УЧЕБНИКИ
ВЫСШЕЙ ШКОЛЫ ЭКОНОМИКИ

ВШЭ
HSE

А. О. Крыштановский

АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

с помощью пакета SPSS

Допущено Министерством образования и науки Российской Федерации в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки «Социология»

Издательский дом ГУ ВШЭ

Москва 2006

УДК 303.4:004.9
ББК 60.5в7
К85

Подготовлено при содействии НФПК —
Национального фонда подготовки кадров в рамках
программы «Совершенствование преподавания
социально-экономических дисциплин в вузах»

Ответственные редакторы:
доктор социологических наук, профессор Ю.Н. Толстова, А.В. Рыжова

Рецензент:
доктор социологических наук, профессор Г.Г. Татарова

ISBN 5-7598-0373-5

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В учебном пособии изложен курс лекций по анализу данных, читавшийся автором в течение ряда лет студентам-социологам II и III курса Государственного университета — Высшей школы экономики (ГУ ВШЭ) (бакалавриат направления «Социология»).

В книге рассматриваются методы, используемые социологом на практике: построение и анализ одномерных и двумерных частотных таблиц; анализ взаимосвязи качественных и количественных переменных с помощью теста Стьюдента и модели однофакторного дисперсионного анализа; построение моделей регрессии; поиск латентных переменных методами факторного анализа, главных компонент, многомерного шкалирования; получение многомерных группировок с помощью кластерного анализа. Подробно описывается, каким образом эти методы могут быть реализованы с помощью пакета SPSS — одной из самых распространенных в мире систем статистической обработки данных социальных исследований.

В последние годы появился целый ряд работ, посвященных описанию того, как можно анализировать статистические данные с помощью пакета SPSS[1]. Среди них некоторые адресованы социологам.

[1] Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. М: DiaSoft, 2002; Наследов А.Д. SPSS: компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005; Плис А.И., Сливина Н. А. Практикум по прикладной статистике в среде SPSS. Ч. 1. Классические процедуры статистики. М.: Финансы и статистика, 2004; Ростовцев П.С, Ковалева Г. Д. Анализ социологических данных с применением статистического пакета SPSS: Учеб.-метод, пособие. Новосиб. гос. ун-т, 2001; Таганов Д.Н. SPSS: статистический анализ в маркетинговых исследованиях. СПб.: Питер, 2003; Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 2002.

Тем не менее, предлагаемая книга имеет шанс занять достойное место в отечественной литературе, на наш взгляд, потому, что автор аккумулировал свой богатый опыт исследователя-социолога, специалиста по анализу данных и педагога, долгие годы работавшего со студентами-социологами и обладавшего талантом хорошо объяснять содержательную сущность математических построений.

В книге, во-первых, каждый из методов анализируется с точки зрения возможности решения тех или иных социологических задач. Приводятся многочисленные примеры использования рассматриваемых методов для анализа данных конкретных социологических исследований, с соответствующей точки зрения изучается специфика каждого метода. Обращается внимание на особенности интерпретации результатов анализа социологических данных. Тщательно разбираются ограничения каждого метода, по мере возможности даются рекомендации по их преодолению.

Во-вторых, при рассмотрении любого метода подробно рассматривается, каким образом на каждом шаге его реализации может использоваться пакет SPSS.

Отметим, что, читая отраженный в книге курс, автор для успешного освоения студентами технических приемов работы с компьютерными программами после каждой лекции предлагал слушателям небольшое задание для самостоятельной работы. Выполнение таких заданий контролировалось на семинарских занятиях. Практика показала эффективность подобного подхода: на базе работы с конкретными социологическими данными у студентов формировались практические навыки использования компьютерных программ при решении социологических задач.

Хочется надеяться, что социолог, прочитавший книгу, при решении стоящей перед ним задачи сумеет выбрать наиболее адекватный метод, определить и обосновать вид соответствующей математической модели, проанализировать (и преодолеть) ее ограничения, выполнить расчеты модели на компьютере, проанализировать математико-статистический смысл полученных результатов, дать соответствующую социологическую интерпретацию.

Книга рассчитана на студентов, прослушавших базовые курсы математики (математический анализ и линейная алгебра), информатики, теории вероятностей и математической статистики, а также основ социологии и методов социологических исследований. Кроме того, может быть полезна социологам для эффективного анализа имеющейся у них информации.

В подготовке книги принимали участие сотрудники и студенты кафедры методов сбора и анализа социологической информации факультета социологии ГУ ВШЭ.

1
глава

ПОДГОТОВКА К АНАЛИЗУ ДАННЫХ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1.1

Социяльное исследовяние и янялиз дянных: основные понятия

Анализ информации, собираемой в процессе эмпирических социологических исследований, представляет собой не просто совокупность технических приемов и методов, позволяющих в той или иной форме визуализировать полученные данные. Анализ данных является ключевым этапом всего исследования, в ходе которого Происходит непосредственная проверка соответствия собранной информации тем моделям социальных явлений, которые, явно или латентно, имеются у социологов. И более того, в ходе анализа формулируются и проверяются новые модели, адекватно отражающие те закономерности, которые есть в собранных данных.

Очевидно, что в случае простой визуализации собранной информации мы имеем дело лишь с *обработкой* социологических данных. Если ставятся задачи построения определенной модели изучаемого социального явления и проверки соответствия этой модели имеющимся данным, можно говорить именно об *анализе* данных.

В ходе как обработки, так и анализа данных часто используют одни и те же технические и математические приемы, однако с гносеологической точки зрения это два разных подхода к данным. В первом случае социолог использует стандартный набор средств (как правило — это одномерные распределения, таблицы, гистограммы и графики) для наиболее наглядной демонстрации полученных данных, которые, при удачном подборе технических средств, вроде бы говорят сами за себя. Во втором случае исследователь выдвигает определенную модель социального явления, демонстрирует соответствие (либо противоречие) данных этой модели и ведет дальнейшую разработку именно модели, отвлекаясь от самих данных.

При работе с социологическими данными используются два основополагающих понятия:

- единица анализа (анкета, случай);
- переменная.

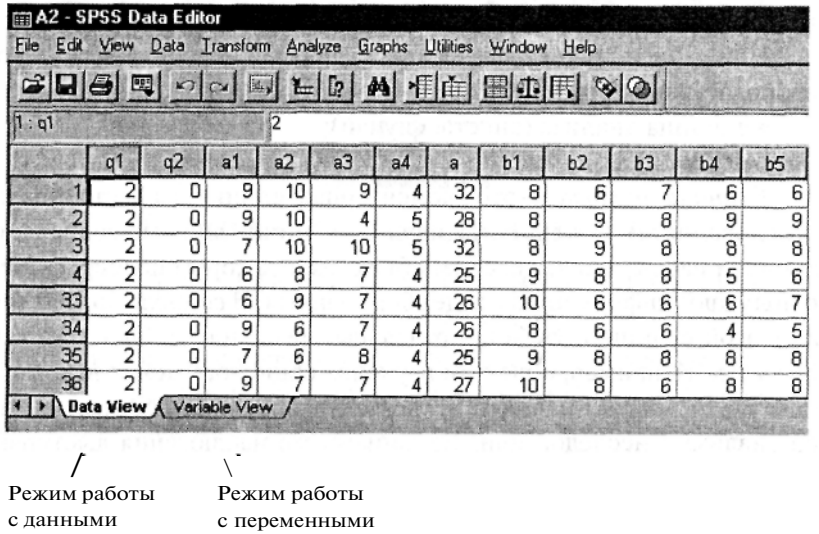
Единица анализа — это элементарная, единичная часть объекта исследования. В большинстве случаев единица анализа совпадает с единицей наблюдения, т.е. с тем объектом, о котором непосредственно получают информацию в ходе сбора данных. В социологии, как правило, этой единицей является отдельный респондент. Однако это не всегда так. Например, объектом изучения социолога может выступать семья как целостная единица и, следовательно, она выступает единицей анализа в исследовании. Единицами же наблюдения выступают члены семей, т.е. отдельные респонденты, о которых, собственно, и собирается информация. Преобразование информации, собранной о единицах наблюдения, в информацию о единицах анализа является самостоятельным и не только техническим этапом исследования.

Переменная — это элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа. Простейшими переменными являются, скажем, пол или зарплата респондента. Ключевыми характеристиками переменной является то, что, с одной стороны, для каждой единицы анализа она имеет одно, вполне определенное значение, а с другой стороны — то, что не все единицы анализа имеют одинаковое значение переменной.

1.2

Представление данных в пакете SPSS

Матрица, в которой представляются данные в программной системе SPSS, изображена на рис. 1.1. Редактор данных состоит из двух частей: таблицы для работы собственно с данными (рис. 1.2) и таблицы работы с переменными (рис. 1.3).



	q1	q2	a1	a2	a3	a4	a	b1	b2	b3	b4	b5
1	2	0	9	10	9	4	32	8	6	7	6	6
2	2	0	9	10	4	5	28	8	9	8	9	9
3	2	0	7	10	10	5	32	8	9	8	8	8
4	2	0	6	8	7	4	25	9	8	8	5	6
33	2	0	6	9	7	4	26	10	6	6	6	7
34	2	0	9	6	7	4	26	8	6	6	4	5
35	2	0	7	6	8	4	25	9	8	8	8	8
36	2	0	9	7	7	4	27	10	8	6	8	8

Рис. 1.1. Представление данных в пакете SPSS

Каждая строка в матрице данных содержит информацию по одной единице анализа. В примере (см. рис. 1.1) в качестве единицы анализа выступает анкета, содержащая ответы одного респондента. Все единицы анализа в матрице данных автоматически нумеруются. Номера располагаются в первой колонке матрицы данных, в остальных колонках — соответствующие значения переменных.

Прежде всего рассмотрим простейшие количественные методы анализа данных. В зависимости от решаемых задач разделим их на три основных типа.

1. Одномерный описательный анализ раскрывает некоторые характеристики частотных распределений.

2. Двумерный описательный анализ связан с описанием формы и силы взаимосвязи между переменными, а также со сравнением значений некоторой переменной в разных социальных группах.

3. Объяснительный анализ направлен на выявление силы влияния переменных друг на друга.

1.3

Построение частотных распределений

Анализ частотных распределений результатов количественного социологического исследования — это первый шаг при обработке собранной информации. Во многих случаях этот анализ не является, строго говоря, анализом данных, а выполняет функции получения общих представлений об изучаемых социальных группах.

Первый шаг одномерного описательного анализа для объяснения какого-то явления — его описание. Результаты любого массового опроса содержат ответы большого числа респондентов на широкий круг анкетных вопросов. Даже в рамках только одного вопроса анкеты объем исходной информации достаточно велик для того, чтобы можно было охватить его одним взглядом и каким-то образом суммировать. Именно задачу сжатия исходной информации, компактного ее представления для дальнейшего осмысления и решают методы одномерного описательного анализа.

Одномерный описательный анализ решает поставленную задачу взаимодополняющими методами:

- построения частотных распределений;
- графического представления поведения анализируемой переменной;

• получения статистических характеристик распределения анализируемой переменной.

В табл. 1.1 представлен фрагмент данных по результатам социологического опроса[1].

Таблица 1.1. Фрагмент матрицы данных из трех переменных в формате SPSS, содержащий результаты социологического опроса

File Edit View Data Transform Analyze Graphs Utilities Window Help

4 :

	q9	q9a	q10	var	var	var	var
1	3	3	3				
2	2	3	4				
3	5	6	4				
4	5	4	4				
5	2	2	4				
6	2	3	3				
7	2	2	3				
8	2	2	2				
9	5	5	4				
10	3	4	4				
11	2	3	3				
12	3	5	5				
13	2	3	3				
14	2	3	3				
15	3	5	5				
16	2	5	3				
17	2	5	5				
18	2	3	3				
19	2	3	4				

Переменная q9, представленная во второй колонке матрицы, содержит ответы респондентов на вопрос анкеты:

q9 Что вы могли бы сказать о своем настроении в последние дни?

1. Прекрасное настроение.
2. Нормальное, ровное состояние.
3. Испытываю напряжение, раздражение.
4. Испытываю страх, тоску.
5. Затрудняюсь ответить.

