую модель, используя соответствующую диаграмму путей, где переменные могут быть расположены в определенном порядке — слева направо, если говорить о направлении их влияния друг на друга[1], а взаимосвязь между переменными изображается соединяющей их линией. Давайте рассмотрим простейший случай трех переменных и будем считать, что эти три переменные суть академическая успеваемость ребенка, его интерес к учебе и заинтересованность родителей в школьных успехах своего ребенка, измеряемые одновременно. Существует множество разных моделей, которые можно задать, используя три переменные. При этом необходимо выбрать одну или несколько переменных, которые мы хотим объяснить. В случае трех переменных нас может интересовать объяснение одной переменной или любой пары из них. Например, можно считать, что и интерес ребенка к учебе, и родительская заинтересованность в его школьных успехах находятся под влиянием академической успеваемости ребенка, а также, что интерес ребенка к учебе и родительская заинтересованность в его академических успехах связаны друг с другом. Данную модель можно изоб-

[1] Соблюдение такого порядка не является строго обязательным. При изображении диаграммы скорее обращают внимание на ее наглядность.

Рис. 6.1. Диаграмма путей, показывающая влияние академической успеваемости на интерес ребенка и заинтересованность родителей и ковариацию двух последних параметров

разить с помощью диаграммы путей (рис. 6.1). Академическая успеваемость ребенка находится слева, поскольку мы считаем, что она влияет на интерес ребенка к учебе и родительскую заинтересованность в его школьных успехах. Две последние переменные расположены правее. Направление влияния показано прямыми стрелками, направленными вправо и соединяющими академическую успеваемость ребенка как с его интересом к учебе, так и с родительской заинтересованностью в его академических успехах. Тот факт, что интерес ребенка и заинтересованность родителей связаны друг с другом, отражен на диаграмме с помощью соединяющей эти переменные дугообразной линии со стрелками на обоих концах. В общем случае двусторонняя дугообразная стрелка и односторонние прямые стрелки называются параметрами, их значения показывают с помощью коэффициентов, которые могут быть стандартизированы так, что их значения лежат в пределах от 0 до +1,00.[1]

Количество параметров, которые предполагается вычислить в модели, ограничено числом измеренных переменных, которые могут использоваться для этих вычислений. В случае трех переменных могут быть вычислены только три параметра[2]. Такая модель называется *однозначно определенной*, поскольку у нас как раз имеется необходимое и достаточное количество переменных, чтобы вычислить три данных параметра. Если бы в модели предполагалось вычислить четыре или более параметров, то мы столкнулись бы с ситуацией недоопределенной модели, поскольку не было бы достаточного количества измеренных переменных, чтобы определить или оценить все параметры. Одна из недоопределенных моделей представлена на рис. 6.2, где отражена взаимно-обратная или

[1] Правильнее было бы говорить, что стандартизованные значения параметров лежат в диапазоне от –1 до +1.

[2] Такое ограничение выполняется для множественной регрессии. В структурном моделировании, как будет показано далее, число параметров, которые можно оценить, также ограничено, но там действует другая формула.

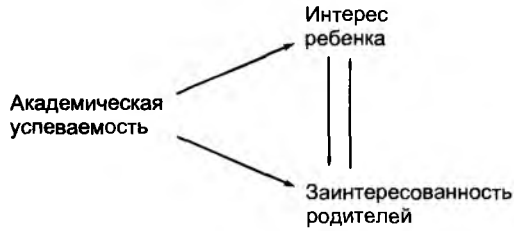


Рис. 6.2. Диаграмма путей, показывающая влияние академической успеваемости на интерес ребенка к учебе и родительскую заинтересованность в его школьных успехах, при взаимном влиянии двух последних параметров

двусторонняя взаимосвязь между интересом ребенка к учебе и родительской заинтересованностью в школьных успехах ребенка, при которой интерес ребенка влияет на родительскую заинтересованность, а та, в свою очередь, влияет на интерес ребенка к учебе[1]. Поэтому в данном случае невозможно оценить влияние интереса ребенка на родительскую заинтересованность независимо от влияния родительской заинтересованности на интерес ребенка к учебе, поскольку при оценке обоих параметров используется одна и та же переменная — академическая успеваемость ребенка.

Модели, в которых число свободных параметров меньше числа переменных, являются переопределенными, поскольку какие-то из возможных параметров предполагаются уже заранее известными. Например, модель, изображенная на рис. 6.1, была бы переопределенной, если бы интерес ребенка к учебе и родительская заинтересованность в его успеваемости не коррелировали между собой[2]. Хотя ценность подобных[3] моделей может быть и не совсем очевидной в случае всего лишь трех переменных, однако они представляют все больший интерес по мере увеличения числа переменных, поскольку позволяют определить, какая из моделей с наименьшим числом параметров лучше всего объясняет экспериментальные данные.

Анализ путей иногда называют каузальным анализом (например, L. R. James и др., 1982), или каузальным моделированием (например, H. B. Asher, 1983). Подобные термины могут вводить в

---

[1] Если сравнивать множество параметров, подлежащих оценке модели, приведенной на рис. 6.1, и модели, приведенной на рис. 6.2, то легко заметить, что двусторонняя стрелка между двумя зависимыми переменными (1 параметр) заменилась двумя односторонними стрелками, т.е. число параметров, подлежащих оценке, увеличилось на 1 и стало больше числа переменных, равного 3. Поэтому мы и говорим о модели, изображенной на рис. 6.2, как недоопределенной.

[2] Т.е. ковариация между этими двумя переменными была заранее определена как равная нулю.

[3] Переопределенных.

заблуждение, если их интерпретировать таким образом, будто анализ путей может применяться для того, чтобы определить, приводят ли изменения одной переменной к изменениям другой. Анализ путей лишь позволяет выяснить, связаны ли две или более переменных, и не может определить, является ли выявленная связь именно детерминирующей. Адекватной процедурой для установления детерминаций является правильный экспериментальный план, в котором можно изменять значения детерминирующей, или независимой, переменной и измерять эффект такого воздействия на зависимую переменную, при условии того, что значения остальных переменных остаются постоянными.

Когда переменные просто измеряются в один и тот же момент времени, как в нашем примере с академической успеваемостью, то невозможно определить их детерминирующий порядок. В терминах взаимосвязи между академической успеваемостью ребенка, например, и его интересом к учебе возможны следующие варианты: 1) интерес ребенка к учебе определяет его школьные успехи; 2) школьная успеваемость определяет интерес ребенка к учебе; 3) оба параметра влияют друг на друга; 4) непосредственная взаимосвязь параметров фиктивна и обусловлена влиянием какой-то другой переменной, например социально-экономическое положение ребенка. Будем считать, что нас интересует в качестве объясняемой переменной не интерес ребенка к учебе или заинтересованность родителей в его школьных успехах, а академическая успеваемость ребенка.

Проиллюстрируем применение анализа путей с помощью трех моделей, изображенных на рис. 6.3, *а—в*. Первую модель назовем «коррелирующей прямой», поскольку в ней предполагается, что интерес ребенка к учебе коррелирует с заинтересованностью родителей в его успехах, и обе переменные влияют на академическую успеваемость ребенка. Вторую модель будем называть «прямой и косвенной», поскольку на академическую успеваемость ребенка прямо влияет родительская заинтересованность в его школьных успехах, а родительская заинтересованность косвенно влияет на его академическую успеваемость через интерес ребенка к учебе, на который она оказывает непосредственное влияние. Третью модель назовем «косвенной», поскольку в ней родительская заинтересованность лишь косвенно влияет на академическую успеваемость ребенка через его интерес к учебе, на который она оказывает непосредственное воздействие.

Необходимо остановиться на двух моментах, касающихся этих моделей. Во-первых, они не исчерпывают все варианты моделей, которые можно проверить. Например, можно проверить модель, в которой не разрешена корреляция между интересом ребенка к учебе и родительской заинтересованностью в его школьных успехах, или такую модель, где интерес ребенка оказывает

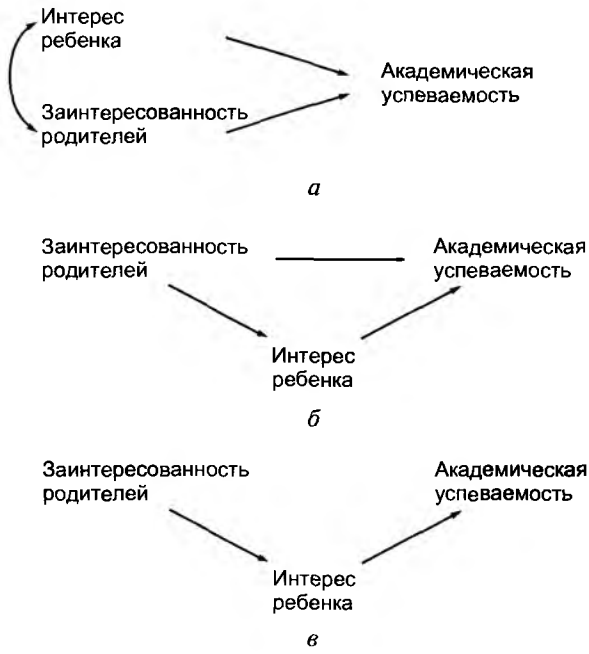


Рис. 6.3. Три модели путей:

а — «коррелирующая прямая»; *б* — «прямая и косвенная»; *в* — «косвенная»

влияние на родительскую заинтересованность. Во-вторых, нельзя определить, какая из моделей — первая («коррелирующая прямая») или вторая («прямая и косвенная») — обеспечивает более адекватное представление данных, поскольку обе модели полностью определены, и поэтому обе смогут полностью объяснить экспериментальные данные. Однако есть возможность определить, дает ли третья модель («косвенная») настолько же адекватное описание данных, насколько это позволяет сделать вторая модель, так как третья модель является подмножеством последней[1].

Приведем расчеты, включаемые в анализ путей, с помощью придуманных данных из табл. 6.1. Три переменные — академическая успеваемость ребенка (АУ), интерес ребенка к учебе (ДИ) и заинтересованность родителей в его школьных успехах (РИ) — состоят из трех показателей, изменяющихся в пределах от 1 до 9. Более высокие баллы означают более высокий уровень выражен-

[1] В литературе, посвященной анализу путей, часто используется также термин «вложенная модель» (nested model). Третья модель получается из второй за счет фиксирования одного из параметров. Влияние интереса родителей на академическую успеваемость полагается равным нулю. Термин «вложенная модель» уже использован в гл. 2.

Таблица 6.1. **Баллы по девяти показателям, оценивающим три переменные**

Наблюдения	АУ*1	АУ2	АУ3	ДИ**1	ДИ2	ДИ3	РИ***1	РИ2	РИ3
1	3	5	3	4	5	2	3	2	4
2	2	3	4	3	2	2	2	4	2
3	3	4	5	4	6	3	5	2	4
4	4	5	5	6	5	7	4	3	3
5	5	5	4	4	3	5	2	4	4
6	3	6	5	4	3	3	5	3	4
7	4	3	4	5	5	4	4	3	6
8	6	7	6	6	5	5	7	8	9
9	4	7	4	4	6	5	7	8	5
10	3	2	3	3	4	3	5	3	4
11	7	5	4	3	3	2	3	2	3
12	9	6	6	5	6	5	6	4	5
13	7	6	5	8	8	6	3	4	3
14	5	4	4	8	5	6	6	4	5

*АУ — академическая успеваемость; **ДИ — интерес ребенка к учебе (детский интерес); ***РИ — родительская заинтересованность в школьных успехах ребенка (родительский интерес).

ности этих качеств. Средний балл для каждой из этих трех переменных приведен в табл. 6.2. Коэффициент корреляции между академической успеваемостью и интересом ребенка к учебе равен 0,53, между академической успеваемостью и родительской заин-

Таблица 6.2. **Средние баллы для трех переменных**

Наблюдения	АУ*	ДИ**	РИ***
1	3,67	3,67	3,00
2	3,00	2,33	2,67
3	4,00	4,33	3,67
4	4,67	6,00	3,33
5	4,67	4,00	3,33
6	4,67	3,33	4,00
7	3,67	4,67	4,33

Окончание табл. 6.2

Наблюдения	АУ*	ДИ**	РИ***
8	6,33	5,33	8,00
9	5,00	5,00	6,67
10	2,67	3,33	4,00
11	5,33	2,67	2,67
12	7,00	5,33	5,00
13	6,00	7,33	3,33
14	4,33	6,33	5,00

*АУ — академическая успеваемость; **ДИ — интерес ребенка к учебе (детский интерес); ***РИ — родительская заинтересованность в школьных успехах ребенка (родительский интерес).

тересованностью — 0,45, между интересом ребенка к учебе и родительской заинтересованностью в его школьных успехах — 0,38.

Множественная регрессия

Простейший способ проведения анализа путей состоит в использовании множественной регрессии. Поскольку стандартизированный коэффициент регрессии между независимой (предиктором) и зависимой (критерием) переменными равен коэффициенту корреляции между ними в случае, когда нас интересуют только эти две переменные, множественная регрессия оказывается необходимой для вычисления стандартизированных регрессионных коэффициентов с учетом наличия одной или нескольких дополнительных независимых переменных. Путевые коэффициенты для трех моделей, изображенных на рис. 6.3, представлены на рис. 6.4. В первых двух моделях имеются пути, где мы должны учесть наличие другой переменной, и это пути между родительской заинтересованностью и академической успеваемостью[1] и между интересом ребенка к учебе и его успеваемостью[2]. В обоих случаях подходящие путевые коэффициенты — стандартизованные частные коэффициенты множественной регрессии, где зависимой переменной является академическая успеваемость, а независимыми — интерес ребенка к учебе и родительская заинтересованность в его школьных успехах. Остальные путевые коэффициенты в этих моделях, такие, как коэффициент между роди-

[1] Дополнительным учитываемым предиктором является интерес ребенка.

[2] Дополнительным учитываемым предиктором является заинтересованность родителей.

тельской заинтересованностью и интересом ребенка к учебе, не учитывают влияние других переменных и равны по величине коэффициентам корреляции.

Как видно из рис. 6.4, величины путевых коэффициентов для первых двух моделей совпадают, и не существует способа провести различие между двумя этими моделями в терминах используемых статистик. Дает ли третья модель более простое и приемлемое описание данных по сравнению со второй моделью, зависит от того, значим ли статистически стандартизированный коэффициент регрессии между заинтересованностью родителей и акаде-

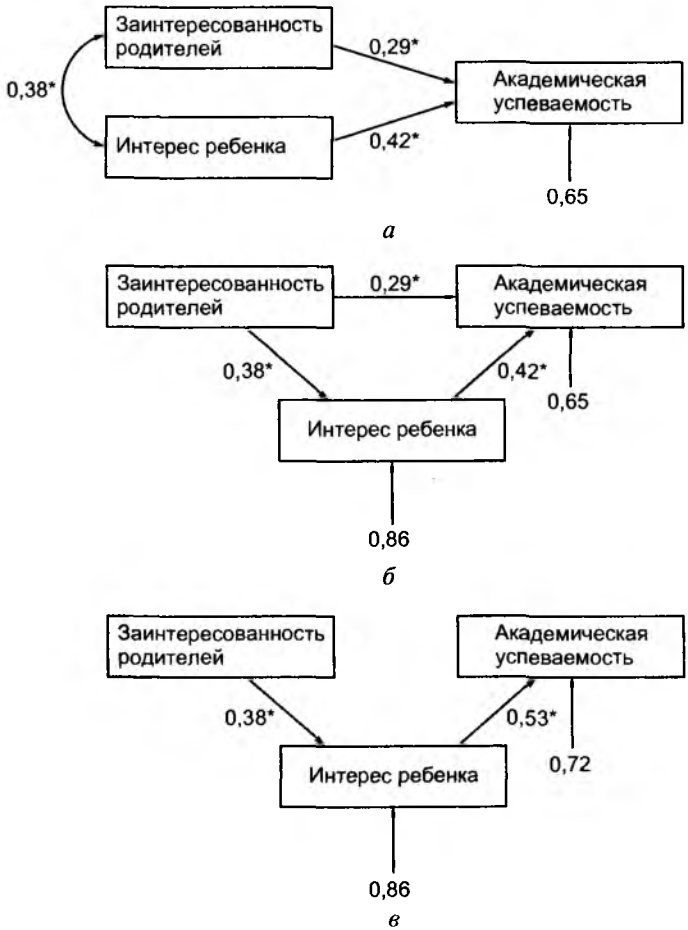


Рис. 6.4. Три путевые модели с коэффициентами и долями необъясненной дисперсии:

а — «коррелирующая прямая»; *б* — «прямая и косвенная»; *в* — «косвенная» [$n = 140$; $^*p < 0{,}001$ (двусторонний критерий)]

мической успеваемостью ребенка во второй модели[1]. Если этот коэффициент статистически значим, то вторая модель является более адекватной. Если данный коэффициент статистически не значим, то третья модель является более простой. Статистическая значимость путевых коэффициентов зависит от размера выборки. В случае 14 наблюдений ни один из путевых коэффициентов не является статистически значимым на уровне 0,05 двустороннего критерия. Использование такой малой выборки неприемлемо в исследованиях подобного рода, где минимальный объем выборки должен быть равен, по крайней мере, 30 наблюдениям. Если размер выборки равен 140 наблюдениям, то все путевые коэффициенты оказываются статистически значимыми на выбранном уровне значимости.

Статистически значимые путевые коэффициенты между родительской заинтересованностью и интересом ребенка к учебе и между интересом ребенка и его академической успеваемостью означают, что интерес ребенка к учебе опосредует взаимосвязь между родительской заинтересованностью в школьных успехах ребенка и его академической успеваемостью. Величина этой косвенной связи между заинтересованностью родителей и академической успеваемостью ребенка может быть вычислена путем умножения путевого коэффициента между родительской заинтересованностью и интересом ребенка к учебе на путевой коэффициент между интересом ребенка к учебе и его академической успеваемостью: 0,16 ($0,38 \times 0,42 = 0,16$) — косвенный эффект для второй модели на рис. 6.3 и 0,20 ($0,38 \times 0,53 = 0,20$) — то же, для третьей модели.

Доля необъясненной дисперсии также часто отображается на диаграммах путей. Эта доля показана на рис. 6.4 числами, расположенными непосредственно под направленными вверх стрелками[2]. В первой модели есть только одна переменная, объясняемая другими переменными, — академическая успеваемость. Следовательно, имеется только одна стрелка, направленная вверх. Доля дисперсии академической успеваемости, не объясняемая заинтересованностью родителей и интересом ребенка к учебе, составляет 0,65. Эта величина получается путем вычитания из единицы неуточненного[3] (обычного) квадрата множественной корреляции ($1 - 0,346 = 0,654$) регрессии академической успеваемости по заинтересованности родителей и интересу ребенка к учебе[4]. Во вто-

---

[1] Т.е. является ли он статистически значимо отличным от нуля.

[2] В правилах построения моделей путевого анализа полагается, что необъясненная дисперсия имеется только у зависимых переменных.

[3] Различие между уточненной и неуточненной множественной корреляцией см. в конце гл. 4.

[4] См. гл. 5: «Квадрат множественной корреляции равен сумме по всем предикторам [стандартизированный частный коэффициент регрессии × коэффициент корреляции]», поэтому квадрат множественной корреляции академической успеваемости можно вычислить так: $(0,29 \times 0,45) + (0,42 \times 0,53) = 0,13 + 0,22 = 0,35$.

рой и третьей моделях имеется также направленная вверх стрелка для интереса ребенка к учебе, который объясняется заинтересованностью родителей в его школьных успехах[1]. Доля необъясненной дисперсии академической успеваемости во второй модели рассчитывается так же, как и для первой модели, и составляет поэтому 0,65. Доля необъясненной дисперсии интереса ребенка к учебе одинакова для второй и третьей моделей и вычисляется как разность единицы и квадрата коэффициента корреляции между заинтересованностью родителей и интересом ребенка к учебе ($0,38^2 = 0,14$), что дает 0,86 ($1 - 0,14 = 0,86$). Доля необъясненной дисперсии академической успеваемости в третьей модели вычисляется аналогично, путем вычитания квадрата коэффициента корреляции между интересом ребенка к учебе и его академической успеваемостью ($0,53^2 = 0,28$) из единицы, что равно 0,72 ($1 - 0,28 = 0,72$).

Одной из проблем выполнения анализа путей с помощью множественной регрессии является отсутствие показателя согласия модели с экспериментальными данными, характеризующего степень воспроизведения исходных корреляций экспериментальных данных с помощью модели. Следовательно, невозможно сравнить степень соответствия данным модели, являющейся подмножеством другой модели — как второй, так и третьей моделей, представленных на рис. 6.4. Вторая проблема состоит в том, что в измерениях переменных обычно имеется ошибка, уменьшающая величины связи между ними. Сравнение величин путевых коэффициентов становится проблематичным, когда ошибка измерения в модели изменяется от переменной к переменной. Структурное моделирование (моделирование структурными уравнениями) дает и меры согласия, а также учитывает ошибку измерения. В следующей главе будет объяснено, как это делается. Однако структурное моделирование может также использоваться для вычисления путевых коэффициентов в моделях, изображенных на рис. 6.4, что будет показано далее. Чтобы познакомиться с компьютерной программой и результатами проведения структурного моделирования, полезно вначале осуществить простейший анализ путей, не учитывающий ошибки измерений. Более того, следует лучше осознавать разницу между анализом путей с учетом и без учета ошибки измерения.

Одним из критериев согласия модели с данными является критерий хи-квадрат, вычисленный с использованием теории наименьших квадратов. Так как первые две модели на рис. 6.4 полностью (однозначно) определены, они обеспечивают идеальное согласие с данными, при котором критерий хи-квадрат равен нулю

[1] Т.е. является зависимой переменной.

и не имеет степеней свободы[1]. При расчете критерия хи-квадрат переопределенной модели учитывается также и объем выборки. В случае 14 наблюдений критерий хи-квадрат для третьей модели равен 1,31 и имеет одну степень свободы. Это значение не является статистически значимым, а вероятность его реализации при случайном раскладе больше 0,05. Однако такая малая выборка не позволяет делать достоверных выводов. В случае более адекватной поставленной задаче выборки из 140 наблюдений число степеней свободы по-прежнему равно единице, но критерий хи-квадрат равен уже 13,96, что является статистически значимым при $p < 0{,}001$. Поскольку третья модель является подмножеством второй, можно сравнить показатели согласия этих моделей, вычитая значения критерия хи-квадрат и числа степеней свободы второй модели из соответствующих значений третьей модели. Поскольку вторая модель обеспечивает идеальное соответствие данным[2], разность между критерием хи-квадрат и числом степеней свободы для данных двух моделей будет совпадать с критерием хи-квадрат (например, 13,96 – 0,00 = 13,96) и числом степеней свободы (1 – 0 = 1) третьей модели, а значимость — со статистической значимостью для третьей модели. При объеме выборки, равном 140 наблюдениям, разность между критерием хи-квадрат и числом степеней свободы будет статистически значимой, показывая, что третья модель значимо хуже согласуется с данными в сравнении со второй моделью[3].

Число степеней свободы модели определяется путем вычитания количества оцениваемых параметров из числа дисперсий и ковариаций наблюдаемых переменных. Число дисперсий и ковариаций наблюдаемых переменных задается общей формулой: $n(n+1)/2$, где n — количество наблюдаемых переменных. Число дисперсий и ковариаций наблюдаемых переменных для всех трех моделей, представленных на рис. 6.4, равно 6 [3(3 + 1)/2 = 6]. Коли-

---

[1] Как уже упоминалось в гл. 2, число параметров, которые можно оценить в модели, не должно превышать числа дисперсий и ковариаций всех измеряемых переменных, т.е. если в модель включено n измеряемых переменных, то максимальное число оцениваемых параметров не должно быть больше $n(n+1)/2$. Число степеней свободы вычисляется как разность между $n(n+1)/2$ и числом оцениваемых параметров. Если эта разность больше нуля, то модель называется *переопределенной*, если разность в точности равна нулю, то модель является *однозначно определенной*, в случае когда разность меньше нуля, модель является *недоопределенной*.

[2] Для нее критерий хи-квадрат и число степеней свободы равны нулю.

[3] То, что третья модель не соответствует экспериментальным данным, следует уже из того, что хи-квадрат равен 13,96 при одной степени свободы, т.е. уровень значимости нулевой гипотезы $p < 0{,}01$, и ее можно отвергнуть. Нулевая гипотеза в случае проверки структурной модели утверждает, что модель соответствует экспериментальным данным. Поэтому проверять по критерию разностей вложенных моделей необходимости нет.

чество оцениваемых параметров равно сумме числа гипостазируемых (устанавливаемых в модели) взаимосвязей и числа остаточных членов, соответствующих ошибкам измерения для каждой переменной в моделях, приведенных на рис. 6.4.[1] Для первых двух моделей оно равно шести, а для третьей — пяти.

Отчет о результатах

Было бы более уместно проанализировать пример, рассматриваемый в настоящей главе, с помощью структурного моделирования, учитывающего ошибки измерений, как это будет сделано в следующей главе. Если бы это не было возможно, краткий способ описания одной из интерпретаций результатов этого довольно простого примера состоит в следующем: «Академическая успеваемость оказалась статистически значимо положительно связанной непосредственно с родительской заинтересованностью ($\beta = 0{,}29$; *df* (число степеней свободы) = 138[2]; *p* (двусторонний критерий) < 0,001) и опосредованно через интерес ребенка к учебе ($\beta = 0{,}20$; $df = 138$; $p < 0{,}05$). Интерес ребенка к учебе был значимо положительно связан как с академической успеваемостью ($\beta = 0{,}42$; $df = 138$; $p < 0{,}001$), так и с родительской заинтересованностью (*r* (коэффициент корреляции) = 0,38; $df = 138$; $p < 0{,}001$)».

Реализация процедуры множественной регрессии в программе SPSS для Windows

Чтобы получить стандартизированные коэффициенты регрессии для моделей, представленных на рис. 6.4, сначала введите данные из табл. 6.2 в **Редактор данных (Data Editor)** с соответствующими весовыми коэффициентами, равными 10, так, как это было описано в гл. 4.

---

[1] Большинство авторов предпочитают различать ошибки измерения для зависимых переменных — необъясненная дисперсия и ошибки измерения для независимых переменных — истинная ковариация. Если не учитывать этого замечания, то, согласно, например, модели *а* на рис. 6.4, можно насчитать лишь 4 оцениваемых параметра. Чтобы прийти к согласию с автором, нужно добавить еще два параметра, соответствующие истинным ковариациям двух независимых переменных: родительской заинтересованности и интереса ребенка. В моделях *б* и *в* к изображенным параметрам нужно добавить ковариацию одной независимой переменной (родительской заинтересованности).

[2] Чтобы предупредить возможное недоумение читателя, еще раз укажем, что в данном случае речь идет о числе степеней свободы при проверке значимости регрессионного коэффициента (140 – 2), а не путевой модели.

Выберите в строке меню в верхней части окна программы пункт **Анализ (Analyze)**, а в ниспадающем меню — пункт **Регрессия (Regression)** и затем — **Линейная регрессия (Linear)**, чтобы открыть диалоговое окно **Линейная регрессия (Linear Regression)**, изображенное на рис. 4.3.

Выберите **АУ** (академическая успеваемость) и нажмите верхнюю кнопку ▸, чтобы переместить данную переменную в поле **Зависимая переменная (Dependent)**.

Выберите переменные **ДИ** (интерес ребенка) и **РИ** (родительская заинтересованность) и с помощью второй сверху кнопки ▸ переместите их в список **Независимые переменные (Independent(s))**.

Нажмите на кнопку **Статистики (Statistics)** и откройте дочернее диалоговое окно **Линейная регрессия: Статистики (Linear Regression: Statistics)**, изображенное на рис. 4.4.

Установите флажок **Описательные статистики (Descriptives)**, чтобы вывести коэффициенты корреляции между тремя переменными и их статистическую значимость.

Нажмите **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно.

Нажмите **ОК** и проведите анализ.

Выводимая SPSS информация по результатам работы

Из всех результатов, выдаваемых программой, приведем лишь табл. 6.3 регрессионных коэффициентов и их статистических значимостей. Стандартизированные коэффициенты регрессии академической успеваемости ребенка (АУ — AcaAch) по его интересу к учебе (ДИ — ChiInt) и родительской заинтересованности (РИ — ParInt) равны 0,417 и 0,286 соответственно.

Таблица 6.3. **Выводимые SPSS частные регрессионные коэффициенты (partial regression coefficients)**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,044	0,316		6,468	0,000
	ДИ	0,358	0,064	0,417	5,584	0,000
	РИ	0,231	0,060	0,286	3,834	0,000

a. Dependent Variable: АУ.

Процедура LISREL для анализа путей без учета ошибок измерения

Применим программу структурного моделирования LISREL для иллюстрации процесса получения результатов для второй и третьей моделей. Запустив LISREL, выберем пункт **File** в строке меню в верхней части окна программы, а в ниспадающем меню — подпункт **New**, открывающий диалоговое окно **New**. Выберем **Syntax only** и откроем окно **Syntax**, в котором будем набирать команды. После окончания набора команд выберем пункт **File**, а в ниспадающем меню — подпункт **Run** (выполнить) **LISREL**. Если мы правильно ввели команды, программа выдаст окно результатов, в противном случае сначала будет отображена диаграмма путей. Чтобы увидеть другие результаты, сопровождающие данную диаграмму, выберите пункт **Window** в строке меню в верхней части окна программы, а в нем — файл с расширением **OUT**.

Команды, выделенные жирным шрифтом, необходимы для выполнения процедуры, позволяющей получить результаты анализа третьей модели:

```
PA: Indirect Model without Error
DAta NInputvar=3 NObserv=140
LAbels
ChiInt AcaAch ParInt
KMatrix
1,00
0,53   1,00
0,38   0,45   1,00
MOdel NYvar=2 NXvar=1 GAmma=FRee BEta=SD PSi=Diagonal
Fixed GAmma (2,1)
PDiagram
OUtput EFfects
```

Опишем кратко назначение этих команд. Строка, начинающаяся с **PA** (сокращение для анализа путей — **P**ath **a**nalysis), выводит заголовок в верхней части каждой «страницы» вывода результатов. Названия команд, таких, как **DA**ta или **LA**bels, могут быть сокращены до двух символов, которые выделены заглавным регистром, чтобы их можно было отличить от остальной части имени команды.

Строка, начинающаяся с **DA**, содержит информацию об анализируемых данных. Она констатирует, что количество (**N**umber) входных (**I**nput) переменных (variables), включаемых в анализ, равно трем, а число (**N**umber) наблюдений (**O**bservation) равно 140.

Строка, начинающаяся с **LA**, информирует нас о том, что имена меток, которые приписаны этим трем переменным, нахо-

дятся в списке на следующей строке. На печать выводятся только первые восемь символов названия метки. Переменные, на которые не направлено ни одной стрелки (заинтересованность родителей), следуют последними. Их часто называют *экзогенными переменными*. Переменные, на которые направлены какие-то стрелки (интерес ребенка к учебе и академическая успеваемость), идут первыми в списке. Их обычно называют *эндогенными переменными*.

Строка, начинающаяся с **KM**, показывает, что данные будут вводиться как корреляционная матрица. Переменные в нижнетреугольной матрице перечислены в том же порядке, что и их метки. Например, первая строка матрицы представляет собой коэффициент корреляции интереса ребенка к учебе с самим собой и, конечно же, равен единице.

Строка, начинающаяся с **MO**, определяет тестируемую модель, которая состоит из двух эндогенных[1] и одной экзогенной[2] переменных. Параметры модели содержатся в ряде матриц. Коэффициенты пути между экзогенными и эндогенными переменными определяются **Gamma** матрицей, в то время как коэффициенты пути между двумя эндогенными переменными определяются **BEta** матрицей. Путевые коэффициенты в Gamma матрице являются свободными (**FRee**) и могут варьировать[3]. Поскольку в данной модели отсутствует путь между родительской заинтересованностью и академической успеваемостью ребенка[4], соответствующий путевой коэффициент должен быть зафиксирован **FIxed**, т.е. не может свободно меняться. Это достигается путем использования строки, начинающейся с **FI**. Нужный путевой коэффициент определяется своим положением в **GA**mma матрице и находится во второй строке первого и, в нашем случае, единственного столбца. Чтобы провести анализ второй модели, просто удалим эту строку, начинающуюся с команды **FI**, освобождая, таким образом, данный параметр.

Матрица Beta определена таким образом, что единственный коэффициент пути между интересом ребенка к учебе и его академической успеваемостью является свободным. Матрица Beta задается как нижняя треугольная матрица с нулевой диагональю (**SubD**iagonal).

PSi матрица задастся как диагональная матрица (Diagonal) и содержит доли необъясненной дисперсии двух эндогенных переменных (интереса ребенка к учебе и его академической успеваемости).

---

[1] Измеряемые эндогенные (зависимые переменные модели) обозначаются *Y*.

[2] Измеряемые экзогенные (независимые от модели переменные) обозначаются *X*.

[3] Столбцы Gamma матрицы соответствуют экзогенным (независимым) переменным, а строки — эндогенным (зависимым) переменным, поэтому в нашем случае матрица состоит из одного столбца и двух строк.

[4] Т.е. значение этого путевого коэффициента фиксировано равным нулю.

Строка, начинающаяся с **PD**, выводит диаграмму путей для данной модели (**P**ath **D**iagram).

Строка, начинающаяся с **OU**, определяет вывод результатов программы (**Ou**t).

Совокупные и косвенные эффекты выводятся путем добавления **EF** (**Ef**fect).

Выводимая LISREL информация по результатам работы

Для краткости представим и прокомментируем только часть выводимых программой результатов. Выводимая диаграмма путей похожа на диаграмму для третьей модели, изображенную на рис. 6.4, за исключением того, что на ней отсутствуют звездочки, указывающие на статистическую значимость, и присутствует подробная информация о критерии хи-квадрат с нормально распределенными весовыми коэффициентами наименьших квадратов и корне квадратном из среднеквадратической ошибки приближения (RMSEA). Коэффициенты путей также отражены в матрицах, приведенных в табл. 6.4. Эти путевые коэффициенты были стандартизированы, поскольку вводилась корреляционная, а не ковариационная матрица. Коэффициент пути между интересом ребенка к учебе (ChiInt) и его академической успеваемостью (AcaAch) выведен в Бета матрице и равен 0,53. Величина в круглых скобках непосредственно под ним представляет собой его стандартную ошибку, равную 0,07. Далее следует значение t-критерия, определяющего статистическую значимость данного путевого коэффициента. В нашем случае $t = 7{,}34$. Для выборки объемом в 140

Таблица 6.4. **Выводимые LISREL результаты для анализа путей в случае третьей модели без учета ошибок измерения**

LISREL Estimates (Maximum Likelihood)

BETA

	ChiInt	AcaAch
ChiInt	- -	- -
AcaAch	0,53	- -
	(0,07)	
	7,34	

GAMMA

	ParInt
ChiInt	0,38
	(0,08)
	4,83
AcaAch	- -

наблюдений критическое значение t для двустороннего критерия равно ±1,98 при уровне значимости 0,05; ±2,62 — при уровне значимости 0,01 и ±3,37 — при уровне значимости 0,001. Так как 7,34 > 3,37, данный путевой коэффициент статистически значимо отличен от нуля на уровне ниже 0,001 двустороннего критерия Стьюдента[1]. Коэффициент пути между родительской заинтересованностью (ParInt) и академической успеваемостью (AcaAch) выведен в Gamma матрице и равен 0,38. При значении, соответствующем t, равному 4,83, этот коэффициент также значим на уровне ниже 0,001 двустороннего критерия Стьюдента.

Значения доли необъясненной дисперсии для двух эндогенных переменных — интереса ребенка к учебе (ChiInt) и его академической успеваемости (AcaAch) — содержатся в PSI матрице в табл. 6.5 и равны 0,86 и 0,72 соответственно.

Косвенный эффект влияния родительской заинтересованности (ParInt) на академическую успеваемость ребенка (AcaAch) показан в матрице из табл. 6.6 и равен 0,20. При $t = 4,03$ данный коэффициент является статистически значимым на уровне ниже 0,001 двустороннего критерия Стьюдента.

Таблица 6.5. **Выводимые LISREL доли необъясненной дисперсии в случае третьей модели без учета ошибок измерения**

```
PSI
Note: This matrix is diagonal.

          ChiInt      AcaAch
          ------      ------
            0,86        0,72
           (0,10)      (0,09)
            8,31        8,31
```

Таблица 6.6. **Выводимое LISREL значение косвенного эффекта в случае третьей модели без учета ошибок измерения**

```
Indirect Effects of X on Y

          ParInt
          -------
ChiInt        -
AcaAch      0,20
           (0,05)
            4,03
```

[1] Или с уровнем доверия более 0,999.

Рекомендуемая литература

Bryman, A. and Cramer, D. (2001) *Quantitative Data Analysis with SPSS Release 10 for Windows*. Hove: Routledge.

Pedhazur, E.J.(1982) *Multiple Regression in Behavioral Research: Explanation and Prediction*, 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Дополнительиая литература, рекомендуемая научным редактором

Митина О.В. Структурное моделирование: состояние и перспективы // Вестн. Пермского гос. пед. ун-та. Сер. 1. Психология. — 2005. — № 2. — С. 3—15.

Глава 7

АНАЛИЗ ПУТЕЙ С УЧЕТОМ ОШИБКИ ИЗМЕРЕНИЯ

В гл. 7 рассмотрены два основных способа обращения с ошибками измерения, когда переменные измеряются с помощью двух или более показателей, или индикаторов, как в примере из гл. 6. Тот же пример будет использован в гл. 7, чтобы можно было сравнить результаты применения двух различных методов.

Надежность измерения

Более простой метод учета ошибок измерения состоит в том, чтобы измерить надежность оценки переменных и использовать это в модели. Один из показателей, который можно применить для этой цели, — альфа Кронбаха (1951), или показатель внутренней устойчивости. Чтобы вычислить коэффициент надежности альфа, производят всевозможные разбиения множества индикаторов на две группы, затем для каждого наблюдения вычисляют усредненные баллы по каждой из двух групп и определяют корреляцию между этими двумя наборами. Альфа Кронбаха — средняя величина по всем таким корреляциям, каждая из которых соответствует определенному разбиению множества индикаторов на два подмножества. В случае всего лишь трех индикаторов для каждой из трех переменных в нашем примере одна половина будет состоять из одного индикатора, в то время как другая — из двух оставшихся индикаторов. Таким образом, имеется три возможных половинных коэффициента надежности (индикатор 1 против агрегированного показателя индикаторов 2 и 3; индикатор 2 против агрегированного показателя индикаторов 1 и 3; индикатор 3 против агрегированного показателя индикаторов 1 и 2). Коэффициенты надежности альфа изменяются в пределах от нуля до единицы. Значения 0,80 и выше считаются показателями надежной меры[1].

Коэффициенты надежности альфа для трех переменных — родительской заинтересованности (РИ), интереса ребенка к уче-

[1] Здесь нет однозначно принятой границы, когда считать измерения надежными. Некоторые исследователи считают хорошим показателем и 0,7 (см.: E. J. Pedhazur, L. P. Schmelkin).

бе (ДИ) и его академической успеваемости (АУ) — соответственно равны 0,83; 0,85 и 0,73. С учетом этих показателей надежности стандартизированные путевые коэффициенты для трех наших моделей путей приведены на рис. 7.1. Индикатор, или измеренный показатель, называется *наблюдаемой*, или *явной*, *переменной* и изображается прямоугольником или квадратом, в то время как сама переменная называется *латентной* и изображается с помощью круга или эллипса. Стрелки направлены от латентных переменных к наблюдаемым и показывают, что индикаторы являются отражением лежащих в их основе латентных переменных. Например, бо́льшая родительская заинтересованность будет отражаться на используемом для ее измерения индикаторе. Однако помимо стрелок, исходящих из латентных переменных, в каждую наблюдаемую переменную входит еще одна дополнительная стрелка. Величина, указываемая рядом с этой стрелкой, представляет собой долю специфической дисперсии, или дисперсии, связанной с ошибкой измерения соответствующего индикатора. Она вычисляется путем вычитания коэффициента надежности из единицы. Так, доля дисперсии ошибки родительской заинтересованности составляет 0,17 ($1 - 0{,}83^1 = 0{,}17$).

Поскольку здесь учитывается дисперсия ошибки, значения путевых коэффициентов на рис. 7.1 выше, чем на рис. 6.4. Так, коэффициент пути между родительской заинтересованностью и академической успеваемостью на рис. 7.1 равен 0,34 (по сравнению с 0,29 на рис. 6.4). Величина косвенного влияния родительского интереса на академическую успеваемость вычисляется совершенно так же, как и в случае, когда анализ путей проводился с помощью множественной регрессии. Путевой коэффициент между родительской заинтересованностью и интересом ребенка к учебе умножается на путевой коэффициент между интересом ребенка к учебе и его академической успеваемостью, что дает приближенное значение косвенного эффекта, равное 0,23 ($0{,}45 \times 0{,}52 = 0{,}23$) для второй модели и 0,34 ($0{,}49 \times 0{,}70 = 0{,}34$) для третьей модели (см. рис. 7.1). В случае выборки из 14 наблюдений ни один из путевых коэффициентов в трех моделях путей, приведенных на рис. 7.1, не является статистически значимым. Анализ путей на такой маленькой выборке проводить нельзя. При более разумном объеме выборки в 140 наблюдений все путевые коэффициенты являются статистически значимыми.

Первые две модели (см. рис. 7.1) являются однозначно определенными и обеспечивают идеальное или предельное согласие с данными при минимальном значении критерия хи-квадрат, равном нулю, и отсутствии степеней свободы. Критерий хи-квадрат

---

[1] Эти показатели вычисляются как корень квадратный из соответствующего показателя надежности (см. далее).