[1] Опрос проводился ВЦИОМ «Мониторинг общественного мнения» // Мониторинг общественного мнения: экономические и социальные перемены. 2002. № 6.

В матрице данных ответы представлены в виде числовых кодов. Поскольку полностью вся матрица содержит ответы 2407 респондентов, просто просмотр ответов всех опрошенных либо на экране компьютера, либо в распечатанном виде на листах бумаги не дает возможности понять, каково было настроение опрошенных. Получить обобщенную, агрегированную картину ответов на данный вопрос позволяет таблица одномерного частотного распределения, представленная в табл. 1.2.

Таблица 1.2. Одномерное частотное распределение переменной q9

	Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Прекрасное настроение	158	6,6	6,6	6,6
Нормальное, ровное состояние	1185	49,2	49,2	55,8
Испытываю напряжение, раздражение	752	31,2	31,2	87,0
Испытываю страх, тоску	163	6,8	6,8	93,8
Затрудняюсь ответить	149	6,2	6,2	100,0
Total	2407	100,0	100,0	

Построение одномерного частотного распределения в рамках пакета SPSS выполняется с помощью команды *Frequencies,* расположенной в блоке команд *Descriptives* (рис. 1.2). На рис. 1.3 представлено меню команды *Frequencies.*

Таблица 1.2 демонстрирует одномерное частотное распределение переменной q9 в том виде, как это распределение вычисляется командой *Frequencies* пакета SPSS. Рассмотрим информацию, которую дает таблица одномерного частотного распределения.

Колонка *Frequency* (частота) содержит частоты, т.е. то количество респондентов, которые выбрали тот или иной вариант ответа. Таким образом, из табл. 1.2 видно, что вариант «1» выбрали 158 респондентов, вариант «2» — 1185 респондентов и т.д. Последняя стро-

ка в табл. 1.2 — *Total* — в колонке *Frequency* показывает общее количество опрошенных, иными словами — объем выборки.

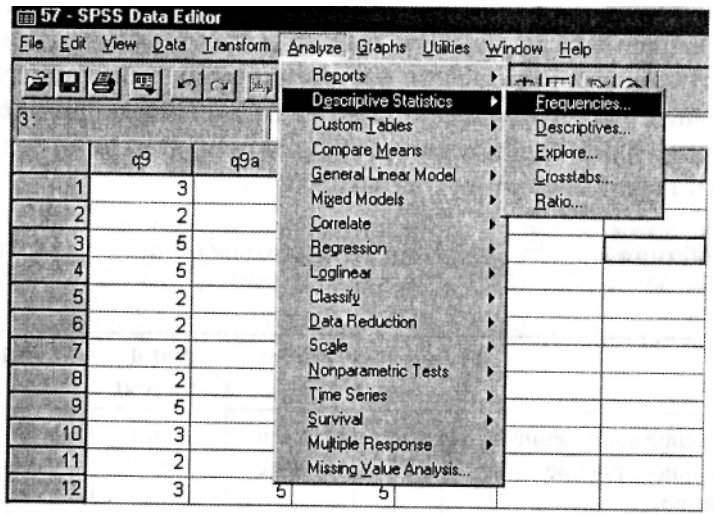


Рис. 1.2. Метод вызова команды построения одномерных частотных распределений в пакете программ SPSS

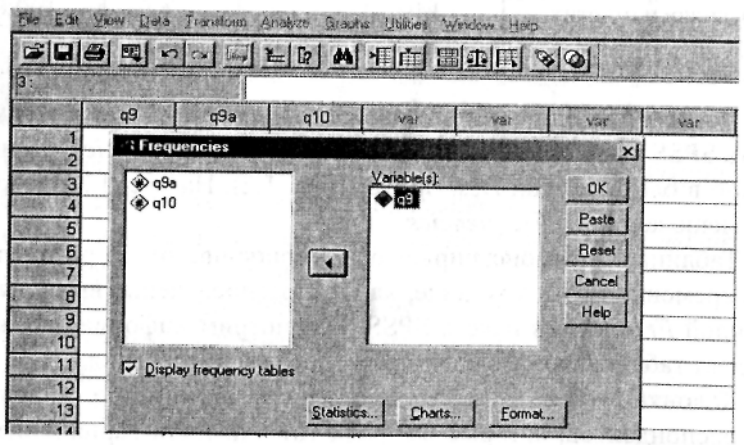


Рис. 7.5. Меню команды построения одномерных частотных распределений

Делать выводы о том, много или мало респондентов отметили при опросе ту или иную градацию в вопросе, опираясь на значения в колонке *Frequency,* невозможно, поскольку необходимо постоянно соотносить эти числа с общим количеством опрошенных. Поэтому удобнее использовать колонку *Percent* (процент), которая содержит процентные значения для каждой из частот. В результате, базируясь на значениях этой колонки, можно сказать, что более распространенным ответом является «нормальное, ровное состояние», поскольку этот вариант отметили 49,2% респондентов.

Колонка *Valid Percent* связана с такой важной в социологической практике характеристикой, как «Отсутствие ответа». Мы знаем, что в ходе любого массового опроса какая-то часть опрашиваемых не отвечает на поставленные вопросы. Причины такого рода «неответов» различны. Это и просто нежелание людей давать информацию по тем или иным показателям. Это и отсутствие собственного мнения по определенным вопросам. Возможности преодоления проблемы «неответов» на этапе сбора социологической информации достаточно подробно рассматриваются у разных авторов, однако очевидно, что эту проблему нельзя решить полностью.

На этапе работы с собранными данными проблема «неответов» может быть сформулирована следующим образом: как анализировать ту информацию, которая может быть квалифицирована как «отсутствие ответа».

Необходимо отметить, что на этот вопрос нет однозначного ответа. В зависимости от характера решаемых задач существуют разные подходы к анализу информации, которая соответствует «неответам». Отметим, что числовые коды, связанные с «неответами», называют *пропущенные данные* (Missing values).

Первый подход к рассмотрению кодов пропущенных данных рассматривает эти коды как равноправные остальным числовым кодам, которые приписаны всем другим типам ответов. Одномерное частотное распределение (см. табл. 1.2) представляет именно такой подход. Действительно, числовой код «5» приписанный варианту «Затрудняюсь ответить», т.е. фактически коду пропущенных данных, представ-

лен точно так же, как и остальные числовые коды. В результате табл. 1.2 демонстрирует нам, что затруднились ответить на поставленный вопрос 149 человек, или 6,2% общего числа опрошенных. При этом и все остальные проценты в табл. 1.2 рассчитаны от *числа опрошенных.*

Альтернативным вариантом построения таблицы одномерного частотного распределения выступает возможность исключения из дальнейшего анализа тех респондентов, которые затруднились дать ответ. Действительно, какие у нас основания рассматривать тех 149 респондентов, которые не дали ответа на поставленный вопрос точно так же, как и тех, кто дал содержательный ответ? Простейшим выходом в рамках данной модели рассуждений является приписывание коду «5» статуса пропущенных данных и исключение из дальнейшего анализа тех, кто дал такой ответ. В табл. 1.3 представлено одномерное частотное распределение, в котором коду «5» приписан статус пропущенных данных.

Таблица 1.3. Одномерное частотное распределение переменной q9

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. Прекрасное настроение	158	6,6	7,0	7,0
	2. Нормальное, ровное состояние	1185	49,2	52,5	59,5
	3. Испытываю напряжение, раздражение	752	31,2	33,3	92,8
	4. Испытываю страх, тоску	163	6,8	7,2	100,0
Total		2258	93,8	100,0	
Missing	5. Затрудняюсь ответить	149	6,2		
Total		2407	100,0		

Таблица 1.3 отличается от табл. 1.2 тем, что код «5» помечен как код пропущенных данных. Колонка *Percent,* как и раньше, содержит процентное *распределение всех опрошенных,* а колонка *Valid Percent* — процентное распределение от того числа респондентов, которые дали ответы, не помеченные кодами пропущенных данных. Иными словами, колонка *Valid Percent* представляет одномерное частотное *распределение от числа ответивших респондентов.*

Вопрос о том, какой из показателей — процент опрошенных, либо процент ответивших необходимо использовать для выявления определенных социологических закономерностей, некорректен. Оба показателя несут определенную информацию и, как правило, используются одновременно, однако их интерпретация существенно различна. Например, если в ходе опроса, за кого собираются голосовать респонденты на предстоящих выборах, мы получим, что за кандидата А собирается голосовать 20% опрошенных и 40% ответивших, то оба этих числа представляют интерес. Действительно, первое число говорит нам, что 20% общего количества взрослого населения собирается поддержать кандидата А на будущих выборах. Поскольку коды пропущенных данных в такого рода опросах получают, как правило, те респонденты, которые говорят, что не будут участвовать в выборах, то число 40% говорит нам о том, сколько процентов может набрать кандидат А в ходе голосования.

Присвоение кода пропущенных данных для переменных в пакете SPSS выполняют по таблице «Описание переменных». На рис. 1.4 приведены таблица «Описание переменных» и показатель, с помощью которого задаются коды пропущенных данных. В нашем примере у переменной q9 задан код пропущенных данных «5».

На рис. 1.5 представлено меню, которое позволяет задавать коды пропущенных данных для выбранной переменной, а также показывает три варианта задания кодов пропущенных данных:

• не определять коды пропущенных данных для переменной *{No missing values);*

• задать несколько (от 1 до 3) значений кодов пропущенных данных (по одному значению в каждом из открытых окон — *Discrete missing values);*

• задать интервал значении кодов пропущенных данных и одно значение кода пропущенных данных *{Range plus one optional discrete missing value)*.

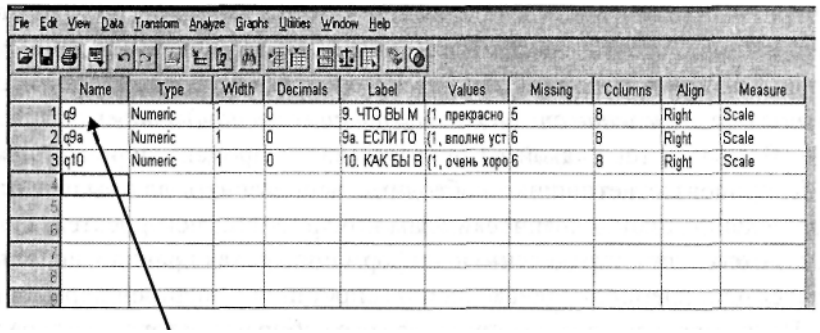


Рис. 1.4. Таблица «Описание переменных» в пакете программ SPSS

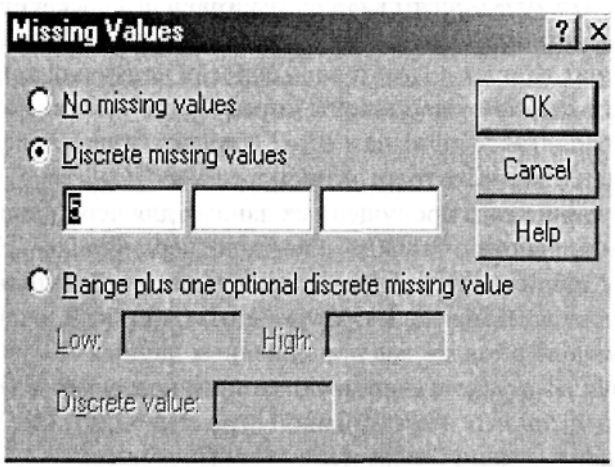


Рис. 1.5. Меню задания кодов пропущенных данных

По какой причине может потребоваться задание не одного, а нескольких кодов пропущенных данных? Эта возможность отражает реальные ситуации, по которым в ходе опроса мы нередко имеем несколько причин того, что респондент не отвечает на вопрос анкеты. Рассмотрим вопрос, в котором присутствует несколько вариантов, каждый из которых объясняет причину, почему респондент не дает ответа.

За кого из кандидатов вы голосовали на прошедших выборах?

1. За кандидата А.
2. За кандидата В.

7. Я не участвовал в голосовании.
8. Я голосовал, но не помню за кого.
9. Я участвовал в голосовании, но не хочу говорить, за кого отдал свой голос.