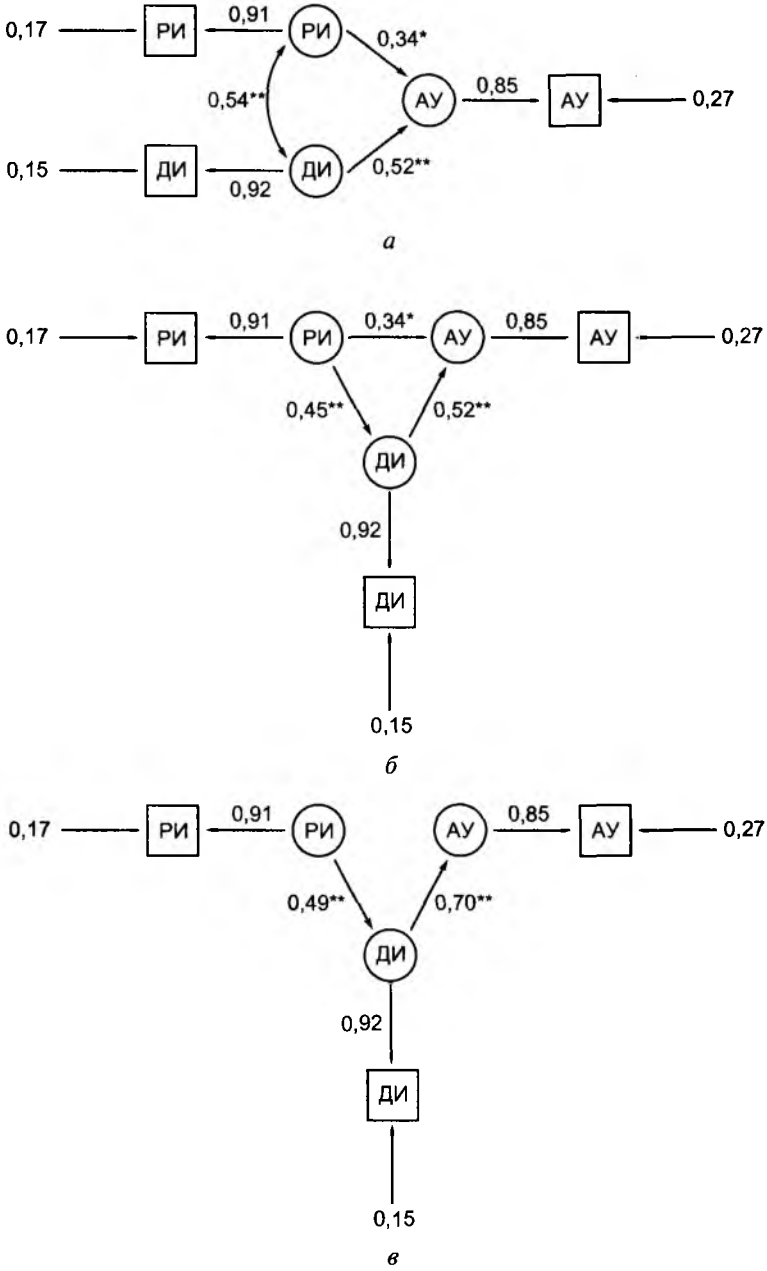


Рис. 7.1. Три модели путей с учетом ошибок измерения:

а — «коррелирующая прямая»; *б* — «прямая и косвенная»; *в* — «косвенная» [n = = 140; *$p < 0{,}05$; **$p < 0{,}001$ (двусторонний критерий)].

для переопределенной модели будет зависеть от объема выборки. При 14 наблюдениях критерий хи-квадрат для третьей модели равен 1,00 и обладает одной степенью свободы. Это значение не является статистически значимым, $p > 0{,}05$. В случае 140 наблюдений число степеней свободы по-прежнему равно единице, но критерий хи-квадрат теперь равен 10,65, что статистически значимо с $p < 0{,}001$. Поскольку третья модель является подмножеством или вложенной моделью по отношению ко второй модели[1], можно сравнить степень соответствия данным третьей и второй моделей, вычитая критерий хи-квадрат и число степеней свободы второй модели из соответствующих показателей третьей. Так как вторая модель обеспечивает идеальное соотвстствие данным, разность значений критерия хи-квадрат и числа степеней свободы будет совпадать с критерием хи-квадрат (например, 10,65 – 0,00 = 10,65), числом степеней свободы (1 – 0 = 0) и статистической значимостью третьей модели. При объеме выборки в 140 наблюдений разность величин хи-квадрат является статистически значимой, показывая, что третья модель хуже согласуется с данными, чем вторая[2].

Число степеней свободы модели определяется путем вычитания числа оцениваемых параметров из числа всевозможных дисперсий и ковариаций наблюдаемых переменных. Число дисперсий и ковариаций наблюдаемых переменных вычисляется по общей формуле: $n(n + 1)/2$, где n — число наблюдаемых переменных. Число дисперсий и ковариаций для всех трех моделей, приведенных на рис. 7.1, равно 6 [3(3 + 1)/2 = 6]. Число оцениваемых параметров равно 6 для первых двух моделей и 5 — для третьей модели. Оцениваемые параметры отображаются в выходном файле программы LISREL для структурного моделирования.

Измерительные модели

Более сложный метод учета ошибки измерения в структурном моделировании в случае, когда переменные измеряются с помощью двух или более индикаторов, состоит в том, чтобы представить переменную как фактор по аналогии с конфирматорным факторным анализом. Индикаторы конкретной переменной ха-

---

[1] Один параметр фиксирован равным нулю.

[2] Как и в предыдущей главе, сделаем замечание: тот факт, что третья модель не соответствует экспериментальным данным, следует уже из того, что критерий хи-квадрат равен 10,65 при одной степени свободы, т.е. уровень значимости нулевой гипотезы $p < 0{,}01$, и ее можно отвергнуть. Нулевая гипотеза в случае проверки структурной модели утверждает, что модель соответствует экспериментальным данным. Поэтому проверять по критерию разностей вложенных моделей необходимости нет.

рактеризуются через нагрузки на фактор, отражающий эту переменную. Для того чтобы модель не оказалась недоопределенной, необходимо фиксировать значение одного из параметров для каждого фактора[1]. Это обычно осуществляется путем приписывания путевому коэффициенту между одним из индикаторов переменной и ее фактором единичного значения, что задает измерительную шкалу латентной переменной. Стандартизированные путевые коэффициенты для моделей показаны на рис. 7.2, где путевым коэффициентам между каждым фактором и первым индикатором этого фактора приписывалось фиксированное единичное значение (АУ1, ДИ1 и РИ1). Значения показателей согласия для данной модели, а также путевых коэффициентов между латентными переменными одни и те же, независимо от того, нагрузка на какой из трех индикаторов по каждому фактору выбирается в качестве фиксированного значения. Косвенный путевой коэффициент между родительской заинтересованностью и академической успеваемостью равен произведению путевых коэффициентов между родительской заинтересованностью и интересом ребенка к учебе и между интересом ребенка к учебе и его академической успеваемостью. Он приближенно равен 0,18 ($0,41 \times 0,43 = 0,18$) для второй модели и 0,28 ($0,44 \times 0,63 = 0,28$) — для третьей модели. При размере выборки в 14 наблюдений ни один из этих путевых коэффициентов не является значимым, в то время как при объеме выборки, равном 140 наблюдениям, все эти путевые коэффициенты являются значимыми.

Поскольку число оцениваемых параметров во всех трех моделях меньше суммарного числа дисперсий и ковариаций наблюдаемых переменных, все три модели являются переопределенными. Как и ранее, соответствие первых двух моделей данным одинаково и не может сравниваться[2]. Для переопределенных моделей значение критерия хи-квадрат зависит от объема выборки, который в нашем случае целесообразно довести до 140 наблюдений. Так как третья модель является подмножеством второй, можно сравнить их с точки зрения лучшего соответствия экспериментальным данным. Это осуществляется путем вычитания критерия хи-квадрат второй модели из критерия хи-квадрат третьей модели ($107,81 - 90,27 = 17,54$) и выяснения статистической значимости этой разности с учетом того, что число степеней свободы равно разности числа степеней свободы данных двух моделей, в нашем случае это 1 ($25 - 24 = 1$). Обладая одной степе-

---

[1] Значение одного из параметров для каждого фактора фиксируют не для того, чтобы преодолеть ситуацию недоопределенности, а чтобы задать масштаб измерений каждой латентной переменной. Собственно говоря, этот же факт указывается в следующем предложении.

[2] Точнее, ни одна из двух моделей не является вложенной в другую.

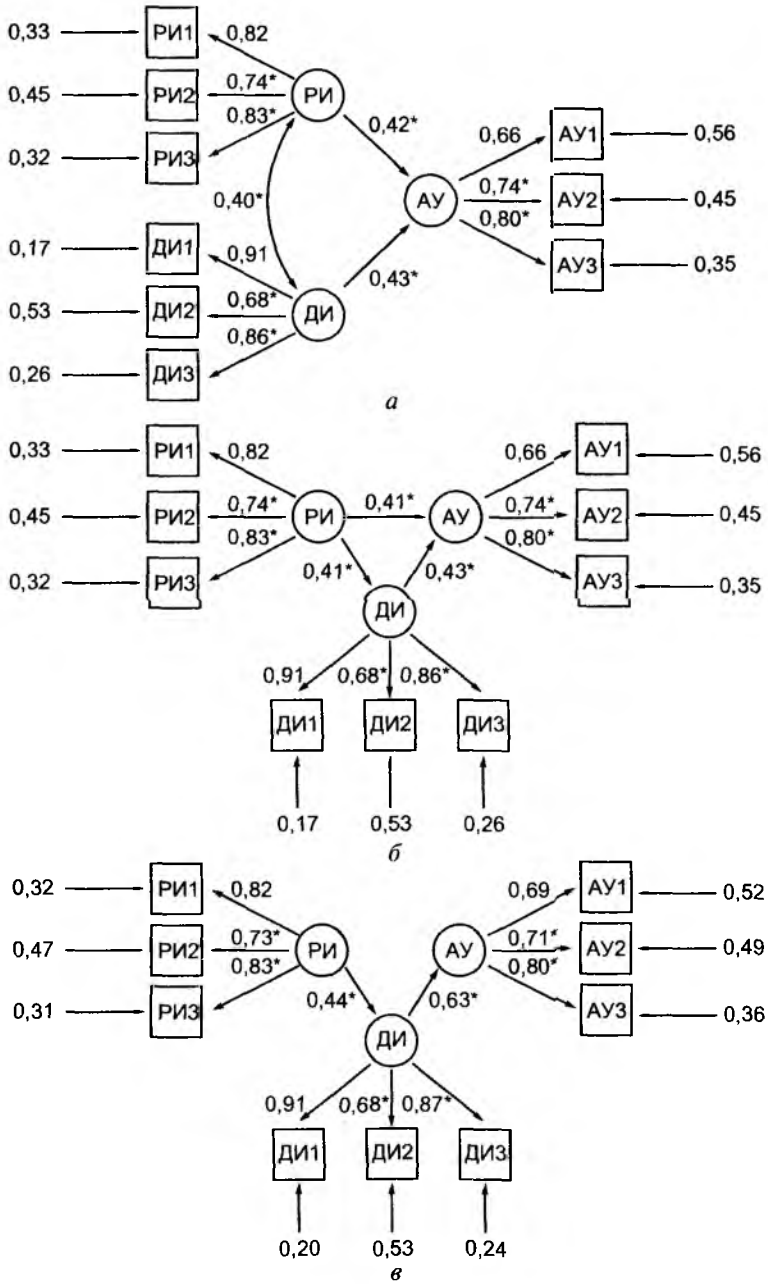


Рис. 7.2. Три модели путей с измерительными моделями:

а — «коррелирующая прямая» (χ^2 (хи-квадрат) = 90,27; *df* (число степеней свободы) = = 24; $p < 0{,}001$); *б* — «прямая и косвенная» ($\chi^2 = 90{,}27$; $df = 24$; $p < 0{,}001$); *в* — «косвенная» ($\chi^2 = 107{,}81$; $df = 25$; $p < 0.001$) [$n = 140$; $^*p < 0{,}001$ (двусторонний критерий)]

нью свободы, критерий хи-квадрат должен иметь значение 3,84, чтобы быть значимым согласно двустороннему критерию на уровне 0,05, что имеет место в нашем случае. Следовательно, третья модель обеспечивает значимо худшее согласие с данными по сравнению со второй моделью.

Однако тот факт, что критерий хи-квадрат для первой и второй моделей является статистически значимым, означает, что эти две модели не обеспечивают достаточного согласия с данными, если использовать критерий хи-квадрат как меру согласия. К настоящему времени разработано множество различных показателей (критериев) согласия. И каждая компьютерная программа, реализующая структурное моделирование, в том числе и LISREL, выдает целый ряд таких показателей[1]. Описание этих индексов можно найти в других книгах (например, J.C.Loehlin, 1998). Выдаваемые LISREL показатели свидетельствуют о том, что согласие наших моделей с данными ниже желаемого. В этих обстоятельствах может оказаться целесообразным освободить один или более параметров, чтобы улучшить соответствие модели данным.

Отчет о результатах

Обычно отчет о результатах структурного моделирования включает довольно длинные объяснения выполненных операций, приводятся диаграммы путей проверяемых моделей. Если нас интересуют исключительно результаты этих моделей, то на диаграммах путей также отображаются путевые коэффициенты и доли необъясненной дисперсии ошибки, как показано на рис. 7.1 и 7.2. Там, где модели с наилучшим согласием отличаются от первоначальных, результаты отображаются в дополнительных диаграммах путей. Обычно для проверяемых моделей приводят значения нескольких индексов согласия.

Процедура LISREL для анализа с учетом надежности измерения

Приведенные ниже команды, выделенные жирным шрифтом, следует выполнить, чтобы получить соответствующие результаты для третьей модели на рис. 7.1:

```
PA: Indirect model with measurement reliability
DAta NInputvar=3 NObserv=140
LAbels
```

[1] В англоязычной литературе для этих показателей используется термин «index» (индекс).

```
CI AA PI
KMatrix
1,00
0,53   1,00
0,38   0,45   1,00
MOdel NYvar=2 NXvar=1 NKvar=1 NEvar=2 c
LY=Diagonal LX=Diagonal c
TEpsilon=Fixed TDelta=Fixed c
GAmma=FRee BEta=SD PSi=Diagonal
Fixed GAmma (2,1)
LEta
CL AA
LKsi
PI
MAtrix LY
*
0,92   0,85
MAtrix LX
*
0,91
MAtrix TEpsilon
0,15   0,27
MAtrix TDelta
*
0,17
PDiagram
OUtput Effect
```

Опишем кратко назначение этих команд. Ключевые слова, такие, как **DAta** (данные) и **LAbels** (метки, имена переменных), могут быть сокращены до первых двух символов, которые для этого выделяются использованием заглавного регистра.

Первая строка, озаглавленная **PA** (**P**ath **A**nalysis — Путевой анализ), обеспечивает вывод заголовка на каждой странице окна вывода результатов.

Строка, начинающаяся с **DA**, указывает число (Number) входных переменных (**Input vari**ables) (в нашем случае оно равно трем) и число (Number) наблюдений (**o**bservations) (равно 140).

Строка, начинающаяся с **LA**, означает, что следующая строка содержит имена вводимых переменных. Имена переменных перечислены в том же порядке, что и переменные в корреляционной матрице, которая помещена на следующей строке после строки, начинающейся с **KM**. Первыми перечислены эндогенные переменные (в нашем случае Детский **И**нтерес к учебе и его **А**кадемическая **У**спеваемость). Экзогенные переменные (у нас — **Р**одительский **И**нтерес) перечисляются последними.

Строки в программе LISREL могут содержать до 127 символов, включая пробелы. Строка, начинающаяся с **MO,** была разбита на

четыре строчки для сокращения числа последовательно отображаемых знаков в строке вывода на странице и лучшего восприятия текста (неразбитая строка содержала 119 символов и пробелов). Для обозначения продолжения строки используется символ «**c**» в конце после завершения определенной частной команды.

Число (Number) измеряемых переменных **Y**, соответствующих эндогенным латентным переменным, равно двум. Число (Number) латентных эндогенных переменных **E** также равно двум. Число (Number) измеряемых переменных **X**, соответствующих экзогенных латентным переменным, равно одному. Число (Number) латентных экзогенных переменных **K** также равно одному[1].

Параметры модели содержатся в нескольких матрицах. Матрица для каждой из латентных переменных (**LY** и **LX**) составлена только из параметров, расположенных на диагонали (**DIagonal**) матрицы, так как имеется только один путевой коэффициент для каждой наблюдаемой и латентной переменной. Этим путевым коэффициентам были присвоены фиксированные значения, равные корням квадратным из коэффициентов надежности альфа для наблюдаемых переменных, с помощью команд **MAtrix LY** и **LX**. Например, путевому коэффициенту между первой эндогенной переменной (**ДИ**) и ее индикатором было присвоено фиксированное значение 0,92 ($\sqrt{0,85} = 0,92$), а коэффициенту пути между второй эндогенной переменной (**АУ**) и ее индикатором — фиксированное значение 0,85 ($\sqrt{0,73} = 0,85$).

---

[1] Здесь необходимо пояснить ситуацию, ибо с формальной точки зрения измеряемая переменная **X** является эндогенной: зависит от латентной переменной **E**. Однако никакой ошибки тут нет, если вспомнить, как объяснял все эти понятия сам автор программы LISREL и разработчик всей терминологии Карл Йореског (K.G.Jöreskog). Он использовал термины «эндогенная» и «экзогенная» для характеристики латентных переменных: независимые латентные переменные называются экзогенными, а зависимые — эндогенными. Путевая диаграмма, содержащая только латентные переменные, называется *структурной моделью*. Согласно концепции Йореского, латентные переменные проявляются только через измеряемые. С одной стороны, установленные непосредственно значения измеряемых переменных позволяют определить значения латентных переменных, измерить которые напрямую нельзя. Соответствующие каждой латентной переменной измеряемые переменные называются *индикаторами* этой латентной переменной. С другой стороны, эти латентные переменные детерминируют индикаторы. Таким образом, индикаторы бывают и у экзогенных, и у эндогенных переменных. Если же возникает ситуация модели, в которой вместо эндогенных и (или) экзогенных латентных переменных рассматриваются наблюдаемые переменные (как это сделано в гл. 6), то предлагается вводить фиктивные латентные переменные для каждой наблюдаемой переменной с фиксированным коэффициентом детерминации, равным 1, и фиксированным остаточным членом, равным нулю. Статистические показатели модели при этом никак не меняются, а соответствующие фиктивные латентные переменные оказываются либо экзогенными, либо эндогенными.

Матрицы долей дисперсии ошибок индикаторов экзогенных (**TDelta**[1]) и эндогенных (**TEpsilon**) латентных переменных являются диагональными по умолчанию, а соответствующим элементам этих матриц могут быть присвоены фиксированные значения с помощью команд **MA**trix **TE**psilon и **TD**elta. Эти значения получаются путем вычитания коэффициента надежности соответствующей переменной из единицы. Например, доле дисперсии ошибки индикатора первой эндогенной переменной (**ДИ**) было присвоено фиксированное значение, равное 0,15 (1 – 0,85 = 0,15), а доле дисперсии ошибки индикатора второй эндогенной переменной (**АУ**) — значение 0,27 (1 – 0,73 = 0,27).

Пути между экзогенными и эндогенными переменными определяются **GA**mma матрицей, в то время как пути между парами эндогенных переменных — **BE**ta матрицей. В **GA**mma матрице путевые коэффициенты могут варьировать. Поскольку в данной модели отсутствует путь между родительской заинтересованностью и академической успеваемостью ребенка[2], соответствующий путевой коэффициент должен быть фиксированным (**FI**xed), а не свободным. Это указывается строкой, начинающейся с **FI**. Данный коэффициент пути определяется своим положением в **GA**mma матрице во второй строке первого столбца. Для анализа второй модели просто удаляем данную строку[3].

BEta матрица задается таким образом, что единственный коэффициент пути между интересом ребенка к учебе и его академической успеваемостью является свободным параметром.

Матрица **PSi** задается как диагональная матрица, в которой указываются оценки ошибок[4] двух эндогенных переменных (интереса ребенка к учебе и его академической успеваемости).

Строка, начинающаяся с **LE**, задает имена (метки **L**abel) двух **e**ta-, или латентных эндогенных, переменных — **д**етского **и**нтереса к учебе (**ДИ**) и **а**кадемической **у**спеваемости ребенка (**АУ**). Сами эти имена приводятся в следующей строке.

Строка, начинающаяся с **LK**, определяет имя (метку **L**able) **k**si-, или латентной экзогенной, переменной **р**одительского **и**нтереса (**РИ**). Имя этой переменной приведено в следующей строке.

Строка, начинающаяся с **PD**, задает вывод диаграммы пути (**p**ath **d**iagram) для данной модели.

Строка, начинающаяся с **OU**, задает спецификации выводимых результатов. Величины прямых и косвенных эффектов выводятся с помощью добавления **EF**.

---

[1] Для того чтобы не путаться во всей этой терминологии матриц, обозначаемой разными буквами, мы отсылаем читателя к основной схеме, предложенной К.Йореского м в приложении.

[2] Т.е. он принудительно приравнен нулю.

[3] Т.е. освобождаем, позволяем ему варьировать.

[4] Свободными по умолчанию (см. прил.).

Выводимая LISREL информация по результатам анализа путевой модели с учетом надежности измерений

Для краткости будут приведены и прокомментированы лишь некоторые фрагменты выводимых результатов. Выводимая диаграмма путей аналогична приведенной на рис. 7.1 для третьей модели с той лишь разницей, что на ней отсутствуют звездочки, означающие статистическую значимость, и присутствуют значения критерия хи-квадрат, вычисленного по методу наименьших квадратов, и корень квадратный из среднеквадратичной ошибки приближения (RMSEA). Путевые коэффициенты также приведены в матрицах табл. 7.1. Например, путевой коэффициент между явной и латентной переменными интереса ребенка к учебе (**CI**[1]) отражен в **Lambda-Y** матрице и равен 0,92. Путевой коэффициент между латентными эндогенными переменными интереса ребенка к учебе (**CI**) и его академической успеваемости (**AA**[2]) приведен в BETA матрице и равен 0,70. Величина в круглых скобках непосредственно под ним представляет собой его стандартную ошибку, равную 0,09. Под ней приведено значение *t*-критерия (критерия Стьюдента) для определения статистической значимости данного путевого коэффициента ($t = 7{,}65$). Для выборки объемом в 140 наблюдений при использовании двустороннего критерия Стьюдента критическое значение *t* для уровня значимости 0,05 приближенно равно +1,98, для уровня значимости 0,01 — ±2,62, для уровня значимости 0,001 — ±3,37. Поскольку 7,65 > 3,37, данный путевой коэффициент статистически значим на уровне ниже 0,001.

Таблица 7.1. **Выводимые LISREL результаты для путевых коэффициентов модели с учетом надежности измерений**

LISREL Estimates (Maximum Likelihood)

LAMBDA-Y

	CI	AA
CI	0,92	--
AA	--	0,85

LAMBDA-X

	PI
PI	0,91

[1] В русской версии ДИ.
[2] В русской версии АУ.

BETA

	CI	AA
CI	--	--
AA	0,70 (0,09) 7,65	--

GAMMA

	PI
PI	0,49 (0,09) 5,27
AA	--

Доля дисперсии ошибок индикаторов также представлена в матрицах табл. 7.2. Например, доля дисперсии ошибки для индикатора интереса ребенка к учебе (CI) выведена в **THETA-EPS** матрице и равна 0,15.

Косвенный эффект влияния латентной переменной родительской заинтересованности (CI) на академическую успеваемость ребенка (AA) показан в матрице табл. 7.3 и равен 0,34. Имея соответствующее значение *t*, равное 4,47, данный эффект является статистически значимым по двустороннему критерию Стьюдента на уровне ниже 0,001.

Таблица 7.2. **Выводимые LISREL результаты для оценок ошибок дисперсий третьей модели с учетом надежности измерений**

THETA-EPS

CI	AA
0,15	0,27

Squared Multiple Correlations for Y-Variables

CI	AA
0,85	0,73

THETA-DELTA

PI
0,17

Squared Multiple Correlations for X-Variables

PI
0,83

Таблица 7.3. **Выводимые LISREL результаты для косвенного эффекта в третьей модели с учетом надежности измерений**

```
Indirect Effects of KSI on ETA

                PI
              --------
CI            ========
AA              0,34
               (0,08)
                4,47
```

Процедура LISREL для анализа путей с учетом измерительной модели

Чтобы получить результаты для третьей модели, приведенные на рис. 7.2, необходимо выполнить команды, приведенные ниже и выделенные жирным шрифтом:

```
PA: Indirect model with measurement model
DAta NInputvar=9 NObserv=140
LAbels
cil ci2 ci3 aa1 aa2 aa3 pil pi2 pi3
KMatrix
1,00
0,64  1,00
0,79  0,54  1,00
0,42  0,40  0,41  1,00
0,29  0,37  0,38  0,49  1,00
0,41  0,35  0,45  0,54  0,59  1,00
0,25  0,39  0,33  0,19  0,39  0,41  1,00
0,23  0,21  0,42  0,19  0,58  0,34  0,59  1,00
0,27  0,24  0,26  0,25  0,35  0,38  0,69  0,61  1,00
MOdel NYvar=6 NXvar=3 NKvar=1 NEvar=2 c
GAmma=FRee BEta=SD PSi=DIagonal
Fixed GAmma (2,1)
FRee LX (2,1) LX(3,1)c
LY(2,1) LY(3,1) LY(5,2) LY(6,2)
STartval 1 LX (1,1) c
LY(1,1) LY(4,2)
LEta
CI AA
LKsi PI
PDiagram
OUtput SS EF
```

Прокомментируем лишь различия между данным и предыдущим набором команд.

Количество вводимых переменных равно девяти, а их имена (метки) — от CI1 до PI3 соответственно. Корреляционная матрица содержит коэффициенты корреляции этих переменных.

Имеется шесть эндогенных (**NY**) и три экзогенных (**NX**) явных (наблюдаемых) переменных[1].

Прямоугольные или квадратные матрицы латентных переменных (**LX** и **LY**) по умолчанию содержат фиксированные параметры. Нам необходимо сделать свободными путевые коэффициенты, или нагрузки, вторых и третьих индикаторов для каждой латентной переменной, что мы осуществляем с помощью строки, начинающейся с **FR**. Нам необходимо задать фиксированные единичные коэффициенты путей для первых индикаторов каждой латентной переменной, что мы выполняем с помощью команды, начинающейся с **ST**.

Первое число в круглых скобках соответствует строке, а второе — столбцу. Поскольку у нас есть только одна экзогенная латентная переменная (**LX**), то и столбец тоже только один, с тремя строками. У нас имеется две эндогенные латентные переменные, а значит, для них заданы два столбца с шестью строками. В первых трех строках первого столбца расположены индикаторы интереса ребенка к учебе, в то время как в последних трех строках второго столбца расположены индикаторы академической успеваемости.

Четыре строки, начинавшиеся с **MA** в предыдущем наборе команд, здесь опущены, поскольку данные величины будут вычисляться для каждого из индикаторов латентных переменных.

Чтобы вывести стандартизированные коэффициенты, в строку, начинающуюся с **OU**, были добавлены символы **SS**.

Как и в предыдущем случае, чтобы провести анализ для второй модели, необходимо исключить строку, начинающуюся с инструкций **FIxed GAmma**.

Выводимая LISREL информация по результатам анализа модели путей с учетом измерительной модели

На диаграмме путей приводятся нестандартизированные путевые коэффициенты для рассматриваемой модели. Поскольку они могут изменяться в зависимости от масштаба используемых измерений, в случае, когда нас интересуют относительные величины коэффициентов, полезнее видеть стандартизированные показатели. Это можно изменить с помощью последовательного выбора пункта **View** строки меню в верхней части окна программы, пункта

[1] Правильнее было бы сказать: имеется 6 наблюдаемых переменных, относящихся к латентным эндогенным переменным, и 3 наблюдаемые переменные, относящиеся к латентным экзогенным переменным.

Estimations — в появившемся ниспадающем меню и пункта **Standardized Solution** — в появляющемся дочернем ниспадающем меню второго порядка. Альтернативным вариантом является вывод стандартизированных показателей в окне вывода программы под заголовком **Standardized Solution**, как это показано в табл. 7.4. Например, стандартизированный коэффициент пути между родительской заинтересованностью (PI[1]) и интересом ребенка (CI[2]) отображается в GAMMA матрице и равен 0,44.

Статистическая значимость путевых коэффициентов определяется по значениям соответствующих *t*-величин. Они могут быть отображены на диаграмме путей при выборе пункта **t-Values** в уже упоминавшемся дочернем меню второго порядка. Альтернативным способом вывода является третья строка соответствующих матриц под общим заголовком **LISREL estimates** в окне вывода результатов LISREL. Например, *t*-величина для путевого коэффициента между родительской заинтересованностью (PI) и интересом ребенка (CI) приведена в GAMMA матрице и равна 4,58 (табл. 7.5). Это статистически значимо по двустороннему критерию при уровне значимости меньше 0,001.

Таблица 7.4. **Выводимые LISREL результаты для стандартизированных путевых коэффициентов третьей модели с учетом измерительной модели**

Standardized Solution

LAMBDA-Y

	CI	AA
ci1	0,90	--
ci2	0,68	--
ci3	0,87	--
aa1	--	0,69
aa2	--	0,71
aa3	--	0,80

LAMBDA-X

	PI
pi1	0,82
pi2	0,73
pi3	0,83

BETA

	CI	AA
CI	--	--
AA	0,63	--

[1] В русской версии РИ.
[2] В русской версии ДИ.

GAMMA

	PI
CI	0,44
AA	--

Таблица 7.5. **Выводимые LISREL результаты для нестандартизированных путевых коэффициентов третьей модели с учетом измерительной модели**

LAMBDA-Y

	CI	AA
ci1	1,00	--
ci2	0,76	--
	(0,08)	
	9,01	
ci3	0,97	--
	(0,08)	
	12,14	
aa1	--	1,00
aa2	--	1,03
		(0,15)
		6,90
aa3	--	1,16
		(0,16)
		7,26

LAMBDA-X

	PI
pi1	1,00
pi2	0,89
	(0,10)
	8,58
pi3	1,01
	(0,11)
	9,32

BETA

	CI	AA
CI	--	--
AA	0,49	--
	(0,08)	
	5,82	

GAMMA

	PI
CI	0,48
	(0,10)
	4,58
AA	--

Таблица 7.6. **Выводимые LISREL результаты для стандартизированного косвенного эффекта третьей модели с учетом измерительной модели**

Standardized Indirect Effects of KSI on ETA

	PI
CI	--
AA	0,28

Таблица 7.7. **Выводимые LISREL результаты для нестандартизированного косвенного эффекта третьей модели с учетом измерительной модели**

Indirect Effects of KSI on ETA

	PI
CI	--
AA	0,23
	(0,06)
	3,73

Стандартизированный косвенный эффект латентной переменной родительской заинтересованности (PI) на латентную переменную академической успеваемости ребенка (AA) показан в матрице табл. 7.6 и равен 0,28. Величина *t* для данного путевого коэффициента приведена в третьей строке матрицы табл. 7.7 и равна 3,73. Данный эффект является статистически значимым по двустороннему критерию Стьюдента на уровне ниже 0,001.

Рекомендуемая литература

Byrne, B.M. (1998) *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hair, J.F., Jr., Anderson, R.E., Tatham, R.L. and Black, W.C. (1998) *Multivariate Data Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction,* 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Дополнительная литература, рекомендуемая научным редактором

Митина О.В. Структурное моделирование: состояние и перспективы // Вестн. Пермского гос. пед. ун-та. Сер. 1. Психология. — 2005. — № 2. — С. 3—15.

ЧАСТЬ IV
ВЕРОЯТНОСТЬ РЕАЛИЗАЦИИ БИНАРНОЙ ПЕРЕМЕННОЙ

Глава 8
БИНАРНАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Предисловие научного редактора

Если и зависимая и независимая переменные не количественные, а номинативные, то для установления связей между ними используют логистическую регрессию. В нашей стране этот вид анализа практически не используется. Однако стоит напомнить, что существует метод анализа связей между номинативными переменными, разработанный российским математиком С.Чесноковым, и называется он «детерминационный анализ» (О.В.Митина, 2004).

Логистическая, или логит-множественная, регрессия применяется в том случае, если надо определить, какие переменные сильнее всего связаны с вероятностью реализации конкретной категории другой переменной. Эта категория может быть одним из значений дихотомической, или бинарной, переменной, содержащей всего две категории (два варианта ответа), например сдача или провал экзамена, наличие или отсутствие определенного заболевания в результате проведения диагностики, признание виновным или невиновным и т.п. Или же категория может быть одним из вариантов ответов полихотомической мультиномиальной переменной, предусматривающей три или более вариантов ответов. Например, в результате диагностики психического состояния пациента может быть установлено наличие: 1) депрессии; 2) тревоги; 3) того и другого; 4) ни того, ни другого. Логистическая множественная регрессия называется *бинарной логистической множественной регрессией*, когда зависимая переменная, или переменная-отклик, является дихотомической, и *мультиномиальной логистической множественной регрессией*, когда зависимая переменная является мультиноминальной. В данной главе будет рассмотрена лишь бинарная логистическая множественная регрессия, которую для краткости будем называть логистической регрессией. Мультиномиальная логистическая множественная

регрессия рассматривается в других источниках (например, D.W.Hosmer and S.Lemeshow, 1998).

Пытаясь понять логистическую регрессию, полезно сравнить ее с множественной регрессией. Оба метода похожи во многих отношениях и могут применяться сходным образом. Однако статистики, используемые в логистической регрессии, более сложные. Чтобы разобраться в них, полезно провести сравнение с соответствующими статистиками в множественной регрессии. Логистическая регрессия считается более адекватным методом анализа данных, чем множественная регрессия, в том случае, когда зависимая переменная, или отклик, является качественной, а не количественной. Причины этого будут объяснены позже. Независимые переменные могут вводиться в определенном порядке, как в случае иерархической множественной регрессии, или в соответствии со статистическими критериями, как в случае пошаговой множественной регрессии. Можно считать иерархическую логистическую регрессию средством определения того, значимо ли одна или более независимых переменных максимизируют вероятность того, что зависимая переменная принимает или не принимает определенное значение (т.е. реализуется определенная категория).

Проиллюстрируем применение логистической регрессии на данных из табл. 4.1 с той разницей, что непрерывная переменная академической успеваемости ребенка будет преобразована в бинарную переменную успешной сдачи или провала на экзаменах. Дети, у которых академическая успеваемость 2 или ниже, считаются провалившими экзамены и кодируются нулем. Те, кто набрал 3 и более баллов, считаются сдавшими экзамены и кодиру-

Таблица 8.1. **Академическая успеваемость ребенка, закодированная как бинарная переменная**

Наблюдения	Успеваемость ребенка	Способности ребенка	Интерес ребенка	Заинтересованность родителей	Заинтересованность учителей
1	0	1	1	2	1
2	0	3	3	1	2
3	0	2	3	3	4
4	1	3	2	2	4
5	1	4	4	3	2
6	1	2	3	2	3
7	1	3	5	3	4
8	1	4	2	3	2
9	1	3	4	2	3

ются единицей (табл. 8.1). Независимые переменные могут быть количественными или качественными, но здесь ограничимся демонстрацией лишь количественных переменных и будем сравнивать пошаговую множественную регрессию и логистическую регрессию.

Прогнозируемые значения

Результаты пошаговой множественной регрессии, выполняемой в данном случае, отличаются от результатов, приведенных в гл. 4, поскольку изменились значения академической успеваемости ребенка, что привело к изменению коэффициентов корреляции между академической успеваемостью и независимыми переменными. Предиктором, объясняющим наибольшую долю дисперсии академической успеваемости ребенка, являются его умственные способности, за которыми следует заинтересованность учителей, а затем родительская заинтересованность. Можно вычислить величину академической успеваемости ребенка на основании трех независимых переменных с помощью следующего уравнения регрессии:

[Значение прогнозируемой академической успеваемости] = = [Значение умственных способностей ребенка] + [Значение заинтересованности учителей] + [Значение родительской заинтересованности] + [Константа уравнения регрессии],

где значения независимых переменных умножаются на соответствующие нестандартизированные частные коэффициенты регрессии. Эти коэффициенты приближенно равны 0,28; 0,11 и 0,09 соответственно.

Константа уравнения регрессии равна –0,62. Используя эти величины, можно вычислить значение прогнозируемой академической успеваемости для каждого из девяти наблюдений (см. табл. 8.1). Например, прогнозируемое значение –0,05 академической успеваемости для первого наблюдения, где величина умственных способностей ребенка равна 1, заинтересованность учителей равна 1, а родительская заинтересованность равна 2, вычисляется следующим образом:

$$(0{,}28 \times 1) + (0{,}11 \times 1) + (0{,}09 \times 2) + (-0{,}62) = \\ = 0{,}28 + 0{,}11 + 0{,}18 - 0{,}62 = -0{,}05.$$

Реальное значение этой переменной равно нулю, так что прогнозируемое значение очень близко к реальности. Прогнозируемые значения для каждого из девяти наблюдений приведены в табл. 8.2.

Таблица 8.2. **Величины для вычисления квадрата множественной корреляции**

Наблюдения	Реальное значение	Прогнозируемое значение	(Прогноз – среднее)^2	(Реальность – среднее)^2
1	0	–0,05	$(-0,05 - 0,67)^2 = 0,52$	$(0 - 0,67)^2 = 0,45$
2	0	0,53	$(0,53 - 0,67)^2 = 0,02$	$(0 - 0,67)^2 = 0,45$
3	0	0,65	$(0,65 - 0,67)^2 = 0,00$	$(0 - 0,67)^2 = 0,45$
4	1	0,84	$(0,84 - 0,67)^2 = 0,03$	$(1 - 0,67)^2 = 0,11$
5	1	0,99	$(0,99 - 0,67)^2 = 0,10$	$(1 - 0,67)^2 = 0,11$
6	1	0,45	$(0,45 - 0,67)^2 = 0,05$	$(1 - 0,67)^2 = 0,11$
7	1	0,93	$(0,93 - 0,67)^2 = 0,07$	$(1 - 0,67)^2 = 0,11$
8	1	0,99	$(0,99 - 0,67)^2 = 0,10$	$(1 - 0,67)^2 = 0,11$
9	1	0,73	$(0,73 - 0,67)^2 = 0,00$	$(1 - 0,67)^2 = 0,11$
Сумма	6		0,89	2,00
Среднее	0,67			

Квадрат множественной корреляции

Исходя из прогнозируемых и фактических значений (баллов) можно вычислить, какая доля дисперсии академической успеваемости объясняется данными тремя независимыми переменными. Эта доля задается квадратом множественной корреляции, вычисляемым по следующей формуле:

$$[\text{Квадрат множественной корреляции}] = \frac{[\text{Сумма (предсказанное значение} - \text{среднее значение})]^2}{[\text{Сумма (фактическое значение} - \text{среднее значение})]^2}.$$

Иными словами, доля объясняемой дисперсии определяется отношением дисперсии прогнозируемых значений к дисперсии фактических значений. Эти дисперсии приведены в табл. 8.2. На самом деле дисперсии в 30 раз больше, поскольку объем выборки равен 270, а не 9, но это не играет роли, так как множитель 30 появляется как в числителе, так и в знаменателе, и потому сокра-

щается. Квадрат множественной корреляции приближенно равен 0,445 (0,89/2,00 = 0,445). Используя данный метод, можно вычислить квадрат множественной корреляции для любого множества независимых переменных и определить величину статистической значимости добавляемых предикторов, используя *F*-отношение (см. гл. 6).

Прогнозируемая вероятность

В случае логистической регрессии мы вычисляем прогнозируемую вероятность реализации категории, а не прогнозируемую величину признака. Вероятность появления категории представляет собой отношение числа реализаций этой категории к общему числу случаев и может быть выражена следующей формулой:

$$[\text{Вероятность категории}] = \frac{[\text{Частота категории}]}{[\text{Частота всех категорий}]}.$$

Вероятность изменяется в интервале от 0 до 1. Вероятность того, что ребенок успешно сдаст экзамен, без учета независимых переменных, есть отношение количества детей, успешно сдавших экзамен, к общему количества детей, и в данном случае составляет около 0,67 (6/9 = 0,667).

Введем понятие шанс события как отношение частоты возникновения события к частоте его отсутствия. В примере со сдачей экзамена шанс ребенка успешно сдать экзамен равен 6:3, что можно упростить и представить как 2:1. Тогда вероятность события может быть выражена через его шанс по следующей формуле:

$$[\text{Вероятность категории}] = \frac{[\text{Шанс события}]^{1}}{[1+\text{шанс события}]}.$$

Вернемся к примеру с экзаменом. Если подставить шанс успешно сдать экзамен в формулу, приведенную выше, то можно видеть, что вероятность успешной сдачи экзамена равна 0,67 [2,0/(1 + 2,0) = 0,667]. Основанием определения вероятности через

---

[1] Действительно, пусть $w_у$ — частота наступления события; $w_н$ — частота отсутствия события, тогда сумма ($w_у + w_н$) равна совокупной частоте всех возможных в данном опыте событий, а вероятность события $P = \frac{w_у}{w_у + w_н}$.

Эту формулу можно преобразовать, разделив числитель и знаменатель на $w_н$:

$P = \frac{\frac{w_у}{w_н}}{\frac{w_у}{w_н} + 1}$, но ведь $\frac{w_у}{w_н}$ и есть шанс события.

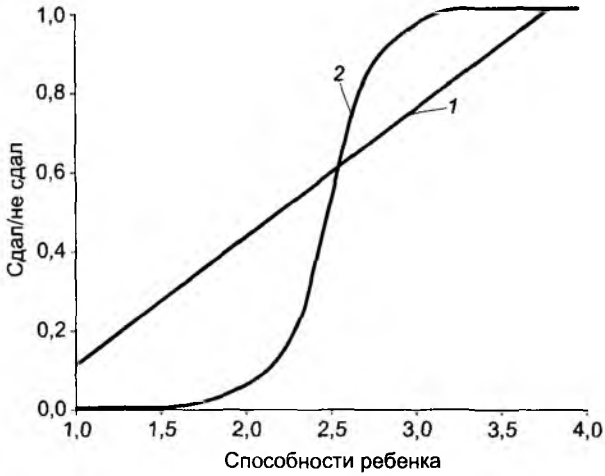


Рис. 8.1. Графики множественной (*1*) и логистической (*2*) регрессии

шанс является тот факт, что формула, используемая в случае логистической регрессии для вычисления вероятности появления события, выражается подобным образом[1].

Множественная линейная регрессия предполагает, что связь между зависимой переменной и независимыми признаками лучше всего может быть представлена прямой линией. На рис. 8.1 изображена прямая регрессии для зависимого признака (успешной сдачи экзамена) и единственного независимого признака (умственных способностей ребенка). Ограничимся одной независимой переменной, чтобы упростить объяснение. Нестандартизированный коэффициент регрессии для умственных способностей ребенка приближенно равен 0,31. Это означает, что для любого изменения умственных способностей ребенка на единицу соответствующее изменение признака успешной или неудачной сдачи экзамена составляет 0,31. Данное соотношение остается одним и тем же как для низкого, так и для высокого уровня умственных способностей ребенка. Логистическая регрессия предполагает, что кривая зависимости между объясняемой переменной и объясняющими переменными имеет S-образную, или сигмоидальную, форму, как показано на рис. 8.1. Это означает, что взаимосвязь зависимого и независимого признаков сильнее всего выражена в области средней точки 0,5 кривой между провалом на экзамене и его успешной сдачей и слабее всего — на концах кривой. Заданное изменение независимой переменной будет иметь гораздо больший эффект в середине кривой, чем на

[1] $y = \frac{x}{x+1}$.

любом из ее концов. Когда зависимый признак является бинарным, точки, отвечающие значениям зависимой и независимой переменой, будут расположены ближе к такой S-образной кривой, чем к прямой линии.

Логарифмический шанс

Эта S-образная зависимость выражается в терминах натурального логарифма шанса или, как его называют различные авторы, логарифмического шанса, или логит-преобразования. Такую взаимосвязь можно проиллюстрировать, вычислив натуральный логарифм шанса для 15 значений вероятности в интервале от 0,001 до 0,999 (табл. 8.3) и изобразив эти вероятности как функцию натурального логарифма соответствующих шансов (рис. 8.2).