С точки зрения социологического анализа результатов голосования, коды «7», «8» и «9» должны быть определены как коды пропущенных данных, поскольку эти ответы не содержат информации о том, за кого голосовал респондент. Однако с точки зрения социологических задач весьма интересным может быть изучение, например, характеристик тех респондентов, кто не участвовал в голосовании. Для осуществления анализа этой социальной совокупности мы можем отменить задание кода «7» как кода пропущенных данных и сосредоточиться на анализе данной группы респондентов.

Третий вариант задания кодов пропущенных данных чаще всего встречается в ситуации, когда анализируемая переменная выражена количественно. Иногда в ответах респондентов на вопросы, например, о размере получаемых доходов встречаются данные, которые, строго говоря, не могут быть признаны ошибочными, однако, скорее всего, являются недостоверными. Например, если респондент сказал, что у него 15 детей или что его зарплата 5 млн. руб., эти ответы едва ли корректны. Иными словами, для многих показателей мы можем указать границы, допустимых значений, а те данные, которые выходят за эти границы целесообразно признать пропущенными данными. В меню определения кодов пропущенных данных в разделе *Range plus one optional discrete missing value* (интервал и, возможно, одно значение пропущенных данных) можно задать верхнюю и нижнюю границы

интервала, все значения внутри будут являться пропущенными данными. В этом разделе наряду с интервалом можно задать одно точное значение кода пропущенных данных.

1.4

Графическое представление поведения анализируемой переменной

Наряду с табличным представлением одномерное частотное распределение можно визуализировать в графической форме. Наиболее популярные формы — это столбиковые и круговые диаграммы. На рис. 1.6 и 1.7 представлены эти виды диаграмм для одномерного частотного распределения табл. 1.2.

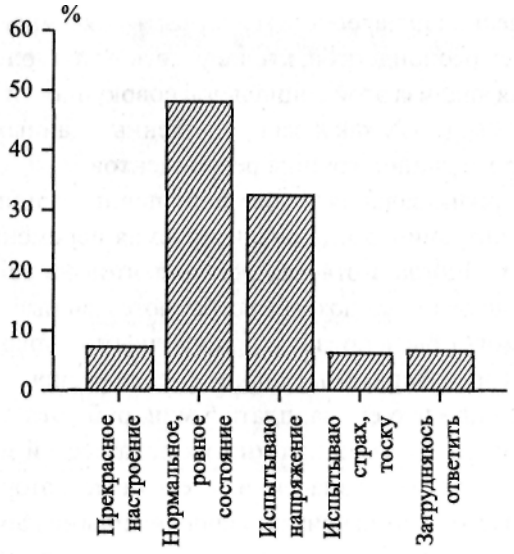


Рис. 1.6. «Что вы могли бы сказать о своем настроении в последние дни?»

Рис. 1.7. «Что вы могли бы сказать о своем настроении в последние дни?»

Диаграммы на рис. 1.6 и 1.7 построены с помощью графических возможностей пакета программ SPSS. Команды для построения графических диаграмм могут выполняться либо непосредственно из модуля вычисления одномерных частотных распределений (команда *Frequencies),* либо из специального блока команд *Graphs,* в котором представлены возможности графического анализа пакета программ SPSS.

В нижней части меню команды *Frequencies* (см. рис. 1.3) есть педаль Charts..., нажатие на которую приводит к вызову меню построения диаграмм одномерного частотного распределения (рис. 1.8).

Графические диаграммы в качестве метода построения одномерных частотных распределений повышают наглядность полученных закономерностей и могут использоваться, прежде всего, для презентации результатов социологических исследований. Какой из видов диаграмм выбрать для каждого конкретного случая — зависит от эстетических пристрастий и существенного значения не имеет.

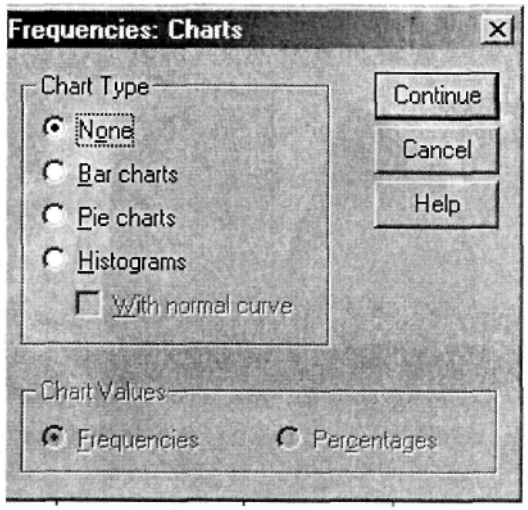


***Рис.** 1.8.* **Меню Charts команды Frequencies**

1.5

Использование статистических характеристик для анализа одномерных распределений

Одной из важнейших характеристик при описании поведения отдельных переменных является показатель средней тенденции. В курсе «Методы социологического исследования» подробно обсуждаются вопросы уровней измерения, используемые в социологических анкетах, а также рассматриваются возможности применения различных мер центральной тенденции для показателей с разным уровнем измерения[2].

[2] См.: Ядов В.А. Социологическое исследование: методология, программа, методы. Самара: Изд-во Самарского ун-та, 1995. С. 98—109.

Возможности использования различных мер средней тенденции для шкал различного типа приведены в табл. **1.4.**

Таблица 1.4. Возможности использования различных мер средней тенденции для шкал различного типа

№ п/п	Уровень измерения	Допустимые меры средней тенденции
1	Номинальный	Мода
2	Порядковый	Мода, медиана
3	Метрический	Мода, медиана, среднее арифметическое

Рассмотрим специфику использования мер средней тенденции для анализа социологических данных на примере среднего арифметического. Среднее арифметическое широко используется в повседневной жизни и не нуждается в дополнительных рекомендациях. Вместе с тем использование *только* среднего арифметического для описания значений переменной таит определенную опасность.

Говоря о среднем значении некоторой переменной мы, по сути дела, заменяем рассмотрение всей совокупности значений этой переменной единственным показателем, фактически предполагая, что значение этого показателя достаточно хорошо описывает поведение анализируемой переменной. Очевидно, что в данном случае среднее значение выступает в качестве определенной модели значений переменной.

Несомненно, что среднее арифметическое переменной представляет совокупность значений этой переменной неполно и с возможными ошибками. Зная, например, среднее значение зарплаты среди совокупности опрошенных, мы не можем достаточно точно определить зарплату того или иного респондента. Только в том случае, когда все значения переменной одинаковы, среднее значение абсолютно точно отражает поведение переменной. Во всех других случаях среднее арифметическое как модель переменной является моделью неточной. Следовательно, для нас важно знать не только значение данной модели, но и степень точности, качества этой модели.

Рассмотрим данные о заработной плате пяти респондентов, полученные в ходе социологического исследования (табл. 1.5).

Таблица 1.5. Данные о средней заработной плате, среднее значение заработной платы, расхождение среднего и фактических данных

№ п/п	Значение заработной платы, руб.	Среднее значение, руб.	Расхождение реальной зарплаты и среднего значения, руб.
1	17 000	15 500	1500
2	13 000	15 500	-2500
3	18 000	15 500	2500
4	15 000	15 500	500
5	14 500	15 500	-1000

Данные, приведенные в табл. 1.5, можно представить в виде условной формулы:

Реальные данные = Модель + Остаток.

Расхождение реальных данных и модели в этой формуле называется остатком.

В каком случае модель средней зарплаты будет с небольшой погрешностью описывать реальные данные? Ключевым вопросом при анализе данных с помощью какой бы то ни было модели является оценка того, насколько хороша модель. Остатки дают нам эффективный инструмент для оценки качества модели: очевидно, что модель тем лучше, чем меньше остатки.

Таким образом, наряду со средней характеристикой, которая удобна тем, что дает нам картину (вернее, часть картины) поведения значений переменной, целесообразно иметь и еще одно число, которое оценивало бы качество средней как модели. Функции такой характеристики выполняют *меры разброса,* наиболее известна среди них *дисперсия*.

Фактически дисперсия представляет собой не что иное, как сумму квадратов остатков, деленную на количество наблюдений:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (1.1)$$

где x_i — значение переменной x для г'-го респондента; x — среднее значение переменной x; n — количество опрошенных респондентов.

Недостатком дисперсии является то, что эту величину трудно оценить интуитивно. Данные, представленные в табл. 1.5, имеют понятные нам единицы измерения — рубли. Поэтому мы сразу можем оценить, что за величина остатка, скажем, у респондента 4 — 500 руб. Понятна нам и размерность среднего показателя — 15 500 руб. Мы можем интерпретировать это значение, соотнося его с нашим знанием социальной действительности.

В то же время значение дисперсии для данных табл. 1.5 составляет 4 000 000. Едва ли мы можем, хотя бы на качественном уровне, оценить, большая эта величина или маленькая. Это значение не дает нам ответа на главный вопрос — хороша ли наша модель среднего арифметического, т.е. средней зарплаты. Причина того, что дисперсия плохо приспособлена для ответа на вопрос о качестве модели среднего, в том, что остатки берутся в квадрате. Для того чтобы преодолеть это затруднение, используют два производных от дисперсии показателя — стандартное отклонение и стандартная ошибка среднего.

Стандартное отклонение — это корень квадратный из дисперсии. Стандартное отклонение для данных табл. 1.5 — 2000.

$$S = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{n}}. \qquad \mathbf{(1.2)}$$

Стандартная ошибка среднего (со. *x)* тоже широко используется для решения задачи оценки качества среднего как модели с несколько иной стороны: она дает возможность соотнести величину x с генеральным математическим ожиданием. Последнее с вероятностью 0,95 лежит в интервале (x ± 2с.о.x).

$$\text{с.о.}\bar{x} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n}. \qquad \textbf{(1.3)}$$

По табл. 1.5 значение стандартной ошибки среднего составляет 894. Таким образом, можно утверждать, что с вероятностью 0,95 математическое ожидание зарплаты должно лежать в интервале 15 500 ± ± 2 х 894, или от 13 712 до 17 288 руб. (см. п. 1.7).

Подводя итог, необходимо подчеркнуть, что использование среднего арифметического без указания одного из показателей качества среднего как модели (дисперсии, стандартного отклонения, либо стандартной ошибки среднего) не дает возможности удовлетворительной интерпретации полученного среднего.

Проведенные рассуждения о необходимости дополнения характеристики средней тенденции показателем качества этой модели справедливо и в отношении тех переменных, которые измерены на номинальном или порядковом уровне. Для номинальных переменных мерой центральной тенденции может выступать только мода, т.е. наиболее часто встречающееся значение переменной. Мода не имеет какого-то показателя разброса. Определенной характеристикой может считаться лишь само процентное значение модальной величины. В качестве примера рассмотрим табл. 1.6, в которой приведено одномерное частотное распределение респондентов, проживающих в населенных пунктах разного типа.

В табл. 1.6 модальным значением является «2». Тот факт, что на эту градацию приходится 53,7% всех опрошенных респондентов, говорит о том, что на все остальные градации приходится лишь 46,7%, что может указывать на разброс значений. Однако данное указание достаточно слабо, поскольку не показывает, как именно разбросаны данные по другим градациям анализируемой переменной.

Для переменных, измеренных на порядковом уровне, основной мерой центральной тенденции является медиана. Рассчитаем медиану для переменной q23: *Насколько вы удовлетворены состоянием своего здоровья?,* которая фиксирует ответы респондентов по 7-балльной порядковой шкале (табл. 1.7).

Таблица 1.6. Одномерное частотное распределение переменной CITY «Тип населенного пункта»

№ п/п	**Населенный** пункт	**Frequency**	**Percent**	**Valid Percent**	**Cumulative Percent**
1	Москва	520	21,5	21,5	21,5
2	Областной центр	1300	53,7	53,7	75,2
3	Малый город в области	350	14,5	14,5	89,7
4	Сельский населенный пункт	250	10,3	10,3	100,0
Total		2420	100,0	100,0	

Медиана является такой точкой на шкале, которая делит всю совокупность опрошенных на две равных части — тех, кто отметил градации меньше этой точки (либо равные ей), и тех, кто отметил градации больше этой точки. Из табл. 1.7 видно, что в вопросе q23 градации 1, 2, 3 и 4 отметили 50,4% респондентов, и, следовательно, градация «4» является медианой.

Таблица 1.7. Одномерное частотное распределение переменной q23

№ п/п		**Frequency**	**Percent**	**Valid Percent**	**Cumulative Percent**
1	Полностью удовлетворен	336	12,2	12,2	12,2
2		355	12,9	12,9	25,1
3		388	14,1	14,1	39,2
4		308	11,2	11,2	50,4
5		322	11,7	11,7	62,1
6		360	13,1	13,1	75,2
7	Совершенно неудовлетворен	685	24,9	24,9	100,0
Total		2754	100,0	100,0	

Наиболее распространенным показателем, характеризующим разброс значений переменной, измеренной на порядковом уровне, является *квартильное отклонение*. Чтобы понять смысл этого показателя, необходимо уяснить значение понятия *квартиля*.

Квартиль является естественным развитием медианы, с той разницей, что квартильное разбиение делит всех респондентов не на 2, а на 4 части. Первый квартиль — это такая точка на шкале, значения меньше (либо равные) которой отметили 25% опрошенных. Второй квартиль — точка, меньше которой отметили 50% опрошенных (следовательно, второй квартиль совпадает с медианой). Наконец, третий квартиль — точка, градации меньше которой отметили 75% опрошенных.

В примере табл. 1.7 первый квартиль — это градация «2» переменной q29, поскольку градации «1» или «2» отметили 25,1% опрошенных. Второй квартиль (медиана) — «4», а третий квартиль — градация «7». Квартильное отклонение — это разница между третьим и первым квартилями. В рассматриваемом примере квартильное отклонение равно 5. При том что в целом рассматриваемая переменная q23 имеет 7 градаций, квартильное отклонение, равное 5, может рассматриваться как достаточно большое, если рассматривать шкалу как метрическую, можно сделать вывод, что модель средней тенденции (в данном случае — медиана) неточно отражает поведение переменной, поскольку много респондентов имеют значения переменной, существенно отличающиеся от медианы.

Обдумывая логику разбиения совокупности значений переменной на 2 (медиана), либо на 4 (квартили) равнонаполненных части, вполне можно поставить задачу разбиения и на 5, и на 10, и вообще на любое количество равных частей. Действительно, при анализе социологических данных иногда используются *квинтильное* (на 5 равных частей) и *децильное* (на 10 равных частей) разбиения. Соответственно применительно к таким разбиениям можно использовать такие меры разброса, как квинтильное и децильное отклонения.

Вызов блока вычисления мер средней тенденции и разброса в рамках команды построения одномерных частотных распределений проводится с помощью кнопки *Statistics* в нижней части главного меню команды *Frequencies* (см. рис. 1.3). Нажатием этой кнопки вызывают на экран меню Statistics (рис. 1.9).

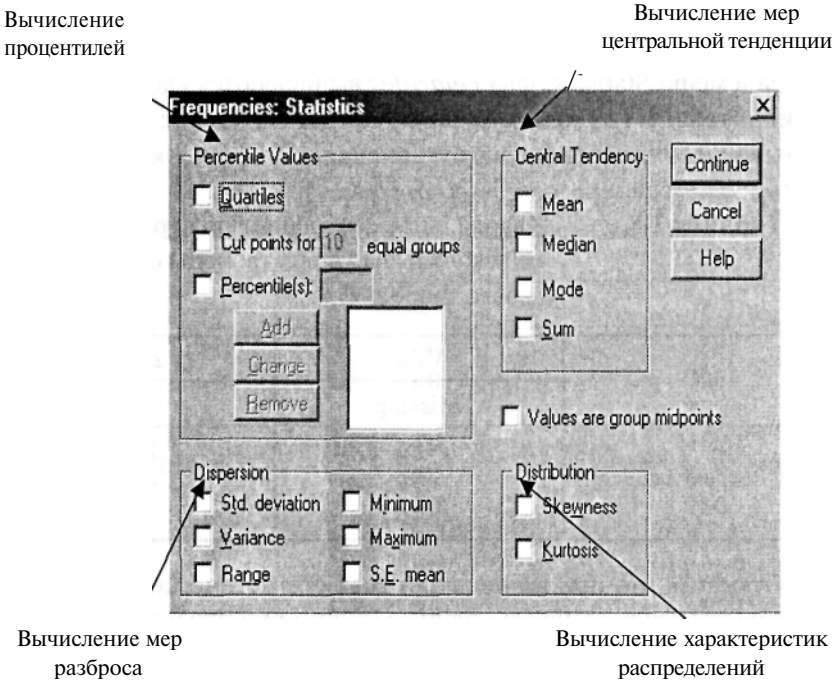


Рис. 1.9. Меню Statistics команды вычисления одномерных частотных распределений

Меню Statistics состоит из четырех отдельных блоков:

- вычисление мер центральной тенденции (Central Tendency);
- вычисление процентилей (квартили, квинтили и т.п.) (Percentile Values);
- вычисление мер разброса (Dispersion);
- вычисление характеристик распределений (Distribution).

Выбор необходимых окон в каждом из блоков приводит к вычислению соответствующих статистических показателей. Отметим, что в рамках меню Statistics команды *Frequencies* невозможны вычисления Показателей квартильного (квинтильного, децильного и т.п.) отклонения. Вычисляются только сами точки процентильного разбиения.

Проиллюстрируем это вычислением квартилей для переменной q23, одномерное частотное распределение см. в табл. 1.7. В случае выбора в меню Statisics окна *Quartiles* вычисляются следующие статистики (табл. 1.8). Данные табл. 1.8 представляют все необходимое для вычисления квартильного отклонения.

Таблица 1.8. Квартильное разбиение для переменной «Насколько вы удовлетворены состоянием своего здоровья?»

N	Valid	2754
	Missing	0
Percentiles	25	2,00
	50	4,00
	75	7,00

Отметим, что при вычислении всех статистических характеристик только значения, не отмеченные кодами пропущенных данных.

Полезным и нередко используемым показателем при анализе количественных переменных является *децильное отношение*. Продемонстрируем использование данного показателя на примере. В ходе социологического исследования, проведенного в сентябре 2003 г. ВЦИОМ, респондентам, в частности, задавался вопрос о размере их заработной платы на основном месте работы. При анализе данного показателя возникла потребность изучить, насколько высока неоднородность значений получаемой респондентами заработной платы.

В качестве первого шага для решения этой задачи было построено децильное разбиение исследуемого показателя (табл. 1.9).

О чем говорят материалы табл. 1.9? О том, что заработную плату до 1800 руб. получают 10% опрошенных (граница первого дециля), а также о том, что 10% опрошенных получают зарплату в размере 15 000 руб. и выше (граница десятого дециля).

Децильное отношение — это отношение десятого дециля к первому. Этот показатель демонстрирует, во сколько раз больше получают 10% наиболее высокооплачиваемых респондентов по сравнению с 10% наименее оплачиваемых. В нашем примере децильное отношение составляет 8,3, что показывает степень неоднородности заработной платы.

Таблица 1.9. Децильное разбиение для переменной «Размер вашего заработка за последний месяц»

N	Valid	1079
	Missing	0
Percentiles	10	1800
	20	3000
	30	3600
	40	4500
	50	6000
	60	7500
	70	9000
	80	11 100
	90	15 000

1.6

Стандартизация показателей

Одной из задач, возникающих при одномерном анализе социологических данных, является сопоставление значения определенной переменной для конкретного респондента со средним значением этой переменной в какой-то социальной группе. Например, если результаты опроса показали, что некий респондент за последний месяц потратил 70 руб. на покупку хлеба, и не зная средней величины затрат на покупку данного вида товаров в том регионе, где проживает респондент, мы не можем сказать, много или мало денег потратил респондент на хлеб. Величина «70 рублей» может быть осознана и проинтерпретирована только в сравнении с затратами других респонден-

тов. Для того чтобы сразу оценить относительную величину того или иного количественного показателя для конкретного респондента, используется метод стандартизации исходных данных.

Существует несколько различных подходов к стандартизации данных, но самый распространенный — это так называемая Z-стандартизация. Вычисление стандартизованной величины Z_{xj} для значения переменной *x* для i'-го респондента проводится по формуле

$$Z_{xi} = \frac{x_i - \overline{x}}{S}, \qquad (1.4)$$

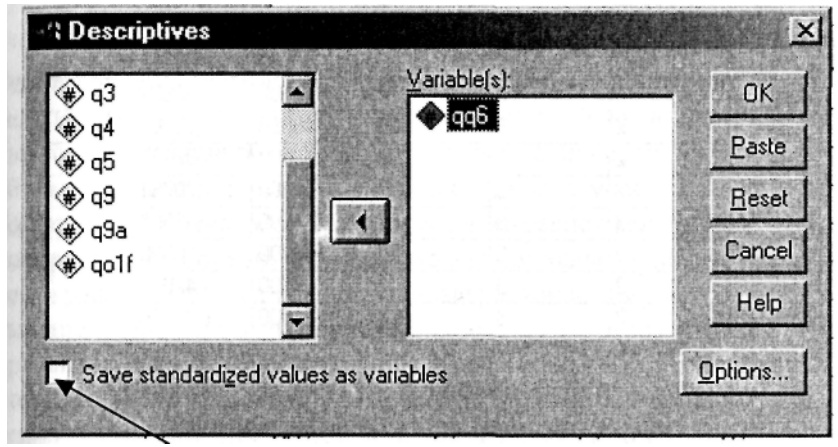
где *x.*— значение переменной для i'-го респондента; x — среднее значение переменной *x; S* — стандартное отклонение для переменной *x.*

Значение показателя Z_{xj} для i'-го респондента более информативно с точки зрения задачи относительного положения данного респондента, чем значение исходной переменной *x.*. Действительно, из формулы (1.4) следует, что если для i'-го респондента Z_{xj} положительно, данный респондент имеет значение переменной *x.* большее, чем средний опрошенный респондент. Таким образом, знак Z. сразу говорит нам о положении респондента (по переменной *x)* относительно других опрошенных.

После того как мы выяснили, большее или меньшее значение по переменной *x* имеет данный респондент по сравнению с другими опрошенными, необходимо узнать, насколько это значение больше или меньше, чем у других респондентов. Из свойств стандартного нормального распределения следует, что 68% Z_{xi} должны лежать в интервале от -1 до 1, а 95% — в интервале от -2 до 2. Таким образом, если по модулю значение % меньше единицы, мы можем сказать, что значение переменной *x* для данного респондента вполне типично. Если значение Z_{xi} по модулю находится от 1 до 2, можно говорить, что данный респондент по рассматриваемому показателю значительно отличается от среднего респондента. Наконец, если Z_{x}. по модулю превосходит 2, можно утверждать, что данный респондент резко отличается от среднего[3].

[3] Все последние утверждения, строго говоря, справедливы лишь в ситуации, когда распределение исходной переменной *x* не сильно отличается от нормального. Вместе с тем практика показывает, что в абсолютном большинстве случаев это именно так.

В блоке команд *Descriptives statistics* есть команда, с помощью которой можно провести Z-стандартизацию для отобранных переменных. Это команда *Descriptives* (рис. 1.10).



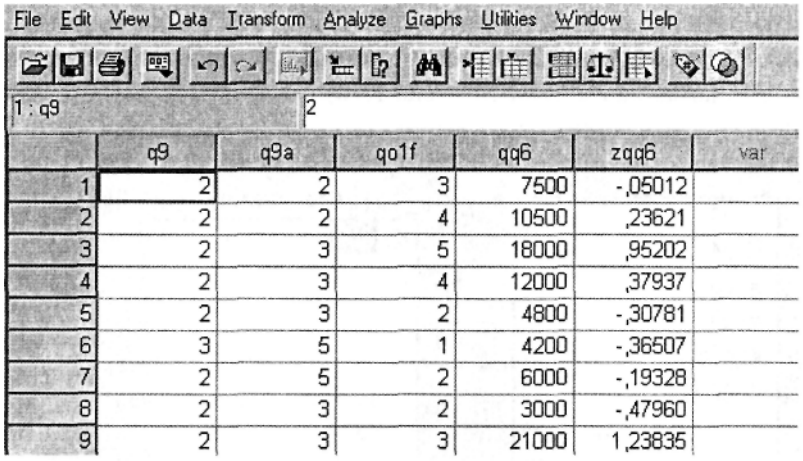
Выбрать окно, для получения стандартизованных значений отобранных переменных

Рис. 1.10. Главное меню команды Descriptives

Команда *Descriptives* в значительной степени дублирует функции команды *Frequencies,* поскольку отвечает за вычисление одномерных статистических характеристик (мер средней тенденции и мер разброса) для выбранных переменных. Важной задачей, которую решает команда *Descriptives,* является вычисление Z-стандартизованных значений. В нижней части главного меню команды находится окно, при выборе которого SPSS автоматически вычисляет стандартизованные значения для выбранных переменных.