Шанс вычисляется как отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Вероятность того, что событие не произойдет, равна разности между единицей и вероятностью того, что событие произойдет. Таким образом, если вероятность осуществления события равна

Таблица 8.3. **Значения вероятностей, шансов и логит-преобразований**

Вероятность	1 – вероятность	Шанс	Логит-преобразование
0,0001	0,9999	0,0001	–9,2102
0,0010	0,9990	0,0010	–6,9068
0,0100	0,9900	0,0101	–4,5951
0,1000	0,9000	0,1111	–2,1972
0,2000	0,8000	0,2500	–1,3863
0,3000	0,7000	0,4286	–0,8473
0,4000	0,6000	0,6667	–0,4055
0,5000	0,5000	1,0000	0,0000
0,6000	0,4000	1,5000	0,4055
0,7000	0,3000	2,3333	0,8473
0,8000	0,2000	4,0000	1,3863
0,9000	0,1000	9,0000	2,1972
0,9900	0,0100	99,0000	4,5951
0,9990	0,0010	999,0000	6,9068
0,9999	0,0001	9999,0000	9,2102

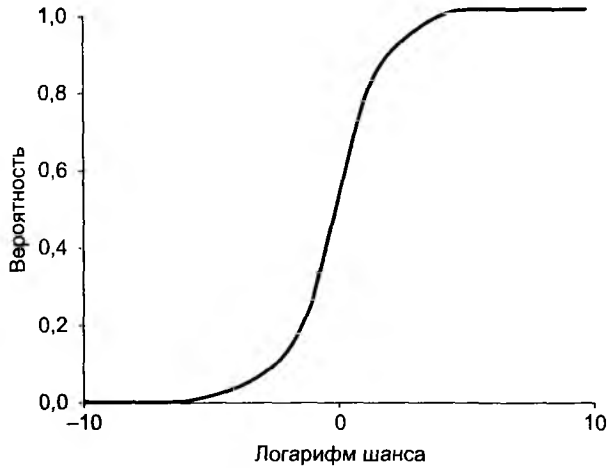


Рис. 8.2. График вероятности по отношению к логарифму шанса

0,7, то вероятность его неосуществления равна 0,3 (1 – 0,7 = 0,3). Шанс появления события приближенно равен 2,33 (0,7/0,3 = 2,33). Чтобы получить логит-преобразование признака, вычисляем натуральный, или неперов, логарифм его шанса[1].

Например, натуральный логарифм числа 2,33 приближенно равен 0,85. Чтобы осуществить обратное преобразование логит-функции в шанс, необходимо возвести число e в степень, равную численному значению логит-функции[2].

Так, для значения логистической функции, равного 0,85, соответствующее значение шанса приближенно равно 2,33 ($2{,}718^{0{,}85}$ = = 2,33). В нашем примере шанс ребенка успешно сдать экзамен равен примерно 2,00 (0,667/0,333 = 2,00), а его логарифмический шанс приближенно равен 0,69 (натуральный логарифм от 2,00 равен 0,693). Логарифмический шанс, равный 0,69, дает значение соответствующего шанса, равное 2,00 ($2{,}718^{0{,}693} = 2{,}00$).

В то время как при множественной линейной регрессии регрессионные коэффициенты и свободный член используются для вычисления прогнозируемого значения зависимого признака, при логистической регрессии их применяют для вычисления его логарифмического шанса.

Затем логарифмический шанс преобразуется в обычный шанс наступления события, который используется в формуле для вычисления прогнозируемой вероятности события.

Вероятность категории определяют по следующей формуле:

---

[1] $\mathrm{Ln} = \log_e$ (логарифм по основанию e — известная из курса математического анализа константа, приближенно равная 2,718).

[2] Это правило базируется на основном логарифмическом тождестве: $b = a^{\log_a b}$.

$$[\text{Вероятность категории}] = \frac{2{,}718^{\text{натуральный логарифм шанса}}}{1 + 2{,}718^{\text{натуральный логарифм шанса}}} =$$

$$= \frac{e^{\text{натуральный логарифм шанса}}}{1 + e^{\text{натуральный логарифм шанса}}} = \frac{[\text{Шанс события}]}{[1 + \text{шанс события}]}.$$

Проиллюстрируем вычисление вероятности успешной сдачи ребенком экзамена для набора баллов, которые приведены в первой строке выборочной матрицы. Нестандартизированные коэффициенты логистической регрессии для умственных способностей ребенка, родительской и учительской заинтересованности приближенно равны 2,24; 0,60 и 0,77 соответственно, а свободный член для данной выборки — (–8,86). Логарифмический шанс успешной сдачи экзамена детьми, у которых величина умственных способностей равна 1, родительской заинтересованности — 2, учительской заинтересованности — 1, приближенно равен (–4,65):

$$(2{,}24 \times 1) + (0{,}60 \times 2) + (0{,}77 \times 1) + (-8{,}86) =$$
$$= 2{,}24 + 1{,}20 + 0{,}77 - 8{,}86 = -4{,}65.$$

Отрицательное значение логистической функции означает, что у детей с такими значениями независимых переменных нет шанса успешно сдать экзамен. Положительное значение логит-преобразования указывает на то, что у детей есть благоприятный шанс сдать экзамен, а нулевое значение логистического преобразования говорит о том, что шансы успешно сдать или провалить экзамен равны. Значение логистической функции, равное –4,65, дает значение шанса 0,01 ($2{,}718^{-4{,}65} = 0{,}01$), которое, в свою очередь, даст значение прогнозируемой вероятности, приближенно равное 0,01 ($0{,}01/(1 + 0{,}01) = 0{,}01$). Прогнозируемые вероятности для девяти испытуемых представлены в третьем столбце табл. 8.4.

Коэффициент регрессии для умственных способностей ребенка, равный 2,24, означает, что для каждого единичного приращения умственных способностей ребенка соответствующее приращение логарифмического шанса равно 2,24. Это означает увеличение шанса успешно сдать экзамен в 9,39 раз ($2{,}718^{2{,}24} = 9{,}39$). Можно показать это, увеличив значение умственных способностей ребенка с 1 до 2, что дает значение натурального логарифма шанса, равное –2,41:

$$(2{,}24 \times 2) + (0{,}60 \times 2) + (0{,}77 \times 1) + (-8{,}86) =$$
$$= 4{,}48 + 1{,}20 + 0{,}77 - 8{,}86 = -2{,}41.$$

Логит-преобразование величины –2,41 дает значение шанса, равное 0,09 ($2{,}718^{-2{,}41} = 0{,}09$), что примерно в 9,39 раз больше 0,01 ($0{,}01 \times 9{,}39 = 0{,}09$). Отношение шансов есть число, на которое мы умножаем шанс наступления события (реализации категории) при

Таблица 8.4. **Прогнозируемые вероятности и вычисление величины логарифмического правдоподобия для модели с тремя независимыми переменными**

Наблюдения	Исход (И)	Прогнозируемая вероятность (P)	Log P	И × log P	1 – И	1 – P	Log(1 – P)	(1 – И) × log(1 – P)	Логарифм правдоподобия
1	0	0,01	–4,61	0,00	1	0,99	–0,01	–0,01	–0,01
2	0	0,50	–0,69	0,00	1	0,50	–0,69	–0,69	–0,69
3	0	0,62	–0,48	0,00	1	0,38	–0,97	–0,97	–0,97
4	1	0,89	–0,12	–0,12	0	0,11	–2,21	0,00	–0,12
5	1	0,97	–0,03	–0,03	0	0,03	–3,51	0,00	–0,03
6	1	0,30	–1,20	–1,20	0	0,70	–0,36	0,00	–1,20
7	1	0,94	–0,06	–0,06	0	0,06	–2,81	0,00	–0,06
8	1	0,97	–0,03	–0,03	0	0,03	–3,51	0,00	–0,03
9	1	0,80	–0,22	–0,22	0	0,20	–1,61	0,00	–0,22
Сумма									–3,33

единичном приращении одной из независимых переменных, в то время как остальные независимые переменные остаются постоянными.

Логарифмическое правдоподобие

В качестве статистики, которая используется для того, чтобы определить, насколько хорошо модель согласуется с данными, применяется логарифмическое правдоподобие, обычно умножаемое на –2, так что оно приближенно принимает форму распределения хи-квадрат. Точное соответствие данным достигается при нулевом значении статистики, в то время как с увеличением значения уменьшается согласие с данными. Поскольку значение логарифмического правдоподобия, умноженное на –2, зависит от объема выборки: оно тем больше, чем больше этот объем[1],

[1] Т.е. при достаточно большом объеме выборки в этом случае значение будет достаточно большим, чтобы считать, что теоретическая модель не соответствует экспериментальным данным, даже если эта модель и достаточно точна.

влияние объема выборки следует исключить. Это можно сделать, вычитая –2 значение логарифмического правдоподобия для модели, содержащей независимые переменные, из –2 значения логарифмического правдоподобия для модели, содержащей только корректирующую константу уравнения регрессии. Прогнозируемая вероятность осуществления категории для всех наблюдений в модели, содержащей только корректирующую константу, есть полная вероятность осуществления данной категории в нашей выборке. В нашем примере интересующая категория — факт сдачи экзамена. Вероятность ее реализации равна 0,67[1] (табл. 8.5). Значение логарифмического правдоподобия, умноженное на –2, для нашего примера, где три независимые переменные — умственные способности ребенка, родительская заинтересованность и заинтересованность учителей — включены в уравнение прогноза того, сдаст ли ребенок экзамены, приближенно равно 199,80 ($-2 \times -3{,}33 \times 30 = 199{,}80$)[2]. Значение логарифмического правдоподобия, умноженное на –2, для модели, содержащей только корректирующую регрессионную константу, приближенно равно 343,80 ($-2 \times -5{,}73 \times 30 = 343{,}80$). Разность между соответствующими значения логарифмического правдоподобия, умноженными на –2, дает значение критерия хи-квадрат, приближенно равное 144,00 (343,80 – 199,80 = 144,00). Число степеней свободы модели равно количеству независимых переменных в данной модели. Число степеней свободы в модели, содержащей только регрессионную константу, равно нулю, а в модели, содержащей три предиктора, соответственно, трем. Разность числа степеней свободы для двух данных моделей равна 3 (3 – 0 = 3). Чтобы быть значимой на уровне 0,05, критерий хи-квадрат должен принимать значение 7,82 или больше. Поскольку значение 144,00 критерия хи-квадрат больше, чем 7,82, оно является статистически значимым. Следовательно, можно сделать вывод о том, что модель, содержащая три независимые переменные, достаточно хорошо согласована с данными.

Значение логарифмического правдоподобия равно сумме вероятностей, связанных с прогнозируемыми и фактическими результатами для каждого наблюдения, и может быть вычислено по следующей формуле:

[Логарифмическое правдоподобие] =

---

[1] Это вероятность того, что ребенок сдаст экзамен, так как по результатам данных нашей выборки шесть детей из девяти экзамен сдали, то это соотношение как раз и равняется 0,67 = 6/9. Заметим, что эта вероятность является априорной и никаким образом не зависит от реального исхода — сдачи или провала экзамена, поэтому она одинакова как для детей, сдавших экзамен, так и детей, не сдавших его.

[2] На 30 мы умножили из-за того, что изначально предположили, что в нашей гипотетической выборке число испытуемых в 30 раз больше (270), чем представлено в табл. 8.5.

= [Фактический исход зависимой переменной × логарифм прогнозируемой вероятности] + [(1 – фактический исход зависимой переменной) × логарифм (1 – прогнозируемая вероятность)].

Этапы данного вычисления для девяти наблюдений приведены в табл. 8.4 для модели с тремя независимыми переменными и в табл. 8.5 для модели, содержащей только константу. Поскольку девять представленных наблюдений повторяются каждое по 30 раз, суммарную величину логарифмического правдоподобия необходимо умножить на тридцать. Чтобы получить значение логарифмического правдоподобия, умноженное на –2, необходимо умножить итоговое значение на –2. Для модели с тремя независимыми переменными приближенное значение равно 199,80 (–3,33 × 30 × –2 = 199,80). Соответствующее значение для модели, содержащей только константу, будет приближенно равно 343,80 (–5,73 × 30 × –2 = 343,80).

Используя критерий хи-квадрат, модель с заданным числом независимых переменных можно сравнить с моделью, где один или даже несколько из рассматриваемых предикторов опущены.

Таблица 8.5. **Прогнозируемые вероятности и вычисление величины логарифмического правдоподобия для модели, содержащей только константу**

Наблюдения	Исход (И)	Прогнозируемая вероятность (P)	Log P	И × log P	1 – И	1 – P	Log(1 – P)	(1 – И) × log(1 – P)	Логарифм правдоподобия
1	0	0,67	–0,40	0,00	1	0,33	–1,11	–1,11	–1,11
2	0	0,67	–0,40	0,00	1	0,33	–1,11	–1,11	–1,11
3	0	0,67	–0,40	0,00	1	0,33	–1,11	–1,11	–1,11
4	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
5	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
6	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
7	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
8	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
9	1	0,67	–0,40	–0,40	0	0,33	–1,11	0,00	–0,40
Сумма									–5,73

Таблица 8.6. **Сравнение степени согласия моделей логистической регрессии с данными**

Сравниваемые модели	Разность логарифмов правдоподобия	χ^2	Разность степеней свободы	Значимость p
3 и 4 предиктора	199,80 – 198,20 =	1,60	4 – 3 = 1	Незначима
2 и 3 предиктора	207,45 – 199,80 =	7,65	3 – 2 = 1	0,05

Например, модель с тремя независимыми переменными (умственные способности ребенка, родительская и учительская заинтересованность) можно сравнить с моделью, содержащей все четыре предиктора, или с моделью, содержащей две независимые переменные (умственные способности ребенка и заинтересованность родителей в его успехах). Раньше было показано, что значение логарифмического правдоподобия, умноженное на –2, для модели с тремя независимыми переменными приближенно равно 199,80. Сравниваемые величины для модели с четырьмя и модели с двумя предикторами соответственно равны 198,20 и 207,45. Разность данных величин между моделями с тремя и четырьмя независимыми переменными дает значение критерия хи-квадрат, приближенно равное 1,60 (199,80 – 198,20 = 1,60), обладающее одной степенью свободы (4 – 3 = 1). Чтобы быть статистически значимым на уровне 0,05, критерий хи-квадрат с одной степенью свободы должен принимать значение 3,84 или выше. Поскольку 1,60 < 3,84, модель с четырьмя независимыми переменными, содержащая дополнительный предиктор интереса ребенка к учебе, не дает статистически значимого улучшения согласия с данными. Разность соответствующих значений между моделями с тремя и двумя независимыми переменными дает приближенное значение критерия хи-квадрат, равное 7,65 (207,45 – 199,80 = 7,65), с одной степенью свободы (3 – 2 = 1). Поскольку критерий хи-квадрат равен 7,65, что больше, чем 3,84, модель с тремя независимыми переменными, содержащая дополнительную переменную учительской заинтересованности, обеспечивает статистически значимое увеличение степени согласия с данными по сравнению с двухпредикторной моделью, не содержащей заинтересованности учителей. Статистики для проведенных сравнений даны в табл. 8.6.

Пошаговый выбор

Для ввода или удаления предикторов из логистического регрессионного анализа могут применяться различные статистические критерии. В использованном здесь методе прямого выбора первым вводился предиктор с наиболее значимой величиной эффектив-

ного показателя статистики Рао (1973). Этот статистический показатель является мерой связи в случае логистической регрессии. Чем больше его значение, тем сильнее связь. Если ни один из данных статистических показателей не является статистически значимым, процедура останавливается и показывает, что ни один из предикторов не обеспечивает хорошего согласия с данными. В нашем примере наиболее значимое значение статистики Рао достигалось для умственных способностей ребенка, которые и были введены первыми в уравнение логистической регрессии. Соответствующий показатель был приближенно равен 97,28, что является статистически значимым на уровне менее 0,001.

Если разность умноженных на –2 значений логарифмического правдоподобия между моделью с данным предиктором и моделью без него является статистически значимой на уровне менее 0,10, то данный предиктор сохраняется в уравнении логистической регрессии[1]. Если же указанная разность не является статистически значимой на данном уровне, то процедура останавливается и показывает, что ни один из предикторов не обеспечивает хорошего согласия с данными. Разность умноженных на –2 значений логарифмического правдоподобия между этой моделью и моделью, содержащей только регрессионную постоянную, приближенно равна 110,47 (343,80 – 233,33 = 110,47), обладает одной степенью свободы и является значимой на указанном уровне. Таким образом, умственные способности ребенка являются первой независимой переменной, включаемой в модель. Соответствующие статистики для данного и последующих шагов алгоритма приведены в табл. 8.7.

Следующим в уравнение логистической регрессии включается предиктор со второй по статистической значимости величиной статистики Рао. Им оказывается заинтересованность родителей в школьных успехах ребенка, для которой соответствующий показатель приближенно равен 26,27. Разность умноженных на –2 значений логарифмического правдоподобия между моделями с данным предиктором и без него приближенно равна 25,88 (233,33 – 207,45 = 25,88), обладает одной степенью свободы и является статистически значимой на уровне 0,10. Следовательно, родительская заинтересованность является второй переменной, включаемой в модель. Критерий разности умноженных на –2 значений логарифмического правдоподобия применяется и к первой независимой переменной умственных способностей ребенка[2]. Соответствующая разность между моделью, содержащей дан-

[1] Значимость проверяется по критерию хи-квадрат с одной степенью свободы.

[2] Т.е. сравнивается модель, в которой участвуют оба предиктора, с моделью, в которой участвует только второй предиктор (интерес родителей), а первый (способности ребенка) опущен.

Таблица 8.7. **Основные показатели статистик для прямого выбора предикторов в случае логистической регрессии**

Этапы	Предикторы	Критерий ввода		Критерий исключения			
		Величина статистики	p	Разность логарифмов правдоподобия	χ^2	Разность степеней свободы	p
1	Способности ребенка	97,28	0,001	343,80 - 233,33 =	110,47	1 - 0 = 1	0,001
	Интерес ребенка	45,00	0,001				
	Заинтересованность родителей	33,75	0,001				
	Заинтересованность учителей	25,12	0,001				
2	Способности ребенка			309,50 - 207,45 =	102,05	2 - 1 = 1	0,001
	Заинтересованность родителей	26,27	0,001	233,33 - 207,45 =	25,88	2 - 1 = 1	0,001
	Заинтересованность учителей	23,86	0,001				
	Интерес ребенка	8,84	0,01				
3	Способности ребенка			295,27 - 199,80 =	95,47	3 - 2 = 1	0,001
	Заинтересованность родителей			204,88 - 199,80 =	5,08	3 - 2 = 1	0,05
	Заинтересованность учителей	6,33	0,05	207,45 - 199,80 =	7,65	3 - 2 = 1	0,01
	Интерес ребенка	2,41	Незначима				
4	Интерес ребенка	3,39	Незначима				

ный предиктор, и моделью без него приближенно равна 102,05 (309,50 – 207,45 = 102,05), обладает одной степенью свободы и является статистически значимой на уровне 0,10. Таким образом, умственные способности ребенка остаются в рассматриваемой модели.

Следующей по порядку независимой переменной со статистически значимой величиной статистики Рао является заинтересованность учителей, и для нее этот показатель составляет примерно 6,33. Разность умноженных на –2 значений логарифмического правдоподобия между содержащей данный предиктор моделью и моделью, не содержащей его, приближенно равна 7,65 (207,45 – 199,80 = 7,65), обладает одной степенью свободы и является статистически значимой на уровне 0,10. Следовательно, заинтересованность учителей является третьей переменной, вводимой в модель. Умственные способности ребенка и родительская заинтересованность в его школьных успехах остаются в модели, поскольку умноженные на –2 значения разностей величин логарифмического правдоподобия, соответственно равные 95,47 (295,27 – 199,80 = 95,47) и 5,08 (204,88 – 199,80 = 5,08)[1], являются значимыми на уровне 0,10.

Поскольку величина статистики Рао для четвертой переменной интереса ребенка к учебе не является статистически значимой, анализ останавливается. Таким образом, три предиктора (умственные способности ребенка, родительская и учительская заинтересованность) обеспечивают статистически значимую согласованность с данными.

Распределение прогнозируемых вероятностей

Один из способов оценки улучшения согласованности модели с данными состоит в рассмотрении разности прогнозируемых вероятностей осуществления и неосуществления категории. В случае точной модели прогнозируемая вероятность осуществления категории (например, успешной сдачи экзаменов для тех, кто действительно сдал их) равна единице, а прогнозируемая вероятность ее неосуществления (например, провала на экзаменах для тех, кто на самом деле сдал) равна нулю. Разность прогнозируемых вероятностей осуществления и неосуществления категории достигает максимума, равного единице. В случае модели, содержащей только корректирующую регрессионную постоянную, прогнозируемая вероятность осуществления категории (сдачи экзаменов) одна и та же, как для

[1] Это два показателя логарифмического правдоподобия с опущенными в качестве предикторов способностями ребенка и родительской заинтересованностью соответственно.

тех, кто на самом деле провалил экзамен, так и для тех, кто их сдал, и в нашем примере равна 0,67. Разность между прогнозируемой вероятностью осуществления и неосуществления категории равна минимально возможному нулевому значению. Модели, обеспечивающие лучшее согласие с данными, будут увеличивать разность между прогнозируемой вероятностью осуществления категории, с одной стороны, для тех, кто ее фактически осуществил, а с другой — для тех, кто осуществить ее реально не смог.

Это можно увидеть в нашем примере. Прогнозируемые вероятности трех рассмотренных моделей вместе с моделью, содержащей только регрессионную постоянную, и точной моделью представлены в табл. 8.8. Средние значения прогнозируемых вероятностей для наблюдений, соответствующим детям, успешно сдавшим экзамены, и тем, кто их провалил, приведены вместе с разностями соответствующих средних величин. Эта разность увеличивается от 0,37 для модели с одним предиктором (умственными способностями ребенка) до 0,43 в случае трех предикторов (умственных способностей ребенка, родительской и учительской заинтересованности). Иными словами, модель с тремя независимыми переменными лучше всего различает тех, кто успешно сдал экзамены, и тех, кто их провалил.

Таблица 8.8. **Прогнозируемые вероятности для сдавших экзамены и провалившихся их в случае пяти моделей логистической регрессии**

Наблюдения	Исход	Модели				
		Константа	1	2	3	Точная модель
1	0	0,67	0,08	0,04	0,01	0,00
2	0	0,67	0,81	0,53	0,50	0,00
3	0	0,67	0,38	0,61	0,62	0,00
Среднее		0,67	0,42	0,39	0,38	0,00
4	1	0,67	0,81	0,80	0,90	1,00
5	1	0,67	0,97	0,99	0,97	1,00
6	1	0,67	0,38	0,30	0,30	1,00
7	1	0,67	0,81	0,94	0,94	1,00
8	1	0,67	0,97	0,99	0,97	1,00
9	1	0,67	0,81	0,80	0,80	1,00
Среднее		0,67	0,79	0,80	0,81	1,00
Разность		0,00	0,37	0,41	0,43	1,00

Отчет о результатах

Выбор того или иного способа описания результатов анализа, освещенного в данной главе, зависит от конкретных задач и целей такого анализа. Очень краткий отчет может быть представлен следующим образом: «В ходе анализа данных была проведена прямая пошаговая бинарная логистическая регрессия. Независимые переменные вводились в анализ на основании наибольшей статистической значимости показателя статистики Рао с уровнем значимости 0,05 и ниже и исключались из анализа, если значение вероятности критерия разности умноженных на –2 значений логарифмического правдоподобия превышало 0,10. Первой была включена переменная умственных способностей ребенка (критерий хи-квадрат с одной степенью свободы равен 110,47; $p < 0{,}001$), второй — переменная родительской заинтересованности (критерий хи-квадрат с одной степенью свободы равен 25,88; $p < 0{,}001$) и третьей — переменная заинтересованности учителей в школьных успехах ребенка (критерий хи-квадрат с одной степенью свободы равен 7,65; $p < 0{,}05$). Ввод переменной «Интерес ребенка к учебе» не дал статистически значимого улучшения согласия модели с данными. Вероятность успешной сдачи экзаменов была связана с более высокими умственными способностями ребенка, большей родительской и учительской заинтересованностью в его школьных успехах».

Процедура логистической регрессии в SPSS для Windows

Чтобы осуществить логистическую регрессию, описанную в данной главе, следуйте указаниям, приведенным ниже.

Если данные из табл. 4.1 были сохранены в отдельном файле, откройте этот файл в **Редакторе данных (Data Editor)**, выбрав пункты **Файл (File)**, **Открыть (Open)**, **Данные (Data)** в соответствующих меню и подменю и имя файла в диалоговом окне **Открытие файла (Open File)** и нажав на кнопку **Открыть (Open)**. В противном случае введите данные так, как показано на рис. 4.1.

Измените значения переменной «Успеваемость ребенка» таким образом, что если первоначально баллы были меньше или равны двум, то новое значение равно нулю, а если первоначально баллы были больше или равны трем, то новое значение равно единице, как показано в табл. 8.1.

Установите весовые коэффициенты с помощью процедуры **Установка весовых коэффициентов (Weight cases)**, описанной в гл. 4.

Выберите пункт **Анализ (Analyze)** строки меню в верхней части окна программы, в ниспадающем меню — пункт **Регрессия**

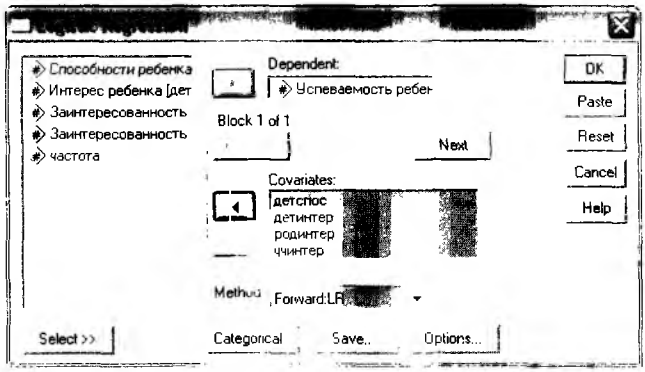


Рис. 8.3. Диалоговое окно **Логистическая регрессия (Logistic Regression)**

(Regression), а в появившемся подменю — пункт **Бинарная логистическая регрессия (Binary Logistic)**, в результате чего откроется диалоговое окно **Логистическая регрессия (Logistic Regression)**, изображенное на рис. 8.3.

Выберите переменную **Успеваемость ребенка** и с помощью нажатия первой сверху кнопки ▸ переместите эту переменную в поле **Зависимая переменная (Dependent)**.

Выберите переменные, начиная с умственных способностей ребенка (**«детспос»**) и закончив заинтересованностью учителей (**«учинтер»**), и с помощью второй сверху кнопки ▸ переместите их в список **Ковариаты (Covariates)**.

В раскрывающемся списке **Метод (Method)** выберите пункт **Включение (Enter)**, а в появившемся ниспадающем меню — пункт **Прямой (Forward: LR)** для прямой пошаговой логистической регрессии, в которой критерий удвоенного отношения логарифмических правдоподобий[1] используется для решения вопроса об исключении введенных предикторов из анализа.

Если вы хотите вывести значения прогнозируемых вероятностей для данной модели (как это показано в табл. 8.4 и 8.8), нажмите на кнопку **Сохранить (Save)**, открывающую дочернее диалоговое окно **Логистическая регрессия: Сохранение новых переменных (Logistic Regression: Save New Variables)**, изображенное на рис. 8.4. Установите флажок **Вероятности (Probabilities)** в группе **Прогнозируемые значения (Predicted Values)**, а затем нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно. Когда анализ будет осуществлен, эти прогнозируемые вероятности бу-

[1] Это утверждение не противоречит рассмотренному выше критерию разности логарифмических правдоподобий, так как, используя свойства логарифма, разность можно выразить через отношение: $\log a - \log b = \log (a/b)$.

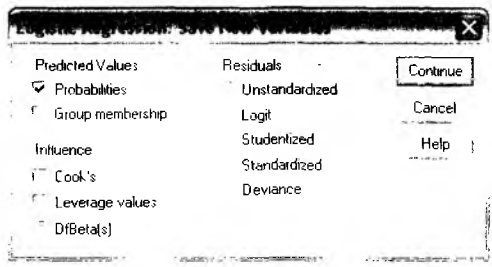


Рис. 8.4. Дочернее диалоговое окно **Логистическая регрессия: Сохранение новых переменных (Logistic Regression: Save New Variables)**

дут помещены в столбец рядом с частотами (переменная «частота») в **Редакторе данных (Data Editor)**.

Нажмите **OK**, чтобы выполнить логистический регрессионный анализ.

Выводимая SPSS информация по результатам работы

Результаты для всех таблиц будут приведены отдельно от первых четырех таблиц. Эти таблицы будут представлены в том же порядке, в каком они следуют в окне вывода результатов.

В табл. 8.9 приведены показатели статистики Рао для четырех предикторов и их статистические значимости, необходимые для решения вопроса о включении переменных в анализ на первом этапе, который обозначен нулевым номером. Наибольшую величину 97,297 имеет показатель статистики для умственных способностей ребенка, а соответствующая вероятность (уровень значимости) равна 0,000 (т.е. меньше, чем 0,001).

В табл. 8.10 отражены значения критерия хи-квадрат для трех этапов, на каждом из которых включение нового предиктора обес-

Таблица 8.9. **Выводимые SPSS показатели статистики для трех независимых переменных на первом шаге**

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ДЕТСПОС	97,279	1	0,000
		ДЕТИНТЕР	45,000	1	0,000
		РОДИНТЕР	33,750	1	0,000
		УЧИНТЕР	25,116	1	0,000
	Overall Statistics		120,399	4	0,000

Таблица 8.10. **Выводимые SPSS значения критерия хи-квадрат для трех этапов логистического анализа**

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	110,386	1	0,000
	Block	110,386	1	0,000
	Model	110,386	1	0,000
Step 2	Step	25,884	1	0,000
	Block	136,270	2	0,000
	Model	136,270	2	0,000
Step 3	Step	5,847	1	0,016
	Block	142,117	3	0,000
	Model	142,117	3	0,000

печивает значимое улучшение согласия модели с данными. Например, на третьем этапе значение критерия хи-квадрат шага равно 5,847 для добавления заинтересованности учителей к предшествующей модели, содержащей умственные способности ребенка и родительскую заинтересованность в его школьных успехах. Этот показатель немного отличается от значения, которое мы приводили ранее при ручном счете в предыдущих подразделах (7,65), из-за ошибок округления. Данное значение представляет собой разность умноженных на –2 значений логарифмического правдоподобия двух рассматриваемых моделей. Значение критерия хи-квадрат блока и модели, равное 142,117, представляет собой разность умноженных на –2 значений логарифмического правдоподобия между данной моделью с тремя предикторами и моделью, содержащей только регрессионную постоянную. Вычисленное нами ранее значение данной величины равно 144,00 (343,80 – 199,80 = 144,00).

В табл. 8.11 отражены значения умноженной на –2 величины логарифмического правдоподобия для модели на каждом из трех шагов. Например, значение умноженной на –2 величины логарифмического правдоподобия на третьем шаге равно 201,600, а вычисленное нами ранее значение этой величины равно 199,80.

В табл. 8.12 представлены результаты классификации для трех этапов логистического анализа. В этой таблице прогнозируемые вероятности, приведенные в табл. 8.8, распределены по категориям, т.е. всем вероятностям, ниже порогового значения (cutvalue) 0,500, приписано значение нуль (провал на экзаменах), а вероятностям, превосходящим или равным 0,500, — единичное значе-

Таблица 8.11. **Выводимые SPSS значения умноженной на –2 величины логарифмического правдоподобия для каждого из трех этапов логистического регрессионного анализа**

Model Summary

Step	–2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	233,331	0,336	0,466
2	207,447	0,396	0,550
3	201,600	0,409	0,568

ние (успешная сдача экзаменов). Она отражает число наблюдений, попадающих в эти две прогнозируемые категории. Например, на третьем этапе 90 человек провалили экзамены. Из них для 30 человек этот провал предсказывался (в таблице это наблюдение 1, но исходя из нашего учебного допущения за одним наблюдением стоит 30 испытуемых с такими же баллами по всем переменным), а для 60 учеников прогнозировалась успешная сдача (наблюдения 2 и 3); 180 учеников экзамен сдали успешно, из них 150 учеников сдали экзамены успешно в соответствии с прогнозом (наблюдения 4, 5, 7, 8 и 9), а 30 — несмотря на прогнозиро-

Таблица 8.12. **Выводимые SPSS результаты классификации для трех этапов логистического регрессионного анализа**

Classification Table[a]

			Predicted		
			Успеваемость ребенка		
	Observed		0	1	Percentage Correct
Step 1	Успеваемость ребенка	0	60	30	66,7
		1	30	150	83,3
	Overall Percentage				77,8
Step 2	Успеваемость ребенка	0	30	60	33,3
		1	30	150	83,3
	Overall Percentage				66,7
Step 3	Успеваемость ребенка	0	30	60	33,3
		1	30	150	83,3
	Overall Percentage				66,7

a. The cut value is 0,500.

вавшийся провал (наблюдение 6). Таким образом, было сделано 66,7 % верных прогнозов [(30 + 150) × 100/270 = 66,67]. Обратите внимание, что в терминах данного порогового значения процент правильных прогнозов уменьшается с 77,8 % для первой модели до 66,7 % для последующих моделей. Как следует из табл. 8.8, наблюдается небольшое, но значимое увеличение прогнозируемой вероятности успешной сдачи экзаменов для тех, кто на самом деле сдал их, при переходе от модели с одним предиктором к трехпредикторной модели. Таким образом, данные классификационные таблицы могут вводить в заблуждение относительно адекватности модели.

В табл. 8.13 приведены предикторы, выбираемые на каждом этапе, и значения их нестандартизированных регрессионных коэффициентов. Умственные способности ребенка были включены на первом шаге, родительская заинтересованность — на втором, а заинтересованность учителей — на третьем, завершающем, этапе. На третьем этапе значения регрессионных коэффициентов, или *B*, для умственных способностей ребенка, родительской заинтересованности и заинтересованности учителей равны 2,240; 0,600 и 0,770 соответственно. Регрессионная константа равна –8,859. Коэффициент регрессии, равный 2,240 для умственных способностей ребенка, означает, что увеличение значения умственных способностей ребенка на единицу (один пункт пятибалльной шка-

Таблица 8.13. **Выводимые SPSS значения нестандартизированных коэффициентов регрессии для предикторов на каждом этапе логистического регрессионного анализа**

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ДЕТСПОС	1,906	0,237	64,822	1	0,000	6,724
	Constant	–4,295	0,625	47,163	1	0,000	0,014
Step 2[b]	ДЕТСПОС	2,265	0,323	49,050	1	0,000	9,635
	РОДИНТЕР	1,290	0,270	22,901	1	0,000	3,635
	Constant	–7,974	1,183	45,450	1	0,000	0,000
Step 3[c]	ДЕТСПОС	2,240	0,326	47,144	1	0,000	9,397
	РОДИНТЕР	0,600	0,345	3,027	1	0,082	1,822
	УЧИНТЕР	0,770	0,332	5,392	1	0,020	2,161
	Constant	–8,859	1,357	42,588	1	0,000	0,000

a. Variable(s) entered on step 1: ДЕТСПОС.

b. Variable(s) entered on step 2: РОДИНТЕР.

c. Variable(s) entered on step 3: УЧИНТЕР.

Таблица 8.14. **Выводимые SPSS изменения значений умноженной на –2 величины логарифмического правдоподобия**

Model if Term Removed

Variable		Model Log Likelihood	Change in –2 Log Likelihood	df	Sig. of the Change
Step 1	ДЕТСПОС	–171,859	110,386	1	0,000
Step 2	ДЕТСПОС	–154,751	102,054	1	0,000
	РОДИНТЕР	–116,666	25,884	1	0,000
Step 3	ДЕТСПОС	–147,636	93,673	1	0,000
	РОДИНТЕР	–102,442	3,283	1	0,070
	УЧИНТЕР	–103,724	5,847	1	0,016

лы) приводит к увеличению логарифмического шанса успешной сдачи экзаменов на 2,240 и увеличивает шансы ребенка успешно сдать экзамен в $e^{2,240}$, или 9,397, раз.

В табл. 8.14 показаны изменения значений умноженной на –2 величины логарифмического правдоподобия при удалении предикторов из модели на каждом шаге и статистическая значимость этих изменений. Например, на третьем этапе удаление заинтересованности учителей из модели приводит к изменению значения умноженной на –2 величины логарифмического правдоподобия на 5,847 (ранее, вычисляя вручную, мы получили 7,65), что при

Таблица 8.15. **Выводимые SPSS показатели статистики и их статистическая значимость на каждом этапе для предикторов, не включаемых в модель**

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	ДЕТИНТЕР	8,840	1	0,003
		РОДИНТЕР	26,273	1	0,000
		УЧИНТЕР	23,857	1	0,000
	Overall Statistics		31,936	3	0,000
Step 2	Variables	ДЕТИНТЕР	2,411	1	0,120
		УЧИНТЕР	6,329	1	0,012
	Overall Statistics		11,081	2	0,004
Step 3	Variables	ДЕТИНТЕР	3,385	1	0,066
	Overall Statistics		3,385	1	0,066

одной степени свободы статистически значимо на уровне не более 0,016. В данной таблице также отображена величина логарифмического правдоподобия модели без каждого из предикторов. Например, на третьем этапе величина логарифмического правдоподобия для модели без заинтересованности учителей равна –103,724, что при умножении на –2 даст значение 207,448.

Наконец, в табл. 8.15 приведены статистические баллы и их статистическая значимость для предикторов, не включенных в модель на каждом этапе. Например, на третьем этапе в уравнение были включены умственные способности ребенка, родительская заинтересованность и заинтересованность учителей, а интерес ребенка к учебе включен в модель не был. Статистический балл для этой переменной равен 3,385 и имеет вероятность 0,066, что не является статистически значимым, и поэтому данная переменная не вводится в модель на четвертом этапе.

Рекомендуемая литература

Menard, S. (1995) *Applied Logistic Regression Analysis.* Thousand Oaks, CA: Sage.

Pample, EC. (2000) Logistic Regression: A Primer. Thousand Oaks, CA: Sage.

SPSS Inc. (2002) *SPSS 11.0 Regression Models.* Upper Saddle River, NJ: Prentice-Hall.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

ЧАСТЬ V
ПРОВЕРКА РАЗЛИЧИЙ МЕЖДУ ГРУППОВЫМИ СРЕДНИМИ

Глава 9
ВВЕДЕНИЕ В ДИСПЕРСИОННЫЙ И КОВАРИАЦИОННЫЙ АНАЛИЗ

Предисловие научного редактора

В ч. V проверяются различия между групповыми средними. Для этого применяют различные методы дисперсионного (вариационного) и ковариационного анализа. Эти методы можно использовать, когда зависимая переменная количественная, а среди независимых обязательно должны быть качественные переменные. Использование анализа ковариаций в нашей стране практически отсутствует в отличие от достаточно проработанного дисперсионного анализа (ANOVA и MANOVA).

Дисперсионный анализ (для его обозначения обычно используется сокращение ANOVA[1]) и ковариационный анализ (его обычно сокращенно обозначают ANCOVA[2]) являются параметрическими статистическими методами для определения того, значимо ли отличается дисперсия количественной переменной от ожидаемой в условиях случайной реализации при изменениях качественной переменной или взаимодействии этих двух количественной и качественной переменных с другими качественными переменными. Если дисперсия при этом различается, а качественная переменная состоит лишь из двух групп значений или категорий, это означает, что средние величины для этих групп различны. Там, где качественная переменная состоит более чем из двух групп значений, значимая разница определяет, что средние величины для двух или более из этих групп, вероятнее всего, будут различаться. Выяснение того, какие из средних величин будут различаться, требует дальнейшей статистической проверки.

Качественную переменную часто называют независимой переменной или фактором. В случае дисперсионного анализа может

[1] Буквально Analysis Of Variance.
[2] Буквально Analysis Of Covariance.

быть один или более факторов. Дисперсионный анализ в случае одного фактора называется *однофакторным дисперсионным анализом*. Дисперсионный анализ в случае двух факторов называется *двухфакторным дисперсионным анализом*, при наличии трех факторов — *трехфакторным* и т.д. Группы значений, или категории, внутри фактора часто называются *уровнями*. Когда в дисперсионном анализе участвуют два или более факторов, вместо количества факторов может приводиться число уровней для каждого фактора. Например, двухфакторный дисперсионный анализ с двумя группами значений для каждого фактора может называться 2×2 дисперсионным анализом. Трехфакторный дисперсионный анализ с двумя группами значений для первого фактора и тремя группами значений для каждого из двух других факторов может называться $2 \times 3 \times 3$ дисперсионным анализом и т.д.

Количественная переменная обычно называется зависимой переменной. Таких зависимых переменных может быть несколько, но в ходе дисперсионного анализа каждый раз выбирают одну и работают с ней. Многомерный дисперсионный (MANOVA) и ковариационный (MANCOVA) анализ относится к двум и более зависимым переменным, изучаемым одновременно. Однофакторный MANOVA решает те же задачи, что и дискриминантный анализ, который будет рассмотрен в гл. 12. Ковариационный анализ используется для контроля влияния одной или нескольких количественных переменных, которые коррелируют с зависимой переменной. Значения зависимой переменной для групп при разных уровнях фактора могут происходить из одних и тех же случаев (наблюдений) или же из различных наблюдений. Там, где эти значения берутся из различных наблюдений, фактор называется *несвязанным фактором*[1]. Пол испытуемого (который в данном случае и есть случай, или наблюдение) является примером несвязанного фактора[2]. Там, где значения берутся из тех же согласованных, или связанных, наблюдений, фактор называется *связанным фактором*. Одна и та же переменная, измеряемая в различные моменты времени для одних и тех же наблюдений, является примером связанного фактора. Дисперсионный анализ может также классифицироваться в соответствии с тем, содержит ли он только несвязанные факторы (несвязный дисперсионный анализ), только связанные факторы (связный дисперсионный анализ) или

---

[1] В терминах испытуемых, играющих роль наблюдений, различия в данных одной и той же выборки, подвергнутой влиянию разных условий, это пример влияния связанного фактора на разных уровнях, а различия в данных, полученных на выборках, для каждой из которых реализовывался только один из возможных уровней, — пример влияния несвязанного фактора.

[2] Невозможно создать в ходе эксперимента такие условия, чтобы испытуемый изменил свой пол.

комбинацию связанных и несвязанных факторов (смешанный дисперсионный анализ).

Задача дисперсионного анализа состоит в том, чтобы определить, какие факторы и взаимодействия этих факторов объясняют значимую долю общей дисперсии переменной. Простейший дисперсионный анализ — однофакторный дисперсионный анализ для двух независимых групп — эквивалентен критерию Стьюдента для двух несвязанных групп при допущении равенства дисперсий этих двух групп. Поскольку мы будем использовать критерий Стьюдента для несвязанных групп при таком экспериментальном плане, проиллюстрируем вычисления и применение дисперсионного анализа в данной главе на примере следующего по сложности дисперсионного анализа, а именно однофакторного дисперсионного анализа для трех несвязанных групп. Рассмотрим однофакторный ковариационный анализ в следующей главе, а затем 2×2 дисперсионный анализ для несвязанных факторов.