На рис. 1.11 приведена матрица данных, которая получается после выполнения стандартизации переменной qq6[4].

[4] Переменная qq6 содержит ответы на вопрос: каков был размер вашего заработка, доходов от основной работы, полученных в прошлом месяце (после вычета налогов).



File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : q9 | 2

	q9	q9a	qo1f	qq6	zqq6	var
1	2	2	3	7500	-,05012	
2	2	2	4	10500	,23621	
3	2	3	5	18000	,95202	
4	2	3	4	12000	,37937	
5	2	3	2	4800	-,30781	
6	3	5	1	4200	-,36507	
7	2	5	2	6000	-,19328	
8	2	3	2	3000	-,47960	
9	2	3	3	21000	1,23835	

Рис. 1.11. Матрица данных, содержащая значения переменной qq6 и стандартизованной переменной zqq6

Переменная zqq6 представляет собой стандартизованное значение переменной qq6. Использование стандартизованной переменной позволяет сказать, что величина зарплаты у первого респондента приблизительно равна среднему значению по массиву опрошенных. А вот размер заработной платы у респондента номер 9 значительно выше, чем у среднего респондента.

Использование стандартизованных переменных весьма полезно и при решении задачи сопоставления показателей, измеренных в разных единицах. Например, в нашем распоряжении есть данные по опросам в России и США, и получается, что у российского респондента А средняя зарплата составляет 9000 руб. в мес, а у американского респондента В — 2000 долл. в мес. Очевидно, что, не зная значений средней зарплаты в России и США, мы не можем сказать, выше ли респондент А респондента В, с точки зрения средней заработной платы, в их социальном кругу.

Если у нас есть возможность сопоставлять не исходные данные о величинах зарплат, а соответствующие стандартизованные показатели, мы легко можем ответить на поставленный вопрос.

1.7

Интервальное оценивание

Одномерное частотное распределение позволяет констатировать определенные закономерности в той совокупности респондентов, которые были опрошены в ходе проведенного исследования. Однако объектом социологического исследования выступает, в абсолютном большинстве случаев, не та совокупность респондентов, которая непосредственно опрашивается, а какая-то социальная либо социально-демографическая группа. Опрошенные респонденты выступают лишь как представители этой группы, как выборка, которая призвана репрезентировать поведение группы в целом. Поэтому возникает закономерный вопрос: как соотносится одномерное распределение, характеризующее поведение той или иной переменной в выборочной совокупности, с поведением этой переменной во всей анализируемой социальной общности? Иными словами, как можно перенести результат, полученный для выборки, на всю изучаемую генеральную совокупность?

Поскольку размер обследованной выборочной совокупности существенно меньше, чем генеральная совокупность, то перенесение результатов с выборочной совокупности на генеральную возможно лишь с определенной точностью. Иными словами, если в ходе опроса получено, что в выборочной совокупности 6,9% опрошенных ответили, что они «в целом довольны своей жизнью», это вовсе не значит, что во всей генеральной совокупности своей жизнью довольны именно 6,9% населения. Выборочный метод дает нам правило, которое позволяет, зная значение определенного параметра в выборочной совокупности, оценить возможное значение этого параметра в генеральной совокупности[5].

[5] Подробно вопросы генерализации результатов социологических опросов см.: Батыгин Г.С. Лекции по методологии социологических исследований. М.: Аспект-Пресс, 1995. С. 145—189.

Теоремы математической статистики говорят нам, что если выборка исследования реализуется с соблюдением определенных требований, результаты, полученные на выборке, могут быть перенесены на генеральную совокупность доверительных интервалов. Таким образом, если в выборочной совокупности оказалось 6,9% респондентов, довольных своей жизнью, в генеральной совокупности таких респондентов будет (6,9 ± Д)%. Величина А называется максимальной ошибкой выборки, а интервал (6,9 - А, 6,9 + А) — доверительным интервалом; А вычисляется по формуле

$$\Delta = z\sqrt{\frac{S^2}{n}}, \qquad (1.5)$$

где z — критические точки нормального распределения; S^2 — дисперсия анализируемого показателя; n — объем выборки.

Нетрудно видеть, что $\sqrt{\frac{S^2}{n}} = \text{с.о.}\bar{x}$ **(см. 1.3).**

2 глава

ВЗАИМОСВЯЗЬ ПЕРЕМЕННЫХ

Описанная в первой главе обработка данных отдельно по каждой из переменных является, как правило, первым, исходным этапом анализа собранной информации. Вместе с тем наиболее интересные вопросы, занимающие социологов, связаны с одновременным анализом значений более одной переменной.

Обычный подход к анализу собранных данных предполагает формирование моделей типа: «социальные группы с разным уровнем образования (уровнем дохода, местом жительства и т.п.) отличаются по характеру проведения досуга (политическим предпочтениям, степени удовлетворенности жизнью и т.п.)». Другими словами, допускается, что существует переменная (скажем, принадлежность к определенной социальной группе), которая объясняет поведение других переменных. Таким образом, в этой модели у нас есть причина и есть следствие. В традиционной терминологии объясняющие переменные называются *независимыми,* а объясняемые переменные—*зависимыми*.

В простейшем случае анализа двух переменных модель влияния представлена на рис. 2.1. Здесь влияние одной независимой переменной ставится в центр изучения, а влияние других переменных на зависимую переменную выступает в качестве причины, формирующей остатки, т.е. не объясняемую данной моделью часть поведения зависимой переменной. Если остаток невелик, можно считать, что наша модель описания поведения зависимой переменной с помощью независимой переменной достаточно точно объясняет собранные данные.

Функцию меры качества модели взаимосвязи переменных выполняют *коэффициенты связи*. Ниже мы подробно остановимся на коэффи-

циентах связи, их особенностях и методах вычисления, но подход одинаков — чем выше коэффициент, тем больше взаимосвязь переменных, тем выше качество модели, и тем, соответственно, меньше остаток.

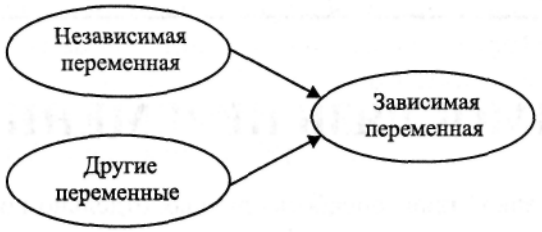


Рис. 2.1. Объясняющая модель поведения зависимой переменной

2.1

Двумерные таблицы

К наиболее часто используемым инструментам изучения взаимосвязи двух переменных относятся методы анализа *таблицы сопряженности*. Анализ таблицы является весьма простым и наглядным, и вместе с тем эффективным инструментом изучения одновременно двух переменных. Двумерная таблица сопряженности для переменных ql 2 и q2 (табл. 2.1) составлена по данным исследования «Мониторинг социальных и экономических перемен в России», которые получены из ответов на вопросы:

qlO Как бы вы оценили в настоящее время материальное положение вашей семьи?

1. Хорошее, очень хорошее.
2. Среднее.
3. Плохое, очень плохое.
4. Затрудняюсь ответить.

ql2 Как бы вы оценили в целом политическую обстановку в России?

1. Благополучная, спокойная.
2. Напряженная.
3. Критическая, взрывоопасная.
4. Затрудняюсь ответить.

Таблица 2.1. Таблица сопряженности для переменных q10nq12

ql0KaK6biBbi оценили в настоящее время материальное положение вашей семьи?	ql2 Как бы вы оценили в целом политическую обстановку в России?				Всего
	благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	12	48	47	17	124
Среднее	20	478	666	138	1302
Плохое, очень плохое	11	160	701	И	953
Затрудняюсь ответить	0	6	15	7	28
Всего	43	692	1429	243	2407

В табл. 2.1 на пересечении строк и столбцов находятся числа, показывающие, какое количество единиц анализа (в данном случае — респондентов) обладают одновременно данными градациями по переменным qlO и q12. Например, на пересечении первой строки и второго столбца стоит число 48 — это значит, что градацию «1» переменной q 10 (считают материальное положение своей семьи хорошим или очень хорошим) и градацию «2» переменной q12 (считают политическую обстановку в России напряженной) одновременно отметили 48 человек.

Внизу таблицы сопряженности располагаются суммарные данные по всем колонкам, а с правого края таблицы — аналогичные сум-

мы по всем строкам. Иными словами, сбоку справа и снизу находятся одномерные частотные распределения для переменных, использованных в таблице.

Можно ли по данным табл. 2.1 сразу дать ответ на вопрос о наличии зависимости между переменными qlO и q12? По всей вероятности, нет — стоящие в клетках таблицы числа ничего особенного не демонстрируют. Поставим вопрос иначе — а что, собственно, мы ищем? По всей видимости, при наличии зависимости между переменными qlO и q12 при разных значениях переменной q10 поведение данных по переменной q12 будет различным. Если говорить о примере табл. 2.1 — это значит, что респонденты, по-разному оценивающие свое материальное положение, будут по-разному оценивать политическую обстановку в России.

Если бы количество респондентов, имеющих различные значения переменной qlO, было одинаковым, в табл. 2.1 можно было бы сравнивать между собой строки и оценить, насколько схожи значения в клетках, располагающихся в одной колонке. Однако количество респондентов по строкам сильно разнится, поэтому для такого сравнения построим таблицу, в клетках которой располагаются не абсолютные количества единиц анализа, а процент от сумм по строкам. Другими словами, число респондентов в каждой строке берется за 100% и от этого числа считается процент в каждой клетке таблицы. Таким образом, мы как бы нормируем каждую строку таблицы и получаем возможность сравнения распределений по строкам (табл. 2.2).

Таблица 2.2 показывает, что оценка политической ситуации в России значительно отличается по группам респондентов, по-разному оценивающих материальное положение своей семьи, и, следовательно, имеется определенная зависимость между переменными qlO и q12.

При анализе зависимостей двух переменных важнейшим является вопрос о том, какую из переменных считать зависимой, т.е. подверженной влиянию, а какую — независимой, т.е. влияющей. В табл. 2.1 и в последующих рассуждениях предполагалось, что оценка материального положения семьи — независимая переменная, иными словами, она влияет на оценку политической ситуации, которая, следовательно, выступает зависимой переменной. Если мы поменяем местами переменные в модели и будем считать, что оценка политической ситуации оказывает влияние на оценку материального положения семьи, целесообразно изменить таблицу и проводить нормирование не от сумм по строкам, а от сумм по колонкам. Таблица 2.3 построена именно таким образом, т.е. использованы данные табл. 2.1, но нормированные по колонкам.

Таблица 2.2. Таблица сопряженности переменных ql0 и q12, %

qlO Как бы вы оценили в настоящее время материальное положение вашей семьи?	q12 Как бы вы оценили в целом политическую обстановку в России?				Всего i
	благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	9,7	38,7	37,9	13,7	100,0
Среднее	1,5	36,7	51,2	10,6	100,0
Плохое, очень плохое	1,2	16,8	73,6	8,5	100,0
Затрудняюсь ответить	0	21,4	53,6	25,0	100,0
Всего	1,8	28,7	59,4	10,1	100,0

Очевидно, что при решении вопроса о зависимости между переменными qlO и q12 при анализе табл. 2.3 необходимо сравнивать распределения по разным колонкам таблицы, а не по строкам, как при анализе таблицы, представленной на рис. 2.2. Такое сравнение показывает, что среди респондентов, оценивающих политическую ситуацию в России как критическую, материальное положение своей семьи оценивают как плохое 49,1% респондентов (колонка 3, строка 3 табл. 2.3). В то же время среди оценивающих политическую ситуацию оптимистичнее, как напряженную, материальное положение своей семьи считают плохим 23,1% респондентов (колонка 3, строка 2 табл. 2.3).