Однофакторный дисперсионный анализ для несвязанного фактора

Предположим, что у нас есть три группы, состоящие из двух, четырех и трех человек, как показано в табл. 9.1. Эти три группы лиц могут быть депрессивными больными, которым назначен один из трех видов лечения — отсутствие терапии, фармакотерапия и психотерапия. Применение дисперсионного анализа не зависит от того, были ли больные случайно распределены по данным терапевтическим группам или нет. В настоящем примере можно сказать, что три группы отражают соответственно три различных семейных статуса — разведенные, состоящие в браке и никогда не состоявшие в браке[1]. Зависимая переменная представляет собой тяжесть депрессии, измеренную по девятибалльной шкале, в которой более высокие значения соответствуют более выраженной депрессии. Средний балл депрессии для разведенных — самый высокий (6,00), затем следуют никогда не состоявшие в браке (средний балл депрессии 4,00) и замыкают список состоящие в браке (средний балл 3,00), как показано в табл. 9.2. Мы хотим знать, действительно ли разница между средними этих групп (межгрупповые различия) значимо выше, чем можно было бы ожидать при случайном распределении данного признака. Если различия имеют место, то важно знать, каковы они: какие из средних величин отличаются от других.

[1] Чтобы предвосхитить возможное недопонимание читателя, еще раз подчеркнем, что независимым фактором в данном случае является семейный статус, а не вид лечения.

Таблица 9.1. **Баллы выраженности депрессии для трех групп наблюдений**

Наблюдение	Группа	Депрессия
1	1	7
2	1	5
3	2	2
4	2	3
5	2	4
6	2	3
7	3	5
8	3	3
9	3	4

Групповое среднее может рассматриваться как величина, отражающая «истинное» значение определяемого показателя для данной группы, в то время как разброс баллов вокруг среднего для данной группы — как случайность, ошибка или необъясненный разброс значений. В дисперсионном анализе мы сравниваем оценку межгрупповой дисперсии с оценкой внутригрупповой дисперсии. Если оценка межгрупповой дисперсии значительно выше, чем оценка внутригрупповой дисперсии, то различия между средними вряд ли объясняются случайностью или ошибкой. Отношение оценки межгрупповой дисперсии к оценке внутригрупповой дисперсии известно как F-отношение, названное так в честь Сэра Рональда Фишера, разработавшего основы дисперсионного анализа:

$$F = \frac{[\text{Оценка межгрупповой дисперсии}]}{[\text{Оценка внутригрупповой дисперсии}]}.$$

Чем больше величина отношения оценки межгрупповой дисперсии к оценке внутригрупповой дисперсии, тем больше величина F и тем вероятнее, что средние между группами значимо различаются между собой.

Чтобы вычислить оценку межгрупповой дисперсии, предполагаем, что истинным значением изучаемой величины для испытуемого в группе является внутригрупповое среднее для данной группы. Из каждого из этих истинных значений вычитаем общее среднее по всей выборке, которое в нашем случае равно 4,00, затем возводим в квадрат эти разности, или отклонения, и складываем,

чтобы получить сумму квадратов (краткое название для суммы квадратов отклонений). Разделив эту сумму квадратов на число межгрупповых степеней свободы, равное количеству групп, минус единица, получаем среднее квадратичное (краткое название для среднеквадратичного отклонения, являющегося оценкой дисперсии). Как следует из табл. 9.2, межгрупповая сумма квадратов, число степеней свободы и среднеквадратичное отклонение равны 12,00; 2 и 6,00 соответственно.

Чтобы вычислить оценку внутригрупповой дисперсии, вычитаем групповое среднее из индивидуального балла для каждого наблюдения в этой группе, затем возводим в квадрат эти отклонения и складываем их, чтобы получить сумму квадратов отклонений от внутригрупповых средних по выборке в целом. Разделив

Таблица 9.2. **Вычисление *F*-отношения для однофакторного дисперсионного анализа**

Наблюдение	Группа	Депрессия	Межгрупповые квадратные отклонения	Внутригрупповые квадратные отклонения
1	1	7	$(6-4)^2=4$	$(6-7)^2=1$
2	1	5	$(6-4)^2=4$	$(6-5)^2=1$
Среднее по группе		12/2 = 6,00		
3	2	2	$(3-4)^2=1$	$(3-2)^2=1$
4	2	3	$(3-4)^2=1$	$(3-3)^2=0$
5	2	4	$(3-4)^2=1$	$(3-4)^2=1$
6	2	3	$(3-4)^2=1$	$(3-3)^2=0$
Среднее по группе		12/4 = 3,00		
7	3	5	$(4-4)^2=0$	$(4-5)^2=1$
8	3	3	$(4-4)^2=0$	$(4-3)^2=1$
9	3	4	$(4-4)^2=0$	$(4-4)^2=0$
Среднее по группе		12/3 = 4,00		
Среднее по выборке		36/9 = 4,00		
Сумма квадратов (*SS*)			12	6
Число степеней свободы (*df*)			3 – 1 = 2	9 – 3 = 6
Среднее квадратичное (*MS*)			12/2 = 6,00	6/6 = 1,00
F			6,00/1,00 = 6,00	

Таблица 9.3. **Таблица для однофакторного дисперсионного анализа**

Источник вариации	Сумма квадратов (*SS*)	Число степеней свободы (*df*)	Среднее квадратичное отклонение (*MS*)	*F*	*p*
Межгрупповой	12,00	2	6,00	6,00	0,05
Внутригрупповой	6,00	6	1,00		
Общий	18,00	8			

эту сумму квадратов на внутригрупповое число степеней свободы, получаем среднеквадратичное отклонение, или оценку дисперсии. Внутригрупповое число степеней свободы равно количеству наблюдений минус число групп. Это то же самое, что и вычитание единицы из числа наблюдений в каждой группе с последующим суммированием полученных величин по всем группам. Как показано в табл. 9,2, внутригрупповая сумма квадратов, число степеней свободы и среднеквадратичное отклонение равны 6,00; 6 и 1,00 соответственно.

F-отношение есть частное от деления оценки межгрупповой дисперсии на оценку внутригрупповой дисперсии, и в нашем случае равно 6,00 (6,00/1,00 = 6,00). Статистическую значимость данного *F*-отношения можно проверить по таблице критических значений распределения Фишера, хотя она будет вычислена SPSS. Отношение Фишера с двумя степенями свободы в числителе (для оценки межгрупповой дисперсии) и шестью степенями свободы в знаменателе (для оценки внутригрупповой дисперсии) должно быть больше или равно 5,15, чтобы быть статистически значимо по двустороннему критерию на уровне 0,05. Поскольку 6,00 больше данного критического значения, делаем вывод о том, что средние значения тяжести депрессии значимо различаются в данных трех группах больных с различным семейным положением.

Результаты дисперсионного анализа часто представляются в общем виде в таблице, подобной табл. 9.3. Обратите внимание на то, что общая сумма квадратов и число степеней свободы для выборки равны сумме соответствующих величин для межгруппового и внутригруппового источников разброса.

Общая сумма квадратов может быть вычислена независимо путем вычитания общевыборочного среднего из индивидуальных значений измеряемой величины для каждого наблюдения, последующего возведения полученных отклонений в квадрат и сложения полученных величин.

Общее число степеней свободы равно количеству наблюдений минус единица.

Однородность дисперсии

Дисперсионный анализ основывается на допущении о том, что генеральные совокупности, из которых осуществляются выборки, подчиняются нормальному распределению и имеют равные или однородные дисперсии. Существуют статистические критерии для определения того, значимо ли отличается асимметрия (перекос) и эксцесс (крутизна) распределения от нуля. Они описаны в других источниках (D.Cramer, 1998)[1]. Если дисперсии одной или нескольких групп значительно больше, чем дисперсии остальных, то их бóльшая величина увеличит внутригрупповую дисперсию, что уменьшит вероятность того, что отношение Фишера окажется статистически значимым. Средние величины для групп с малыми дисперсиями могут отличаться друг от друга, но эти различия будут скрыты за счет вклада групп с бóльшими дисперсиями в величину внутригрупповой дисперсии.

Один из критериев оценки однородности дисперсий — тест Левина, представляющий собой однофакторный дисперсионный анализ абсолютных отклонений индивидуальных значений от средних величин для данной группы. В нашем примере абсолютные отклонения равны нулю или единице. Например, абсолютное отклонение для первого наблюдения есть абсолютное значение разности между средним для группы, равном 6, и индивидуальным значением рассматриваемой величины, равном 7, что дает единицу. Отношение Фишера для этих абсолютных отклонений вычислено в табл. 9.4 ($F \approx 0{,}61$). Поскольку 0,61 меньше критического значения 5,15, значимых различий между группами в средних величинах абсолютных отклонений нет, и дисперсии являются однородными. Если бы дисперсии значимо различались, мы могли бы попытаться их уравнять путем преобразования исходных значений депрессии, например, извлекая корень квадратный или логарифмируя их по основанию натуральных логарифмов.

Сравнение средних

Значимость величины отношения Фишера для фактора, состоящего только из двух групп, означает, что средние этих групп различны. В том случае, когда фактор состоит из трех или более групп, нам необходимо определить, какие групповые средние различаются, сравнивая группы попарно. В случае трех групп необходимо осуществить три сравнения (группу 1 с группой 2; группу 1 с группой 3; группу 2 с группой 3). Там, где еще до сбора данных имеются веские основания предполагать, что какие-то из

[1] См. также А.Д. Наследов, 2004.

Таблица 9.4. **Вычисление теста Левина для однофакторного дисперсионного анализа**

Наблюдение	Группа	Депрессия	Абсолютные отклонения	Межгрупповые квадратные отклонения	Внутригрупповые квадратные отклонения
1	1	7	6 - 7 = 1	$(1{,}00 - 0{,}67)^2 = 0{,}11$	$(1{,}00 - 1{,}00)^2 = 0{,}00$
2	1	5	6 - 5 = 1	$(1{,}00 - 0{,}67)^2 = 0{,}11$	$(1{,}00 - 1{,}00)^2 = 0{,}00$
Среднее по группе		12/2 = 6,00	2/2 = 1,00		
3	2	2	3 - 2 = 1	$(0{,}50 - 0{,}67)^2 = 0{,}03$	$(0{,}50 - 1{,}00)^2 = 0{,}25$
4	2	3	3 - 3 = 0	$(0{,}50 - 0{,}67)^2 = 0{,}03$	$(0{,}50 - 0{,}00)^2 = 0{,}25$
5	2	4	3 - 4 = 1	$(0{,}50 - 0{,}67)^2 = 0{,}03$	$(0{,}50 - 1{,}00)^2 = 0{,}25$
6	2	3	3 - 3 = 0	$(0{,}50 - 0{,}67)^2 = 0{,}03$	$(0{,}50 - 0{,}00)^2 = 0{,}25$
Среднее по группе		12/4 = 3,00	2/4 = 0,50		
7	3	5	4 - 5 = 1	$(0{,}67 - 0{,}67)^2 = 0{,}00$	$(0{,}67 - 1{,}00)^2 = 0{,}11$
8	3	3	4 - 3 = 1	$(0{,}67 - 0{,}67)^2 = 0{,}00$	$(0{,}67 - 1{,}00)^2 = 0{,}11$
9	3	4	4 - 4 = 0	$(0{,}67 - 0{,}67)^2 = 0{,}00$	$(0{,}67 - 0{,}00)^2 = 0{,}45$
Среднее по группе		12/3 = 4,00	**2/3 = 0,67**		
Среднее по выборке			**6/9 = 0,67**		
Сумма квадратов (*SS*)				0,34	1,67
Число степеней свободы (*df*)				3 - 1 = 2	9 - 3 = 6
Среднее квадратичное отклонение (*MS*)				0,34/2 = 0,17	1,67/6 = 0,28
F				0,17/0,28 = 0,61	

средних будут различаться, можно проверить наличие подобных различий, используя априорный критерий, например односторонний критерий Стьюдента для несвязанных групп. Значения статистики Стьюдента, число степеней свободы[1] и уровни значимости для этих трех сравнений приведены в табл. 9.5. Различия имеются лишь между средними первой и третьей групп. Средняя тяжесть депрессии для группы разведенных (6,00) значимо выше, чем для группы состоящих в браке (3,00).

Там, где нет веских оснований для прогноза, какие из средних величин могут различаться, можно использовать различные апостериорные (*post-hoc*) критерии для определения того, какие средние различны, если таковые различия вообще имеются. Эти критерии учитывают тот факт, что чем больше сравнений сделано, тем выше вероятность, что некоторые средние будут различны в силу случайных причин. Например, на уровне значимости 0,05 можно предполагать, что в одном из 20 сравнений средние будут различны на этом уровне значимости в силу случайных причин. В случае 100 сравнений этот показатель возрастет до пяти. Один из наиболее общих и универсальных *post-hoc*-критериев — критерий Шеффе, который может использоваться для групп разного размера. Этот критерий является консервативным в том смысле, что он с меньшей вероятностью обнаруживает факт значимости (неслучайности) различий между средними.

Критерий Шеффе представляет собой критерий Фишера, равный отношению квадрата разности между средними двух сравниваемых групп к внутригрупповому среднеквадратичному отклонению для всех групп, умноженному на весовой коэффициент, отражающий число наблюдений в двух сравниваемых группах (n_1 и n_2):

$$F = \frac{[\text{Среднее группы 1} - \text{среднее группы 2}]^2}{[\text{Внутригрупповое среднеквадратичное отклонение} \times (n_1 + n_2)/(n_1 \times n_2)]}.$$

Значимость этого отношения Фишера сравнивается с соответствующим критическим значением величины F, которая взвешена или умножена на межгрупповое число степеней свободы (т.е. число групп минус единица). Критическое значение статистики F с двумя степенями свободы в числителе и шестью степенями свободы в знаменателе[2] для уровня значимости 0,05 приближенно равно 5,15. Следовательно, соответствующее критическое

[1] Число степеней свободы для критерия Стьюдента равно сумме наблюдений в обеих группах минус 2.

[2] Число степеней свободы числителя равно числу групп минус 1, а число степеней свободы знаменателя — общему числу измерений минус число групп.

Таблица 9.5. **Сравнение средних по критерию Стьюдента и критерию Шеффе для несвязанных групп**

Сравнение	Критерий Стьюдента t	Число степеней свободы df	Значимость p	Критерий Шеффе S	df	p
Группы 1↔2	3,46	4	0,026	12,00	2,6	0,037
Группы 1↔3	1,90	3	Незначима	4,80	2,6	Незначима
Группы 2↔3	1,46	5	Незначима	1,72	2,6	Незначима

значение отношения Фишера для данного критерия Шеффе равно 10,30 [5,15 × (3 – 1) = 10,30].

Отношение Фишера для критерия Шеффе для сравнения средних величин выраженности депрессии у разведенных и состоящих в браке равно 12,00:

$$\frac{(6,00-3,00)^2}{1,00\times(2+4)(2\times4)}=\frac{3,00^2}{1,00\times0,75}=\frac{9,00}{0,75}=12,00.$$

Поскольку 12,00 > 10,30, средние для этих двух групп различаются значимо. Отношение Фишера для критерия Шеффе, число степеней свободы и значимости для рассматриваемых трех сравнений приведены в табл. 9.5. Только средние двух рассмотренных групп различаются значимо. Хотя это различие является значимым и по критерию Стьюдента для независимых групп, и по критерию Шеффе, точное значение значимости выше для критерия Шеффе (0,037) по сравнению со значением (0,026) по двустороннему критерию Стьюдента для независимых групп. Что и следовало ожидать в случае более консервативного критерия.

Общая линейная модель

Вычисления для дисперсионного анализа могут быть проведены с помощью множественной регрессии. Оба эти метода могут быть выведены из общей линейной модели. В случае многомерной регрессии квадрат множественной корреляции (R^2) представляет собой долю дисперсии зависимой переменной, или переменной-отклика, объясняемую независимыми переменными, или предикторами. Дисперсия есть сумма квадратов отклонений, деленная на число степеней свободы, равное количеству наблюдений минус единица. Поскольку число степеней свободы при дисперсионном

анализе одно и то же для трех источников дисперсии[1], можно не обращать на него внимание. Следовательно, квадрат множественной корреляции может быть получен как отношение межгрупповой суммы квадратов к общей сумме квадратов:

$$R^2 = \frac{\text{[Межгрупповая сумма квадратов]}}{\text{[Общая сумма квадратов]}}.$$

В дисперсионном анализе эта статистика называется эта-квадрат (η^2), или корреляционное отношение. В нашем примере квадрат коэффициента множественной корреляции, или корреляционное отношение, приближенно равен 0,67 (12,00/18,00 = 0,667).

Как отмечалось в гл. 5, один из способов вычисления квадрата множественной корреляции в случае многомерной регрессии состоит в том, чтобы сложить произведения стандартизированных частных регрессионных коэффициентов по всем предикторам (независимым переменным) на коэффициенты их корреляции с зависимой переменной:

R^2 = сумма по всем предикторам [стандартизированный коэффициент регрессии × коэффициент корреляции].

Хотя семейное положение представляет собой единый фактор, это категориальная переменная. Следовательно, надо ввести дополнительные переменные-предикторы таким образом, чтобы каждая из них задавала определенную категориальную группу в отдельности. Число необходимых для этого предикторов всегда на единицу меньше числа категорий. Таким образом, чтобы задать три группы с различным семейным статусом, необходимо два предиктора. Простейший способ определить группу — ввести такую фиктивную переменную, которая принимает единичное значение для всех наблюдений данной группы и нулевое значение — для наблюдений из всех остальных групп. Например, группа разведенных может быть закодирована единицами, в то время как группы состоящих в браке и никогда не состоявших в браке — нулями (табл. 9.6). Эта новая переменная представляет собой фиктивную переменную. Вторая фиктивная переменная, необходимая для того, чтобы идентифицировать все три группы, принимает значение 1 для состоящих в браке и значение 0 — для разведенных и никогда не состоявших в браке. Используя две эти фиктивные переменные, видим, что группа разведенных определяется единичным значением первой переменной и нулевым значением второй переменной, группа состоящих в браке — нулевым значением первой переменной и единичным значением второй перемен-

[1] Поскольку каждая из трех переменных измерена на одном и том же числе наблюдений.

Таблица 9.6. **Значения фиктивных переменных для трех групп**

Наблюдение	Группа	Депрессия	Фиктивные переменные	
			Разведенные	Состоящие в браке
1	1	7	1	0
2	1	5	1	0
3	2	2	0	1
4	2	3	0	1
5	2	4	0	1
6	2	3	0	1
7	3	5	0	0
8	3	3	0	0
9	3	4	0	0

ной, а группа никогда не состоявших в браке задается нулевыми значениями обеих переменных.

Если провести регрессию выраженности депрессии по этим двум фиктивным переменным, стандартизированный частный коэффициент регрессии для первой фиктивной переменной составит 0,59, а для второй — (–0,35). Коэффициент корреляции выраженности депрессии с первой фиктивной переменной равен 0,76, а со второй — (–0,63).

Таким образом, квадрат коэффициента множественной корреляции равен 0,67:

$$(0{,}59 \times 0{,}76) + (-0{,}35 \times -0{,}63) = 0{,}45 + 0{,}22 = 0{,}67.$$

F-отношение (отношение Фишера) для квадрата коэффициента множественной корреляции (см. гл. 5) равно:

$$F = \frac{\dfrac{[\text{Изменение } R^2]}{[\text{Число добавленных предикторов}]}}{\dfrac{[1 - R^2]}{[N - \text{число предикторов} - 1]}}.$$

Если использовать достаточное количество десятичных разрядов после запятой, отношение Фишера в нашем примере будет равным 6,00, как это имеет место в случае дисперсионного анализа:

$$\frac{\dfrac{0{,}6667}{2}}{\dfrac{(1 - 0{,}6667)}{(9 - 2 - 1)}} = \frac{0{,}3334}{\dfrac{0{,}3333}{6}} = \frac{0{,}3334}{0{,}0556} = 5{,}996.$$

Для того чтобы быть статистически значимым на уровне 0,05, отношение Фишера с двумя степенями свободы в числителе и шестью степенями свободы в знаменателе должно быть больше или равно 5,15, что в нашем случае выполняется.

Можно назвать три причины, почему полезно знать, как соотносятся между собой множественная регрессия и дисперсионный анализ. Во-первых, категориальные переменные, такие, как «Семейное положение», необходимо кодировать фиктивными переменными и вводить их одновременно на одном этапе множественной регрессии. В этом случае доля объясняемой фиктивными переменными дисперсии равна эта-квадрату для дисперсионного анализа. Во-вторых, дисперсионный анализ часто проводится в статистических пакетах с использованием множественной регрессии. В дисперсионном анализе с несколькими факторами, где количество наблюдений для разных факторов не одно и то же или непропорционально, результаты могут различаться в зависимости от метода, используемого для ввода переменных. Понимание этих различий может помочь в выборе наиболее адекватного метода для анализа. В-третьих, вычисления, проводимые в ковариационном анализе, проще объяснить в терминах многомерной регрессии.

Отчет о результатах

Одной из форм краткого отчета о результатах дисперсионного анализа в настоящем примере может быть такая: «Однофакторный дисперсионный анализ выявил, что семейное положение значимо влияет на выраженность депрессии ($F_{2,6} = 6{,}00$; $p < 0{,}05$)». Если бы у нас имелись веские основания для прогноза того, что выраженность депрессии окажется выше в группе разведенных по сравнению с группой состоящих в браке, мы могли бы добавить утверждение следующего рода: «Как прогнозировалось, выраженность депрессии оказалась значимо выше (коэффициент Стьюдента с четырьмя степенями свободы для несвязанных групп $t_4 = 3{,}45$, значимость по одностороннему критерию $p < 0{,}05$) среди разведенных (среднее $M = 6{,}00$, СКО[1] = 1,41), чем среди состоящих в браке (среднее $M = 3{,}00$; СКО = 0,82). Между другими парами средних значимых различий выявить не удалось». Если у нас не было веских оснований для прогнозов, то можно добавить следующее: «Критерий Шеффе показал, что средняя выраженность депрессии значимо отличалась (Шеффе $F_{2,6} = 12{,}00$; $p < 0{,}05$) лишь для группы разведенных и группы состоящих в браке. Средняя выраженность депрессии была выше для разведенных (сред-

[1] СКО — среднее квадратичное отклонение (*SD*).

нее $M = 6{,}00$, СКО = 1,41) по сравнению с состоящими в браке (среднее $M = 3{,}00$; СКО = 0,82)».

Процедура SPSS для Windows

Приведем алгоритм проведения дисперсионного анализа, описанного в данной главе.

Введите данные с помощью **Редактора данных (Data Editor)**, как показано на рис. 9.1. Первый столбец с именем «**группа**» содержит номера трех групп наблюдений. Второй столбец с названием «**депресс**» содержит значения общей выраженности тяжести депрессии. Задайте названия групп, приписав им соответствующие имена (метки) — «**Разведенные**», «**Состоящие в браке**» и «**Никогда не состоявшие в браке**».

В строке меню в верхней части окна программы выберите пункт **Анализ (Analyze)**, в появившемся ниспадающем меню — пункт **Общие линейные модели (General Linear Model)**, а затем — подпункт **Одномерная модель (Univariate)**, чтобы открыть диалоговое окно **Одномерная модель (Univariate)**, представленное на рис. 9.2.

Выберите «**депресс**» и с помощью первой верхней кнопки ▸ переместите эту переменную в поле **Зависимая переменная (Dependent)**.

Выберите «**группа**» и с помощью второй сверху кнопки ▸ переместите эту переменную в список **Фиксированные фактор(ы) (Fixed Factor(s))**.

Чтобы осуществить проверку постфактум, нажмите на кнопку **Критерии постфактум (Post Hoc)**, открывающую дочернее диалоговое окно **Одномерная модель: Множественные сравнения наблюдаемых средних постфактум (Univariate: Post Hoc Multiple Comparisons for Observed Means)**, изображенное на рис. 9.3.

В списке **Факторы (Factors)** выберите переменную «**группа**» и с помощью кнопки ▸ переместите ее в список «**Критерии постфактум для**» **(Post Hoc Tests for)**.

Установите флажок **Шеффе (Scheffe)**, чтобы осуществить проверку по критерию Шеффе, и нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в основное диалоговое окно **Одномерная модель (Univariate)**.

	группа	депресс
1	1	7
2	1	5
3	2	2
4	2	3
5	2	4
6	2	3
7	3	3
8	3	4
9	3	5

Рис. 9.1. Данные для проведения однофакторного дисперсионного анализа в **Редакторе данных (Data Editor)**

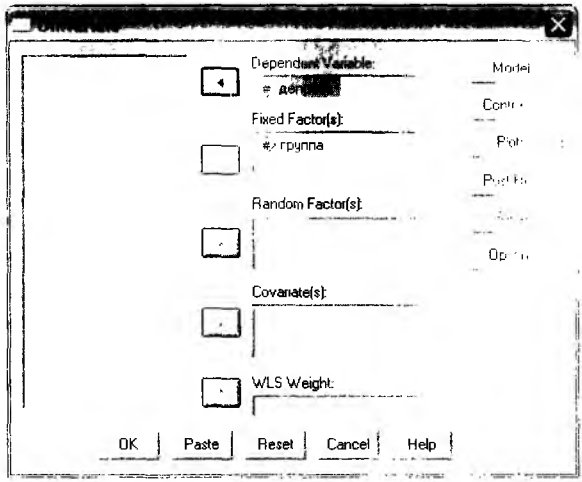


Рис. 9.2. Диалоговое окно **Одномерная модель (Univariate)**

Нажмите на кнопку **Опции (Options)** и откройте дочернее диалоговое окно **Одномерная модель: Опции (Univariate: Options)**, изображенное на рис. 9.4.

Установите флажок **Описательные статистики (Descriptive statistics)**, чтобы вывести средние и стандартные отклонения для трех рассматриваемых групп.

Установите флажок **Критерии однородности (Homogeneity tests)**, чтобы выполнить тест Левина на однородность дисперсий.

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно.

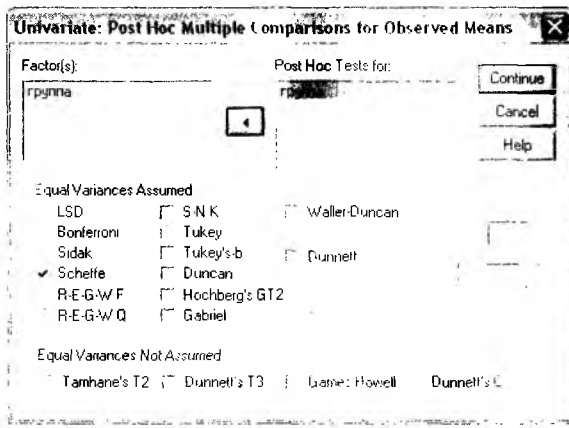


Рис. 9.3. Дочернее диалоговое окно **Постфактум сравнения для измеренных средних (Univariate: Post Hoc Multiple Comparisons for Observed Means)**

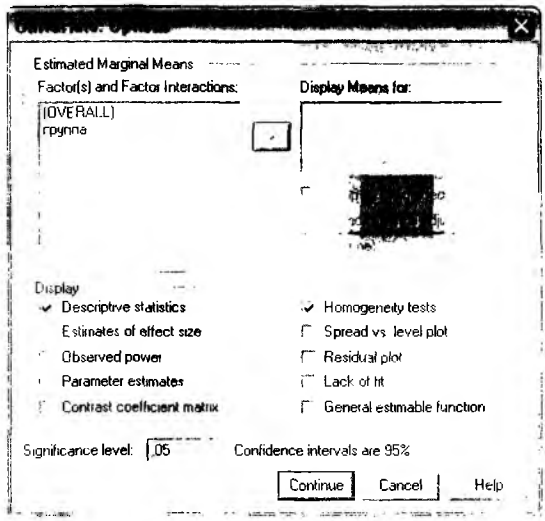


Рис. 9.4. Дочернее диалоговое окно **Одномерная модель. Опции (Univariate: Options)**

Нажмите **OK** для выполнения однофакторного дисперсионного анализа.

Чтобы провести регрессионный анализ, введите значения двух фиктивных переменных в третий и четвертый столбцы матрицы **Редактора данных (Data Editor)**, как показано в табл. 9.6. Выполните регрессию переменной «**депресс**» по этим двум переменным.

Выводимая SPSS информация по результатам работы

При выводе результатов первой отображается таблица средних и стандартных (среднеквадратичных) отклонений, которую здесь не приводим. За ней следует таблица с результатами теста Левина (табл. 9.7). Дисперсии значимо не различаются, поскольку уровень значимости отношения Фишера, равного 0,600, составляет 0,579, что превосходит 0,05.

Затем выводятся показатели дисперсионного анализа (табл. 9.8). Для нас актуальны только три из шести перечисленных показателей, а именно: ГРУППА (GROUP), Ошибка (Error) и Общее исправление (Corrected Total), соответствующие трем представленным в табл. 9.3 источникам разброса[1]. Величина отношения Фишера, равная 6,000, имеет уровень значимости 0,037, что является статистически значимым, поскольку данное значение меньше 0,05.

[1] Межгрупповой, Внутригрупповой, Общий соответственно.

Таблица 9.7. **Выводимые SPSS результаты теста Левина для однофакторного дисперсионного анализа**

Levene's Test of Equality of Error Variances[a]

Dependent Variable: ДЕПРЕСС

F	df1	df2	Sig.
0,600	2	6	0,579

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design:Intercept + ГРУППА.

Последними отображаются результаты проверки по критерию Шеффе. Первая из двух выводимых программой таблиц представлена в табл. 9.9, поскольку интересует нас в наибольшей степени. Сравнения в этой таблице повторяются дважды. Например, первая строка сравнивает средние группы разведенных и состоящих в браке, в то время как третья строка сравнивает средние группы состоящих в браке и разведенных. Для этих сравнений приводятся только величины значимостей, которые в данном случае составляют 0,037. Значения величин отношения Фишера и количества степеней свободы могут быть вычислены так, как описывалось ранее.

В табл. 9.10 приведены стандартизированные частные регрессионные коэффициенты для множественной регрессии с двумя фиктивными переменными. Первый из коэффициентов равен 0,588, второй — (–0,351).

Таблица 9.8. **Выводимые SPSS показатели однофакторного дисперсионного анализа**

Tests of Between-Subjects Effects

Dependent Variable: ДЕПРЕСС

Sorrected Model	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12,000[a]	2	6,000	6,000	0,037
Intercept	156,000	1	156,000	156,000	0,000
ГРУППА	12,000	2	6,000	6,000	0,037
Error	6,000	6	1,000		
Total	162,000	9			
Corrected Total	18,000	8			

a. R Squared = 0,667 (Adjusted R Squared = 0,556).

Таблица 9.9. **Выводимые SPSS результаты проверки по критерию Шеффе для однофакторного дисперсионного анализа**

Multiple Comparisons

Dependent Variable: ДЕПРЕСС
Scheffe

(I) ГРУППА	(J) ГРУППА	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Разведенные	Состоящие в браке	3,00*	0,866	0,037	0,22	5,78
	Никогда не состоявшие в браке	2,00	0,913	0,171	–0,93	4,93
Состоящие в браке	Разведенные	–3,00*	0,866	0,037	–5,78	–0,22
	Никогда не состоявшие в браке	–1,00	0,764	0,471	–3,45	1,45
Никогда не состоявшие в браке	Разведенные	–2,00	0,913	0,171	–4,93	0,93
	Состоящие в браке	1,00	0,764	0,471	–1,45	3,45

Based on observed means.

* The mean difference is significant at the 0,05 level.

Таблица 9.10. **Выводимые SPSS стандартизированные частные коэффициенты регрессии для двух фиктивных переменных**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,000	0,577		6,928	0,000
	ФИКТИВ 1	2,000	0,913	0,588	2,191	0,071
	ФИКТИВ 2	–1,000	0,764	–0,351	–1,309	0,238

a. Dependent Variable: ДЕПРЕСС.

Рекомендуемая литература

Cohen, J. (1968) Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70: 426-43.

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows*. London: Routledge.

Diekhoff, G. (1992) *Statistics for the Social and Behavioral Sciences*. Dubuque, IA: Wm. C. Brown.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction*, 2nd edn. New York: Holt, Rinehart & Winston.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package*. Upper Saddle River, NJ: Prentice-Hall.

Глава 10

НЕСВЯЗНЫЙ ОДНОФАКТОРНЫЙ КОВАРИАЦИОННЫЙ АНАЛИЗ

В ковариационном анализе (ANCOVA) доля дисперсии, совместная для зависимой и одной или нескольких количественных переменных[1], исключается, а затем определяется, может ли оставшаяся дисперсия быть объяснена одним или несколькими факторами и их взаимодействиями. Совместную для двух переменных дисперсию называют *ковариацией*. Количественные переменные[2], имеющие долю общей дисперсии с зависимой переменной и поэтому исключаемые из анализа, называются *ковариатами*. Другими словами, ковариаты коррелируют с зависимой переменной. Если средние значения какой-либо ковариаты одинаковы для различных уровней, или групп фактора, то средние значения отклика (зависимой переменной) для этих уровней не нужно корректировать для данной ковариаты. В таком случае доля дисперсии, объясняемой фактором, не отличается от величины, получаемой при дисперсионном анализе, а величина ошибки меньше, поскольку часть дисперсии ошибки будет объясняться ковариатой.

Следовательно, объясняемая фактором доля дисперсии с большей вероятностью будет статистически значимой при ковариационном, чем при дисперсионном анализе, показывая, что в первом случае вероятность существования различий между групповыми средними выше, чем во втором. Там, где средние значения ковариаты различаются для разных уровней фактора, необходима коррекция средних значений зависимой величины с учетом ковариаты. Чем сильнее различия средних значений ковариаты, тем больше должна быть коррекция средних значений отклика. В результате скорректированные средние значения могут различаться меньше, чем нескорректированные. Более того, могут различаться соотношения величин скорректированных и нескорректированных средних значений. Крайний случай такой ситуации проявляется тогда, когда группа с наибольшим нескорректированным средним значением имеет наименьшее скорректированное среднее.

Наибольшая вероятность того, что средние значения ковариаты окажутся равными, — в истинном эксперименте[3], где наблю-

---

[1] Независимых.

[2] Независимые.

[3] В отличие от квази-эксперимента.

дения или участники случайным образом приписываются различным уровням фактора. Случайное приписывание используется, чтобы гарантировать тот факт, что наблюдения вряд ли различаются кроме как по уровню фактора, или способу экспериментального воздействия, который был для них выбран.

Например, пациенты, участвующие в эксперименте по сравнению различных видов лечения депрессии, скорее всего, различаются по степени выраженности депрессии. При этом у одних групп испытуемых перед началом эксперимента депрессия была выражена сильнее, чем у других. Больные с большей выраженностью депрессии перед началом терапии могут меньше поддаваться лечению и, таким образом, демонстрировать меньшее улучшение состояния здоровья, чем пациенты с менее выраженной депрессией к началу терапии. С другой стороны, они могут демонстрировать большее улучшение, поскольку допустимое возможное улучшение у них больше.

Следовательно, результаты подобного исследования будет труднее интерпретировать, если больные, распределенные в разные группы по способу лечения, различаются по степени выраженности депрессии перед началом терапии. Если выраженность депрессии до начала терапии примерно одинакова у пациентов из разных групп и если выраженность депрессии до начала терапии положительно коррелирует с выраженностью депрессии непосредственно после окончания лечения, ковариационный анализ может обеспечить более чувствительный критерий эффективности различных видов терапии за счет устранения различий в выраженности депрессии до начала терапии внутри различавшихся по видам лечения групп больных.

Однако соблюдение требований случайности распределения по группам не всегда гарантирует одинаковость наблюдений во всех отношениях, особенно если число используемых наблюдений относительно невелико.

Например, может оказаться, что среднее значение выраженности депрессии до начала терапии в каких-то группах было выше, чем в других. Имеются различные способы выхода из этого положения, ни один из которых нельзя признать полностью удовлетворительным. Многие исследователи в сфере социальных наук полагают, что предпочтительным методом в случае различий до начала терапии является ковариационный анализ (например, J. Stevens, 1996). При таком анализе предполагают равенство средних значений до начала терапии, а средние значения после лечения корректируются с учетом исходных различий до начала терапии, что затрудняет интерпретацию результатов.

Средние значения ковариат для наблюдений, которые не были случайным образом распределены по различным уровням фактора, вероятнее всего, будут различаться по группам. Например,

разведенные, состоящие в браке и никогда не состоявшие в браке, скорее всего, различаются по ряду показателей, одним из которых может быть возраст. Разведенные могут быть в целом старше состоящих в браке, которые, в свою очередь, могут быть старше никогда не состоявших в браке. Если выяснится, что возраст коррелирует с зависимой переменной — выраженностью депрессии, то влияние возраста можно будет проконтролировать, проведя ковариационный анализ. Ковариационный анализ часто используется в таких случаях, хотя следует помнить, что одним из допущений (условий применимости) данного метода является равенство средних значений ковариаты по группам[1].

Проиллюстрируем однофакторный ковариационный анализ, используя данные из табл. 9.1 так, что результаты ковариационного анализа можно будет сопоставить с результатами проведенного в предыдущей главе дисперсионного анализа. Как и прежде, единственный фактор имеет три уровня, или группы: разведенные, состоящие в браке и никогда не состоявшие в браке. Наш анализ будет включать одну ковариату, которую будем называть физическое нездоровье. Как и зависимая переменная выраженности депрессии, эта переменная измеряется по девятибалльной шкале, где большее значение указывает на более выраженное нездоровье. Хотя на практике маловероятно, чтобы средние значения ковариаты были одинаковы для всех уровней фактора, мы искусственно создали набор данных, где это условие выполняется, и другой набор данных, в котором средние различны, чтобы иметь возможность сравнить эти два случая. Эти два набора данных приведены в табл. 10.1 вместе со значениями зависимой переменной и средними величинами всех трех переменных для каждой из трех групп и всей выборки. Средние значения для выраженности нездоровья в первом примере одинаковы для всех групп и равны 5, в то время как во втором примере средние по группам соответственно равны 8, 5 и 3.

Чтобы провести ковариационный анализ, необходимо вычислить коэффициент корреляции между ковариатой и зависимой переменной. Зависимая переменная в обоих случаях одна и та же. Коэффициент корреляции в случае равных средних значений ковариаты составил 0,39, а в случае неравных средних значений — 0,31. Результаты ковариационного анализа можно представить в табличной форме так же, как и результаты дисперсионного анализа, приведенные в табл. 9.3. Однако часто бывает полезным показать вариацию, или сумму квадратов, зависимой переменной, объясняемую за счет ковариаты, как это сделано в табл. 10.2. Основные статисти-

[1] Предварительную проверку на равенство средних проводить нужно, однако тот факт, что они различаются, не является препятствием для проведения ковариационного анализа.

Таблица 10.1. **Сырые (первичные) и средние значения для ковариационного анализа**

Наблюдения	Группа	Депрессия	Физическое нездоровье	
			Равные средние	Неравные средние
1	1	7	6	7
2	1	5	4	9
Среднее по группе		12/2 = 6	10/2 = 5	16/2 = 8
3	2	2	4	5
4	2	3	6	6
5	2	4	5	4
6	2	3	5	5
Среднее по группе		12/4 = 3	20/4 = 5	20/4 = 5
7	3	3	5	4
8	3	4	4	2
9	3	5	6	3
Среднее по группе		12/3 = 4	15/3 = 4	9/3 = 3
Среднее по выборке		36/9 = 4	45/9 = 5	45/9 = 5

ки дисперсионного анализа и двух ковариационных анализов для данных из табл. 10.1 представлены для сравнения в табл. 10.2.

Видим, что отношение Фишера F для ковариационного анализа с равными средними (9,00) превосходит аналогичное значение для дисперсионного анализа (6,00), поскольку значительная доля суммы квадратов ошибки зависимой переменной (2,67/6,00 = 0,45) объясняется ковариатой. Исключение этой объясняемой суммы квадратов в ковариационном анализе уменьшает общую величину суммы квадратов (6,00 – 2,67 = 3,33), снижая таким образом значение среднеквадратичного отклонения (3,33/5 = 0,67) и повышая величину F[1]. Обратите внимание на то, что число степеней свободы ошибки в случае ковариационного анализа на единицу меньше. Это означает, что для того, чтобы быть статистически значимым, отношение Фишера F в случае ковариационного анализа должно быть несколько выше, чем в случае дисперсионного анализа. С двумя степенями свободы в числителе и пятью степенями свободы в знаменателе отношение Фишера F должно быть не меньше 5,79, чтобы быть статистически значимым на уровне 0,05, в отличие от дисперсионного анализа, где оно должно быть не ме-

[1] $F = MS_{\text{группы}} / MS_{\text{ошибки}} = 6/0{,}67 \approx 9$.

Таблица 10.2. **Сравнение дисперсионного и ковариационного анализа**

	Дисперсионный анализ[1]				Ковариационный анализ							
					Равные средние				Неравные средние			
	Сумма квадратов (*SS*)	Число степеней свободы (*df*)	Среднее квадратичное отклонение (*MS*)	*F*	*SS*	*df*	*MS*[2]	*F*	*SS*	*df*	*MS*	*F*
Ковариата					2,67	1	2,67	4,00	2,67	1	2,67	4,00
Группа	12,00	2	6,00	6,00*	12,00	2	6,00	9,00*	12,89	2	6,44	9,67*
Ошибка	6,00	6	1,00		3,33	5	0,67		3,33	5	0,67	
Общая	18,00	8			18,00	8			18,00	8		

* $p < 0{,}05$.

нее 5,15. Однако увеличение реального значения отношения Фишера существенно превосходит увеличение критического значения *F* так, что вероятность случайного осуществления данного события в случае ковариационного анализа ($p = 0{,}022$) меньше, чем в случае дисперсионного анализа ($p = 0{,}037$). Поскольку средние значения ковариаты для всех трех групп в точности равны, средние значения зависимой переменной остаются без изменения: выраженность депрессии у разведенных выше всего (6,00), затем следуют никогда не состоявшие в браке (4,00) и состоящие в браке (3,00).

Отношение Фишера *F* для ковариационного анализа с неравными средними (9,67) несколько выше, чем для анализа с равными средними (9,00), а потому и более значимо ($p = 0{,}019$). Поскольку средние значения ковариаты для трех групп различаются, средние значения зависимой переменной должны быть скорректированы. Скорректированное среднее значение выше всего в группе разведенных (8,00), затем следуют состоящие в браке (3,00), а за ними — никогда не состоявшие в браке (2,67). Когда производится корректировка средних значений с учетом различий средних значений ковариаты, никогда не состоявшие в браке оказываются менее депрессивными (2,67), чем состоящие в браке (3,00),

[1] Это практически табл. 9.3.

[2] Как и в гл. 9: $MS = SS/df$

в то время как в случае отсутствия корректировки выраженность депрессии у никогда не состоявших в браке (4,00) несколько выше, чем у состоящих в браке (3,00). Другими словами, порядок значений скорректированных средних несколько отличается от порядка значений нескорректированных средних. Один из способов вычисления этих скорректированных средних значений будет описан позже.

Сравнение средних

Как и в случае дисперсионного анализа, там, где фактор состоит более чем из двух групп, необходимы дополнительные способы для выявления того, какие из групповых средних различны. Одним из них является тест защищенных наименьших значимых различий Фишера[1] (B.E.Huitema, 1980), рассчитываемый по следующей формуле:

$$t = \frac{\left[\begin{array}{l}\text{Скорректированное среднее группы 1} - \\ - \text{скорректированное среднее группы 2}\end{array}\right]}{\sqrt{\left[\begin{array}{l}\text{Скорректированная} \\ \text{среднеквадратичная} \\ \text{ошибка}\end{array}\right]\left\{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)+\left[\frac{\left[\begin{array}{l}\text{Среднее значение} \\ \text{ковариаты группы 1} - \\ - \text{среднее значение} \\ \text{ковариаты группы 2}\end{array}\right]^2}{\left[\begin{array}{l}\text{Сумма квадратов} \\ \text{ошибки ковариаты}^2\end{array}\right]}\right]\right\}}}.$$

Проиллюстрируем применение этой формулы на сравнении разведенных и состоящих в браке, где средние значения не равны. Сумма квадратов ошибки ковариаты равна 6,00 как для случая неравных средних значений, так и для случая равных средних величин. Подставив соответствующие значения в формулу, получаем:

$$t = \frac{8,00-3,00}{\sqrt{0,67\times\left\{\left(\frac{1}{2}+\frac{1}{4}\right)+\left[\frac{(8,00-5,00)^2}{6,00}\right]\right\}}} = \frac{5,00}{\sqrt{0,67\times(0,75+1,50)}} =$$

$$= \frac{5,00}{\sqrt{1,51}} = \frac{5,00}{1,23} = 4,07.$$

[1] В англоязычной литературе этот тест называется Protected Least Significant Difference (PLSD).

[2] Доля дисперсии, не объясняемая ковариатой.

Таблица 10.3. **Сравнение пар средних с помощью теста Фишера защищенных наименьших значимых различий**

Сравнения	Равные средние			Неравные средние		
	t[1]	df[2]	p[3]	t	df	p
Группы 1↔2	4,23	5	0,01	4,07	5	0,01
Группы 1↔3	2,68	5	0,05	2,91	5	0,05
Группы 2↔3	1,60	5	Незначима	0,36	5	Незначима

[1] Критерий Стьюдента.
[2] Число степеней свободы.
[3] Значимость.

Число степеней свободы для данного t-критерия равно числу степеней свободы скорректированной среднеквадратической ошибки, которое в свою очередь равно количеству наблюдений минус количество групп минус единица. Для нашего случая $df = 5$ (9 – 3 – 1 = 5). Если нет веских оснований для прогнозирования того, какие из средних значений будут различаться, будем использовать двусторонний критерий с критическим значением 0,05; если же имеются веские основания для прогноза, какая из средних величин будет больше, — односторонний критерий с критическим значением 0,05. В случае, когда число степеней свободы равно 5, критическое значение двустороннего t-критерия для уровня 0,05 равно 2,58 и 2,02 для одностороннего критерия. Поскольку значение t, равное 4,07, больше каждого из этих критических значений, можно сделать вывод о том, что у разведенных депрессия значимо более выражена, чем у состоящих в браке. Результаты вычислений по двустороннему критерию для всех трех сравнений с равными и неравными средними значениями ковариаты приведены в табл. 10.3. В отличие от методов анализа, проведенных в предыдущей главе, показавших, что разведенные по выраженности депрессии значимо не отличаются от никогда не состоявших в браке, данные результаты показывают, что депрессия у разведенных также статистически более выражена, чем у никогда не состоявших в браке.

Однородность дисперсии и регрессии

Как и дисперсионный анализ, ковариационный анализ основан на допущении о том, что внутригрупповые дисперсии примерно равны и значимо не различаются. Если же они различаются, то преобразование соответствующих значений (баллов) зави-

симой переменной может уменьшить различия. Одним из способов проверки однородности дисперсий является тест Левина, который в данном случае представляет собой просто однофакторный дисперсионный анализ абсолютных отклонений значений (баллов) зависимой переменной от значений, прогнозируемых групповым фактором и ковариатой. Число степеней свободы для межгрупповой суммы квадратов равно количеству групп минус единица, а число степеней для внутригрупповой суммы квадратов равно числу наблюдений минус число групп. Для рассматриваемого примера число степеней свободы для межгрупповой суммы квадратов равно двум (3 – 1 = 2), а для внутригрупповой суммы квадратов — шести (9 – 3 = 6). Критическое значение критерия Фишера с данными степенями свободы для уровня значимости 0,05 равно 5,15. Поскольку величина *F*-отношения для данного критерия (0,516) меньше 5,15, указанное допущение выполняется. Метод вычисления данных абсолютных отклонений будет описан далее.

Другое допущение, выполнение которого необходимо для ковариационного анализа, состоит в том, что линейное соотношение между зависимой переменной и ковариатой, подсчитываемое отдельно для каждой группы, не должно значимо различаться между группами. Если же оно отличается для одной или более групп, то коррекция средних значений зависимой переменной, осуществляемая для этих групп, будет неточной. Это линейное соотношение выражается регрессионными коэффициентами и данное допущение известно как однородность регрессии. Она проверяется с помощью критерия Фишера, где скорректированная внутригрупповая сумма квадратов разделяется на межрегрессионную [between-regressions] сумму квадратов и остаточную сумму квадратов. *F*-отношение представляет собой частное от деления среднего значения межрегрессионной суммы квадратов на среднее значение остаточной суммы квадратов. Среднее значение межрегрессионной суммы квадратов равно межрегрессионной сумме квадратов, деленной на ее число степеней свободы, которое равно числу групп минус единица. Среднее значение остаточной суммы квадратов есть остаточная сумма квадратов, деленная на ее число степеней свободы, которое равно количеству наблюдений минус удвоенное число групп. Для настоящего примера число степеней свободы межрегрессионной суммы квадратов равно двум (3 – 1 = 2), а остаточной суммы квадратов — трем [9 – (2 × 3) = 3]. Критическое значение критерия Фишера с данными степенями свободы для уровня значимости 0,05 равно 9,55. Поскольку величина *F* для ковариаты с равными (0,167) и неравными (0,167) средними значениями меньше, чем 9,55, данное допущение выполняется. Один из способов вычисления *F*-отношения будет кратко описан.

Ковариационный анализ и множественная регрессия

Возможно, самый простой способ получения основных статистик для ковариационного анализа состоит в использовании множественной регрессии. Из гл. 9 известно, что фактор семейного положения может быть представлен двумя фиктивными переменными, приведенными в табл. 9.6. Ковариата остается без изменений. Чтобы вычислить значение F-критерия для проверки однородности регрессии, необходимо создать две дополнительные фиктивные переменные, которые представляют взаимодействие между ковариатой и фактором. Если значения регрессионного коэффициента значимо различаются между группами, значительная доля дисперсии будет объясняться этими фиктивными переменными, которые получаются путем умножения каждой из первых двух фиктивных переменных на значение ковариаты (табл. 10.4). Тем, кто не желает рассматривать подробные выкладки, связанные с данным анализом, следует переходить к следующему подразделу, посвященному скорректированным средним значениям.

Как было показано в гл. 9, F-критерий может быть получен из следующего равенства:

$$F = \frac{\dfrac{[\text{Изменение } R^2]}{[\text{Число добавленных предикторов}]}}{\dfrac{[1 - R^2]}{[N - \text{число предикторов} - 1]}}.$$

Квадрат множественной корреляции (R^2) представляет собой объясненную долю общей дисперсии зависимой переменной. Если известна общая сумма квадратов, то можно вычислить сумму квадратов для различных членов в таблице ковариационного анализа. В гл. 9 было установлено, что общая сумма квадратов для выраженности депрессии равна 18,00. Если известен квадрат множественной корреляции для ковариаты и для ковариаты и первых двух фиктивных переменных, то можно вычислить значение F, сумму квадратов и среднеквадратичное отклонение (см. табл. 10.2). Если известен квадрат коэффициента множественной корреляции для ковариаты и четырех фиктивных переменных, то можно вычислить F-отношение для однородности регрессии. Квадрат множественной корреляции для этих трех наборов предикторов равен 0,099; 0,815 и 0,833 соответственно.

Доля дисперсии, объясняемой группами, равна изменению, или разности, между квадратом множественной корреляции для ковариаты и двух фиктивных переменных (0,815) и квадратом множественной корреляции для ковариаты (0,099), что равняется 0,716 (0,815 – 0,099 = 0,716). Доля остающейся дисперсии, счита-

Таблица 10.4. **Фиктивные переменные для ковариационного анализа с неравными средними значениями ковариаты**

Наблюдения	Группа	Депрессия	Ковариата	Фиктивные переменные			
				Разведенные	Состоящие в браке	Разведенные × Ковариата	Состоящие в браке × Ковариата
1	1	7	7	1	0	7	0
2	1	5	9	1	0	9	0
3	2	2	5	0	1	0	5
4	2	3	6	0	1	0	6
5	2	4	4	0	1	0	4
6	2	3	5	0	1	0	5
7	3	3	4	0	0	0	0
8	3	4	2	0	0	0	0
9	3	5	3	0	0	0	0

ющейся ошибкой, равна разности единицы и квадрата множественной корреляции для ковариаты и двух фиктивных переменных ($1 - 0{,}815 = 0{,}185$). Другими словами, коэффициент ошибки для семейного положения, взятого само по себе ($1 - 0{,}716 = 0{,}284$), сокращается при учете влияния ковариаты ($1 - 0{,}815 = 0{,}185$). Количество предикторов, связанных с изменением квадрата коэффициента множественной корреляции, — это две фиктивные переменные, представляющие три группы. Общее число предикторов, связанных с коэффициентом ошибки, — это ковариата и две фиктивные переменные. Следовательно, число степеней свободы числителя равно двум, а знаменателя — пяти ($9 - 3 - 1 = 5$). Подстановка данных значений в формулу F-отношения дает величину F, равную 9,68, что практически совпадает со значением 9,67, приведенным в табл. 10.2:

$$\frac{0{,}716/2}{0{,}185/5} = \frac{0{,}368}{0{,}037} = 9{,}68.$$

Чтобы вычислить сумму квадратов для групп, умножим изменение квадрата множественной корреляции (0,716) на общую сумму квадратов значений зависимой переменной (18,00) и получим

12,888 (0,716 × 18,00 = 12,888), что с точностью до округления равно 12,89 (см. табл. 10.2). Чтобы вычислить сумму квадратов ошибки, умножим квадрат множественной корреляции, представляющий ошибку (0,185), на общую сумму квадратов зависимой переменной (18,00) и получим 3,33 (0,185 × 18,00 = 3,33), что в точности равно аналогичному значению в табл. 10.2.

F-отношение для проверки однородности регрессии включает долю дисперсии или суммы квадратов, объясняемой взаимодействием между ковариатой и двумя фиктивными переменными, которое представлено последними двумя фиктивными переменными. Эта доля равна разности между квадратом множественной корреляции для ковариаты и четырех фиктивных переменных (0,833) и квадратом множественной корреляции для ковариаты и первых двух фиктивных переменных (0,815), т.е. 0,018 (0,833 – – 0,815 = 0,018). Доля оставшейся дисперсии представляет собой разность между единицей и квадратом множественной корреляции для ковариаты и четырех фиктивных переменных (0,833), что дает 0,167 (1 – 0,833 = 0,167). Число предикторов, связанных с изменением квадрата множественной корреляции, представляет собой две фиктивные переменные, отражающие взаимодействие между ковариатой и тремя группами[1]. Количество предикторов, связанных с остаточной дисперсией, измеряется ковариатой и четырьмя фиктивными переменными[2]. Следовательно, число степеней свободы числителя равно двум, а знаменателя — трем (9 – 5 – – 1 = 3). Подставив эти значения в формулу *F*-отношения, получим, что *F* = 0,161, что приближенно равно значению 0,167, приведенному ранее:

$$\frac{0,018/2}{0,167/3} = \frac{0,009}{0,056} = 0,161.$$

Значение *F*-отношения для ковариаты часто не включается в таблицы ковариационного анализа, но оно приведено в табл. 10.2, чтобы показать, каким образом уменьшается коэффициент ошибки. *F*-отношение определяет, будет ли значимой доля дисперсии, объясняемая ковариатой, после того как учитывается групповой фактор. Эта доля равна разности между квадратом множественной корреляции для ковариаты и первых двух фиктивных переменных (0,815) и квадратом множественной корреляции для первых двух фиктивных переменных (0,667), т.е. 0,148 (0,815 – 0,667 = 0,148). Доля оставшейся дисперсии представляет собой разность единицы и квадрата множественной корреляции для ковариаты и первых двух фиктивных переменных (0,815), что дает 0,185 (1 – 0,815 = 0,185).

---

[1] Т.е. количество добавленных предикторов равно двум.

[2] Т.е. количество предикторов, участвующих в модели, равно пяти.

Число степеней свободы числителя равно единице (ковариата), а знаменателя — пяти (9 – 3 – 1 = 5).[1] Подстановка этих значений в формулу для F-отношения дает значение F, равное 4,00, совпадающее с аналогичной величиной в табл. 10.2:

$$\frac{0,148/1}{0,185/5} = \frac{0,148}{0,037} = 4,00.$$

Обратите внимание на то, что в случае, когда средние значения ковариаты для разных групп в точности совпадают, сумма квадратов для группового фактора (12,00) равна соответствующему значению для дисперсионного анализа, так как средние значения зависимой переменной остаются без изменений. Когда же средние значения ковариаты не совпадают, сумма квадратов для группового фактора может оказаться больше или меньше соответствующего значения для дисперсионного анализа. В нашем случае она несколько больше (12,89). Общая сумма квадратов представляет собой сумму квадратов зависимой переменной. Когда средние значения ковариаты не равны между собой, сумма этих отдельных сумм квадратов уже не будет равна общей сумме квадратов.

Скорректированные средние

Один из способов вычисления скорректированного среднего значения для каждой из групп состоит в использовании следующей формулы:

[Скорректированное групповое среднее] = [Нескорректированное групповое среднее] – [Нестандартизированный коэффициент регрессии для ковариаты × (Среднее ковариаты для группы – Общее среднее ковариаты)].

Нестандартизированный коэффициент регрессии для ковариаты равен –0,67. Среднее значение зависимой переменной для групп и ее общее среднее значение, а также аналогичные величины для ковариаты представлены в табл. 10.1. Подстановка в данную формулу соответствующих значений для разведенных дает скорректированное среднее значение, равное 8,01, что с учетом ошибки округления практически совпадает с приведенным ранее:

$$6,00 - [-0,67 \times (8,00 - 5,00)] = 6,00 - (-2,01) = 8,01.$$

[1] Общее число предикторов равно трем (ковариата плюс две фиктивные переменные).

Один из альтернативных, но менее наглядных способов вычисления скорректированного среднего значения для группы состоит в применении следующей общей формулы:

[Скорректированное групповое среднее] = [(Значение первой фиктивной переменной × Нестандартизированный коэффициент регрессии переменной) + (Нестандартизированный коэффициент регрессии ковариаты × Общее среднее значение ковариаты) + Постоянная].

Нестандартизированные коэффициенты регрессии для двух фиктивных переменных, представляющих фактор, равны 5,33 и 0,33 соответственно. Значение постоянной равно 6,00.[1] Подстановка соответствующих значений для разведенных дает приближенное значение скорректированного среднего значения, равное 7,98, что, с точностью до ошибки округления, совпадает с 8,00:

$$(5{,}33 \times 1) + (0{,}33 \times 0) + (-0{,}67 \times 5{,}00) + 6{,}00 =$$
$$= 5{,}33 + 0 + (-3{,}35) + 6{,}00 = 7{,}98.$$

Из этой формулы становится яснее, почему вычисление скорректированного группового среднего требует лишь знания общего среднего значения ковариаты, а не среднего значения ковариаты для группы. Другими словами, вычисление скорректированных средних значений предполагает, что средние значения ковариаты совпадают.

Абсолютные отклонения для теста Левина

Чтобы получить абсолютные отклонения для теста Левина, необходимо вычислить прогнозируемые значения для всех испытуемых, основываясь на том, к какой группе они принадлежат и каково значение их ковариаты, и вычесть это прогнозируемое значение из фактического значения зависимой переменной. Прогнозируемая выраженность депрессии представляет собой сумму постоянной уравнения регрессии и произведений нестандартизированных коэффициентов регрессии на значения каждого из предикторов в уравнении регрессии. В данном примере уравнение регрессии содержит две фиктивные переменные и ковариату. Нестандартизированные коэффициенты регрессии для этих трех уравнений равны 5,333; 0,333 и –0,667 соответственно, а постоянная — 6,000. Подставив в данное уравнение соответствующие

[1] Значение первой фиктивной переменной для разведенных равно 1, а второй — 0.

Таблица 10.5. **Фактические значения, прогнозируемые значения и абсолютные отклонения для выраженности депрессии**

Наблюдения	Группа	Фактические значения	Прогнозируемые значения	Абсолютные отклонения
1	1	7	6,66	7 – 6,66 = 0,34
2	1	5	5,33	5 – 5,33 = 0,33
3	2	2	3,00	2 – 3,00 = 1,00
4	2	3	2,33	3 – 2,33 = 0,67
5	2	4	3,67	4 – 3,67 = 0,33
6	2	3	3,00	3 – 3,00 = 0,00
7	3	3	3,33	3 – 3,33 = 0,33
8	3	4	4,67	4 – 4,67 = 0,67
9	3	5	4,00	5 – 4,00 = 1,00

значения, находим, что прогнозируемое значение для первого наблюдения приближенно равно 6,66:

$$6{,}000 + (5{,}333 \times 1) + (0{,}333 \times 0) + (-0{,}667 \times 7) =$$
$$= 6{,}000 + 5{,}333 + 0 - 4{,}669 = 6{,}664.$$

Так как выраженность депрессии у данного больного равна 7, абсолютное отклонение для него приближенно равно 0,34. Выраженность депрессии, прогнозируемое значение и абсолютное отклонение для каждого из девяти испытуемых приведены в табл. 10.5. Тест Левина представляет собой однофакторный дисперсионный анализ этих абсолютных отклонений, что дает величину F-отношения, приближенно равную 0,516.

Отчет о результатах

Рассмотрим один из способов краткого изложения результатов данного анализа для ковариаты с неравными средними значениями, так как этот случай, скорее всего, и будет иметь место. Такой отчет может выглядеть следующим образом: «Чтобы учесть влияние плохого самочувствия, которое коррелировало (коэффициент корреляции = 0,31, незначим) с выраженностью депрессии, был проведен однофакторный ковариационный анализ. Семейное положение оказывало статистически значимое влияние на выраженность депрессии ($F_{2,\ 5} = 9{,}67$; $p < 0{,}05$)». Если у нас не было веских оснований для прогнозирования каких-либо различий, то

можно добавить следующее: «Критерий наименьших значимых различий Фишера показал, что средняя выраженность депрессии была значимо выше у разведенных (скорректированное среднее значение $M = 8{,}00$), чем у состоящих в браке (скорректированное среднее значение $M = 3{,}00$, $t_5 = 3{,}29$, значимость по двустороннему критерию $p < 0{,}01$) или никогда не состоявших в браке (скорректированное среднее значение $M = 2{,}67$; $t_5 = 2{,}91$; значимость по двустороннему критерию $p < 0{,}05$)».

Процедура SPSS для Windows

Приведем алгоритм выполнения ковариационного анализа для ковариаты с неравными средними значениями.

Введите данные в **Редакторе данных (Data Editor)**, как показано на рис. 10.1. Поскольку данные в первых двух столбцах те же самые, что и в предыдущей главе, можно открыть соответствующий файл данных и добавить значения в третий столбец с присвоением ему имени «**нездоров**».

В строке меню в верхней части окна программы выберите пункт **Анализ (Analyze)**, в появившемся ниспадающем меню — пункт **Общие линейные модели (General Linear Model)**, а в нем — подпункт **Одномерный анализ (Univariate)**, который открывает диалоговое окно **Univariate**, показанное на рис. 9.2.

Выберите «**депресс**» и нажмите на первую сверху кнопку ▸, чтобы переместить эту переменную в поле **Зависимая переменная (Dependent: (variable))**.

Выберите «**группа**» и нажмите на вторую сверху кнопку ▸, чтобы переместить эту переменную в список **Постоянные факторы (Fixed Factor(s))**.

Выберите «**нездоров**» и нажмите на четвертую сверху кнопку ▸, чтобы переместить эту переменную в список **Ковариаты (Covariate(s))**.

Нажмите на кнопку **Параметры (Options)**, чтобы открыть дочернее диалоговое окно **Одномерный анализ: Параметры (Univariate: Options)**, представленное на рис. 9.4.

	группа	депресс	нездоров
1	1	7	7
2	1	5	9
3	2	2	5
4	2	3	6
5	2	4	4
6	2	3	5
7	3	3	4
8	3	4	2
9	3	5	3

Рис. 10.1 Данные для проведения однофакторного ковариационного анализа в **Редакторе данных (Data Editor)**

Установите флажок **Описательные статистики (Descriptive statistics)**, чтобы вывести средние значения и стандартные отклонения для трех групп.

Установите флажок **Критерии однородности (Homogeneity tests)**, чтобы провести тест Левина для однородности дисперсий.

Выберите «**группа**» в списке **Факторы и взаимодействия факторов (Factor(s) and Factor Interactions)** и нажмите на кнопку ▸, чтобы переместить эту переменную в список **Отобразить средние значения (Display Means for)** и вывести скорректированные средние значения.

Установите флажок **Сравнить главные эффекты (Compare main effects)** и **LSD (none)** в раскрывающемся списке **Коррекция доверительного интервала (Confidence interval adjustment)**, чтобы задать уровень значимости для критерия наименьших значимых различий Фишера (Fisher's protected LSD test).

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно.

Нажмите **OK**, чтобы провести данный анализ.

Чтобы проверить однородность регрессии, введите переменные *в диалоговом окне* ***Одномерный анализ (Univariate)****, как и* выше, а затем нажмите на кнопку **Модель (Model)**, чтобы открыть дочернее диалоговое окно **Одномерный анализ: Модель (Univariate: Model)**, *изображенное* на рис. 10.2.

Установите переключатель **Настройка (Custom)**.

Выберите «**группа(F)**» и затем нажмите на кнопку ▸ под заголовком **Создать условие(я) (Build term(s))**, чтобы переместить эту переменную в список **Модель (Model)**.

Выделите одновременно «**группа(F)**» и «**нездоров(C)**» в раскрывающемся списке под кнопкой ▸, выберите пункт **Взаимодействие**

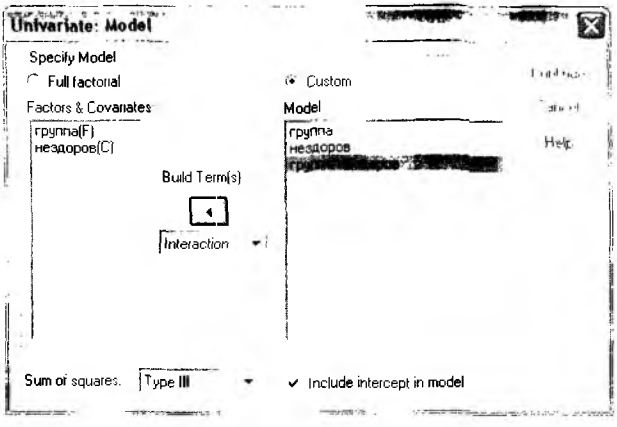


Рис. 10.2. Дочернее окно **Одномерный анализ: Модель (Univariate: Model)**

(**interaction**) и затем нажмите на кнопку ▸, чтобы поместить взаимодействие этих двух переменных в список **Модель (Model)**.

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно.

Нажмите **OK**, чтобы провести данный анализ.

Для проведения регрессионного анализа необходимо ввести значения четырех фиктивных переменных, как показано в табл. 10.4, в столбцы с четвертого по седьмой файла **Редактора данных (Data Editor)**. Выполните регрессию переменной «**депресс**» сначала по ковариате, затем по первым двум фиктивным переменным и, наконец, по последним двум фиктивным переменным.

Выводимая SPSS информация по результатам работы

Не все выводимые SPSS результаты будут представлены и обсуждены. Результаты теста Левина даны в табл. 10.6. Так как величина *F*, равная 0,524, имеет значимость 0,617, то дисперсии значимо не различаются.

В табл. 10.7 содержатся результаты ковариационного анализа. Значения для **Скорректированной модели (Corrected Model)**, **Свободный член (Intercept)** и **Общая сумма (Total)** часто не включаются в подобные таблицы, как и в случае табл. 10.2. Скорректированные средние значения представлены в табл. 10.8.

Уровни значимости для критерия наименьших значимых различий Фишера (Fisher's protected LSD test) приведены в табл. 10.9. Например, уровень значимости для сравнения разведенных и состоящих в браке составляет 0,010.

Критерий однородности регрессии для ковариационного анализа задается значениями для члена **ГРУППА*НЕЗДОРОВ** в таблице ковариационного анализа (табл. 10.10). Величина *F*, равная 0,167, имеет значимость 0,854. Поскольку это значение больше

Таблица 10.6. **Выводимые SPSS результаты теста Левина**

Levene's Test of Equality of Error Variances[a]

Dependent Variable: ДЕПРЕСС

F	df1	df2	Sig.
0,524	2	6	0,617

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design:Intercept + НЕЗДОРОВ + ГРУППА.

Таблица 10.7. **Выводимая SPSS таблица ковариационного анализа**

Tests of Between-Subjects Effects

Dependent Variable: ДЕПРЕСС

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14,667[a]	3	4,889	7,333	0,028
Intercept	12,803	1	12,803	19,204	0,007
НЕЗДОРОВ	2,667	1	2,667	4,000	0,102
ГРУППА	12,889	2	6,444	9,667	0,019
Error	3,333	5	0,667		
Total	162,000	9			
Corrected Total	18,000	8			

a. R Squared = 0,815 (Adjusted R Squared = 0,704).

0,05, *F*-величина является незначимой, что означает, что коэффициент регрессии примерно одинаков для всех групп.

Квадраты множественной корреляции (*R* square) для ковариаты (0,099), ковариаты и первых двух фиктивных переменных (0,815) и для ковариаты и всех четырех фиктивных переменных (0,833) приведены в табл. 10.11.

Постоянная (6,000) и нестандартизированные коэффициенты регрессии для ковариаты (–0,667), первой (5,333) и второй (0,333) фиктивных переменных приведены в табл. 10.12.

Таблица 10.8. **Выводимые SPSS результаты корректировки средних значений**

Estimates

Dependent Variable: ДЕПРЕСС

ГРУППА	Mean	Std. Error	95 % Confidence Interval	
			Lower Bound	Upper Bound
Разведенные	8,000[a]	1,155	5,032	10,968
Состоящие в браке	3,000[a]	0,408	1,951	4,049
Никогда не состоявшие в браке	2,667[a]	0,816	0,568	4,766

a. Covariates appearing in the model are evaluated at the following values: НЕЗДОРОВ = 5,00.

Таблица 10.9. **Выводимые SPSS результаты для критерия наименьших значимых различий Фишера**

Pairwise Comparisons

Dependent Variable: ДЕПРЕСС

(I) ГРУППА	(J) ГРУППА	Mean Difference (I–J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Разведенные	Состоящие в браке	5,000*	1,225	0,010	1,852	8,148
	Никогда не состоявшие в браке	5,333*	1,826	0,033	0,640	10,027
Состоящие в браке	Разведенные	–5,000*	1,225	0,010	–8,148	–1,852
	Никогда не состоявшие в браке	0,333	0,913	0,730	–2,013	2,680
Никогда не состоявшие в браке	Разведенные	–5,333*	1,826	0,033	–10,027	–0,640
	Состоящие в браке	–0,333	0,913	0,730	–2,680	2,013

Based on estimated means

* The mean difference is significant at the 0,05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Таблица 10.10. **Выводимые SPSS результаты для критерия однородности регрессии**

Tests of Between-Subjects Effects

Dependent Variable: ДЕПРЕСС

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15,000[a]	5	3,000	3,000	0,197
Intercept	12,479	1	12,479	12,479	0,039
ГРУППА	2,007	2	1,003	1,003	0,464
НЕЗДОРОВ	2,667	1	2,667	2,667	0,201
ГРУППА*НЕЗДОРОВ	0,333	2	0,167	0,167	0,854
Error	3,000	3	1,000		
Total	162,000	9			
Corrected Total	18,000	8			

a. R Squared = 0,833 (Adjusted R Squared = 0,566).

Таблица 10.11. **Выводимые SPSS квадраты множественной корреляции**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0,314[a]	0,099	−0,030	1,522
2	0,903[b]	0,815	0,704	0,816
3	0,913[c]	0,833	0,556	1,000

a. Predictors: (Constant), НЕЗДОРОВ.

b. Predictors: (Constant), НЕЗДОРОВ, ФИКТИВ 2, ФИКТИВ 1.

c. Predictors: (Constant), НЕЗДОРОВ, ФИКТИВ 2, ФИКТИВ 1, ФИКТИВ 4, ФИКТИВ 3.

Таблица 10.12. **Выводимые SPSS нестандартизированные коэффициенты регрессии для ковариаты и первых двух фиктивных переменных**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,889	1,366		2,114	0,072
	НЕЗДОРОВ	0,222	0,254	0,314	0,876	0,410
2	(Constant)	6,000	1,106		5,427	0,003
	НЕЗДОРОВ	−0,667	0,333	−0,943	−2,000	0,102
	ФИКТИВ 1	5,333	1,826	1,568	2,921	0,033
	ФИКТИВ 2	0,333	0,913	0,117	0,365	0,730

Рекомендуемая литература

Huitema, B.E. (1980) *The Analysis of Covariance and Alternatives.* New York: Wiley.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction,* 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

Глава 11

НЕСВЯЗНЫЙ ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Проведение дисперсионного анализа более чем с одним несвязанным фактором имеет два главных преимущества, первое из которых состоит в том, что метод позволяет исследовать взаимодействие между факторами. Взаимодействие имеет место, когда средний балл по крайней мере одной группы по одному фактору изменяется в соответствии с распределением по двум или более группам по одному или нескольким другим факторам. Например, средний балл выраженности депрессии у женщин по сравнению с мужчинами может значимо различаться в зависимости от того, состоят ли они в браке или нет. Средняя выраженность депрессии у никогда не женатых мужчин может быть значимо выше, чем у женщин, никогда не бывших замужем, но не различаться значимо у женатых мужчин и замужних женщин. В этом случае у нас было бы двустороннее взаимодействие между фактором пола и фактором семейного положения. Если же значимое взаимодействие отсутствует, то любой значимый эффект факторов, составляющих данное взаимодействие, может быть объяснен без учета других факторов, входящих в это взаимодействие. Например, если значимое взаимодействие между семейным положением и полом отсутствует, но имеется значимый эффект, связанный с семейным положением, то можно игнорировать любые влияния, связанные с полом, когда говорим о различиях в выраженности депрессии, связанных с семейным положением. Однако при наличии значимого взаимодействия значимый эффект входящего во взаимодействие фактора должен быть объяснен с учетом других составляющих взаимодействия факторов. Например, если существует значимое взаимодействие между полом и семейным положением, а также значимый эффект, связанный с семейным положением, то мы должны обсуждать различия в выраженности депрессии в связи с обоими факторами — семейным положением и полом, а не с отдельно взятым семейным положением.

Вторым преимуществом дисперсионного анализа с несколькими несвязанными факторами является то, что он обеспечивает более чувствительную проверку влияния фактора в том смысле, что с большей вероятностью окажется статистически значимым, если дисперсия ошибки уменьшается из-за того, что часть ее теперь объясняется другими факторами и их взаимодействиями. На-

пример, фактор семейного положения может стать статистически значимым, если дисперсия, объясняемая полом и/или его взаимодействием с семейным положением, существенно уменьшает коэффициент ошибки.

Дисперсионный анализ с несколькими несвязанными факторами иногда называют факторным дисперсионным анализом. Эффекты, вызванные факторами, могут называться главными эффектами, в противоположность эффектам взаимодействия. Рассмотрим вычисления и интерпретацию факторного дисперсионного анализа на примере простейшего двухфакторного дисперсионного анализа, где оба фактора имеют всего два уровня, или две группы значений. Первый фактор (семейное положение) содержит две группы: состоящие в браке и никогда не состоявшие в браке. Второй фактор (пол) включает две группы — мужчины и женщины. Зависимая переменная представляет собой все ту же измеренную по девятибалльной шкале выраженность депрессии, которая использовалась в примерах двух предыдущих глав. Индивидуальные, или сырые, баллы выраженности депрессии вместе со средними значениями для этих четырех групп, а также средними значениями по факторам и средним значением для всей выборки приведены в табл. 11.1. Число наблюдений в группах различно. Среди них две женщины, никогда не бывшие замужем, шесть никогда не женатых мужчин, четыре замужние женщины и трое женатых мужчин.

Если посмотреть на среднюю выраженность депрессии для респондентов, имеющих тот или иной семейный статус, то она выше для никогда не состоявших в браке (6,00), чем для состоящих в браке (3,14). Если сделать то же самое в отношении пола, то увидим, что она выше для мужчин (5,33), чем для женщин (3,67). При совместном рассмотрении семейного положения и пола выраженность депрессии оказывается выше у замужних женщин (4,00), чем у женщин, никогда не бывших замужем (3,00), но у женатых мужчин (2,00) она оказывается ниже, чем у никогда не женатых мужчин (7,00). Другими словами, оказывается, что между семейным положением и полом имеется взаимодействие, проявляющееся в том, что средняя выраженность депрессии для семейного положения зависит от пола испытуемого. Представим взаимосвязь между двумя или более факторами графически (рис. 11.1). По вертикальной оси, или оси ординат, отложим значения зависимой переменной, в данном случае — средней выраженности депрессии (бо́льшие цифры соответствуют более высоким средним значениям выраженности депрессии). На горизонтальной оси, или оси абсцисс, приведен фактор семейного положения, где левая отметка соответствует никогда не состоявшим в браке, а правая — состоящим в браке. Другой фактор (фактор пола) показан двумя прямыми. При этом не имеет значения, какой из факторов приведен на горизонтальной оси, а какой — в виде прямых. От-

Таблица 11.1. **Индивидуальные баллы и средние значения выраженности депрессии для двухфакторного дисперсионного анализа с несвязанными факторами**

	Женщины	Мужчины	Суммарные показатели по рядам
Никогда не состоявшие в браке	2 4	8 6 8 6 7 7	
Сумма	6	42	48
n	2	6	8
Среднее	3,00	7,00	6,00
Состоящие в браке	3 5 4 4	1 3 2	
Сумма	16	6	22
n	4	3	7
Среднее	4,00	2,00	3,14
Общие показатели по столбцам			
Сумма	22	48	70
n	6	9	15
Среднее	3,67	5,33	4,67

сутствие взаимодействия имеет место, когда прямые, представляющие другой фактор, практически параллельны. Например, если бы пунктирная прямая (Мужчины) была параллельна сплошной прямой (Женщины), то взаимодействие между семейным положением и полом отсутствовало бы. Взаимодействие существует, когда две линии не параллельны, как это имеет место в нашем случае (см. рис. 11.1). Является ли это взаимодействие статистически значимым, зависит от того, будет ли значимой величина F для данного взаимодействия в проводимом дисперсионном анализе.

Двухфакторный дисперсионный анализ выявляет два главных эффекта и эффект взаимодействия, как показано в левом столбце табл. 11.2. Величина F-отношения для каждого эффекта представ-

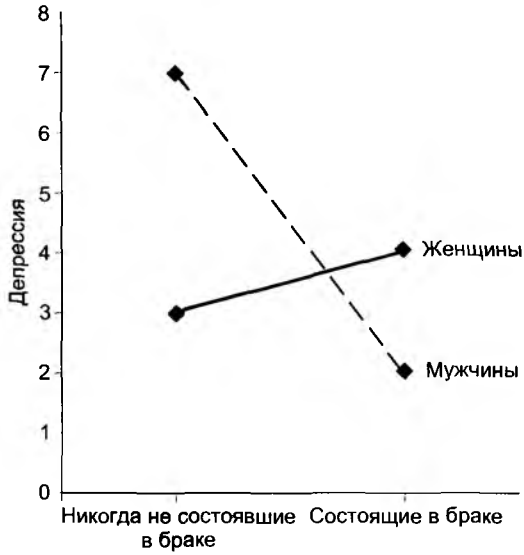


Рис. 11.1. График, показывающий взаимодействие между семейным положением и полом

ляет собой среднюю сумму квадратов для данного эффекта, деленную на средний квадрат для дисперсии, не объясняемой ни одним из этих трех эффектов. Эта последняя среднеквадратичная величина известна под разными названиями: внутригрупповая среднеквадратичная величина, ошибка или остаточная среднеквадратичная величина. Например, величина *F* для взаимодействия есть отношение среднеквадратичной величины взаимодействия (28,80) к среднеквадратичной величине ошибки (0,91), что равно 31,65 (28,80/0,91 = 31,65). Число степеней свободы *df*-эффекта взаимодействия равно произведению числа групп первого фактора минус единица и числа групп второго фактора также минус единица. Поскольку каждый фактор содержит только две группы, число степеней свободы взаимодействия равно единице $[(2 - 1) \times (2 - 1) = 1]$. Число степеней свободы ошибки равно количеству наблюдений минус общее количество групп (произведение количества групп первого и второго факторов). Поскольку в каждом факторе имеется по две группы, общее число групп, или клеток, равно четырем ($2 \times 2 = 4$). Так как количество наблюдений равно пятнадцати, число степеней свободы ошибки равно одиннадцати ($15 - 4 = 11$). Чтобы быть статистически значимым на уровне 0,05, *F*-отношение с одной степенью свободы в числителе и одиннадцатью степенями свободы в знаменателе должно быть не менее 4,84, что в нашем случае выполняется. Следовательно, взаимодействие между семейным положением и полом является статистически значимым.

Неравные и непропорциональные частоты клеток (ячеек)

Когда количество наблюдений в каждой из клеток таблицы факторного дисперсионного анализа различно и не пропорционально, как в нашем случае, три[1] эффекта могут не быть несвязанными или независимыми друг от друга. В этом случае существует три основных метода вычисления эффектов (J. E. Overall and D. K. Spiegel, 1969). Результаты вычислений для взаимодействия (и ошибки) будут одни и те же для всех трех методов, но результаты для главных эффектов могут различаться (см. табл. 11.2). Например, хотя величина F для семейного положения статистически значима для всех трех методов, она равна 14,07 — в случае применения первого из рассматриваемых методов; 24,79 — в случае применения второго метода и 33,49 — в случае применения третьего метода. Когда количество наблюдений одно и то же для каждой ячейки таблицы факторного дисперсионного анализа, три эффекта не связаны и все три метода дают одни и те же результаты для главных эффектов.

Таблица 11.2. **Основные результаты для трех типов факторного дисперсионного анализа с неравными числами в ячейках таблицы**

Источник изменчивости	*df*	Тип I		
		SS	*MS*	*F*
Семейный статус (M)	1	12,80	12,80	14,07*
Пол (S)	1	3,20	3,20	3,52
M × S	1	28,80	28,80	31,65*
Ошибка	11	10,00	0,91	
Сумма эффектов	14	54,80		

Продолжение табл. 11.2

Источник изменчивости	*df*	Тип II		
		SS	*MS*	*F*
Семейный статус (M)	1	22,53	22,53	24,79*
Пол (S)	1	2,06	2,06	2,26
M × S	1	28,80	28,80	31,65*
Ошибка	11	10,00	0,91	
Сумма эффектов	14	63,36		

[1] Имеются в виду два фактора и их взаимодействие между собой.

Продолжение табл. 11.2

Источник изменчивости	*df*	Тип III		
		SS	*MS*	*F*
Семейный статус (M)	1	30,48	30,48	33,49*
Пол (S)	1	2,06	2,06	2,26
M×S	1	28,80	28,80	31,65*
Ошибка	11	10,00	0,91	
Сумма эффектов	14	71,34		

* $p < 0{,}05$.

Все три метода используют множественную регрессию для вычисления суммы квадратов для каждого из эффектов, но способы, которыми это делается, различаются между собой. Первый из представляемых методов в SPSS называется Тип I (Type I) и является методом, используемым по умолчанию (другие названия: Метод 1, регрессия, метод невзвешенных средних, подход с использованием специфичностей). Каждый эффект корректируется с учетом всех остальных эффектов (включая любые ковариаты). Объясняемая каждым эффектом дисперсия специфична для данного эффекта и не объясняется, или не связана ни с каким другим эффектом. Например, дисперсия, связанная с семейным положением, — дисперсия, которая остается после того, как учтен эффект пола и взаимодействие. Это может быть выражено как разность между дисперсией, связанной со всеми тремя эффектами, и дисперсией, связанной с полом и взаимодействием, как показано в табл. 11.3. Дисперсия, связанная с полом, — остаток дисперсии после учета эффекта семейного положения и взаимодействия. Другими словами, это разность между дисперсией, связанной со всеми тремя эффектами, и дисперсией семейного положения и взаимодействия. B. G. Tabachnik и L. S. Fidell (1996) рекомендуют этот подход для экспериментальных планов, предполагающих, что каждая ячейка таблицы одинаково важна. Однако можно привести аргументы в пользу того, что этот подход также возможен и для неэкспериментальных планов, где исследователя интересует специфический эффект каждой переменной.

Второй метод, приведенный в табл. 11.2, называется Тип II (Type II) в SPSS (другие названия: Метод 2, классический экспериментальный подход, метод наименьших квадратов). Главные эффекты корректируются с учетом других главных эффектов (и ковариат), в то время как взаимодействия корректируются с учетом всех остальных эффектов, за исключением взаимодействий

более высокого порядка. Например, дисперсия, связанная с семейным положением, — это не та дисперсия, которая уже объяснена полом, это доля оставшейся дисперсии, объясняемой семейным положением совместно со взаимодействием. Другими словами, это разность между дисперсией, связанной с семейным положением и полом, и дисперсией, связанной с полом (см. табл. 11.3). B.G.Tabachnik и L.S.Fidell (1996) рекомендуют этот подход для неэкспериментальных планов, где больший вес приписывается главным эффектам.

Третий, и последний, из представленных в табл. 11.2 методов в SPSS называется Тип III (Type III) (другие названия: Метод 3, иерархический, или последовательный, подход). Эффекты упорядочены в определенной последовательности согласно замыслу исследователя. Например, исследователь может быть прежде всего заинтересован в изучении влияния семейного положения, затем пола и, наконец, взаимодействия между семейным положением и полом. В этом случае дисперсия, объясняемая семейным положением, может включать долю, которая объясняется полом и взаимодействием между семейным положением и полом. Дисперсия, объясняемая влиянием следующего за семейным положением пола, будет исключать долю, уже объясненную семейным положением. Из табл. 11.3 видно, что влияние пола вычисляется одинаково для методов Типа I и Типа II, и именно поэтому соответствующие результаты для этих двух методов, представленные в табл. 11.2, совпадают.