Таблица 2.3. Таблица сопряженности переменных q10nq12, %

q 10 Как бы вы оценили в настоящее время материальное положение вашей семьи?	q12 Как бы вы оценили в целом политическую обстановку в России?				Всего
	благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	27,9	6,9	3,3	7,0	5,2
Среднее	46,5	69,1	46,6	56,8	54,1
Плохое, очень плохое	25,6	23,1	49,1	33,3	39,6
Затрудняюсь ответить	0	0,9	1,0	2,9	1,2
Всего	100,0	100,0	100,0	100,0	100,0

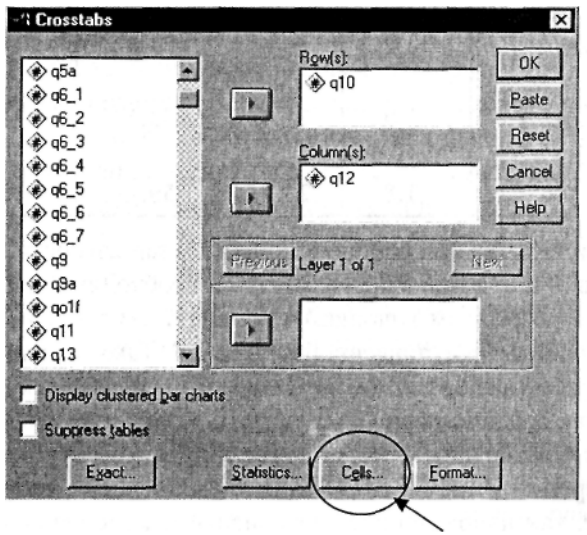


Рис. 2.2. Меню команды Crosstabs пакета SPSS

При анализе таблиц сопряженности крайне важно помнить, что мы, по сути дела, ищем наличие (или отсутствие) определенных *статистических,* а не причинно-следственных зависимостей. Вопрос о том, какая из переменных является причиной, т.е. оказывает влияние, а какая меняется вследствие этой причины, не может быть решен не только с помощью анализа таблиц, но и любым другим формально-статистическим методом. Это вопрос понимания той модели, которую мы проверяем методами построения таблиц либо другими статистическими приемами. Но результатом такой проверки не может быть утверждение: «наша модель верна», либо «наша модель неверна». Утверждать мы можем лишь то, что данные не противоречат (или, наоборот, противоречат) построенной модели, что само по себе отнюдь не является гарантией ее справедливости.

Иллюстрацию этой мысли можно найти у О. Генри. В рассказе «Вождь краснокожих» главный герой предложил изящную модель для ответа на вопрос о том, почему дует ветер — потому, что деревья качаются. Если собрать данные о ветре и поведении деревьев во время ветра, любой статистический метод покажет, что данные ни в коем случае не противоречат этой модели, что, видимо, и послужило Джиму основанием для столь глубокомысленного вывода.

2.2

Обработка данных на компьютере

Построение таблиц сопряженности в пакете программ SPSS осуществляется с помощью команды *Crosstabs.* На рис. 2.2 показано меню этой команды. В списке всех переменных необходимо выбрать те из них, значения которых будут идти по строкам таблиц (окно *Row(s)),* и те, которые пойдут по колонкам (окно *Column(s)).*

Выбранные в меню рис. 2.2 переменные (qlO и q12) определяют те переменные, которые будут представлены в получаемой таблице, но не определяют, каков же именно будет вид таблицы. Действитель-

но, в табл. 2.1, 2.2 и 2.3 использовались переменные q1O и q12, однако таблицы значительно различались. То, какие характеристики будут присутствовать в задаваемой таблице, определяется в меню, которое вызывается нажатием клавиши Cells... (рис. 2.3).

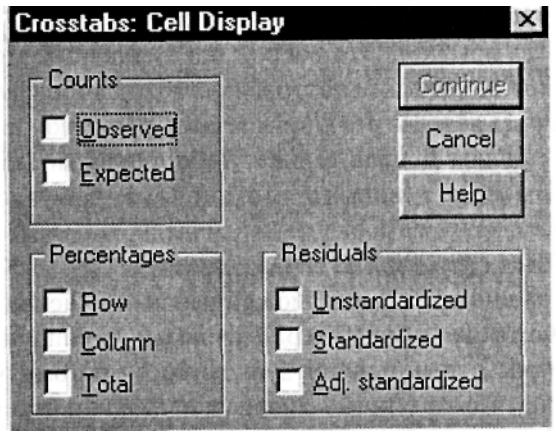


Рис. 2.3. Меню Cells команды Crosstabs

Как видно из рис. 2.3, меню Cells содержит 8 окон, причем каждое определяет параметр, который определит вид получаемой таблицы сопряженности. Таким образом, если выбрать все предлагаемые параметры, можно получить таблицу, в каждой из клеток которой будет восемь разных чисел.

В меню (см. рис. 2.3) все предлагаемые параметры разбиты на три группы:

1) Counts (значения);
2) Percentages (проценты);
3) Residuals (остатки).

Если в блоке *Counts* поставить галочку в окне *Observed,* получим таблицу, в клетках которой показано количество единиц анализа (т.е. табл. 2.1). Если в блоке *Percentages* выбрать окно *Row* (строка), получим таблицу с процентами, нормированными по строкам (табл. 2.2).

Если, наконец, выбрать окно *Column* (колонка), получится табл. 2.3. Подчеркнем, что выбор окон в меню Cells команды *Crosstabs* происходит исключительно исходя из решаемых задач. Можно выбрать любое количество окон, и даже ни одного. Последний вариант, на первый взгляд, кажется странным, ведь в этом случае мы не получим таблицы! Однако в дальнейшем будет показано, что такая ситуация имеет свой смысл.

2.3

Коэффициенты связи для номинальных переменных

В настоящее время существует множество числовых показателей для измерения степени и характера взаимосвязи двух переменных — коэффициентов связи. Наиболее известный из них — коэффициент χ^2.

2.3.1

Коэффициент χ^2

Оказывается, что сформулировать ответ на вопрос: что такое зависимость между ответами на два вопроса анкеты, удается довольно просто — от обратного. Другими словами, «зависимость есть отсутствие независимости». Этот, на первый взгляд, абсолютно не конструктивный ответ сильно продвигает нас вперед, поскольку в теории вероятностей существует строгий подход к определению независимости двух событий.

Два события считаются независимыми в том случае, если вероятность того, что они произойдут одновременно, равна произведению вероятностей того, что произойдет каждое из них.

Поясним последнюю, довольно громоздкую фразу, примером. Мы одновременно бросаем две монеты. В случае, когда обе монеты «правильные» (не деформированные), вероятность выпадения «орла» на каждой из них одинакова и равна $\frac{1}{2}$. Возвращаясь к мысли предыдущего абзаца, можно сказать, что в случае отсутствия зависимости между результатами подбрасывания двух монет вероятность *одновременного* выпадения «орлов» на обеих монетах равна произведению вероятностей выпадению «орла» на каждой из монет: $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$. Таким образом, мы можем провести большую серию опытов, и в том случае, если частота совместного выпадения двух «орлов» сильно отличается от $\frac{1}{4}$, мы можем говорить об отсутствии независимости, т.е. о наличии зависимости между бросанием двух монет.

Перейдем к социологии. Если в массиве данных социологического исследования оказалось $\frac{1}{2}$ мужчин и $\frac{1}{3}$ лиц с высшим образованием, то при *отсутствии зависимости* между полом и образованием мужчин с высшим образованием в массиве должно быть $\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$. Поскольку массив данных уже собран, можно подсчитать, какая в действительности в нашем массиве доля мужчин с высшим образованием (сделать это можно с помощью команды *Crosstabs* — см. разд. 2.2), и, если эта доля сильно отличается от $\frac{1}{6}$, можно говорить, что гипотеза о независимости между полом и наличием высшего образования не подтверждается.

Таким образом, мы получаем некоторый инструмент количественной оценки степени независимости между двумя переменными. Если первый вопрос анкеты имеет три, а второй вопрос — два возможных

варианта ответа, всего возможно шесть комбинаций ответов на эти два вопроса. Для каждой из комбинаций мы можем вычислить вероятность ее (комбинации) появления в случае независимости этих переменных и реальную относительную частоту появления этой комбинации. Далее, находим разность между этими значениями для всех этих шести возможных комбинаций.

Назовем то количество респондентов, которое должно быть в клетке таблицы в случае независимости двух событий, *ожидаемой частотой*. Например, если мы опросили 1000 респондентов, среди которых оказалось $\frac{1}{2}$ мужчин и $\frac{1}{3}$ лиц с высшим образованием, в случае независимости пола и образования ожидаемая частота в клетке «мужчины с высшим образованием» составит: $\frac{1}{2}\times\frac{1}{3}\times 1000 = 166{,}7$.

В меню Cells команды *Crosstabs* (см. рис. 2.3) присутствует окно *Expected*. Если выбрать это окно, в клетках таблицы будут напечатаны ожидаемые частоты, т.е. количества респондентов, которые должны были бы быть в клетках таблицы в случае независимости переменных. В табл. 2.4 представлена таблица для переменных qlO и q12, в клетках которой рассмотрены реальные частоты (окно *Observed* меню Cells) и ожидаемые частоты (окно *Expected*меню Cells).

Как показывают данные табл. 2.4, реальные частоты (Count) и ожидаемые частоты (Expected Count) разные во всех клетках. Следовательно, по нашему мнению, можно сделать вывод о том, что модель независимости переменных qlO и q12 не подтверждается.

Однако в простоте механизма получения такого важного вывода кроется определенная опасность. Ведь мы имеем дело со статистическими данными. Может быть, расхождения между реальными и ожидаемыми частотами носят случайный характер? Действительно, если мы 10 Раз бросим монету и она 6 раз упадет на одну сторону, а 4 — на другую, Даже здравый смысл говорит, что едва ли у нас достаточно оснований Для утверждения о деформированности монеты. Таким образом, когда требуется делать те или иные выводы на основании статистических

данных, нам недостаточно простого сравнения нескольких чисел. Расхождения, равно как и совпадения этих чисел не могут служить достаточным основанием сколь-нибудь серьезных заключений.

Таблица 2.4. Таблицасопряженностипеременныхql0иql2, содержащая реальные и ожидаемые частоты

ql0 Как бы вы оценили в настоящее время материальное положение вашей семьи?		q12 Как бы вы оценили в целом политическую обстановку в России?				Всего
		благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	Count Expected Count	12 2,2	48 35,6	47 73,6	17 12,5	124 124,0
Среднее	Count Expected Count	20 23,3	478 374,3	666 773,0	138 131,4	1302 1302,0
Плохое, очень плохое	Count Expected Count	11 17,0	160 274,0	701 565,8	81 96,2	953 953,0
Затрудняюсь ответить	Count Expected Count	0 0,5	6 8,0	15 16,6	7 2,8	28 28,0
Всего	Count Expected Count	43 43,0	692 692,0	1429 1429,0	243 243,0	2407 2407,0

Механизм проверки гипотезы о независимости переменных несколько сложнее. Во-первых, вычисляется степень суммарного расхождения реальных и ожидаемых частот. При этом необходимо иметь в виду два обстоятельства. Если суммировать просто разности этих частот, с учетом того, что эти разности имеют разные знаки, общая

сумма будет равна нулю. Для того чтобы элиминировать это обстоятельство, предлагается суммировать квадраты разностей. Вторым обстоятельством является следующее. Например, в клетке (1,1) табл. 2.4 квадрат разности частот составит $(12 - 2{,}2)^2 = 96{,}04$, а в клетке (3,3) — $(701 - 565{,}8)^2 = 18279{,}04$. Таким образом, клетка (1,1) даст гораздо меньший вклад в общую сумму, чем клетка (3,3). При этом реальные и ожидаемые частоты в клетке (1,1) различаются более чем в 5 раз, а в клетке (3,3) — приблизительно на 20%. Следовательно, если рассматривать сумму квадратов разностей реальных и ожидаемых частот как показатель их (частот) расхождения, оказывается, что клетки с *относительно меньшим* расхождением могут давать *больший* вклад в значение этого показателя. Чтобы преодолеть эту несообразность, предлагается складывать не абсолютные, а относительные расхождения частот.