Когда количество наблюдений в каждой ячейке не равно, а пропорционально, результаты методов Типа II и Типа I совпадают, поскольку главные эффекты не связаны друг с другом. Экспериментальный план с пропорциональными частотами представлен в табл. 11.4. Отношение частот строк (2:4 и 3:6) равно 1:2. Отношение частот столбцов (2:3 и 4:6) равно 2:3. В общем случае частоты ячеек пропорциональны, когда частота ячейки равна произведению частот строки и столбца, в которых находится данная ячейка, деленному на общую частоту. Это имеет место для всех

Таблица 11.3. **Уравнения регрессии для трех методов дисперсионного анализа**

Эффекты	Тип I	Тип II	Тип III
Семейный статус (M)	M, S, M×S – S, M×S	M, S – S	M
Пол (S)	M, S, M×S – M, M×S	M, S – M	M, S – M
M×S	M, S, M×S – M, S	M, S, M×S – M, S	M, S, M×S – M, S

Таблица 11.4. **Экспериментальный план 2 × 2 с пропорциональными частотами ячеек**

	Женщины	Мужчины	Суммарные показатели по строкам
Никогда не состоявшие в браке	2	3	5
Состоящие в браке	4	6	10
Суммарные показатели по столбцам	6	9	15

четырех ячеек табл. 11.4. Например, частота ячейки, находящейся в первой строке и первом столбце, равна двум, что совпадает с $5 \times 6/15$ ($30/15 = 2$).

Использование фиктивных переменных для кодирования эффектов

Чтобы продемонстрировать вычисления в факторном дисперсионном анализе и тот факт, что эффекты могут коррелировать, когда частоты ячеек не равны и не пропорциональны, необходимо представить главные эффекты и взаимодействие фиктивными переменными, приписывая им значения (кодируя) так, как показано в табл. 11.5. Значения фиктивных переменных приписываются следующим образом: наблюдения, попадающие в какую-либо одну группу, получают значение по этой переменной, равное 1, в другую — (–1), а попадающие во все остальные группы по этому фактору получают значение 0. Чтобы представить фактор, необходимо на одну фиктивную переменную меньше, чем число групп в этом факторе. Поскольку в нашем случае каждый фактор имеет всего две группы, то для представления каждого фактора необходимо только по одной фиктивной переменной. Одна группа кодируется единицей, а другая — минус единицей, при этом выбор абсолютно произволен. Значение 1 по каждому фактору приписывалось испытуемым, попавшим в первую группу этого фактора, и (–1) — испытуемым, попавшим во вторую группу. Так как в каждом факторе всего по две группы, использовать нулевой код не было необходимости. Фиктивная переменная, представляющая взаимодействие, получается путем умножения фиктивных переменных, представляющих оба фактора.

Вычисление коэффициентов корреляции трех фиктивных переменных, представленных в табл. 11.5, между собой дает следующие результаты: коэффициенты корреляции первой фиктивной

Таблица 11.5. **Кодирование эффектов для дисперсионного анализа 2×2**

Семейный статус	Пол	Фиктивные переменные		
		Брак	Пол	Брак × Пол
Никогда не состоявшие в браке	Женщины	1	1	1
То же	То же	1	1	1
Никогда не состоявшие в браке	Мужчины	1	−1	−1
То же	То же	1	−1	−1
»	»	1	−1	−1
»	»	1	−1	−1
»	»	1	−1	−1
»	»	1	−1	−1
Состоящие в браке	Женщины	−1	1	−1
То же	То же	−1	1	−1
»	»	−1	1	−1
»	»	−1	1	−1
Состоящие в браке	Мужчины	−1	−1	1
То же	То же	−1	−1	1
»	»	−1	−1	1

переменной со второй и третьей соответственно равны −0,33 и −0,19; между второй и третьей фиктивными переменными — 0,00. Другими словами, семейное положение коррелирует как с полом, так и со взаимодействием, в то время как пол не коррелирует со взаимодействием. Если мы создадим три новые фиктивные переменные, чтобы представить экспериментальный план с пропорциональными частотами ячеек, приведенный в табл. 11.4, и вычислим коэффициенты корреляции между этими фиктивными переменными, то обнаружим, что коэффициент корреляции между главным эффектом, связанным с семейным положением, и главным эффектом, связанным с полом, равен нулю. Иными словами, эти два главных эффекта не коррелируют. Коэффициенты корреляции между третьей фиктивной переменной и первой и

второй фиктивными переменными равны –0,19 и –0,33 соответственно. Наконец, если мы создадим три новые фиктивные переменные, чтобы отразить экспериментальный план 2×2 с четырьмя наблюдениями в каждой ячейке, то обнаружим, что ни один из эффектов не коррелирует друг с другом.

Множественная регрессия

Рассмотрим вычисление F-отношения и суммы квадратов для метода Типа III, где значения факторов фиксированы и не случайны, как это обычно бывает, и что имеет место в нашем случае. Случайный фактор — это фактор, значения которого выбираются случайным образом из некоторой (генеральной) совокупности значений. Например, если бы нас интересовало влияние уровня фонового белого шума на выполнение какой-то деятельности и мы были ограничены тремя уровнями громкости, не превышающими 90 децибел, то мы могли бы случайным образом выбрать эти три уровня из девяноста возможных целых значений.

Как отмечалось в двух предыдущих главах, критерий Фишера F может быть выведен из следующего равенства:

$$F = \frac{\dfrac{[\text{Изменение } R^2]}{[\text{Число добавленных предикторов}]}}{\dfrac{[1 - R^2]}{[N - \text{число предикторов} - 1]}}.$$

Квадрат множественной корреляции (R^2) представляет собой объясненную долю общей дисперсии, или суммы квадратов, зависимой переменной. Если общая сумма квадратов известна, то можно вычислить сумму квадратов для различных членов таблицы дисперсионного анализа (см. табл. 11.2). Общая сумма квадратов представляет собой сумму квадратов отклонений каждого балла от общего, или выборочного, среднего значения. Выборочное среднее значение в нашем примере равно 4,67 (см. табл. 11.1). Общая сумма квадратов равна 71,33.

Можно вычислить величину F-отношения для семейного положения, если известен квадрат множественной корреляции для пола и взаимодействия, а также для всех трех эффектов. Квадрат множественной корреляции для пола и взаимодействия равен 0,680, а для всех трех эффектов — 0,860. Следовательно, изменение квадрата множественной корреляции для одного предиктора — семейного положения — составляет 0,180 (0,860 – 0,680 = 0,180). Обратите внимание на то, что это изменение дает оценку величины эффекта, которая для семейного положения равна 0,18. Подставив соответствующие значения в формулу для F-отношения, получим:

$F = 14{,}17$, что при условии ошибки округления приближенно равно значению 14,07 (см. табл. 11.2):

$$\frac{0{,}180/1}{(1-0{,}860)/(15-3-1)} = \frac{0{,}180}{0{,}140/11} = \frac{0{,}180}{0{,}0127} = 14{,}17.$$

Сумма квадратов для семейного положения равна квадрату множественной корреляции для семейного положения (0,180), умноженному на общую сумму квадратов (71,33), что приближенно равно 12,84 ($0{,}180 \times 71{,}33 = 12{,}84$) — значению, близкому к величине 12,80 в табл. 11.2.

Чтобы вычислить F-отношение для пола, необходимо знать квадрат множественной корреляции для семейного положения и взаимодействия, который равен 0,815. Следовательно, изменение квадрата коэффициента множественной корреляции для одного предиктора — пола — составляет 0,045 ($0{,}860 - 0{,}815 = 0{,}045$). Подставив соответствующие значения в формулу для F-отношения, получим: $F = 3{,}54$, что приближенно равно значению 3,52, приведенному в табл. 11.2:

$$\frac{0{,}045/1}{(1-0{,}860)/(15-3-1)} = \frac{0{,}045}{0{,}140/11} = \frac{0{,}045}{0{,}0127} = 3{,}54.$$

Сумма квадратов для пола равна квадрату его множественной корреляции (0,045), умноженному на общую сумму квадратов (71,33), что приближенно равно 3,21 ($0{,}045 \times 71{,}33 = 3{,}21$) — значению, близкому к величине 3,20 (см. табл. 11.2).

Чтобы вычислить F-отношение для взаимодействия, необходимо знать квадрат множественной корреляции для семейного положения и пола, который равен 0,456. Изменение квадрата множественной корреляции для одного предиктора — взаимодействия — составляет 0,404 ($0{,}860 - 0{,}456 = 0{,}404$). Подставив соответствующие значения в формулу для F-отношения, получим: $F \approx 31{,}81$, что с учетом ошибки округления примерно совпадает с величиной 31,65 (см. табл. 11.2):

$$\frac{0{,}404/1}{(1-0{,}860)/(15-3-1)} = \frac{0{,}404}{0{,}140/11} = \frac{0{,}404}{0{,}0127} = 31{,}81.$$

Сумма квадратов для взаимодействия равна квадрату его множественной корреляции (0,404), умноженному на общую сумму квадратов (71,33), что приближенно равно 28,82 ($0{,}404 \times 71{,}33 = 28{,}82$), т. е. значению, очень близкому к значению 28,80 (см. табл. 11.2).

Наконец, сумма квадратов (дисперсия) для ошибки есть доля дисперсии, не объясняемая тремя эффектами ($1 - 0{,}860 = 0{,}140$), умноженная на общую сумму квадратов (дисперсию) (71,33), что

приближенно равно 9,99 (0,140 × 71,33 = 9,99), т. е. значению, очень близкому к значению 10,00, приведенному в табл. 11.2.

Уменьшение ошибки

Чтобы показать, что двухфакторный дисперсионный анализ может обеспечить более точную проверку фактора, чем однофакторный дисперсионный анализ, в табл. 11.6 приведены результаты однофакторного дисперсионного анализа для семейного положения вместе с аналогичными результатами двухфакторного дисперсионного анализа. Дисперсия ошибки для двухфакторного дисперсионного анализа (10,00) значительно меньше, чем для однофакторного (40,86), поскольку часть этой дисперсии объясняется полом (3,20) и взаимодействием (28,80). Следовательно, *F*-отношение для семейного положения больше в случае двухфакторного дисперсионного анализа (14,07), чем в случае однофакторного дисперсионного анализа (9,71). Чтобы быть статистически значимой на уровне 0,05, для двухфакторного дисперсионного анализа *F*-отношение с одной степенью свободы в числителе и одиннадцатью степенями свободы в знаменателе должно быть не меньше 4,84. Для однофакторного дисперсионного анализа *F*-отношение с одной степенью свободы в числителе и тринадцатью — в знаменателе должно быть не меньше 4,67, чтобы быть статистически значимой на уровне 0,05. Несмотря на то что критическое значение *F*-отношения для однофакторного дисперсионного анализа ниже, чем для двухфакторного, разница этих критических значений относительно мала (4,84 – 4,67 = 0,17) и гораздо меньше разницы между значениями *F*-отношения для семейного положения (14,07 – 9,71 = 4,36). *Следовательно, влияние семейного положе-*

Таблица 11.6. **Сравнение результатов однофакторного и двухфакторного дисперсионного анализа для определения влияния семейного положения**

Источник изменчивости	Однофакторный дисперсионный анализ			
	SS	*df*	*MS*	*F*
Семейный статус	30,48	1	30,48	9,71*
(Пол)				
(Взаимодействие)				
Ошибка	40,86	13	3,14	
Сумма	71,33	14		

* $p < 0,05$.

Окончание табл. 11.6

Источник изменчивости	Двухфакторный дисперсионный анализ			
	SS	*df*	*MS*	*F*
Семейный статус	12,80	1	12,80	14,07*
(Пол)	3,20	1	3,20	3,52
(Взаимодействие)	28,80	1	28,80	31,65*
Ошибка	10,00	11	0,91	
Сумма	71,33	14		

ния более значимо для двухфакторного дисперсионного анализа ($p = 0{,}003$), чем для однофакторного ($p = 0{,}008$).

Однородность дисперсии

Подобно однофакторному дисперсионному анализу, анализ нескольких факторов дисперсии основан на допущении о том, что генеральные совокупности, из которых извлекаются выборки, имеют нормальное распределение и равные, или однородные, дисперсии. Со способами проверки того, значимо ли асимметрия и эксцесс распределения баллов отличаются от нуля с точки зрения статистики, можно ознакомиться в работе (D. Cramer, 1998)[1]. Одним из способов проверки однородности дисперсии является тест Левина, который представляет собой однофакторный дисперсионный анализ абсолютных отклонений каждого балла от среднего значения для данной группы. Вычисления для этого критерия приведены в табл. 11.7. Величина *F*-отношения для теста Левина равна 0,407. Чтобы быть статистически значимым, значение *F*-отношения с тремя степенями свободы в числителе и одиннадцатью — в знаменателе, должно быть больше или равно 5,59, чего в настоящем случае нет. Если бы величина *F*-отношения оказалась значимой, то следовало бы применить некоторое преобразование исходных баллов, например извлечение корня квадратного из их значений, чтобы сделать дисперсии более однородными.

Сравнение групп

В нашем примере мы имеем значимый эффект для семейного положения и взаимодействия между семейным положением и полом. Так как фактор семейного положения состоит только из двух групп, значимый эффект означает, что выраженность деп-

[1] См. также А.Д. Наследов, 2004.

рессии для никогда не состоявших в браке (6,00) значимо выше, чем для состоящих в браке (3,14). Уровень значимости для *F*-отношения выбран по двустороннему критерию. При отсутствии веских оснований для прогнозирования различий между двумя группами следует использовать двусторонний критерий для оценки уровня значимости этого эффекта, в противном случае, т.е. если есть серьезные основания предполагать, что выраженность депрессии у никогда не состоявших в браке выше, чем у состоящих в браке, можно применить односторонний критерий, уменьшив вдвое значение критической вероятности для двустороннего критерия. При наличии трех или более групп и значимом эффекте необходимо сравнивать попарно средние группы, чтобы выявить, какие групповые средние различаются между собой. Если бы у нас были серьезные основания для конкретных прогнозов, то мы могли бы использовать критерий Стьюдента для независимых групп, чтобы сравнивать пары средних, в противном случае можно использовать какой-нибудь апостериорный критерий, такой как критерий Шеффе, для определения того, какие средние различны.

Эффект взаимодействия учитывает четыре группы. Значимый эффект взаимодействия не позволяет определить, какие средние значения для этих групп отличаются друг от друга. Имея четыре группы, можно провести шесть попарных сравнений, которые показаны в первом столбце табл. 11.8. Если бы у нас были достаточные основания для конкретных предположений о том, какие группы различаются, то мы могли бы использовать односторонний критерий Стьюдента для независимых групп. Результаты для этого критерия приведены в табл. 11.8. Единственные средние значения, которые не отличаются достоверно друг от друга, — это среднее для никогда не бывших замужем женщин (3,00) по сравнению со средними замужних женщин (4,00) и женатых мужчин (2,00).

Если бы у нас не было веских оснований для того, чтобы делать конкретные прогнозы о том, какие группы различаются, то мы могли бы использовать один из апостериорных критериев, таких как критерий Шеффе. Критерий Шеффе представляет собой критерий Фишера, равный отношению квадрата разности между средними значениями для сравниваемых групп к среднеквадратичной ошибке для всех групп, взвешенной с учетом числа наблюдений в сравниваемых группах (n_1 и n_2):

$$F = \frac{[\text{Среднее группы 1} - \text{среднее группы 2}]^2}{[\text{Среднеквадратичная ошибка}] \times [(n_1 + n_2)/(n_1 + n_2)]}.$$

Значимость этого критерия Фишера оценивается по соответствующим критическим значениям *F*, которые умножаются на весовой коэффициент, равный числу степеней свободы взаимодействия. Это число степеней свободы равно количеству групп минус единица, что в нашем случае составляет три (4 – 1 = 3).

Таблица 11.7. **Результаты теста Левина для двухфакторного дисперсионного анализа**

Наблюдения	Группа	Депрессия	Абсолютные отклонения	Квадрат межгруппового отклонения	Квадрат внутригруппового отклонения
1	1,1	2	3 - 2 = 1	$(1,00 - 0,67)^2 = 0,11$	$(1,00 - 1,00)^2 = 0,00$
2	1,1	4	3 - 4 = 1	$(1,00 - 0,67)^2 = 0,11$	$(1,00 - 1,00)^2 = 0,00$
Групповое среднее		6/2 = 3,00	2/2 = 1,00		
3	1,2	8	7 - 8 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
4	1,2	6	7 - 6 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
5	1,2	8	7 - 8 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
6	1,2	6	7 - 6 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
7	1,2	7	7 - 7 = 0	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 0,00)^2 = 0,45$
8	1,2	7	7 - 7 = 0	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 0,00)^2 = 0,45$
Групповое среднее		42/6 = 7,00	4/6 = 0,67		
9	2,1	3	4 - 3 = 1	$(0,50 - 0,67)^2 = 0,03$	$(0,50 - 1,00)^2 = 0,25$
10	2,1	5	4 - 5 = 1	$(0,50 - 0,67)^2 = 0,03$	$(0,50 - 1,00)^2 = 0,25$
11	2,1	4	4 - 4 = 0	$(0,50 - 0,67)^2 = 0,03$	$(0,50 - 0,00)^2 = 0,25$
12	2,1	4	4 - 4 = 0	$(0,50 - 0,67)^2 = 0,03$	$(0,50 - 0,00)^2 = 0,25$
Групповое среднее		16/4 = 4,00	2/4 = 0,50		

13	2,2	1	2 - 1 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
14	2,2	3	2 - 3 = 1	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 1,00)^2 = 0,11$
15	2,2	2	2 - 2 = 0	$(0,67 - 0,67)^2 = 0,00$	$(0,67 - 0,00)^2 = 0,11$
Групповое среднее		6/3 = 2,00	2/3 = 0,67		
Общее среднее			10/15 = 0,67		
Сумма квадратов				0,34	3,01
Число степеней свободы				4 - 1 = 3	15 - 4 = 11
Среднее квадратичное отклонение				0,34/3 = 0,11	3,01/11 = 0,27
F					0,11/0,27 = 0,407

Критическое значение F-отношения с тремя степенями свободы в числителе и одиннадцатью — в знаменателе для уровня значимости 0,05 составляет около 3,59. Следовательно, соответствующее критическое значение F для уровня 0,05 для данного критерия Шеффе составляет 10,77 ($3,59 \times 3 = 10,77$).

Величина F для критерия Шеффе для сравнения средней выраженности депрессии у никогда не бывших замужем женщин и неженатых мужчин равна 26,23:

$$F = \frac{(3,00 - 7,00)^2}{0,91 \times (2 + 6)/(2 \times 6)} = \frac{(-4,00)^2}{0,91 \times 0,67} = \frac{16,00}{0,61} = 26,23.$$

Таблица 11.8. **Сравнение четырех средних значений с помощью критериев Стьюдента и Шеффе для несвязанных групп**

Сравнения	t[1]	df[2]	p[3]	*Scheffé*[4]	df	p
Никогда не состоявшие в браке женщины ↔ Никогда не состоявшие в браке мужчины	4,90	6	0,01	26,23	3,11	0,01
Никогда не состоявшие в браке женщины ↔ Состоящие в браке женщины	1,16	4	Незначима	1,47	3,11	Незначима
Никогда не состоявшие в браке женщины ↔ Состоящие в браке мужчины	0,95	3	Незначима	1,32	3,11	Незначима
Никогда не состоявшие в браке мужчины ↔ Состоящие в браке женщины	5,37	8	0,001	23,68	3,11	0,01
Никогда не состоявшие в браке мужчины ↔ Состоящие в браке мужчины	7,64	7	0,0001	54,35	3,11	0,001
Состоящие в браке женщины ↔ Состоящие в браке мужчины	2,93	5	0,05	7,55	3,11	Незначима

[1] Критерий Стьюдента.
[2] Число степеней свободы.
[3] Значимость.
[4] Критерий Шеффе.

Так как 26,23 > 10,77, средние значения для этих двух групп различаются значимо. Величины F для критерия Шеффе, число степеней свободы и значения p для этих шести сравнений приведены в табл. 11.8. Кроме тех двух попарных сравнениях, где не удалось выявить значимых различий с помощью критерия Стьюдента для независимых групп, средняя выраженность депрессии для замужних женщин (4,00) не отличалась значимо от среднего для женатых мужчин (2,00).

Отчет о результатах

Один из кратких способов представления результатов данного анализа состоит в следующем: «Был проведен несвязный дисперсионный анализ 2 × 2, использующий регрессионный подход. Было обнаружено значимое влияние семейного положения ($F_{1,11} = 14,07$; $p < 0,05$), а также взаимодействия между семейным положением и полом ($F_{1,11} = 31,65$; $p < 0,001$). Выраженность депрессии была значимо выше у никогда не состоявших в браке ($M = 6,00$; СКО = = 2,07), чем у состоящих в браке ($M = 3,15$; СКО = 1,35)». Если у нас не было веских оснований для прогноза эффекта взаимодействия, мы могли бы добавить следующее: «Критерий Шеффе показал, что выраженность депрессии у никогда не женатых мужчин ($M = 7,00$; СКО = 0,89) значимо выше, чем у женатых мужчин ($M = 2,00$; СКО = 1,00), никогда не бывших замужем женщин ($M = 3,00$; СКО = 1,41) и замужних женщин ($M = 4,00$; СКО = 0,82)».

Процедура SPSS для Windows

Алгоритм проведения данного дисперсионного анализа состоит в следующем.

Введите данные в **Редактор Данных (Data Editor)**, как показано на рис. 11.2. Семейное положение, имеющее метку **«семья»**, имеет значение единица для никогда не состоявших в браке и два — для состоящих в браке. **«Пол»** принимает значение единица для женщин и два — для мужчин. Выраженность депрессии была названа **«депресс»**.

В строке **Меню** в верхней части окна программы выберите пункт **Анализ (Analyze)**, в появившемся ниспадающем меню — пункт **Общие линейные модели (General Linear Model)**, а в нем — подпункт **Одномерный анализ (Univariate)**, чтобы открыть диалоговое окно **Одномерный анализ (Univariate)**, изображенное на рис. 9.2.

Выберите переменную **«депресс»** и с помощью верхней кнопки ▸ переместите ее в поле **Зависимая переменная (Dependent (variable))**.

	семья	пол	депресс
1	1	1	2
2	1	1	4
3	1	2	8
4	1	2	6
5	1	2	8
6	1	2	6
7	1	2	7
8	1	2	7
9	2	1	3
10	2	1	5
11	2	1	4
12	2	1	4
13	2	2	1
14	2	2	3
15	2	2	2

Рис. 11.2. Данные в **Редакторе данных (Data Editor)** для двухфакторного дисперсионного анализа

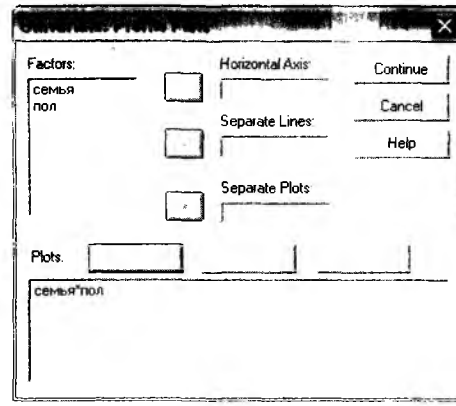


Рис. 11.3. Дочернее диалоговое окно **Одномерный анализ: Диаграммы профиля (Univariate: Profile Plots)**

Выберите переменные **«семья»** и **«пол»** и с помощью второй сверху кнопки ▸ переместите их в список **Постоянные факторы (Fixed Factors)**.

Тип III является методом, используемым по умолчанию, как показано на рис. 10.2. Чтобы выбрать другой метод, нажмите на кнопку **Модель (Model)** и откройте дочернее диалоговое окно **Одномерный анализ: Модель (Univariate: Model)**. Нажмите на направленную вниз стрелку рядом с Тип III (Type III), чтобы отобразить и выбрать в ниспадающем списке нужный вам вариант.

Чтобы вывести график, подобный изображенному на рис. 11.1, нажмите на кнопку **Диаграммы (Plots)** и откройте дочернее диалоговое окно **Одномерный анализ: Диаграммы профиля (Univariate: Profile Plots)**, изображенное на рис. 11.3.

В списке **Факторы (Factors)** выберите переменную **«семья»** и с помощью первой сверху кнопки ▸ переместите ее в поле **Горизонтальная ось (Horizontal Axis)**.

Выберите переменную **«пол»** и с помощью второй сверху кнопки ▸ переместите ее в поле **Отдельные линии (Separate Lines)**.

Нажмите на кнопку **Добавить (Add)**, чтобы поместить **семья* пол** в список **Диаграммы (Plots)**.

Нажмите на кнопку **Продолжить (Continue)**, чтобы закрыть это дочернее диалоговое окно и вернуться в основное диалоговое окно.

Нажмите на кнопку **Параметры (Options)** и откройте дочернее диалоговое окно **Одномерный анализ: Параметры (Univariate: Options)**, изображенное на рис. 9.4.

Установите флажок **Описательные статистики (Descriptive statistics)**, чтобы вывести средние значения и среднеквадратичные отклонения для трех эффектов.

Установите флажок **Критерии однородности (Homogeneity tests)**, чтобы выполнить тест Левина для однородности дисперсий.

Нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в основное диалоговое окно.

Нажмите **OK**, чтобы провести данный анализ.

Для выполнения теста Шеффе для четырех групп, участвующих во взаимодействии, создайте четвертую переменную (например, с именем «**группа**») в **Редакторе Данных (Data Editor)**, значения которой закодированы целыми числами от единицы (никогда не состоявшие в браке женщины) до четырех (состоящие в браке мужчины).

Проведите **Одномерный (Univariate) дисперсионный анализ**, в котором «**депресс**» является **Зависимой переменной (Dependent variable)**, а «**группа**» — **Постоянным фактором (Fixed Factor(s))**.

Нажмите на кнопку **Апостериорные критерии (Post Hoc)**, которая открывает дочернее диалоговое окно **Одномерный Анализ: Апостериорные множественные сравнения для наблюдаемых средних (Univariate: Post Hoc Multiple Comparisons for Observed Means)**, изображенное на рис. 9.3.

Выберите переменную «**группа**» в списке **Факторы (Factors)** и с помощью кнопки ▸ переместите ее в поле **Апостериорные критерии для переменной (Post Hoc Tests for)**.

Выберите **Шеффе (Scheffe)** и нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Одномерный анализ (Univariate)**.

Нажмите **OK**, чтобы провести анализ.

Чтобы провести регрессионный анализ, введите значения для трех фиктивных переменных, как показано в табл. 11.5, в столбцы с четвертого по шестой в **Редакторе Данных (Data Editor)**. Чтобы получить изменение квадрата множественной корреляции, проведите регрессию переменной «**депресс**» по двум последним фиктивным переменным, а затем — по первой.

Выводимая SPSS информация по результатам работы

Не все результаты, выводимые SPSS, будут представлены и обсуждены. Результаты проверки по тесту Левина приведены в табл. 11.9. Поскольку величина F, равная 0,407, имеет уровень значимости 0,751, дисперсии значимо не различаются.

Далее следует таблица дисперсионного анализа (табл. 11.10). Пять из восьми перечисленных элементов представляют для нас интерес, а именно те, что имеют метки «**семья**», «**пол**», «**семья*пол**», **Error (Ошибка)** и **Corrected Total (Скорректированная общая сумма квадратов)** и соответствуют представленным в табл. 11.6 величинам. Например, величина F, равная 14,080, имеет уровень зна-

Таблица 11.9. **Выводимые SPSS результаты теста Левина для двухфакторного дисперсионного анализа**

Levene's Test of Equality of Error Variances[a]

Dependent Variable: ДЕПРЕСС

F	df1	df2	Sig.
0,407	3	11	0,751

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design:Intercept + СЕМЬЯ + ПОЛ + СЕМЬЯ * ПОЛ.

Таблица 11.10. **Выводимые SPSS результаты двухфакторного дисперсионного анализа**

Tests of Between-Subjects Effects

Dependent Variable: ДЕПРЕСС

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	61,333[a]	3	20,444	22,489	0,000
Intercept	204,800	1	204,800	225,280	0,000
СЕМЬЯ	12,800	1	12,800	14,080	0,003
ПОЛ	3,200	1	3,200	3,520	0,087
СЕМЬЯ * ПОЛ	28,800	1	28,800	31,680	0,000
Error	10,000	11	0,909		
Total	398,000	15			
Corrected Total	71,333	14			

a. R Squared = 0,860 (Adjusted R Squared = 0,822).

чимости 0,003 и является статистически значимой, поскольку этот уровень значимости меньше 0,05.

Результаты только первой из двух таблиц для критерия Шеффе отражены в табл. 11.11, поскольку она более актуальна для нас. Сравнения в этой таблице повторяются дважды. Например, в первой строке сравниваются средние значения для группы никогда не состоявших в браке женщин (в таблице — Незамужние женщины (*1*)) и никогда не состоявших в браке мужчин (в таблице — Неженатые мужчины (*2*)), что повторяется в четвертой строке, только с переменой мест сравниваемых величин. Для проведенных сравнений выводится лишь уровень значимости, который в данном случае равен 0,003.

Результаты изменения квадрата множественной корреляции для семейного положения представлены в табл. 11.12. Это изменение

Таблица 11.11. **Выводимые SPSS результаты проверки критерия Шеффе для двухфакторного дисперсионного анализа**

Множественные сравнения

Dependent Variable: ДЕПРЕСС
Scheffe

(I) ГРУППА	(J) ГРУППА	Mean Difference (I–J)	Std. Error	Sig.	95 % Confidence Interval	
					Lower Bound	Upper Bound
Незамужние женщины	Неженатые мужчины	–4,00*	0,778	0,003	–6,55	–1,45
	Замужние женщины	–1,00	0,826	0,697	–3,71	1,71
	Женатые мужчины	1,00	0,870	0,729	–1,86	3,86
Неженатые мужчины	Незамужние женщины	4,00*	0,778	0,003	1,45	6,55
	Замужние женщины	3,00*	0,615	0,004	0,98	5,02
	Женатые мужчины	5,00*	0,674	0,000	2,79	7,21
Замужние женщины	Незамужние женщины	1,00	0,826	0,697	–1,71	3,71
	Неженатые мужчины	–3,00*	0,615	0,004	–5,02	–0,98
	Женатые мужчины	2,00	0,728	0,112	–0,39	4,39
Женатые мужчины	Незамужние женщины	–1,00	0,870	0,729	–3,86	1,86
	Неженатые мужчины	–5,00*	0,674	0,000	–7,21	–2,79
	Замужние женщины	–2,00	0,728	0,112	–4,39	0,39

Based on observed means

* The mean difference is significant at the 0,05 level.

Таблица 11.12. **Выводимые SPSS изменения квадрата множественной корреляции для семейного положения**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F.Change
1	0,825[a]	0,680	0,627	1,378	0,680	12,772	2	12	0,001
2	0,927[b]	0,860	0,822	0,953	0,179	14,080	1	11	0,003

a. Predictors: (Constant), Э1ХЭ2, ЭФФЕКТ2.

b. Predictors: (Constant), Э1ХЭ2, ЭФФЕКТ2, ЭФФЕКТ1.

отражено во второй строке столбца, озаглавленного **R Square Change (Изменение квадрата R)**, и равно 0,179.

Рекомендуемая литература

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows.* London: Routledge.

Overall, J. E. and Spiegel, D. K. (1969) Concerning least squares analysis of experimental data, *Psychological Bulletin*, 72: 311—22.

Pedhazur, E. J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction,* 2nd edn. New York: Holt, Rinehart & Winston.

Pedhazur, E. J. and Schmelkin, L. P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Дополнительная литература, рекомендуемая научным редактором

Гусев А. Н. Дисперсионный анализ в экспериментальной психологии. — М.: УМК «Психология», 2000.

Кулаичев А. П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

ЧАСТЬ VI
РАЗЛИЧЕНИЕ ГРУПП

Глава 12
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Предисловие научного редактора

В гл. 12 описан дискриминантный анализ, используемый в случае, когда на основе статистического массива данных (наблюдения измерены по большому набору переменных) необходимо научиться классифицировать их, т.е. создать некоторую (одну или несколько) композиционную переменную — линейную комбинацию первичных, на основе которых можно классифицировать имеющиеся (обучающая выборка) и вновь получаемые (классифицируемая выборка) наблюдения.

Дискриминантный анализ, или анализ дискриминантных функций, — это параметрический метод, используемый для определения нагрузок количественных переменных (весовых коэффициентов) или предикторов, которые позволяют более качественно описать разделение объектов на группы (две или более). Задаваемое таким образом разделение значительно лучше удовлетворяется экспериментальными данными, чем разделение, в котором весовые коэффициенты выбраны по случайному принципу. Нагрузки переменных формируют новую составную переменную — *дискриминантную функцию*, являющуюся линейной комбинацией нагрузок этих переменных и баллов, которые имеют объекты по переменным. Максимальное число таких функций не должно превосходить число предикторов, с одной стороны, и число групп минус один — с другой. Например, для одного предиктора или двух групп необходимо построить только одну дискриминантную функцию, для трех групп или двух предикторов — две дискриминантные функции. Если дискриминантных функций не менее двух, то они независимы, или ортогональны, друг другу. Каждая функция содержит в качестве составных компонентов линейной комбинации все предикторы, несмотря на то что веса (коэффициенты, с которыми предиктор входит в ту или иную линейную ком-

бинацию) различаются по всем дискриминантным функциям. При этом можно определить точность, с которой дискриминантные функции задают распределение объектов по группам.

По аналогии с множественной регрессией существует три способа введения предикторов в дискриминантный анализ. Согласно стандартному, или прямому, методу, все предикторы вводятся одновременно, несмотря на то что некоторые из них могут играть лишь незначительную роль в распределении по группам. Согласно иерархическому, или последовательному, методу, предикторы вводятся последовательно в порядке, определяемом исследователем. Это позволяет определить вклад каждого предиктора. Например, демографические переменные (возраст, пол и социальный статус) могут быть введены первыми, чтобы проверить, каково воздействие этих переменных. При использовании статистического, или пошагового, метода предикторы-переменные отбираются автоматически (самой программой), согласно показателям, устанавливающим иерархию максимального вклада в дискриминацию. Если два предиктора связаны между собой и имеют почти одинаковую различительную (дискриминантную) силу, то будет выбран предиктор с бо́льшим показателем различения, даже при самой незначительной разнице между различительными силами предикторов.

Проиллюстрируем использование дискриминантного анализа на примере построения функций различения трех групп при на-

Таблица 12.1. **Индивидуальные показатели четырех предикторов для трех групп испытуемых**

Наблюдение	Группа	Тревожный	Беспокойный	Депрессивный	Ощущающий безнадежность
1	1	2	3	1	2
2	1	1	3	3	3
3	1	3	2	2	1
4	1	4	2	3	2
5	1	1	2	1	2
6	2	4	3	3	2
7	2	5	4	2	4
8	2	4	4	2	3
9	2	3	2	3	2
10	2	4	5	1	2
11	3	4	3	5	4
12	3	2	2	3	3
13	3	3	3	5	4
14	3	2	4	4	5
15	3	2	1	4	3

личии четырех предикторов. В эти три группы входят люди (респонденты), у которых были выявлены тревожность, депрессия или не было выявлено ни того, ни другого (группа нормальных). Каждая группа состоит из пяти человек, хотя для дискриминантного анализа не обязательно, чтобы число объектов в каждой группе было одинаковым. В качестве четырех предикторов выступают четыре симптома, выраженность которых оценивается по пятибалльной шкале: переживание тревожности, беспокойства, депрессии и безнадежности. Более высокие баллы определяют бóльшую тяжесть этих симптомов. Оценки предикторов для трех групп представлены в табл. 12.1. Нормальная, тревожная и депрессивная группы обозначены номерами 1, 2 и 3 соответственно. B. G. Tabachnick и L. S. Fidell (1996) считают, что размер меньшей группы должен быть больше числа предикторов. Мы также придерживаемся этого правила в данном случае. Любые выбросы или экстремальные (очень большие или очень маленькие) значения должны быть трансформированы или опущены. В нашем случае таковых нет. Целью является определение симптомов, необходимых для разделения этих групп.

Групповые средние значения и стандартные отклонения

Вначале полезно оценить средние значения и стандартные отклонения показателей четырех предикторов для трех групп, чтобы определить, различаются ли они по трем группам: если да, то каким образом. Также можно провести однофакторный дисперсионный анализ четырех предикторов, чтобы понять, какие из них сами по себе лучше всего позволяют различать три группы с точки зрения уровня значимости[1]. Предикторы с самым высоким уровнем значимости будут лучшими дискриминаторами. Средние значения, стандартные отклонения и уровни значимости для трех групп представлены в табл. 12.2.

Из табл. 12.2 видно, что наиболее значимый предиктор, который лучше всего различает три группы, — это чувство депрессии ($p = 0{,}004$). Средняя выраженность этого показателя у респондентов депрессивной группы — 4,20, тревожной — 2,20 и нормальной — 2,00. Следовательно, возможно, что первая дискриминантная

[1] В данном случае при выполнении однофактороного дисперсионного анализа предполагается установить влияние фактора (того или иного предиктора) на отклик (отнесенность к той или иной группе респондентов). Согласно общей схеме однофакторного анализа, нулевая гипотеза устанавливает отсутствие влияния фактора на отклик. Поэтому, чем ниже p, тем с большей вероятностью можно отвергнуть нулевую гипотезу, не боясь ошибиться и принять альтернативную о наличии влияния фактора на отклик, т.е. можно считать, что чем меньше p, тем более значимо влияние.

функция отделит группу с депрессией от двух других групп, и что предиктором, обладающим наибольшим весом по этой дискриминантной функции, будет симптом чувства депрессии. Следующий по значимости предиктор — чувство безнадежности ($p=0{,}013$). Респонденты депрессивной группы чувствуют себя более безнадежно (3,80), чем тревожной (2,60) и нормальной (2,00) групп. Таким образом, переживание безнадежности может быть предиктором, обладающим вторым по значимости весом по данной дискриминантной функции. Третий по уровню значимости предиктор — это переживание тревоги ($p=0{,}035$). Респонденты тревожной группы в среднем больше испытывают тревогу (4,00), чем депрессивной (2,60) и нормальной (2,00) групп. Поэтому, вероятно, именно чувство тревоги будет обладать наибольшим весом по второй дискриминантной функции, которая отделяет группу с тревожностью от двух других групп.

Необходимо заметить, что если у нас имеются одна дискриминантная функция, отделяющая группу с депрессией от двух других групп, и вторая функция, отделяющая тревожную группу, то третья дискриминантная функция уже не нужна. Первые две функции дают достаточно информации для отделения нормальной группы от двух других. Если, например, высокий балл по первой дискриминантной функции имеют респонденты из группы с депрессией, а высокий балл по второй дискриминантной функции — респонденты из тревожной группы, то нормальная группа будет представлена низкими баллами по обеим функциям. Вот почему максимальное количество дискриминантных функций, определяемое исходя из числа групп, всегда на единицу меньше этого числа.

Таблица 12.2. **Средние и стандартные отклонения и уровни значимости четырех предикторов для трех групп**

Предикторы	Отклонение	Норма	Тревожные	Депрессивные	p
Тревожный	Среднее	2,00	4,00	2,60	0,035
	Стандартное	1,30	0,71	0,89	
Беспокойный	Среднее	2,40	3,60	2,60	0,161
	Стандартное	0,55	1,14	1,14	
Депрессивный	Среднее	2,00	2,20	4,20	0,004
	Стандартное	1,00	0,84	0,84	
Ощущающий безнадежность	Среднее	2,00	2,60	3,80	0,013
	Стандартное	0,71	0,89	0,84	

Дискриминантные функции

Первая дискриминантная функция обеспечивает максимальное или лучшее разделение между группами. Вторая дискриминантная функция дает следующее по значимости разделение между группами и является независимой, или ортогональной, первой дискриминантной функцией и т.д. Если бы у нас было больше трех групп, то имелось бы больше, чем две дискриминантные функции. Дискриминантная функция похожа на регрессионное уравнение (она включает константу и все предикторы со своими весами), поэтому уравнения дискриминантных функций в обоих случаях нашего примера будут иметь следующую общую форму:

[Дискриминантная функция] = [Тревожный] + [Беспокойный] + + [Депрессивный] + [Ощущающий безнадежность] + [Константа].

Значение константы и нестандартизованные весовые значения будут различными для двух функций (табл. 12.3). Из таблицы видно, что «Ощущающий безнадежность» — предиктор, который имеет наибольший вес в первой функции (0,834), затем следует «Депрессивный» (0,723). «Тревожный» — предиктор, который имеет наибольший вес по второй дискриминантной функции (0,753). Эти значения вычисляются с помощью матричной алгебры и в данной книге, чтобы не перегружать читателя, не приведены.

Чтобы получить балл для каждой из двух дискриминантных функций, возьмем значения каждого из предикторов (см. табл. 12.1), умножим их на соответствующие весовые значения и сложим все это с константой. Таким образом, первый случай в табл. 12.4 имеет балл, равный примерно –1,63 по первой дискриминантной функции:

$$(2 \times -0{,}421) + (3 \times -0{,}397) + (1 \times 0{,}723) + (2 \times 0{,}834) + (-1{,}986) =$$
$$= (-0{,}842) + (-1{,}191) + (0{,}723) + (1{,}668) + (-1{,}986) = -1{,}628$$

и балл, равный примерно –1,15 по второй дискриминантной функции:

$$(2 \times 0{,}753) + (3 \times 0{,}304) + (1 \times 0{,}113) + (2 \times 0{,}358) + (-4{,}401) =$$
$$= (1{,}056) + (0{,}912) + (0{,}113) + (0{,}716) + (-4{,}401) = -1{,}154.$$

Баллы по двум дискриминантным функциям для 15 случаев представлены в табл. 12.4.

Средние значения по двум дискриминантным функциям для трех групп представлены в табл. 12.5. По первой дискриминантной функции среднее для депрессивной группы (2,09) больше, чем для нормальной (–0,792) и тревожной (–1,341) групп. Это означает, что эта дискриминантная функция отделяет депрессивную

Таблица 12.3. **Выводимые SPSS веса, или коэффициенты, предикторов и константы для двух дискриминантных функций**

Canonical Discriminant Function Coefficients

	Function	
	1	2
Тревожный	–0,421	0,753
Беспокойный	–0,397	0,304
Депрессивный	0,723	0,113
Ощущающий безнадежность	0,834	0,358
(Constant)	–1,986	–4,401

Unstandardized coefficients.

группу от двух других. Для второй дискриминантной функции среднее тревожной группы (0,887) выше среднего депрессивной (0,184) и нормальной (–1,071) групп, что отделяет тревожную группу от двух других.

Можно показать, что первая дискриминантная функция обеспечивает лучшее различение групп, если проводить однофакторный дисперсионный анализ между баллами по двум дискрими-

Таблица 12.4. **Баллы по двум дискриминантным функциям для 15 случаев**

Наблюдения	Группы	1	2
1	1	–1,63	–1,15
2	1	1,07	–1,32
3	1	–1,76	–0,95
4	1	–0,63	0,28
5	1	–0,81	–2,2
6	2	–1,02	0,58
7	2	–0,90	2,24
8	2	–1,31	1,13
9	2	–0,21	–0,48
10	2	–3,27	0,96
11	3	2,09	1,52
12	3	1,05	–0,87
13	3	2,51	0,77
14	3	2,64	0,57
15	3	2,17	–1,06

Таблица 12.5. **Выводимые SPSS средние двух дискриминантных функций для трех групп**

Functions at Group Centroids

ГРУППА	Function	
	1	2
Норма	–0,752	–1,071
Тревога	–1,341	0,887
Депрессия	2,092	0,184

Unstandardized canonical discriminant functions evaluated at group means.

нантным функциям, результаты чего представлены в табл. 12.6 вместе с двумя другими мерами различения. Межгрупповая сумма квадратов, примерно равная 33,70 для первой дискриминантной функции, больше чем межгрупповая сумма квадратов (9,84) для второй функции. Различительная сила дискриминантных функций может быть также выражена в терминах собственных значений и канонической корреляции.

Собственное значение представляет собой отношение межгрупповой суммы квадратов к внутригрупповой:

$$[\text{Собственное значение}] = \frac{[\text{Межгрупповая сумма квадратов}]}{[\text{Внутригрупповая сумма квадратов}]}.$$

Собственное значение, равное 2,81 (33,70/12,00 = 2,81) для первой дискриминантной функции, больше, чем для второй, рав-

Таблица 12.6. **Результаты однофакторного дисперсионного анализа для двух дискриминантных функций, собственные значения и каноническая корреляция**

	1	2
Межгрупповое среднеквадратичное отклонение	33,70	9,84
Внутригрупповое среднеквадратичное отклонение	12,00	12,00
Общее среднеквадратичное отклонение	45,70	21,84
Собственные значения	33,70/12,00=2,81	9,84/12,00=0,82
Каноническая корреляция	$\sqrt{33,70/45,70} = 0,86$	$\sqrt{9,84/21,84} = 0,67$

нос 0,82 (9,84/12,00 = 0,82). Собственное значение может быть также выражено как процент общей дисперсии баллов дискриминантной функции. Для этого нужно разделить собственное значение дискриминантной функции на сумму собственных значений всех дискриминантных функций и умножить результат на 100. Так, первая дискриминантная функция объясняет 77,41 % [100 × 2,81/(2,81 + 0,82) = = 77,41] всех изменений, тогда как вторая дискриминантная функция объясняет оставшиеся 22,59 % [100 × 0,82/(2,81 + 0,82) = 22,59] изменений.

Каноническая корреляция определяется по следующей формуле:

$$[\text{Каноническая корреляция}] = \sqrt{\frac{[\text{Межгрупповая сумма квадратов}]}{[\text{Общая сумма квадратов}]}}.$$

Каноническая корреляция (0,86) для первой дискриминантной функции больше, чем каноническая корреляция (0,67) для второй.

Статистическая значимость дискриминантных функций

Мы заинтересованы в получении только таких дискриминантных функций, которые распределяют данные по группам на статистически значимом уровне, т.е. значимо лучше, чем это происходило бы в случае, когда разделение по группам осуществлялось случайным образом. Процедура, позволяющая оценить значимость, состоит, во-первых, в определении того, являются ли все дискриминантные функции вместе взятые статистически значимыми. Эта проверка осуществляется с помощью показателя ламбда Уилкса. Для отдельных предикторов и дискриминантных функций показатель ламбда Уилкса равен отношению внутригрупповой суммы квадратов к общей (для всех дискриминантных функций это произведение коэффициентов Уилкса для отдельных дискриминантных функций). Таким образом, общий показатель ламбда Уилкса для двух дискриминантных функций равен примерно 0,144:

$$12,00/45,70 \times 12,00/21,84 = 0,263 \times 0,549 = 0,1444.$$

Показатель ламбда Уилкса изменяется в пределах от 0 до 1 (1 — средние всех групп имеют одно и то же значение, т.е. являются одинаковыми; ≈ 0 — групповые средние различны).

Показатель ламбда Уилкса может быть преобразован в критерий хи-квадрат (J. Stevens, 1996), уровень значимости которого можно определить. Значение критерия хи-квадрат для показателя ламбда Уилкса, равного 0,144, составляет примерно 20,33 (табл.

Таблица 12.7. **Выводимая SPSS информация о статистической значимости дискриминантных функций**

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0,144	20,330	8	0,009
2	0,549	6,289	3	0,098

12.7). Это можно получить в SPSS. Число степеней свободы *df* для данного критерия хи-квадрат определяется числом предикторов, умноженным на число групп, минус 1 [4 × (3 – 1) = 8]. С восемью степенями свободы, чтобы быть статистически значимым при уровне значимости 0,05, критерий хи-квадрат должен иметь значение не меньше 15,51. В нашем случае это значение даже больше. В противном случае анализ следовало бы прекратить, поскольку ни одна из функций не дает статистически значимый результат.

Если критерий хи-квадрат является значимым (как в нашем случае), то показатель лямбда Уилкса для первой дискриминантной функции удаляется и тест проводится для того, чтобы выявить, являются ли оставшиеся дискриминантные функции значимыми. В нашем примере только две дискриминантные функции, поэтому данный тест просто определяет, является ли значимым показатель лямбда Уилкса для второй дискриминантной функции (12,00/21,84 = 0,549), значение критерия хи-квадрат при этом будет примерно равным 6,29. Число степеней свободы для данного критерия хи-квадрат определяется произведением числа предикторов минус 1 на число групп минус 2 [(4 – 1) × (3 – 2) = 3]. С тремя степенями свободы критерий хи-квадрат должен иметь значение ≥ 7,82, чтобы быть значимым при уровне значимости 0,5. В нашем случае этот результат не достигается, следовательно, первая дискриминантная функция является значимой, а вторая — нет. Число степеней свободы для оценки значимости *n* первых дискриминантных функций вычисляется по следующей формуле:

[Число предикторов – *n*] × [Число групп – (*n* + 1)].

Применив данную формулу к нашему примеру, получим три степени свободы: ((4 – 1) × [3 – (1 + 1)]).

Интерпретация дискриминантных функций

Использование нестандартизованных канонических коэффициентов дискриминантной функции (см. табл. 12.3) для интерпретации дискриминантных функций может привести к неверным ре-

Таблица 12.8 **Выводимые SPSS коэффициенты стандартизованных канонических дискриминантных функций**

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Тревожный	–0,421	0,753
Беспокойный	–0,391	0,299
Депрессивный	0,646	0,101
Ощущающий безнадежность	0,681	0,292

зультатам, в частности в случае, если стандартные отклонения предикторов существенно различаются. Интерпретацию дискриминантных функций обычно проводят на основании абсолютных значений стандартизованных коэффициентов (табл. 12.8) и общих внутригрупповых корреляций между предикторами и стандартизованными дискриминантными функциями (табл. 12.9). Результаты применения этих двух методов несколько различаются. Для первой дискриминантной функции наибольший стандартизованный коэффициент имеет чувство ощущения безнадежности (0,681), за которым следует чувство депрессии (0,646), в то время как наибольшая корреляция выявлена для чувства депрессии (0,719), за которым идет чувство ощущения безнадежности (0,538). Для вто-

Таблица 12.9. **Выводимые SPSS общие внутригрупповые корреляции между предикторами и стандартизованными каноническими дискриминантными функциями**

Structure Matrix

	Function	
	1	2
Депрессивный	0,719*	0,330
Ощущающий безнадежность	0,538*	0,537
Тревожный	–0,234	0,849*
Беспокойный	–0,180	0,569*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions.
Variables ordered by absolute size of correlation within function.
* Largest absolute correlation between each variable and any discriminant function.

рой дискриминантной функции чувство тревожности имеет как наибольший стандартизованный коэффициент (0,753), так и наибольшую корреляцию (0,849). Чтобы посчитать стандартизованные коэффициенты, надо умножить нестандартизованный коэффициент на диагональный элемент этого предиктора в общей внутригрупповой ковариационной матрице. Например, для чувства тревожности диагональный элемент равен 1,000, а нестандартизованный коэффициент –0,421 (см. табл. 12.3), следовательно, стандартизованный коэффициент будет также равен –0,421. Вычисление общих внутригрупповых корреляций довольно сложно и здесь рассматриваться не будет.

Классификация

Эффективность дискриминантного анализа может быть определена с помощью процента случаев, правильно отнесенных к своей группе, который сравнивается с процентом возможности случайного правильного распределения случаев по группам. В нашем примере, когда мы имеем равное количество человек в каждой группе, вероятность случайного правильного отнесения респондента к одной из трех групп равна 33 %. Способы использования результатов дискриминантного анализа для распределения случаев по группам слишком сложны и не будут здесь рассматриваться. Принадлежность к группам, предсказанная данными SPSS, представлена в табл. 12.10 вместе с реальным (априорным) распределением по группам исследуемых 15 случаев. Из таблицы видно, что три наблюдения (2, 4 и 9-е) были неправильно отнесены к своим группам. Это означает, что 80 % случаев были правильно идентифицированы ($12/15 \times 100 = 80$).

Данные, представленные в табл. 12.10, могут быть переведены в классификационную таблицу (табл. 12.11), в которой отображены количество и процент правильно и неправильно идентифицированных случаев для каждой группы. Количество и процент правильно идентифицированных случаев расположены по диагонали в клетках таблицы. Например, три респондента, или 60 % ($3/5 \times 100 = 60$), из нормальной группы правильно отнесены к данной группе. Количество и процент случаев, неправильно диагностированных, находятся вне этой диагонали. Например, один респондент, или 20 % ($1/4 \times 100$), из нормальной группы неправильно отнесен к группе тревожных. Из таблицы видно, что процент правильно идентифицированных наиболее высок для депрессивной группы и наиболее низок — для нормальной.

Число случаев в группах часто не является одинаковым при проведении дисперсионного анализа. Это необходимо учитывать при подсчете процента случайного правильного распределения случаев

Таблица 12.10. **Реальное и прогнозируемое распределение по группам для 15 случаев**

Наблюдения	Принадлежность к группе	
	Реальная	Прогнозируемая
1	1	1
2	1	3
3	1	1
4	1	2
5	1	1
6	2	2
7	2	2
8	2	2
9	2	1
10	2	2
11	3	3
12	3	3
13	3	3
14	3	3
15	3	3

Таблица 12.11. **Выводимая SPSS классификационная таблица**

Classification Results[a]

			Predicted Group Membership			
		ГРУППА	Норма	Тревога	Депрессия	Total
Original	Count	Норма	3	1	1	5
		Тревога	1	4	0	5
		Депрессия	0	0	5	5
	%	Норма	60,0	20,0	20,0	100,0
		Тревога	20,0	80,0	0,0	100,0
		Депрессия	0,0	0,0	100,0	100,0

a. 80,0 % of original grouped cases correctly classified.

Таблица 12.12. **Вычисление доли правильно идентифицированных наблюдений в ситуации случайного отнесения**

Группа	Частота (количество в группе)	Вероятность	Правильная частота при случайном распределении
1	2	0,13	0,26
2	5	0,33	1,65
3	8	0,53	4,24
Сумма	15	0,99	6,15
			6,15/15 = 0,41

по группам. Предположим, что мы имеем 2, 5 и 8 случаев в 1, 2 и 3-й группах соответственно (табл. 12.12). Вероятность правильного случайного распределения будет равна 0,13 (2/15 = 0,13); 0,33 (5/15 = 0,33) и 0,53 (8/15 = 0,53) соответственно[1]. Ожидаемое количество случаев в этих группах равно 0,26 (0,13 × 2 = 0,26); 1,65 (0,33 × 5 = 1,65) и 4,24 (0,53 × 8 = 4,24) соответственно. Ожидаемый процент отнесения случаев к данным группам примерно равен 41 [100 × (0,26 + 1,65 + 4,24)/15 = 41,00].

Из табл. 12.12 видно, что три случая (2, 4 и 9-й) были неправильно распределены по группам. Такие ошибки возможны, когда число случаев в группах мало́ и не является одинаковым, а внутригрупповые ковариации неодинаковы, или негомогенны. Равенство, или гомогенность, внутригрупповых ковариаций можно оценить с помощью М-критерия Бокса (Box's M-test). Если он является значимым, то ковариации одинаковы. Если он не является значимым, то ковариации неодинаковы, или гетерогенны. Их различие можно уменьшить, преобразовав значения

[1] Действительно, общее число возможных случаев распределения 15 респондентов по трем группам численностью 2, 3 и 8 человек равно $\frac{15!}{2!5!8!}$. Число возможных случаев, когда респондент из группы, состоящей из двух человек, оказывается правильно отнесенным к своей группе, равно $\frac{14!}{1!5!8!}$. Число возможных случаев, когда респондент из группы, состоящей из пяти человек, оказывается правильно отнесенным к своей группе, равно $\frac{14!}{2!4!8!}$. Соответственно, число возможных случаев, когда респондент из группы, состоящей из пяти человек, оказывается правильно отнесенным к своей группе, равно $\frac{14!}{2!4!7!}$. Вероятность благоприятного исхода вычисляется как отношение числа благоприятных случаев к числу всех возможных случаев.

Таблица 12.13. **Выводимые SPSS результаты М-критерия Бокса равенства ковариационных матриц**

Test Results

Box's M		37,535
F	Approx.	0,907
	df1	20
	df2	516,896
	Sig.	0,578

Tests null hypothesis of equal population covariance matrices.

предикторов, например взяв их корни квадратные или логарифмы. М-Критерий Бокса для нашего примера представлен в табл. 12.13, и мы можем убедиться в том, что уровень значимости 0,578 > 0,5 и, следовательно, не является статистически значимым.

Отчет о результатах

Обычно форма представления результатов зависит от цели анализа. В сжатом виде отчет об анализе, проведенном в этой главе, может быть представлен в следующем виде: «Был проведен прямой дискриминантный анализ с использованием четырех предикторов (чувство тревоги, беспокойства, депрессии и ощущения безнадежности), которые определяли принадлежность человека к одной из трех групп: нормальной; тревожной; депрессивной. Были определены две дискриминантные функции, объясняющие 77 и 23 % дисперсии соответственно. Критерий ламбда Уилкса является значимым для сочетания этих функций ($\chi^2_8 = 20{,}33$; $p < 0{,}01$), но не является значимым после исключения первой функции. Первая функция максимально дифференцировала депрессивную группу от двух других и имела наиболее высокие корреляции с чувством депрессии (0,72) и ощущения безнадежности (0,54). Вторая дискриминантная функция максимально выделяла тревожную группу и наиболее сильно была связана с чувством тревожности (0,85) и беспокойства (0,57). Правильно классифицированы были 80 % случаев по сравнению с 33 %, ожидаемыми как результат правильного распределения по случайности. Все депрессивные случаи были идентифицированы правильно; 80 % тревожных случаев были также определены правильно, а 20 % попали в группу нормальных; 60 % нормальных случаев были правильно отнесены к своей группе, в то время как 20 % были диагностированы как тревожные и 20 % — как депрессивные.

Процедура SPSS для Windows

Приведем алгоритм проведения прямого дискриминантного анализа.

Введите данные из табл. 12.1 в **Редактор данных (Data Editor)**, рис. 12.1.

Предикторам следует присвоить названия **«тревога»**, **«беспокой»**, **«депресс»** и **«безнадеж»**. Номерам групп 1, 2 и 3 присваиваются названия **«норма»**, **«тревога»** и **«депрессия»** соответственно.

Для проведения процедуры дискриминантного анализа выберите **Анализ (Analyze)** на горизонтальной панели инструментов в верхней части окна. В появившемся меню выберите **Классификация (Classify)** и затем — **Дикскриминантный (Discriminant)**, после чего появится диалоговое окно **Дискриминантный Анализ (Discriminant Analysis)**, рис. 12.2.

Выберите переменную **«группа»** и нажмите на первую кнопку ▸, чтобы поместить эту переменную в окно под надписью **Групповая переменная (Grouping Variable:)**.

Нажмите **Определить диапазон (Define Range)**, после чего откроется диалоговое окно **Дискриминантный Анализ: Определить диапазон (Discriminant Analysis: Define Range)**, рис. 12.3.

Впишите «1» в окно напротив надписи **Минимум (Minimum)** и «3» — в окно напротив надписи **Максимум (Maximum)**. Затем нажмите **Продолжить (Continue)** для возвращения в основное окно диалога.

Выделите переменные от **«тревога»** до **«безнадеж»** и затем нажмите на вторую кнопку ▸, чтобы переместить их в окно под надписью **Независимые переменные (Independents)**.

	группа	тревога	беспокой	депресс	безнадеж
1	1	2	3	1	2
2	1	1	3	3	3
3	1	3	2	2	1
4	1	4	2	3	2
5	1	1	2	1	2
6	2	4	3	3	2
7	2	5	4	2	4
8	2	4	4	2	3
9	2	3	2	3	2
10	2	4	5	1	2
11	3	4	3	5	4
12	3	2	2	3	3
13	3	3	3	5	4
14	3	2	4	4	5
15	3	2	1	4	3

Рис. 12.1. Данные примера в **Редакторе данных (Data Editor)**

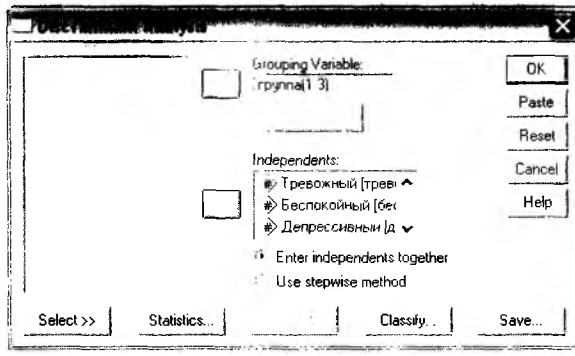


Рис. 12.2. Диалоговое окно **Дискриминантный Анализ (Discriminant)**

Нажмите **Статистика (Statistics)**, после чего откроется диалоговое окно **Дискриминантный Анализ**: **Статистика (Discriminant Analysis: Statistics)**, рис. 12.4.

Здесь можно выбрать следующие полезные опции:

Means (Средние), чтобы получить средние и стандартные отклонения предикторов для трех групп;

Univariate ANOVAs (Одномерный дисперсионный анализ), чтобы получить показатель лямбда Уилкса и показатели статистической значимости однофакторного анализа переменных по четырем предикторам;

Box's M (М-критерий Бокса), чтобы определить, являются ли переменные неравными, когда размеры групп невелики и неравны;

Unstandardized (Нестандартизованные), чтобы получить коэффициенты нестандартизованной канонической дискриминантной функции;

Within-groups covariance (Внутригрупповые ковариации) для проверки вычислений коэффициентов стандартизованной канонической дискриминантной функции.

Нажмите **Продолжить (Continue)**, чтобы вернуться в основное окно диалога.

Для пошагового ввода предикторов выберите **Использование пошагового метода (Use stepwise Method)** и затем **Метод (Method)**, чтобы изменить заданное по умолчанию использование показателя лямбда Уилкса и/или заданные по умолчанию пороговые зна-

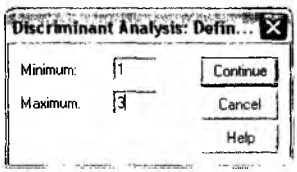


Рис. 12.3. Диалоговое окно **Дискриминантный Анализ**: **Определение диапазона (Discriminant Analysis: Define Range)**

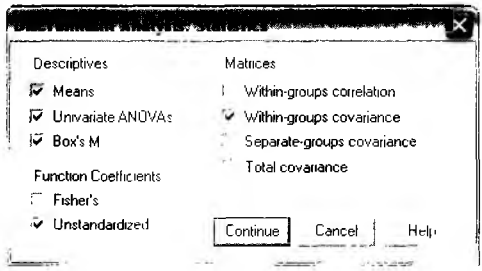


Рис. 12.4. Диалоговое окно **Дискриминантный Анализ**: **Статистика (Discriminant Analysis: Statistics)**

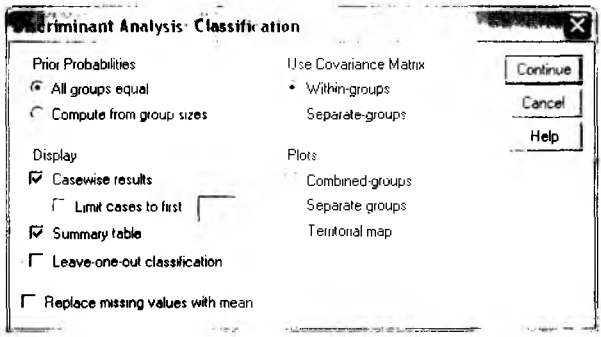


Рис. 12.5. Диалоговое окно **Дискриминантный Анализ**: **Классификация (Discriminant Analysis: Classification)**

чения *F*-критерия (3,84 — для включения предикторов и 2,71 — для их исключения).

Нажмите **Классификация (Classify)**, в результате чего откроется диалоговое окно **Дискриминантный Анализ: Классификация (Discriminant Analysis: Classification)**, представленное на рис. 12.5.

Выберите **Результаты распределения наблюдений (Casewise results)**, чтобы увидеть предсказанное распределение случаев по группам и нестандартизованные баллы дискриминантных функций для каждого случая.

Также выберите **Результирующая таблица (Summary Table)**, чтобы получить классификационную таблицу (см. табл. 12.11).

Для изменения **Исходных Вероятностей (Prior Probabilities)**, когда размеры групп различны, выберите **Вычислить, исходя из размеров групп (Compute from the group sizes)**.

Нажмите **Продолжить (Continue)**, чтобы вернуться в основное окно диалога.

Нажмите **OK** для проведения анализа.

Выводимая SPSS информация по результатам работы

Основные таблицы, которые будут выведены в окне результатов дискриминантного анализа в SPSS, уже были продемонстрированы и обсуждены. Порядок предъявления этих таблиц в окне SPSS несколько отличается от приведенного нами, таблицы будут предъявлены в следующем порядке: табл. 12.13, табл. 12.7, табл. 12.8, табл. 12.9, табл. 12.3, табл. 12.5 и табл. 12.11.

Рекомендуемая литература

Dickhoff, G. (1992) *Statistics for the Social and Behavioral Sciences.* Dubuque, IA: Wm. C. Brown.

Hair, J.F., Jr, Anderson, R.E., Tatham, R.L. and Black, W.C. (1998) *Multivariate Data Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction*, 2nd edn. New York: Holt, Rinehart & Winston.

SPSS Inc. (2002) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics*, 3rd edn. New York: HarperCollins.

Дополнительная литература, рекомендуемая научным редактором

Гусев А.Н. Дисперсионный анализ в экспериментальной психологии. — М.: УМК «Психология», 2000.

Кулаичев А.П. Методы и средства комплексного анализа данных. — М.: Форум — Инфра-М, 2006.

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

ЧАСТЬ VII
АНАЛИЗ ТАБЛИЦ ЧАСТОТ С ТРЕМЯ ИЛИ БОЛЕЕ КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

Глава 13
ЛОГЛИНЕЙНЫЙ АНАЛИЗ

Предисловие научного редактора

В гл. 13 представлен метод, позволяющий одновременно рассматривать три или более качественных переменных — логлинейный анализ. В отечественных исследованиях такой метод еще не используется.

Логлинейный анализ обычно используется для ответа на два типа вопросов о трех или более номинативных переменных и их взаимодействиях. Примером трех таких переменных могли бы быть успешность сдачи детьми экзаменов (success), отсутствие или наличие академических способностей (ability) и наличие или отсутствие интереса к учебе (interest). Самый простой вопрос заключается в том, чтобы установить, значимо ли отличается число наблюдений, с учетом взаимодействия этих качественных переменных, от того, что можно было бы ожидать, исходя из случайной реализации событий[1]. Например, значимо ли различается число детей при разных вариантах взаимодействия между успешностью сдачи экзаменов, способностями и интересом к учебе? Если нас интересует влияние отдельных переменных или взаимодействие двух переменных, то используется критерий хи-квадрат, а не логлинейный анализ. Для анализа влияния трех или более качественных переменных одновременно применяют процедуру логлинейного анализа.

---

[1] Иными словами, каждое наблюдение можно охарактеризовать наличием или отсутствием выраженности того или иного признака (соответствующего той или иной качественной переменной). Возможное сочетание отсутствия/присутствия рассматриваемых признаков называется *паттерном*. Нас интересует анализ распределения всех возможных паттернов, полученных в эксперименте данных, и сравнение его с тем, которое могло бы быть получено, исходя из соображений обычной вероятности.

Другой вопрос состоит в том, какие из переменных и/или их взаимодействие обеспечивают самое простое и экономное (с точки зрения количества объясняющих параметров) объяснение распределения наблюдений. Например, достаточно ли успешности сдачи детьми экзаменов самой по себе, чтобы объяснить распределение показателей, полученных для каждого ребенка? Мы покажем, как с помощью логлинейного анализа можно ответить на оба эти вопроса.

Там, где мы заинтересованы в изучении связи между одной из переменных (т.е. откликом, или зависимой переменной) и другими переменными (т.е. предикторами, или независимыми переменными), возможно, было бы предпочтительнее использовать логистическую регрессию, а не логлинейный анализ, так как последний также изучает взаимосвязь между самими независимыми переменными, которая может нас и не интересовать. Например, возможно, было бы предпочтительнее использовать логистическую регрессию для определения взаимосвязи между зависимой переменной «Успешность сдачи детьми экзаменов» и переменными-предикторами «Способности», «Интерес к учебе», а также взаимодействие между способностями и интересом к учебе. В то время как логлинейный анализ обычно изучает взаимодействия между всеми этими переменными, для логистической регрессии эффекты взаимодействия необходимо создавать, используя фиктивные переменные.

Проиллюстрируем логлинейный анализ с помощью данных, представленных в табл. 13.1, которая показывает количество детей, характеризуемых как успевающие и неуспевающие, способные или неспособные и заинтересованные или незаинтересованные в учебе. Например, в первой строке таблицы указано, что 60 детей характеризуются как успешно сдавшие экзамены, спо-

Таблица 13.1. **Количество детей, характеризуемых в соответствии со способностями, интересом к учебе и успеваемостью**

Ячейка	Способности	Интерес к учебе	Успеваемость	Частота
1	Способный	Заинтересованный	Неуспевающий	60
2	Способный	Заинтересованный	Успевающий	10
3	Способный	Незаинтересованный	Неуспевающий	70
4	Способный	Незаинтересованный	Успевающий	20
5	Неспособный	Заинтересованный	Неуспевающий	80
6	Неспособный	Заинтересованный	Успевающий	30
7	Неспособный	Незаинтересованный	Неуспевающий	90
8	Неспособный	Незаинтересованный	Успевающий	40

собные и заинтересованные в учебе. B.G.Tabachnick и L.S.Fidell (1996) считают, что должно быть по меньшей мере в пять раз больше наблюдений, чем ячеек в таблице, чтобы можно было проводить логлинейный анализ. Поскольку в нашей таблице восемь ячеек, у нас должно быть как минимум сорок наблюдений ($5 \times 8 = 40$), что имеет место. Кроме того, при рассмотрении пары переменных одновременно, например «Способности» и «Успеваемость», ожидаемые частоты во всех ячейках должны быть больше единицы, а для 80 % ячеек — не менее пяти. Далее будет показано, как вычислить ожидаемые частоты для двух переменных. Если ожидаемые частоты меньше, чем названные ранее, то необходимо объединять категории или вообще не включать переменные в рассмотрение.

Логлинейный анализ главным образом сравнивает фактическую частоту наблюдений в ячейке с ожидаемой частотой случайной реализации. Чем больше разность между фактической и ожидаемой частотами, тем более вероятно, что это различие статистически значимо и поэтому вряд ли случайно. Различие между частотами может быть выражено с помощью критерия хи-квадрат Пирсона или хи-квадрат отношения максимального правдоподобия. Второй критерий является предпочтительным, поскольку значения переменных и их взаимодействие с большей вероятностью дают прибавку к общей сумме для модели с равными частотами, чем в случае критерия хи-квадрат Пирсона, особенно для большого числа взаимодействий. Что это означает, станет ясно позднее.

Чтобы понять суть логлинейного анализа, полезно сравнить его с дисперсионным анализом (см. гл. 9—11). Общая дисперсия в дисперсионном анализе (т.е. сумма квадратов разностей между отдельными значениями и средним арифметическим по группе для зависимой переменной) может быть разложена и объяснена за счет независимых переменных и их взаимодействий, называемых *эффектами*. Необъясненная (оставшаяся) дисперсия называется *ошибкой*, или *остаточной ошибкой*. В логлинейном анализе ошибка отсутствует, поскольку для каждой ячейки имеется единственное значение. Зависимой переменной в логлинейном анализе является разность между фактической и ожидаемой[1] частотами. Последняя выражается как критерий хи-квадрат отношения максимального правдоподобия.

Общее значение критерия хи-квадрат максимального правдоподобия, которое необходимо объяснить, предполагает, что все ячейки имеют одну и ту же ожидаемую частоту (т.е. равновероятностны). В нашем примере имеется восемь ячеек и 400 наблюде-

[1] Т.е. частотой, которую можно было бы ожидать исходя из случайной реализации.

Таблица 13.2. **Значения критерия хи-квадрат отношения максимального правдоподобия для главных переменных и их взаимодействий**

Эффект	Число степеней свободы	Критерий хи-квадрат отношения максимального правдоподобия	
Общий	7	132,65***	
Способности (С)	1	16,11***	
Интерес (И)	1	4,01*	
Успеваемость (У)	1	104,65***	
		Граничный	Частный
С× И	1	0,14	0,31
С × У	1	5,70**	5,83**
И × У	1	1,36	1,49
С × И × У	1		0,52
Сумма	7	131,97	132,92

* $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$.

ний, так что ожидаемая частота для каждой ячейки составляет 50,00 (400/8 = 50). Ожидаемая частота выражается десятичной дробью, поскольку может не быть целым числом.

Эта модель с равными частотами предполагает, что ни одна из переменных и/или их взаимодействий не оказывает влияния (не имеет эффекта). Общее значение критерия хи-квадрат максимального правдоподобия может быть разложено на эффекты (влияния) каждой переменной и ее взаимодействий с остальными переменными. Наш пример состоит из трех главных эффектов (способности, интерес к учебе, успеваемость), трех парных взаимодействий (способности × интерес, способности × успеваемость и интерес × успеваемость) и одного тройного взаимодействия (способности × интерес × успеваемость), как показано в табл. 13.2. Модель, содержащая все возможные эффекты, объясняет все значения критерия хи-квадрат максимального правдоподобия и называется полностью насыщенной моделью.

Критерий хи-квадрат отношения максимального правдоподобия можно определить по следующей формуле:

$$\begin{bmatrix}\text{Критерий } \chi\text{-квадрат отношения} \\ \text{максимального правдоподобия}\end{bmatrix} =$$

$$= \{\text{Сумма по всем ячейкам}[\text{Фактическая частота} \times$$

$$\times \ln \frac{\text{Фактическая частота}}{\text{Ожидаемая частота}}]\} \times 2.$$

Отношение фактической частоты ячейки к ее ожидаемой частоте выражается как натуральный, или неперов, логарифм и умножается на фактическую частоту данной ячейки, затем значения, полученные по всем ячейкам, суммируются и умножаются на два.

Общий эффект

Критерий хи-квадрат отношения максимального правдоподобия для модели, предполагающей равенство ожидаемых частот в восьми ячейках, составляет приближенно 132,60. Вычисления для получения данного значения приведены в табл. 13.3. Число степеней свободы для этой модели равно количеству ячеек минус единица (8 – 1 = 7). Чтобы быть значимым на уровне 0,05, значение двустороннего критерия хи-квадрат отношения максимального правдоподобия с семью степенями свободы должно быть ≥ 14,01, что в данном случае выполняется. Следовательно, фактические частоты отличаются от ожидаемых при условии случайной реализации и мы можем теперь определить, какие эффекты объясняют данное различие. Если бы значение критерия хи-квадрат отношения максимального правдоподобия для данной модели было бы незначимо, то ни один из эффектов не был бы значимым.

Главные эффекты

В нашем примере имеется три главных эффекта. Проиллюстрируем вычисления для первого из эффектов — способностей, что отражено в табл. 13.4. Способности имеют две категории (способные и неспособные), поэтому ожидаемая частота составляет 200,00 (400/2 = 200,00) для каждой из категорий. В нашем случае 160 наблюдений (детей) были классифицированы как способные, а 240 — как неспособные. Критерий хи-квадрат отношения максимального правдоподобия приближенно равен 16,00[1], что с учетом ошибки округления совпадает со значением 16,11 (см. табл. 13.2), полученным с помощью SPSS. Число степеней свободы равно количеству категорий минус единица (2 – 1 = 1). Чтобы быть статистически значимой на уровне 0,05, значение двустороннего критерия хи-квадрат отношения максимального правдоподобия должно быть равно 3,84 или выше, что в нашем случае выполняется. Следовательно, фактическая частота наблюдений для способностей значимо отличается от ожидаемой ситуации случайной реализации. В нашей выборке меньше способных детей, чем можно

[1] $[160\ln(160/200) + 240\ln(240/200)]2 = (-35{,}70 + 43{,}76)2 = 16{,}11$.

было бы ожидать в результате случайного распределения наблюдений. Результаты для двух других эффектов представлены в табл. 13.2 и оба являются значимыми.

Таблица 13.3. **Вычисление критерия хи-квадрат отношения максимального правдоподобия для общего эффекта**

Ячейки	Вычисление критерия хи-квадрат отношения максимального правдоподобия
1	$60 \times \ln \frac{60}{50,00} = 60 \times 0,182 = 10,94$
2	$10 \times \ln \frac{10}{50,00} = 10 \times -1,609 = -16,09$
3	$70 \times \ln \frac{70}{50,00} = 70 \times 0,336 = 23,55$
4	$20 \times \ln \frac{20}{50,00} = 20 \times -0,916 = -18,33$
5	$80 \times \ln \frac{80}{50,00} = 80 \times 0,470 = 37,60$
6	$30 \times \ln \frac{30}{50,00} = 30 \times -0,511 = -15,32$
7	$90 \times \ln \frac{90}{50,00} = 90 \times 0,588 = 52,90$
8	$40 \times \ln \frac{40}{50,00} = 40 \times -0,223 = -8,93$
2 × Сумму	$2 \times 66,30 = 132,60$

Таблица 13.4. **Вычисление критерия хи-квадрат отношения максимального правдоподобия для главного эффекта — способностей**

Ячейки	Вычисление критерия хи-квадрат отношения максимального правдоподобия
Способные	$160 \times \ln \frac{160}{200,00} = 160 \times -0,223 = -35,68$
Неспособные	$240 \times \ln \frac{240}{200,00} = 240 \times 0,182 = 43,68$
2 × Сумму	$2 \times 8,00 = 16,00$

Эффекты взаимодействия

В нашей модели имеются три парных эффекта взаимодействия и одно тройное взаимодействие. Существует два способа вычисления критерия хи-квадрат отношения максимального правдоподобия для парного взаимодействия. Более простым методом является граничный критерий, в котором каждое парное взаимодействие анализируется отдельно, без учета других парных взаимодействий и взаимодействий высшего порядка. Более сложным методом является частный критерий, в котором каждое парное взаимодействие корректируется с учетом всех других взаимодействий. В этом методе используется итерационная процедура, при которой ожидаемые частоты становятся все более близкими к фактическим граничным значениям для тех эффектов, которые не анализируются. Мы не будем иллюстрировать эту более сложную процедуру. Результаты для обоих методов приведены в табл. 13.2 и для данного примера в основном совпадают. Если сложить значения критерия хи-квадрат отношения максимального правдоподобия для главных эффектов и взаимодействий, то получим, что общее значение 131,97 для граничного критерия несколько меньше, чем значение 132,65 для модели, предполагающей равные ожидаемые частоты, в то время как значение 132,97 для частного критерия несколько выше указанной величины.

Рассмотрим процедуру вычисления критерия хи-квадрат отношения максимального правдоподобия для парного взаимодействия между способностями и интересом к учебе. Фактические частоты для этого двустороннего взаимодействия представлены в табл. 13.5 вместе с граничными суммами для строк и столбцов. Необходимо вычислить ожидаемые частоты для этих четырех ячеек при условии того, что нам известно, какая доля детей имеет способности, и какая — заинтересована в учебе. Доля способных детей составляет 0,40 (160/400 = 0,40), а доля заинтересованных в учебе равна 0,45 (180/400 = 0,45). Следовательно, можно ожидать, что доля детей, которые одновременно обучаемы и заинтересованы в учебе, составляет 0,18 (0,40 × 0,45 = 0,18), что в расчете на 400 детей дает ожидаемую частоту, равную 72,00 (0,18 × 400 = 72,00).

Этот метод подсчета ожидаемой частоты в ячейке может быть выражен с помощью следующей формулы:

[Ожидаемая частота] = [Сумма по столбцу/Общая сумма] ×
× [Сумма по строке/Общая сумма] × [Общая сумма].

Эту формулу можно упростить, сократив на величину общей суммы в числителе с одним из значений общей суммы в знаменателе следующим образом:

[Ожидаемая частота] = [Сумма по столбцу/Общая сумма] ×
× [Сумма по строке].

Таблица 13.5. **Фактические частоты и граничные суммы для парного взаимодействия между способностями и интересом к учебе**

	Способные	Неспособные	Сумма по строке
Проявляющие интерес к учебе	70	110	180
Не проявляющие интерес к учебе	90	130	220
Сумма по столбцу	160	240	400

Используя упрощенную формулу, можно вычислить ожидаемую частоту для способных и не заинтересованных в учебе детей ($160/400 \times 200 = 88{,}00$), для неспособных и заинтересованных детей ($240/400 \times 180 = 108{,}00$) и для неспособных и не заинтересованных в учебе детей ($240/400 \times 220 = 132{,}00$).

Зная ожидаемые частоты для этой двухвходовой таблицы, можно вычислить значение критерия хи-квадрат отношения максимального правдоподобия (табл. 13.6). Его значение равно 0,10, что с учетом погрешности округления равно 0,14 (см. табл. 13.2). Число степеней свободы для парного взаимодействия равно произведению количества столбцов минус единица на количество строк минус единица $[(2-1) \times (2-1) = 1]$. Чтобы быть значимым на уровне 0,05 по двустороннему критерию, значение критерия хи-квадрат

Таблица 13.6. **Вычисление критерия хи-квадрат отношения максимального правдоподобия для парного взаимодействия между способностями и интересом к учебе**

Ячейки	Вычисление критерия хи-квадрат отношения максимального правдоподобия
Способные, проявляющие интерес к учебе	$70 \times \ln \frac{70}{72{,}00} = 70 \times -0{,}028 = -1{,}96$
Способные, не проявляющие интерес к учебе	$90 \times \ln \frac{90}{88{,}00} = 90 \times 0{,}022 = 1{,}98$
Неспособные, проявляющие интерес к учебе	$110 \times \ln \frac{110}{108{,}00} = 110 \times 0{,}018 = 1{,}98$
Неспособные, не проявляющие интерес к учебе	$130 \times \ln \frac{130}{132{,}00} = 130 \times -0{,}015 = -1{,}95$
$2 \times$ Сумма	$2 \times 0{,}05 = 0{,}10$

Таблица 13.7. **Фактические и ожидаемые частоты для парного взаимодействия между способностями и успеваемостью**[1]

	Сдали экзамен	Не сдали экзамен
Способные	Факт = 130	Obs = 30
	Ожидание = 120,00	Exp = 40,00
Неспособные	Факт = 170	Obs = 70
	Ожидание = 180,00	Exp = 60,00

отношения максимального правдоподобия должно быть ≥ 3,84, что в нашем случае не выполняется. Следовательно, фактическая частота наблюдений не отличается значимо от ожидаемой для обучаемости и интереса к учебе в условиях случайного распределения.

Значение критерия хи-квадрат отношения максимального правдоподобия оказывается значимым только для взаимодействия между способностями и успеваемостью. Чтобы проинтерпретировать это взаимодействие, необходимо взглянуть на фактические и ожидаемые частоты для четырех ячеек табл. 13.7. Из таблицы видно, что число способных детей, успешно сдавших экзамены, больше, чем можно было бы ожидать при случайном распределении. Другими словами, имеется положительная взаимосвязь между способностями и успеваемостью.

Пошаговое обратное исключение

Одной из статистических процедур для определения того, каково наименьшее количество эффектов, необходимых для воспроизведения фактических частот данных, является пошаговое обратное исключение. Оно может начинаться с насыщенной модели[2], которая содержит все эффекты и в силу того, что она точно воспроизводит фактические частоты, не является значимой, в том смысле что обеспечивает точное согласие с данными. На первом шаге рассматривается возможность исключения взаимодействия наивысшего порядка. Если исключение этого взаимодействия приводит к значимому изменению критерия хи-квадрат отношения максимального правдоподобия, то анализ останавливается, поскольку это означает, что данное взаимодействие необходимо для воспроизведения фактических данных. Если же исключение этого эффекта не приводит к значимому изменению величины хи-квад-

[1] В русскоязычной литературе такие таблицы называются двухвходовыми, или сопряженными.

[2] Термин «насыщенность» используется в том смысле, что рассматривается модель, включающая все взаимодействия между переменными.

рат отношения максимального правдоподобия, то данное взаимодействие исключается. На втором шаге рассматриваются следующие по числу участвующих в нем переменных взаимодействия, начиная со взаимодействия, имеющего самую незначимую вероятность осуществления. На заключительном этапе таким же образом могут рассматриваться главные эффекты.

В нашем примере устранение тройного взаимодействия не приводит к значимому изменению критерия хи-квадрат отношения максимального правдоподобия. Это показано в табл. 13.2, где частное значение критерия хи-квадрат отношения максимального правдоподобия равно 0,52 и является незначимым. Следовательно, данный эффект можно исключить из рассмотрения. Далее исследуются взаимодействия второго порядка, начиная со взаимодействия между способностями и интересом к учебе, которое имеет наивысшую незначимую вероятность осуществления. Из табл. 13.2 видно, что частное значение хи-квадрат отношения максимального правдоподобия для этого взаимодействия равно 0,31 и является незначимым. Поскольку исключение данного взаимодействия не приводит к значимому изменению значения хи-квадрат отношения максимального правдоподобия, на втором шаге оно исключается.

На третьем этапе рассматривается возможность исключения незначимого взаимодействия между интересом к учебе и успеваемостью. Частное значение критерия хи-квадрат отношения максимального правдоподобия теперь такое же, как и граничные значения, поскольку не надо учитывать тройное взаимодействие и парные взаимодействия между способностями и интересом к учебе. Так как частное значение критерия хи-квадрат отношения максимального правдоподобия, равное 1,36, не является значимым, данное взаимодействие исключается из рассмотрения. Взаимодействие между способностями и успеваемостью сохраняется, поскольку его исключение привело бы к значимому изменению значения критерия хи-квадрат отношения максимального правдоподобия на 5,70. Сохранение взаимодействия означает, что главные эффекты, которые обусловливают это взаимодействие, также включаются в модель. Таким образом, способности и успеваемость также предполагаются частью модели.

На четвертом шаге рассматриваются главные эффекты, не входящие в состав сохраненных взаимодействий. Единственным подобным главным эффектом является заинтересованность в учебе, которая сохраняется, поскольку ее исключение привело бы к значимому изменению критерия хи-квадрат отношения максимального правдоподобия на 4,01. Здесь анализ останавливается, так как взаимодействия между способностями и успеваемостью нет и главного эффекта заинтересованности в учебе достаточно для воспроизведения значений, которые значимо не отличаются от фактических.