Вычисляемый таким образом показатель, фиксирующий степень расхождения реальных и ожидаемых частот, носит название *коэффициента X^2 (хи-квадрат)* и определяется по формуле

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \qquad \textbf{(2.1)}$$

где *O.* — наблюдаемые частоты; *E.* — ожидаемые частоты; *n* — число клеток в таблице.

По формуле (2.1) определяем, что коэффициент $\%^2$для табл. 2.4 составляет 195.

Полученный результат, однако, не приближает нас к поставленной цели — выяснению того, зависимы или независимы между собой переменные ql0 и q12. Действительно, мы не знаем, величина коэффициента $x^2 = 195$ — это большое или маленькое расхождение ожидаемых и наблюдаемых частот? Конечно, если бы мы получили $\%^2 = 0$, можно было бы однозначно говорить о точном совпадении этих частот, и, следовательно, о том, что модель независимости двух анализируемых переменных точно описывает реальные данные. Для случая же $\%^2 > 0$ хотелось бы найти какое-то точное значение Z, когда мы могли бы сказать: если $\%^2 < Z$, $\%^2$ маленький, можно считать, что откло-

нение наблюдаемых и ожидаемых частот незначительно и данные не противоречат модели независимости.

Сделать же это поможет то, что в математической статистике давно известно *теоретическое распределение* коэффициента $\%^2$ при условии, что в генеральной совокупности признаки независимы. Что дает нам знание такого распределения? Чтобы это понять, сделаем небольшое отступление.

Предположим, что некто в нашем присутствии бросает монету, которая 5 раз подряд падает на одну и ту же сторону. Очевидно, что мы отнесемся к этому факту с некоторым удивлением, но удивление не будет очень велико. Если монета у этого человека упадет на одну и ту же сторону 10 раз подряд, это будет уже достаточно удивительно, а если 20 раз подряд, то, по всей видимости, наше удивление уступит место подозрениям. Почему наша реакция будет таковой? Разумеется, не потому, что мы знаем теоретический закон биномиального распределения (по крайней мере, большинство даже не догадывается о его существовании). Жизненный опыт, в достаточно грубом виде, подсказывает, что падение монеты на одну сторону 10 раз подряд — маловероятное событие, а 20 раз подряд — событие, в некотором смысле, уникальное.

Таблицы биномиального распределения точно указывают, что десятикратное выпадение одной и той же стороны монеты имеет вероятность $(1/2)^{10}$, т.е. меньше 0,001. В жизненных ситуациях нет необходимости точно знать эту вероятность — мы и так догадываемся, что это весьма маловероятно. Таким образом, интуитивное знание подсказывает нам, что число 10 может рассматриваться как критическое для ситуации бросания монеты с выпадением на одну сторону. Точное знание говорит о том, что вероятность достижения (а тем более — превышения) числа 10 составляет менее чем 0,001.

Описанный пример с монетой дает основания для двух очень важных заключений. Во-первых, когда мы имеем дело со статистическими характеристиками, мы можем формулировать выводы лишь с какой-то определенной вероятностью. Действительно, нельзя сказать, что выпадение даже ста раз подряд одной стороны монеты невозможно. Оно просто крайне маловероятно.

Во-вторых, знание теоретического закона распределения определенной характеристики позволяет сказать, с какой вероятностью эта характеристика будет иметь то или иное значение. Возвратимся к обсуждаемому коэффициенту $\%^2$. Наличие теоретического закона его распределения коэффициента позволяет нам сказать, с какой вероятностью возможно встретить определенное значение этого коэффициента.

Каково теоретическое распределение коэффициента $\%^2$? Это теоретическое распределение носит то же название -— *y}*. Таким образом, под одним термином «хи-квадрат» скрываются две совершенно разные сущности — коэффициент, фиксирующий степень расхождения теоретических и эмпирических частот, и закон распределения. На рис. 2.4 показан график плотности функции распределения $\%^2$ с разными *степенями свободы*.

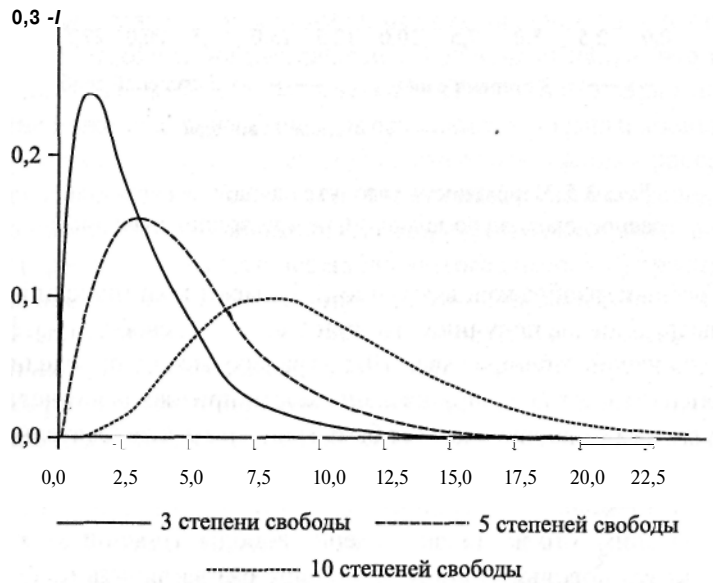


Рис. 2.4. Функция плотности распределения $\%^2$ с разным числом степеней свободы

На рис. 2.5 приведены графики вероятности того, что случайная величина, распределенная по закону χ^2, не превзойдет определенное значение.

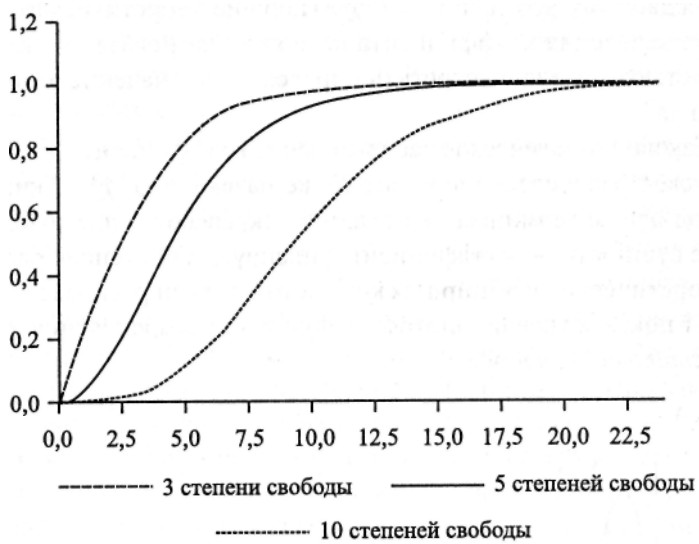


Рис. 2.5. Вероятность того, что случайная величина, распределенная по закону χ^2, не превзойдет значение x

Графики, изображенные на рис. 2.5, имеют практический смысл. Например, если мы получили, что при 5 степенях свободы коэффициент χ^2 для нашей таблицы равен 10, по графику можно определить, что с вероятностью $P = 0{,}92$, при независимости признаков в генеральной совокупности x^2 должен был бы быть меньше 10. Следовательно, при том же предположении вероятность неравенства $\chi^2 > 10$ составляет $P = 1 - 0{,}92 = 0{,}08$.

Объясним, что же такое степени свободы. Поясним, что на самом деле это довольно просто. Если еще раз взглянуть на формулу (2.1), видно, что значение χ^2 зависит от двух показателей — степени расхождения ожидаемых и наблюдаемых частот в клетках таблицы и количества самих клеток. Действительно, мы можем получить боль-

тую величину X^2, если ожидаемые и наблюдаемые частоты сильно расходятся или когда расхождения этих частот маленькие, но самих слагаемых много. Очевидно, что это два совершенно разных случая. В первом у нас, по всей видимости, имеет место случай неподтверждения модели независимости переменных. Во втором — скорее наоборот. Таким образом, одно и то же значение коэффициента χ^2 возможно в двух противоположных ситуациях. Причина тому—разное количество клеток в таблице. Следовательно, когда мы говорим о величине χ^2, необходимо говорить и о числе клеток. Эта характеристика и фиксируется показателем «число степеней свободы». Вычисляется число степеней свободы по формуле: $N = (г - 1)(с-1)$, гдс/У—число степеней свободы; $г$ — число строк в таблице; $с$ — число колонок.

Таким образом, если вернуться к разбору вопроса о наличии связи между переменными qlO и q12 в табл. 2.5, получаем, что для этой таблицы коэффициент X^2 равен 195 при 9 степенях свободы. На рис. 2.5 нет графика вероятности разных значений для 9 степеней свободы. В таблице же распределений χ^2 можно найти, что вероятность того, что $x^1 \wedge 195$, крайне мала ($P < 0{,}0001$). Вряд ли такое событие может реально осуществиться. Целесообразно предположить, что вычисленное нами значение X^2 не имеет отношения к распределению x^2. Из такого предположения будет следовать, что в генеральной совокупности нет независимости между рассматриваемыми переменными (при независимости все наблюдаемые значения x^2 распределены именно по закону x^2 > это теоретически доказано), т.е. что эти переменные зависимы. В таких случаях говорят, что *гипотеза о независимости отвергается на уровне значимости* ос = *0,0001.*

Ограничения использования коэффициента χ^2. Важность метода проверки гипотезы о зависимости между переменными с использованием коэффициента x^2 состоит в том, что в ходе построения этой Модели не делают никаких опущений об уровне измерения самих переменных. Иными словами, можно использовать данный метод применительно к переменным, измеренным на любом уровне. Этот метод является чрезвычайно важным при обработке социологических Данных, поскольку анкетная информация, в подавляющем большинстве случаев, содержит данные, измеренные на разных уровнях.

Однако одно ограничение применения коэффициента χ^2 все-таки есть. Доказано, что коэффициент χ^2 будет иметь теоретическое распределение χ^2 только в случае, когда ожидаемые частоты в таблице имеют значения 5 и более. Это не значит, что если в таблице есть ожидаемые частоты меньше 5, нельзя пользоваться формулой (2.1) для вычисления коэффициента χ^2. Формулой пользоваться можно, однако это становится бессмысленным, поскольку полученное значение коэффициента нельзя проверить на уровень значимости. Иными словами, если нарушено данное ограничение, мы не можем сказать, насколько вероятно то или иное значение и соответственно на каком уровне значимости мы можем принять или отвергнуть гипотезу о независимости анализируемых переменных.

Это ограничение является достаточно болезненным при анализе социологических данных. Действительно, даже в табл. 2.5, несмотря на большое число опрошенных, мы видим, что в трех клетках из 16 ожидаемые частоты имеют значения меньше 5. При этом само ограничение выглядит несколько странным. Действительно, если у нас есть таблица со 100 клетками и только в одной клетке ожидаемая частота меньше 5, неужели распределение будет настолько отличаться от теоретического, что нельзя пользоваться таблицами?

Существует эмпирическое правило, что если в таблице не больше 20% клеток, в которых ожидаемая частота меньше 5, и нет клеток, в которых ожидаемая частота меньше 1, то реальное распределение коэффициента χ^2 достаточно хорошо описывается теоретическим распределением χ^2-[1] В другой работе указывается, что для использования теоретического распределения χ^2 достаточно, чтобы ожидаемые частоты были больше 3^2.

Такая размытость в рекомендациях позволяет сделать вывод: для корректного использования коэффициента χ^2 необходимо стремиться к тому, чтобы клеток с маленькими ожидаемыми частотами было как можно меньше.

[1] См.: Ростовцев П.С, Ковалева Г.Д. Анализ социологических данных с применением статистического пакета SPSS: Учеб.-метод, пособие. Новосибирск: Новосиб. гос. ун-т, 2001. С. 57.

[2] См.: Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 1998. С. 295.