Таблица 13.8. **Ожидаемые частоты для заинтересованных и не заинтересованных в учебе детей в зависимости от способностей и успеваемости**

	Сдали экзамен	Не сдали экзамен
Способные, проявляющие интерес к учебе	0,45 × 130 = 58,50	0,45 × 30 = 13,50
Способные, не проявляющие интерес к учебе	0,55 × 130 = 71,50	0,55 × 30 = 16,50
Неспособные, проявляющие интерес к учебе	0,45 × 170 = 76,50	0,45 × 70 = 31,50
Неспособные, не проявляющие интерес к учебе	0,55 × 170 = 93,50	0,55 × 70 = 38,50

Ожидаемые частоты для этой модели равны произведению фактической частоты детей в зависимости от способностей и успеваемости и суммарной вероятности заинтересованности или незаинтересованности детей в учебе. Суммарная вероятность того, что дети заинтересованы в учебе, равна 0,45 (180/400 = 0,45), а не заинтересованы в учебе — 0,55 (220/400 = 0,55). Имеется 130 детей, одновременно характеризующихся как способные и успешно сдавшие экзамены. Следовательно, ожидаемое количество таких детей, которые при этом должны проявлять интерес к учебе, равно 58,50 (0,45 × 130 = 58,50), а не заинтересованных в учебе — 71,50 (0,55 × 130 = 71,50). Ожидаемые частоты для всех восьми ячеек приведены в табл. 13.8.

Критерий хи-квадрат отношения максимального правдоподобия для этой модели составляет 2,20, как показано в вычислениях, приведенных в табл. 13.9. Число степеней свободы для данной модели равно трем, поскольку из нее были исключены одно тройное и два парных взаимодействия, обладающие каждое одной степенью свободы. Чтобы быть статистически значимым на уровне 0,05, значение двустороннего критерия хи-квадрат отношения максимального правдоподобия должно быть ≥ 7,82, чего в данном случае не наблюдается.

Отчет о результатах

Способ фиксации результатов логлинейного анализа зависит от целей проведения анализа. Если мы определяем, какая из мо-

Таблица 13.9. **Вычисление критерия хи-квадрат отношения максимального правдоподобия для наилучшим образом согласующейся с данными модели, полученной с помощью пошагового обратного исключения**

Ячейки	Вычисление критерия хи-квадрат отношения максимального правдоподобия
1	$60 \times \ln \frac{60}{58,50} = 60 \times 0,025 = 1,50$
2	$10 \times \ln \frac{10}{13,50} = 10 \times -0,300 = -3,00$
3	$70 \times \ln \frac{70}{71,50} = 70 \times 0,021 = 1,48$
4	$20 \times \ln \frac{20}{16,50} = 20 \times 0,192 = 3,84$
5	$80 \times \ln \frac{80}{76,50} = 80 \times 0,045 = 3,60$
6	$30 \times \ln \frac{30}{31,50} = 30 \times -0,049 = -1,46$
7	$90 \times \ln \frac{90}{93,50} = 90 \times -0,038 = -3,42$
8	$40 \times \ln \frac{40}{38,50} = 40 \times 0,038 = 1,52$
2 × Сумма	$2 \times 1,10 = 2,20$

делей обеспечивает наилучшее согласие с данными, то один из кратких способов изложения обнаруженных закономерностей состоит в следующем: «Проводился частотный анализ трехвходовых таблиц в целях построения иерархической логлинейной модели способностей, интереса к учебе и успеваемости у детей. В результате пошагового обратного исключения была получена модель, включающая главный эффект заинтересованности в учебе и парное взаимодействие между способностями и успеваемостью (значение критерия хи-квадрат отношения максимального правдоподобия с тремя степенями свободы, равное 2,20, статистически незначимо), демонстрирующая хорошее согласие между фактическими частотами и порожденными моделью ожидаемыми частотами. Около 45 % детей были характеризованы как за-

интересованные в учебе. Примерно 81 % детей, у которых были диагностированы высокие способности, успешно сдали экзамены, по сравнению с 71 % детей, характеризуемых как имеющие невысокие способности».

Процедура SPSS для Windows

Алгоритм проведения описанного в данной главе логлинейного анализа с помощью пошагового обратного исключения состоит в следующем.

Введите данные с помощью **Редактора Данных (Data Editor)**, как показано на рис. 13.1. Единицы означают наличие способностей, интереса к учебе и успеваемости, тогда как двойки указывают на их отсутствие. Восемь строк представляют восемь ячеек таблицы, которым приписаны фактические веса в соответствии со значениями в столбце, озаглавленном «**частота**». Эта процедура описана в гл. 4. Выберите пункт **Данные (Data)** из строки меню в верхней части окна программы, а в ниспадающем меню — пункт **Весовые коэффициенты (Weight Cases)**, чтобы открыть диалоговое окно **Весовые коэффициенты (Weight Cases)**, показанное на рис. 4.2. Из списка переменных **Весовые коэффициенты (Weight cases by)** выберите «**частота**» и с помощью кнопки ▸ переместите ее в поле **Частоты (Frequency variable)**, а затем нажмите **OK**, чтобы закрыть это диалоговое окно.

Чтобы провести логлинейный анализ, выберите пункт **Анализ (Analyze)** в строке меню в верхней части окна программы, пункт **Логлинейный анализ (Loglinear)** в ниспадающем меню, а затем — подпункт **Выбор модели (Model Selection)**, который открывает диалоговое окно **Выбор модели логлинейного анализа (Model Selection Loglinear Analysis)**, изображенное на рис. 13.2.

Выберите в списке переменные, начиная со способностей ребенка («**способн**») и заканчивая успеваемостью ребенка («**успех**»), и с помощью первой кнопки ▸ переместите эти три переменные в список **Факторы (Factors)**.

	способн	интерес	успех	частота
1	1	1	1	60
2	1	1	2	10
3	1	2	1	70
4	1	2	2	20
5	2	1	1	80
6	2	1	2	30
7	2	2	1	90
8	2	2	2	40

Рис. 13.1. Данные примера в **Редакторе Данных (Data Editor)**

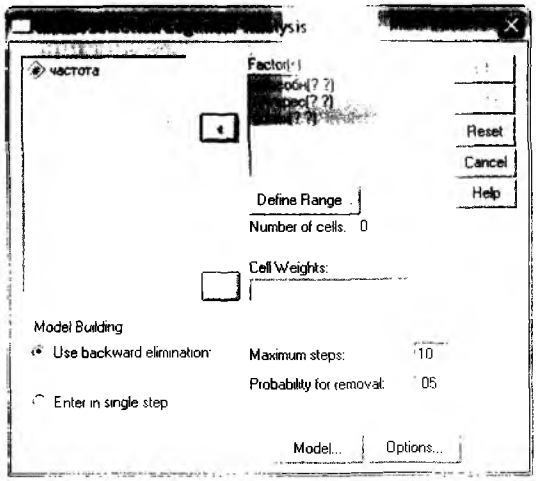


Рис. 13.2. Диалоговое окно **Выбор модели логлинейного анализа (Model Selection Loglinear Analysis)**

Так как каждая из трех переменных имеет по две категории, можно определить множества их значений, нажав на кнопку **Задать диапазон значений (Define Range)** и открыв дочернее диалоговое окно ***Логлинейный анализ*: Задание диапазона значений (Loglinear Analysis: Define Range)**, изображенное на рис. 13.3.

Введите значение «1» в поле **Минимальное значение (Minimum)** и «2» — в поле **Максимальное значение (Maximum)**, а затем нажмите на кнопку **Продолжить (Continue)**, чтобы вернуться в диалоговое окно **Выбор модели логлинейного анализа (Model Selection Loglinear Analysis)**. Если бы каждая переменная имела свое число категорий или переменные были бы закодированы по-разному, то пришлось бы выбирать каждую переменную отдельно и задавать ее диапазон значений.

Если мы хотим вывести частное значение критерия хи-квадрат отношения максимального правдоподобия для насыщенной модели, как показано в табл. 13.2 и 13.10, то следует нажать на кнопку **Параметры (Options)**, которая открывает дочернее диалоговое окно **Логлинейный анализ: Параметры (Loglinear Analysis: Options)**, изображенное на рис. 13.4. Установите флажок **Таблица связей (Association table)** и нажмите на кнопку **Продолжить (Continue)**,

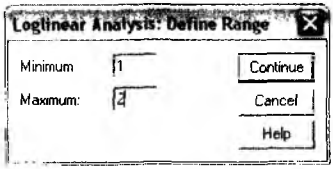


Рис. 13.3. Дочернее диалоговое окно **Логлинейный анализ: Задание диапазона значений (Loglinear Analysis: Define Range)**

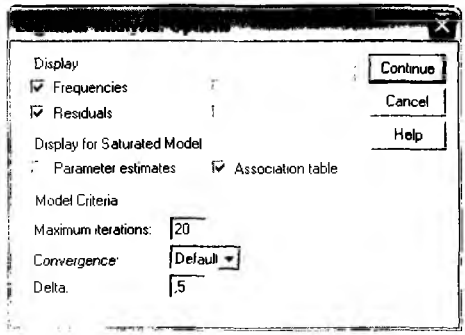


Рис. 13.4. Дочернее диалоговое окно **Логлинейный анализ: Параметры (Loglinear Analysis: Options)**

чтобы вернуться в диалоговое окно **Выбор модели логлинейного анализа (Model Selection Loglinear Analysis)**.

Нажмите **OK**, чтобы провести анализ.

Выводимая SPSS информация по результатам работы

Не все результаты, выводимые программой, будут представлены. Первая часть даст фактические и ожидаемые частоты для насыщенной модели вместе со значениями критерия хи-квадрат Пирсона и отношения максимального правдоподобия. Ожидаемые частоты совпадают с фактическими частотами. К каждой частоте прибавлена величина 0,5, чтобы избежать проблем, которые возникают, когда приходится иметь дело с ячейками, не содержащими наблюдений. Оба значения критерия хи-квадрат равны нулю, что означает полное согласие с данными.

Таблица связей (см. табл. 13.10) содержит частные значения критерия хи-квадрат отношения максимального правдоподобия, которые соответствуют значениям из табл. 13.2, за исключением опущенного в табл. 13.10 значения для тройного взаимодействия.

Таблица 13.10. **Выводимая SPSS таблица связей**

```
Tests of PARTIAL associations.
```

Effect Name	DF	Partial	Chisq	Prob	Iter
СПОСОБН*ИНТЕРЕС	1		0,305	0,5811	2
СПОСОБН*УСПЕХ	1		5,834	0,0157	2
ИНТЕРЕС*УСПЕХ	1		1,491	0,2221	2
СПОСОБН	1		16,108	0,0001	2
ИНТЕРЕС	1		4,007	0,0453	2
УСПЕХ	1		104,650	0,0000	2

Таблица 13.11. **Выводимые SPSS результаты проверки эффектов K-мерных взаимодействий**

```
Tests that K-way and higher order effects are zero.

   K  DF  L.R. Chisq    Prob   Pearson Chisq    Prob  Iteration

   3   1       0,524  0,4691           0,519  0,4711       3
   2   4       7,881  0,0960           7,295  0,1211       2
   1   7     132,645  0,0000         120,000  0,0000       0
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Tests that K-way effects are zero.

   K  DF  L.R. Chisq    Prob   Pearson Chisq    Prob  Iteration

   1   3     124,765  0,0000         112,705  0,0000       0
   2   3       7,357  0,0614           6,776  0,0794       0
   3   1       0,524  0,4691           0,519  0,4711       0
```

Показатели тройного взаимодействия представлены в табл. 13.11. Значение критерия хи-квадрат отношения максимального правдоподобия для тройного взаимодействия равно 0,524 и приведено под заголовком L.R.Chisq (likelihood ratio chi-square test) в первой строке верхней таблицы и еще раз — в третьей строке нижней таблицы.

Значение 132,645 в третьей строке верхней таблицы относится к модели с равными частотами, 7,881 во второй строке этой таблицы представляет собой сумму значений для тройного взаимодействия (0,524) и парных взаимодействий между способностями и заинтересованностью в учебе (0,305), способностями и успеваемостью (5,697) и интересом к учебе и успеваемостью (1,355). Значения двух последних взаимодействий были вычислены без учета тройного и оставшегося парного взаимодействия и приведены в табл. 13.14.

Значение 124,765 в первой строке нижней таблицы представляет собой сумму значений для трех главных эффектов: обучаемости (16,108), интереса к учебе (4,007) и успеваемости (104,650). Значение 7,357 во второй строке этой таблицы есть сумма значений для парных взаимодействий, приведенных выше (0,305; 5,697 и 1,355).

Может создаться впечатление, что вывод результатов заканчивается после таблицы связей. Чтобы посмотреть на оставшуюся часть результатов, нажмите левую клавишу мыши, когда курсор находится в границах уже выведенных результатов. Эти результаты заканчиваются черной чертой. Подведите курсор к маленькому

```
* * * * * * * * HIERARCHICAL LOG LINEAR * * * * * * * *
```

Рис. 13.5. Нижняя часть исходного вывода результатов

черному прямоугольнику в середине черной линии, нажмите и удерживайте левую клавишу мыши, когда на короткое время появится значок песочных часов, а затем — верхняя и нижняя стрелки. Переместите черную линию до конца окна вывода результатов.

В табл. 13.12 приведены результаты первого этапа обратного исключения. Насыщенная модель, представленная тройным взаимодействием, имеет значение критерия хи-квадрат отношения максимального правдоподобия 0,00000. Если тройное взаимодействие исключается из данной модели, то изменение критерия хи-квадрат отношения максимального правдоподобия (L.R.Chisq Change) составляет 0,524, что не является значимым ($p = 0{,}4691$). Следовательно, данное взаимодействие исключается, оставляя модель с тремя парными взаимодействиями.

В табл. 13.13 представлены результаты второго этапа анализа. Самым незначимым с точки зрения вероятности осуществления является взаимодействие между способностями и интересом к учебе ($p = 0{,}5811$). Следовательно, это взаимодействие исключается. Изменение критерия хи-квадрат отношения максимального правдоподобия в результате исключения как данного, так и тройного взаимодействия составляет 0,82858 (0,305 + 0,524 = 0,829), что является незначимым ($p = 0{,}661$).

В табл. 13.14 показаны результаты третьего этапа анализа. Единственное взаимодействие, которое оказывается незначимым, — это взаимодействие между заинтересованностью в учебе и успеваемостью ($p = 0{,}2445$). Таким образом, данное взаимодействие исключается из рассмотрения. Изменение критерия хи-квадрат от-

Таблица 13.12. **Выводимые SPSS результаты первого этапа обратного исключения**

```
Backward Elimination (p = 0,050) for DESIGN 1 with generating
class

   СПОСОБН*ИНТЕРЕС*УСПЕХ

Likelihood ratio chi square =  0,00000   DF = 0    P = -INF

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

If Deleted Simple Effect is  DF L.R.Chisq Change   Prob Iter

СПОСОБН*ИНТЕРЕС*УСПЕХ         1            0,524 0,4691   3

Step 1

   The best model has generating class

   СПОСОБН*ИНТЕРЕС
   СПОСОБН*УСПЕХ
   ИНТЕРЕС*УСПЕХ

Likelihood ratio chi square =  0,52405   DF = 1   P = 0,469
```

Таблица 13.13. **Выводимые SPSS результаты второго этапа обратного исключения**

```
If Deleted Simple Effect is  DF L.R.Chisq Change   Prob Iter

   СПОСОБН*ИНТЕРЕС            1           0,305 0,5811   2
   СПОСОБН*УСПЕХ              1           5,834 0,0157   2
   ИНТЕРЕС*УСПЕХ              1           1,491 0,2221   2

Step 2

   The best model has generating class

      СПОСОБН*УСПЕХ
      ИНТЕРЕС*УСПЕХ

Likelihood ratio chi square =    0,82858  DF = 2  P = 0,661
```

Таблица 13.14. **Выводимые SPSS результаты третьего этапа обратного исключения**

```
If Deleted Simple Effect is  DF L.R.Chisq Change   Prob Iter

   СПОСОБН*УСПЕХ              1           5,697 0,0170   2
   ИНТЕРЕС*УСПЕХ              1           1,355 0,2445   2

******** HIERARCHICAL LOG LINEAR ********
Step 3

   The best model has generating class

      СПОСОБН*УСПЕХ
      ИНТЕРЕС

Likelihood ratio chi square =    2,18332  DF = 3  P = 0,535
```

Таблица 13.15. **Выводимые SPSS результаты четвертого этапа обратного исключения**

```
If Deleted Simple Effect is  DF L.R.Chisq Change   Prob Iter

   СПОСОБН*УСПЕХ              1           5,697 0,0170   2
   ИНТЕРЕС                    1           4,007 0,0453   2

Step 4
   The best model has generating class
      СПОСОБН*УСПЕХ
      ИНТЕРЕС

Likelihood ratio chi square =    2,18332  DF = 3  P = 0,535
```

ношения максимального правдоподобия в результате исключения этого взаимодействия вместе с двумя предыдущими составляет 2,18332 (0,305 + 0,524 + 1,355 = 2,184), что не является значимым (p = 0,535). В табл. 13.15 приведены результаты четвертого этапа анализа. Вероятность исключения как взаимодействия между спо-

собностями и успеваемостью (p = 0,0170), так и главного эффекта интереса к учебе (p = 0,0453) является статистически значимой. Таким образом, ни один из этих двух эффектов исключать нельзя. Два других главных эффекта — способности и успеваемость — не приведены, поскольку их присутствие подразумевается их взаимодействием друг с другом. Следовательно, наиболее согласующаяся с данными модель состоит из взаимодействия между способностями и успеваемостью и главным эффектом заинтересованности в учебе.

После этой таблицы выводятся фактические и ожидаемые частоты для данной модели, а вслед за ними — значения критерия хи-квадрат Пирсона и отношения максимального правдоподобия. Ожидаемые частоты — те же, что и в табл. 13.8. Значение критерия хи-квадрат Пирсона, равное 2,14078, может быть получено по соответствующей формуле после подстановки значений ожидаемых частот из этой таблицы.

Рекомендуемая литература

Howell, D.C. (1997). *Statistical Methods for Psychology*, 4th edn. Belmont, CA: Duxbury.

SPSS Inc. (2002) *SPSS 11.0 Advanced Models*. Upper Saddle River, NJ: Prentice-Hall.

Stevens, J. (1996) *Applied Multivariate Statistics for the Social Sciences*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics*, 3rd edn. New York: HarperCollins.

ГЛОССАРИЙ (СЛОВАРЬ СПЕЦИАЛЬНЫХ ТЕРМИНОВ)

Апостериорный критерий (Post-hoc test) — критерий для определения того, различаются ли две группы в случае, когда у нас нет веских внестатистических оснований предполагать различия между ними.

Априорный критерий (A priori test) — критерий для определения того, различаются ли между собой две группы, когда есть веские основания предполагать различия между ними, исходя из теоретических соображений.

Балльная статистика (Score statistic) — мера связи в логистической регрессии.

Варимакс (Varimax) — процедура вращения, обеспечивающая независимость факторов друг от друга (ортогональность).

Взаимодействие (Interaction) — взаимосвязь между двумя переменными значимо отличается в зависимости от значений одной или нескольких других переменных.

Вращение факторов (Rotation of factors) — совокупность процедур максимизации нагрузок некоторых переменных на определенный фактор с одновременной минимизацией нагрузок других переменных на этот и остальные факторы, с тем чтобы упростить интерпретацию данного фактора.

Выброс (Outlier) — величина, или значение, во множестве величин, или значений, сильно от них отличающаяся.

Главная компонента (Principal component) — переменная, являющаяся линейной комбинацией других переменных, веса которых подобраны таким образом, чтобы объяснить наибольшую долю дисперсии всех переменных, а дисперсии каждой переменной принимаются равными единице.

Гнездовая модель (Nested model) — модель, которая является более простым подмножеством другой модели и может быть выведена из нее за счет фиксирования значения какого-либо параметра, который ранее оценивался.

Дендрограмма (дерево) (Dendrogram) — графическое представление результатов кластерного анализа, в котором отрезки указывают, какие переменные или кластеры объединяются в пары на каждом этапе анализа.

Диаграмма накопления (слияния) (Icicle plot) — графическое представление результатов кластерного анализа, при котором значения абсциссы используются для указания на переменные или кластеры, объединяемые в пары на каждом этапе.

Диаграмма путей (Path diagram) — диаграмма, показывающая, какие переменные предполагаются или фактически являются связанными с другими переменными.

Дискриминантный анализ (Discriminant analysis) — параметрическая статистическая процедура, разработанная для определения того, какие весовые коэффициенты переменных наилучшим образом позволяют провести различия между двумя или более группами.

Дискриминантная функция (Discriminant function) — переменная, представляющая собой линейную комбинацию предикторов, веса которых выбраны таким образом, чтобы максимально различать группы, и независимая от других дискриминантных функций.

Дисперсионный анализ (Analysis of variance (ANOVA)) - параметрический статистический метод для определения того, отличаются ли значимо друг от друга средние значения единственной зависимой переменной, полученные в двух или более группах (в сравнении с тем, каковы были бы различия при случайном формировании групп).

Дисперсия (Variance) — мера изменчивости или разброса значений одной или нескольких количественных переменных, равная среднему арифметическому квадратов отклонений.

Зависимая переменная (Dependent variable) — переменная, на значения которой предположительно влияют одна или более других переменных.

Идентификация (Identification) — та степень, до которой параметры модели структурных уравнений могут быть оценены с помощью исходных данных.

Иерархическая агломеративная кластеризация (Hierarchical agglomerative clustering) — вид кластерного анализа, на каждом шаге которого переменная или кластер объединяются в пару с наиболее близкой переменной или кластером до тех пор, пока не останется единственный кластер.

Иерархический, или последовательный, ввод переменных (Hierarchical or sequential entry) — порядок, в котором переменные включаются в анализ; определяется исследователем. Взаимодействия переменных также могут включаться в анализ.

Квадрат Евклидова расстояния (Squared Euclidean distance) — сумма квадратов разностей значений двух переменных для данной выборки.

Кластерный анализ (Cluster analysis) — различные процедуры для определения способов группировки, или кластеризации, переменных или наблюдений.

Ковариата (Covariate) — переменная, связанная линейным соотношением с зависимой переменной и контролируемая в ковариационном анализе.

Ковариационный анализ (Analysis of covariance (ANCOVA)) — параметрический статистический критерий для определения того, значимо ли отличаются друг от друга средние значения единственной зависимой переменной, полученные в двух или более группах, когда контролируется одна или несколько ковариат[1].

Ковариация (Covariance) — совместная дисперсия двух или более количественных переменных.

Кодирование с помощью фиктивных переменных (Dummy coding) — метод задания фиктивной переменной, при котором две определяемые

---

[1] Т. е. устраняется эффект влияния этих ковариат.

группы кодируются соответственно единичными и нулевыми значениями.

Кодирование эффекта (Effect coding) — метод определения фиктивной переменной, при котором две определяемые группы кодируются значениями «1» и «–1», а остальные группы — нулевыми значениями.

Конфирматорный (подтверждающий) факторный анализ (Confirmatory factor analysis) — вид факторного анализа, используемый для определения вероятности того, что конкретная, или гипостазируемая, факторная структура подтверждается данными.

Корреляция, или коэффициент корреляции (Correlation or correlation coefficient) — показатель силы и направления линейной взаимосвязи между двумя количественными переменными.

Корреляционная матрица (Correlation matrix) — массив или таблица, содержащая корреляции между тремя или более переменными.

Косоугольные факторы (Oblique factors) — факторы, или компоненты, значения которых коррелируют со значениями других факторов, или компонентов.

Критерий Кайзера, или Кайзера — Гутмана (Kaiser or Kaiser—Guttman criterion) — статистический критерий, при котором факторы или компоненты, имеющие собственные значения, не превосходящие единицы, игнорируются.

Критерий Левина (Levene's test) — однофакторный дисперсионный анализ абсолютных отклонений внутри групп для определения того, являются ли дисперсии двух или более групп подобными или однородными.

Критерий собственных значений (следа) (Scree test) — график собственных значений факторов, используемый для того, чтобы определить, сколько исходных факторов должно быть оставлено для вращения.

Критерий хи-квадрат отношения правдоподобия (Likelihood ratio chi-square test) — критерий хи-квадрат, основанный на логарифме фактических и ожидаемых частот.

Критерий хи-квадрат Пирсона (Chi-square test, Pearson's) — критерий хи-квадрат сравнивает фактические частоты категорий одной или двух качественных переменных с их ожидаемыми частотами.

Критерий Шеффе (Scheffe test) — параметрический апостериорный критерий для определения того, различаются ли значимо средние значения двух групп.

Критическое значение (Critical value) — значение статистики, определяющее уровень, при котором и за пределами которого нулевая гипотеза отвергается.

Латентная переменная (Latent variable) — ненаблюдаемая переменная, измеряемая с помощью одной или более явных переменных, или индикаторов.

Линейная связь, или отношение (Linear association or relationship) — связь, при которой значения одной переменной изменяются только в одном направлении, по мере того как увеличиваются значения другой переменной, что при графическом изображении может быть точно представлено наклонной прямой линией.

Логарифм (Logarithm) — количество раз, которое необходимо перемножить само на себя фиксированное число — основание (например, 10 или e), чтобы получить данное положительное число.

Логарифмический критерий отношения максимального правдоподобия (–2 log likelihood (ratio) test) — критерий хи-квадрат для сопоставления степени согласия с данными двух моделей, из которых одна является подмножеством другой. Критерий хи-квадрат равен разности логарифмов значений критериев хи-квадрат максимального правдоподобия двух рассматриваемых моделей, умноженной на –2.

Логарифмическое правдоподобие (Log likelihood) — сумма вероятностей, связанных с фактическими и прогнозируемыми исходами для модели выборочных данных.

Логистическая, или логит-регрессия (Logistic or logit regression) — вид регрессионного анализа, при котором зависимая переменная, или отклик (критерий), является категориальной, или номинальной. График соотношения между зависимой и независимой переменными предположительно имеет форму S-образной кривой с минимальным значением вероятности, равным нулю, и максимальным значением вероятности, равным единице.

Манифестная (явная, наблюдаемая) переменная (Manifest, observable variable) — переменная, являющаяся непосредственным отражением измерения (меры), используемой для ее оценки.

Маргинальные (граничные) суммарные частоты (Marginal totals) — суммарные частоты строк и столбцов в таблице сопряженности.

Метод максимального правдоподобия (Maximum likelihood method) — метод нахождения оценок параметров генеральной совокупности для модели, которые с наибольшей вероятностью обусловливают паттерн наблюдений в выборочных данных.

М-критерий Бокса (Box's M) — статистический критерий для проверки равенства, или однородности, матриц ковариаций независимых переменных для групп, сформированных по уровням значений зависимой переменной.

Многомерный дисперсионный анализ (Multivariate analysis of variance (MANOVA)) — параметрический статистический метод для определения того, отличаются ли средние значения в двух или более группах по двум и более зависимым переменным друг от друга значимо (в сравнении с тем, каковы были бы различия при случайном формировании групп).

Множественная корреляция (Multiple correlation или R) — показатель силы линейной взаимосвязи между зависимой количественной переменной и двумя или более независимыми количественными переменными. В качестве независимых переменных могут также рассматриваться взаимодействия между предикторами.

Множественная регрессия (Multiple regression) — параметрический метод для определения линейной взаимосвязи между зависимой количественной переменной и двумя или более независимыми количественными переменными, учитывающий любые взаимосвязи между предикторами. При этом могут также учитываться взаимодействия между независимыми переменными.

Нагрузка (Loading) — корреляция между переменной и фактором или дискриминантной функцией.

Натуральный, или неперов, логарифм (Natural or Naperian logarithm) — логарифм по основанию e (e ≈ 2,718).

Недоопределенная модель (Under-identified model) — модель структурных уравнений, в которой недостаточно значений данных, чтобы оценить ее параметры.

Независимая переменная (Independent variable) — переменная, значения которой предположительно влияют на одну или несколько других переменных.

Нестандартизированные коэффициенты, или веса (Unstandardized coefficients or weights) — коэффициенты, или веса, независимых переменных в уравнении, выраженные через свои исходные значения.

Общность (Communality) — дисперсия переменной, совместная со всеми остальными переменными, принимающими участие в анализе.

Опосредующий эффект (Moderating or moderator effect) — результат взаимодействия между двумя или более переменными, который изменяется в зависимости от значений одной или нескольких других переменных.

Ортогональные факторы (Orthogonal factors) — факторы, или компоненты, чьи значения не коррелируют со значениями других факторов, или компонент.

Отношение шансов (Odds ratio) — число, на которое умножается шанс осуществления категории при изменении на одну единицу какой-либо независимой переменной, в то время как остальные предикторы в модели остаются неизменными.

Параметр (Parameter) — количественная характеристика генеральной совокупности значений.

Параметрический критерий (Parametric test) — статистический критерий, основанный на предположении о том, что генеральная совокупность, из которой извлекаются выборки, подчиняется закону нормального распределения.

Переопределенная модель (Over-identified model) — модель структурных уравнений, в которой количество известных из наблюдений данных превосходит число оцениваемых параметров, что позволяет определить степень согласия модели с данными.

Показатель согласия (Goodness-of-fit index) — статистический критерий, разработанный для определения степени того, насколько конкретная модель, описывающая набор данных, в состоянии воспроизвести эти данные.

Показатель степени, или степень (Exponent or power) — число, записываемое как верхний индекс для другого числа — основания и показывающее, сколько раз необходимо умножить основание само на себя.

Показатель лямбда Уилкса (Wilks' lambda) — мера (величина), включающая отношение внутригрупповой суммы квадратов к общей сумме квадратов, используемая для определения того, значимо ли различаются средние значения переменных в зависимости от групп.

Последовательность слияния (Agglomeration schedule) — таблица, показывающая, какие переменные или кластеры объединяются в пары на каждом этапе кластерного анализа.

Пошаговый ввод переменных (Stepwise entry) — переменные включаются в анализ с использованием статистических критериев. Взаимодействия между переменными также могут включаться в анализ.

Прямой облимин (Direct oblimin) — процедура вращения, позволяющая факторам коррелировать друг с другом.

Путь (Pathway) — линия на диаграмме путей, обозначающая взаимосвязь между двумя переменными.

Таблица сопряженности (Contingency table) — таблица, содержащая фактические частоты наблюдений в категориях группировки двух или более переменных.

Собственное значение (Eigenvalue) — доля дисперсии, объясняемой фактором или дискриминантной функцией.

Стандартизированные коэффициенты, или веса (Standardized coefficients or weights) — коэффициенты, или веса, независимых переменных в уравнении выражаются через их стандартизированные значения.

Стандартный, или непосредственный, ввод переменных (Standard or direct entry) — все переменные включаются в анализ одновременно. Взаимодействия переменных также могут включаться в анализ.

Степени свободы (Degrees of freedom) — число компонент, участвующих в определении значения статистического показателя, которые могут свободно изменяться и используются для определения подходящего критического значения для данного показателя.

Сумма квадратов (Sum of squares) — сумма квадратов отклонений от среднего.

Таблица сопряженности (Contingency table) — таблица, содержащая фактические частоты наблюдений в категориях группировки двух или более переменных.

Точно определенная модель (Just-identified model) — модель структурных уравнений, в которой данных в точности столько, сколько необходимо для оценки всех ее параметров.

Уровень (Level) — группа или условие по независимой переменной или фактору в дисперсионном анализе.

Фактор в дисперсионном анализе (Factor, analysis of variance) — независимая переменная.

Фактор в факторном анализе (Factor, factor analysis) — общий термин для переменной, представляющей собой линейную комбинацию других переменных и вводимой для объяснения общей дисперсии этих переменных[1].

Фиктивная переменная (Dummy variable) — переменная, создаваемая в регрессионном анализе для представления качественной переменной, содержащей три или более категорий.

Частная корреляция (Partial correlation) — показатель силы и направления линейной взаимосвязи между двумя количественными переменными с выделением или учетом их линейной взаимосвязи с одной или несколькими другими количественными переменными.

Часть корреляции, или получастная корреляция (Part or semi-partial correlation) — показатель силы и направления линейной взаимосвязи

---

[1] Правильнее было бы сказать, что это фактор в анализе главных компонент.

между зависимой количественной переменной и независимой количественной переменной, учитывающий линейную взаимосвязь между рассматриваемым предиктором и одним или несколькими другими количественными предикторами.

Шанс (Odds) — отношение вероятности осуществления категории к вероятности ее неосуществления.

Эндогенная переменная (Endogenous variable) — переменная в анализе путей, объясняемая одной или несколькими другими переменными[1].

Экзогенная переменная (Exogenous variable) — переменная в анализе путей, не объясняемая никакой другой переменной[2].

Эксплораторный (исследовательский) факторный анализ (Exploratory factor analysis) — различные процедуры для определения числа факторов, необходимых для объяснения совместной дисперсии трех и более переменных[3].

***F*-отношение** (*F* ratio) — отношение, которое можно выразить через два набора дисперсий, значимость которого проверяется с помощью *F*-распределения Фишера.

LISREL — название компьютерной программы, разработанной для проведения анализа линейных структурных соотношений, также известного как моделирование структурными уравнениями.

SPSS — название компьютерной программы, разработанной для проведения различных видов статистического анализа и обработки данных.

---

[1] Здесь надо уточнить, что речь идет именно о переменной, будь то измеряемая или латентная, но не об индикаторах латентных экзогенных переменных. Формально индикаторы являются зависимыми от своих латентных конструктов (независимые они или зависимые). Однако в LISREL эти индикаторы (показатели измерения которых не являются самоцелью, а выполняются для нахождения значения, определяющего их латентный конструкт) латентных экзогенных переменных также считаются экзогенными.

[2] Это не совсем корректное определение. Точнее было бы сказать: «Экзогенная переменная — переменная в путевом анализе, которая объясняется влиянием только внешних по отношению к исследуемой модели событий, т.е. либо это независимая в модели переменная (латентная или измеряемая), либо зависимая от независимой измеряемая переменная (индикатор)».

[3] Все-таки это задача метода главных компонент — максимально объяснять дисперсию, а задача эксплораторного факторного анализа — максимально объяснять корреляционную или ковариационную структуру наблюдаемых переменных.

ПРИЛОЖЕНИЕ

Модель структурных уравнений и модели измерений К. Йореского, используемые в LISREL

Основная схема в терминах матричных уравнений выглядит следующим образом:

η, ξ, **y**, **x**, ζ, ε, δ — векторы переменных, т.е. каждый из них представляет собой целую группу переменных, объединенных по следующим правилам:

компоненты вектора η — это эндогенные латентные переменные;

компоненты вектора ξ — это экзогенные латентные переменные;

векторы **x** и **y** содержат измеряемые переменные, соответствующие латентным переменным ξ и η;

векторы ε и δ содержат ошибки измерений переменных **x** и **y**.

Структурная модель: $\eta = \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta$.

Эндогенные и экзогенные латентные переменные связаны между собой системой линейных уравнений (структурная модель) с помощью матриц коэффициентов **B** (Beta) и **Γ** (Gamma) и остаточным вектором ζ.

Измеряемая модель для $\mathbf{y} = \Lambda_y\eta + \varepsilon$.

Измеряемая модель для $\mathbf{x} = \Lambda_x\xi + \delta$.

Соответствие между измеряемыми и латентными переменными устанавливается двумя системами линейных уравнений (измеряемая модель) с коэффициентами, содержащимися в матрицах Λ_x и Λ_y соответственно, и ошибками измерений.

Коэффициенты корреляционной взаимосвязи между экзогенными латентными переменными содержатся в матрице **Φ**.

Коэффициенты корреляционной взаимосвязи между эндогенными латентными переменными содержатся в матрице **Ψ**.

Диаграммы путей задаются с помощью коэффициентов, содержащихся в следующих матрицах

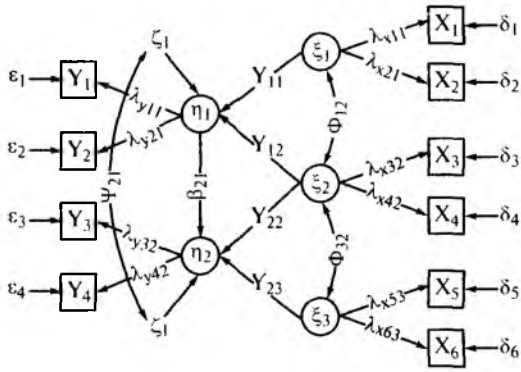
Математическое обозначение	Обозначение в LISREL	Размерность	Форма массива по умолчанию	Содержание массива по умолчанию
ξ	K (Ksi)	NK × 1		
η	E (Eta)	NE × 1		
x	X	NX × 1		
y	Y	NY × 1		

Окончание табл.

Математическое обозначение	Обозначение в LISREL	Размерность	Форма массива по умолчанию	Содержание массива по умолчанию
B	BE (Beta)	NE × NE	Нулевая	Fixed
Γ	GA (Gamma)	NE × NK	Полная четырехугольная матрица	Free
Λ_x	LX (LambdaX)	NX × NK	Полная четырехугольная матрица	Fixed
Λ_y	LY (LambdaY)	NY × NE	Полная четырехугольная матрица	Fixed
Φ	PH(i)	NK × NK	Симметричная относительно главной диагонали	Free
Ψ	PS(i)	NE × NE	Симметричная относительно главной диагонали	Free
Θ_ε	TE (Theta Epsilon)	NY × NY	Диагональная	Free
Θ_δ	TD (Theta Delta)	NX × NX	Диагональная	Free

Fixed — коэффициенты фиксированы.
Free — коэффициенты оцениваются.

Путевая диаграмма модели



СПИСОК ЛИТЕРАТУРЫ

Asher, H.B. (1983). *Causal Modeling*, 2nd edn. Beverly Hills, CA: Sage.

Baron, R.M. and Kenny, D.A. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology*, 51: 1173–82.

Bryman, A. and Cramer, D. (2001) *Quantitative Data Analysis with SPSS Release 10 for Windows*. Hove: Routledge.

Byrne, B.M. (1998) *Structural Equation Modelling with LISREL, PRELIS, and SIMPLIS*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cattell, R.B. (1966) The scree test for the number of factors, *Multivariate Behavioral Research*, 1: 245–76.

Child, D. (1990) *The Essentials of Factor Analysis*. London: Routledge.

Cohen, J. (1968) Multiple regression as a general data-analytic system, *Pyschological Bulletin*, 70: 426–43.

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows*. London: Routledge.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests, *Psychometrika*, 16: 297–334.

Diekhoff, G. (1992) *Statistics for the Social and Behavioral Sciences*. Dubuque, IA: Wm. C. Brown.

Gorsuch, R.L. (1983) *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hair, J.F., Jr, Anderson, R.E., Tatham, R.L. and Black, W.C. (1998) *Multivariate Data Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Hosmer, D.W. and Lemeshow, S. (1998) *Applied Logistic Regression*. New York: Wiley.

Howell, D.C. (1997) *Statistical Methods for Psychology*, 4th edn. Belmont, CA: Duxbury.

Huitema, B.E. (1980). *The Analysis of Covariance and Alternatives*. New York: Wiley.

James, L.R., Mulaik, S.A. and Brett, J.M. (1982). *Causal Analysis: Assumptions, Models, and Data*. Beverly Hills, CA: Sage.

Jöreskog, K.G. and Sörbom, D. (1989) *LISREL 7: A Guide to the Program and Applications*, 2nd edn. Chicago, IL: SPSS Inc.

Kline, P. (1994) *An Easy Guide to Factor Analysis*. London: Routledge.

Lochlin, J.C. (1998). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis,* 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Menard, S. (1995) *Applied Logistic Regression Analysis.* Thousand Oaks, CA: Sage.

Overall, J. E. and Spiegel, D. K. (1969) Concerning least squares analysis of experimental data, *Psychological Bulletin,* 72: 311—22.

Pample, FC. (2000) *Logistic Regression: A Primer.* Thousand Oaks, CA: Sage.

Pedhazur, E.J. (1982) *Multiple Regression in Behavioral Research: Explanation and Prediction.* New York: Holt, Rinehart & Winston.

Pedhazur, E.J. and Schmelkin, L. P. (1991) *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* New York: Wiley.

SPSS Inc. (2002a) *SPSS Base 11.0 User's Guide Package.* Upper Saddle River, NJ: Prentice-Hall.

SPSS Inc. (2002b) *SPSS 11.0 Regression Models.* Upper Saddle River, NJ: Prentice-Hall.

SPSS Inc. (2002c) *SPSS 11.0 Advanced Models.* Upper Saddle River, NJ: Prentice-Hall.

Stevens, J. (1996) *Applied Multivariate Statistics for the Social Sciences,* 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics,* 3rd edn. New York: HarperCollins.

Дополнительная литература, рекомендуемая научным редактором

Кляцкин В. И. Динамика стохастических систем. — М. : Физматлит, 2002.

Митина О. В. Детерминированный анализ: основные понятия, статистические критерии, приемы использования в психологии исследования // Вест. Моск. гос. ун-та. — 2004. — № 4. — С. 46—60.

Митина О. В. Моделирование динамики общественного сознания / О. В. Митина, В. Ф. Петренко // Уч. записки кафедры общей психологии МГУ им. М. В. Ломоносова / под ред. Б. С. Братуся, Д. А. Леонтьева. — Вып. 1. — М. : Смысл, 2002. — С. 236—257.

Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. — СПб. : Речь, 2004.

Наследов А. Д. Методы обработки многомерных данных в психологии. — СПб. : Изд-во СПбУ, 1999.

Bentler P. M. EQS, Structural Equations, Program Manual. CA, Encino: Multivariate Software, Inc., 1995.

Bickel P. J. Sex Bias in Graduate Admissions: Data from Berkeley / P. J. Bickel, E. A. Hammel, J. W. O'Connell // Science, February, 1985, 187. — P. 398—404.

http://www.statsoft.ru/home/ — русская страница компании StatSoft (производитель Statistica)

Joreskog K. G. LISREL 8: User's reference guide (2nd ed.) / K. G. Joreskog, D. Sorbom. Chicago: Scientific Software International, 1996.

Simpson E. H. The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society, Series B, 13, 1951. — P. 238—241.

Wolseley G. The Soldier's Pocket Book. 5th ed. L. : Macmillan, 1986.

ОГЛАВЛЕНИЕ

Часть VII

АНАЛИЗ ТАБЛИЦ ЧАСТОТ С ТРЕМЯ ИЛИ БОЛЕЕ КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

Учебное издание

Крамер Дункан

Математическая обработка данных в социальных науках: современные методы

Учебное пособие

Редактор *Л. В. Честная*
Технический редактор *Е. Ф. Коржуева*
Компьютерная верстка: *Г. Ю. Никитина*
Корректоры *Г. Н. Петрова, Н. В. Савельева*

Изд. № 101107162. Подписано в печать 10.08.2006. Формат 60 × 90/16. Гарнитура «Таймс». Бумага офсетная № 1. Печать офсетная. Усл. печ. л. 18,0. Тираж 3000 экз. Заказ № 3504.

Издательский центр «Академия». www.academia-moscow.ru
Санитарно-эпидемиологическое заключение № 77.99.02.953.Д.004796.07.04 от 20.07.2004.
117342, Москва, ул. Бутлерова, 17-Б, к. 360. Тел./факс: (495)330-1092, 334-8337.

Отпечатано с электронных носителей издательства.
ОАО "Тверской полиграфический комбинат", 170024, г. Тверь, пр-т Ленина, 5.
Телефон: (4822) 44-52-03, 44-50-34, Телефон/факс (4822) 44-42-15
Home page - www. tverpk.ru Электронная почта (E-mail) - sales@tverpk.ru