2.3.2

Коэффициенты связи, основанные на χ^2

В использовании коэффициента χ^2 кроется неудобство, поскольку само по себе значение коэффициента ничего не значит. Действительно, информация о том, что $\chi^2 = 100$, не говорит о наличии либо отсутствии взаимосвязи, поскольку для вывода об этом нужно еще знать число степеней свободы, а после этого необходимо заглянуть в таблицу критических значений распределения χ^2. Хотелось бы иметь такой коэффициент, глядя на значение которого, можно сразу, хотя бы приблизительно оценить наличие либо отсутствие связи.

Эту проблему увидел Пирсон, который предложил коэффициент С, производный от x^2, само значение которого уже говорит о наличии либо отсутствии связи. Этот коэффициент носит название *коэффициента сопряженности Пирсона* (2.2).

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}, \qquad \textbf{(2.2)}$$

где N— число опрошенных.

Как видно из формулы, с ростом значения χ^2 значение коэффициента C возрастает. При этом оно всегда больше нуля и меньше единицы. Недостатком коэффициента сопряженности Пирсона является то, что поскольку его значение зависит от N, сравнивать между собой величины C для разных таблиц, как правило, нельзя.

Более распространен (и, добавим от себя, более удобен) *коэффициент сопряженности Крамера,* обозначаемый обычно как V.

$$V = \sqrt{\frac{\chi^2}{N(K-1)}},$$

где N— число опрошенных; K—наименьшее из чисел (г, с), где $г$ — • число строк; с — число столбцов.

Равно как и коэффициент сопряженности Пирсона С, коэффициент вменяется от нуля до единицы. Оба коэффициента принимают

значение нуль при нулевом значении $\%^2$, т.е. в ситуации, когда анализируемые переменные независимы. Однако, в отличие от коэффициента С, который всегда меньше единицы, коэффициент *V* равен единице в ситуации жестко детерминированной связи между переменными, т.е. в случае, когда одному значению переменной *А* всегда соответствует только одно значение переменной *В.*

Однако два этих граничных значения, с интерпретацией которых есть полная ясность, в практических исследованиях не встречаются. Что же означают те реальные значения коэффициента Крамера, с которыми обычно приходится иметь дело, скажем, 0,3? Ничего особенного это не означает, кроме того, что, по всей видимости, значение $\%^2$ достаточно велико и можно ожидать, что гипотеза о независимости анализируемых переменных не подтвердится. Интересно, что не существует таблиц критических значений для коэффициентов Пирсона или Крамера. Для того чтобы оценить уровень значимости этих коэффициентов, необходимо определить уровень значимости коэффициента x^2, который, собственно, и лежит в их основе.

Как можно проинтерпретировать ситуацию, когда для одной пары переменных коэффициент Крамера равен, например, 0,2, а для другой — 0,5? Можно ли сказать, что вторая пара переменных *сильнее взаимосвязана,* чем первая?

Здесь мы фактически ввели понятие, которое используем в жизни ежедневно и которое, вроде бы, вполне очевидно — сила связи. Так вот, это интуитивное понимание силы связи никак не может быть применено для работы с коэффициентами связи в таблицах сопряженности. Большее значение x^2, равно как и коэффициента Крамера, коэффициента Пирсона, либо какого-то иного[3], означает лишь уменьшение того уровня значимости а, на котором отвергается гипотеза о независимости признаков. О характере же выявленной зависимости и о ее силе обсуждаемые коэффициенты ничего не говорят.

[3] Отметим, что наряду с коэффициентами Крамера и Пирсона существуют и другие коэффициенты, также являющиеся нормировками $\%^2$, например, коэффициент Чупрова. Подробнее см.: Толстова Ю.Н. Анализ социологических данных. М.: Научный мир, 2000. С. 197—201.

2.3.3

Коэффициенты связи, основанные на прогнозе

Для иллюстрации идей, заложенных в коэффициентах, основанных на прогнозе, повторим одну из таблиц сопряженности, которая уже обсуждалась выше (см. табл. 2.2).

Таблица 2.5. Таблица сопряженности переменных **qlO** и q12, %

qlO Как бы вы оценили в настоящее время материальное положение вашей семьи?	q12 Как бы вы оценили в целом политическую обстановку в России?				Всего
	благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	9,7	38,7	37,9	13,7	100,0
Среднее	1,5	36,7	51,2	10,6	100,0
Плохое, очень плохое	1,2	16,8	73,6	8,5	100,0
Затрудняюсь ответить	0	21,4	53,6	25,0	100,0
Всего	1,8	28,7	59,4	10,1	100,0

Модели коэффициентов, основанные на предсказании, строятся на идее, предложенной Л. Гутманом. Предположим, что мы провели исследование и у нас на руках находится один из результатов этого исследования — табл. 2.5. Давайте попробуем на основе этой таблицы предсказать, как оценит политическую обстановку в России по предложенной шкале случайно взятый респондент. Поскольку большинство опрошенных (59,4%) оценивает политическую обстановку в России как «критическую, взрывоопасную», наше предсказание

будет соответствующим: «Случайно взятый респондент охарактеризует политическую обстановку в России как критическую, взрывоопасную». При этом мы сразу можем оценить вероятность того, что наш прогноз будет ошибочным — 1 - 0,594 = 0,406. Таким образом, мы видим, что наш прогноз может быть ошибочным с вероятностью более 40%.

Строя прогноз, мы основывались не на всей информации, содержащейся в табл. 2.5, а только на последней строке этой таблицы, т.е. на одномерном частотном распределении одной из переменных. Как можно использовать всю информацию табл. 2.5 для более качественного прогноза? Обратите внимание на то, что, строя прогноз ответов на вопрос q12 (Оценка политической обстановки в России), мы не использовали информацию об ответах на вопрос qlO (Оценка материального положения семьи). Если мы будем знать, как случайно взятый респондент ответил на вопрос qlO, наш прогноз его ответа на вопрос q12 будет иным. Действительно, если знать, что этот, случайно взятый респондент на вопрос о материальном положении семьи сказал, что он оценивает его как хорошее, либо очень хорошее, то скорее он оценит политическую обстановку в России как «напряженную». Такое предсказание ответа на вопрос q12 будет наилучшим, поскольку данное предсказание даст вероятность ошибки 1 - 0,387 = 0,613. Любой другой прогноз ответа на вопрос q12 (при таком выборе ответа на вопрос qlO) будет иметь большую вероятность ошибки и, следовательно, будет хуже.

Таблица 2.5 показывает, что при других ответах на вопрос о материальном положении семьи наилучшее предсказание ответа на вопрос q12 — «критическая, взрывоопасная».

Подводя итог рассуждениям о предсказании ответов на вопрос q12 об экономическом положении России, можно сделать следующее заключение. Если мы не знаем (либо не используем) информацию об ответах случайно взятого респондента на вопрос qlO, мы будем предсказывать, что, скорее всего, отвечая на вопрос q12, он выберет вариант «критическая, взрывоопасная». Если мы знаем, как этот случайно взятый респондент ответил на вопрос qlO, наше предсказание ответов на q12 будет иным. А именно: если в вопросе qlO респондент

выбрал вариант «хорошее, очень хорошее», вероятнее всего ответ на вопрос q12 будет «напряженная». Если ответ на вопрос qlO будет любым другим, отвечая на q12, респондент вероятнее всего выберет вариант «критическая, взрывоопасная».

Таким образом, в зависимости от того, знаем мы ответы случайно отобранного респондента на вопрос qlO или нет, наш прогноз ответов на вопрос q12 будет разным. Естественно, во втором случае наш прогноз будет точнее. После всего выше сказанного можно сделать два вывода. Во-первых, если знание значений одной переменной улучшает предсказание значений другой переменной, эти две переменные взаимосвязаны. Во-вторых, то, насколько улучшается качество предсказания одной переменной в ситуации знания значения другой, по сравнению с незнанием, может выступать показателем взаимосвязи этих переменных.

Поскольку «предсказание» в обыденной жизни ассоциируется, прежде всего, с предсказанием погоды, проиллюстрируем все вышесказанные примеры из этой области. Предположим, вероятность того, что в Москве будет идти снег в случайно выбранный день года, составляет 0,06. Однако зимой эта вероятность составляет уже примерно 0,2. Таким образом, зная значение переменной «время года» для случайно выбранного дня, мы можем гораздо точнее предсказывать вероятность того, что в этот день пойдет снег.

Логика коэффициента, фиксирующего улучшение предсказания значений одной переменной на основании значений другой переменной, весьма проста. Если назвать прогноз на основе значений только одной переменной *первым прогнозом,* а прогноз на основе двух переменных — *вторым прогнозом,* предлагаемые коэффициенты называются *коэффициентами, основанными на модели прогноза:*

$$\frac{\text{Ошибка при первом прогнозе - Ошибка при втором прогнозе}}{\text{Ошибка при первом прогнозе}}$$

Пока мы обсуждаем коэффициенты, основанные на прогнозе более часто встречающегося значения. Это так называемый *прогноз модального значения.* Коэффициенты для такого прогноза называются Я (лямбда), их предложил Л. Гутман в 1941 г.

Что такое «первый прогноз» при модальном прогнозе? Это модальное значение предсказываемой переменной, обозначим его как Л, а процент, который соответствует значению *A,* — как *Рг А.* При таком обозначении ошибка при первом прогнозе будет $P_t = 1 - Рг A$.

При втором прогнозе мы анализируем по очереди каждую строку таблицы и выбираем в каждой строке модальную частоту. Пусть модальное значение в каждой строке будет *А.,* а соответствующий процент — *Рг А* .. Соответственно ошибка при предсказании значения в i-й строке составит $P = 1 - Рг A$.. Таким образом, ошибка при втором прогнозе будет средней ошибкой предсказания по каждой из строк таблицы:

$$P_2 = \sum_{i=1}^{r} \frac{P_{2i}}{r}. \qquad \textbf{(2.5)}$$

Формула коэффициента, фиксирующего улучшение прогноза теременной, значение которой располагаются по столбцам таблицы, выглядит следующим образом:

$$\lambda_b = \frac{\sum_{i=1}^{c} (\max n_{ij} - \max n_{\bullet j})}{n - \max_{\bullet j}}. \qquad \textbf{(2.6)}$$

У обсуждаемого коэффициента есть одна особенность, отличающая его от коэффициента $\%^2$. В вычислении X_b строки и столбцы участвуют не симметрично. Разумеется, таблицу можно повернуть на 90% и с точки зрения содержащейся в таблице информации от этой операции ничего не изменится. При таком повороте не изменятся значения коэффициентов $\%^2$ и коэффициентов, основанных на $\%^2$. Однако значение коэффициента X_b изменится. Это связано с тем, что в модели коэффициента X_b мы предсказываем значение одной переменной на основании значений другой и переменные включены в модель не симметрично. Фактически одна переменная рассматривается как причина, а другая как следствие.

В этой связи наряду с переменной X_b, которая фиксирует предсказание переменной, расположенной по колонкам таблицы, существует и переменная X_a, которая отражает улучшение предсказания

переменной, расположенной по строкам на основании переменной, расположенной по столбцам. Наконец, когда мы не можем четко сказать, какая из переменных может рассматриваться как причина, а какая как следствие, существует так называемая X_s , т.е. «лямбда симметричная», представляющая полусумму X_a и X_y

Поскольку коэффициенты *X,* так же как и $\%^2$ — статистические меры, то в их отношении встает задача *оценки уровня значимости.* Действительно, если для некоторой таблицы был получен коэффициент, скажем, $X_b = 0{,}1$, можем ли мы утверждать, что некоторая связь между переменными действительно есть, и это значение не есть просто статистическая случайность. Другими словами, требуется проверить статистическую гипотезу $X_b = 0$ на основании полученного выборочного значения $X_b = 0{,}1$.

Логика проверки данной статистической гипотезы совершенно аналогична логике проверки гипотезы о равенстве нулю коэффициента X^2. Нам требуется знание теоретического распределения коэффициента *X,* которое покажет нам, насколько вероятно то или иное значение коэффициента *X.* При вычислении коэффициентов А, в пакете SPSS в команде *Crosstabs* одновременно проводится вычисление уровней значимости а этих коэффициентов. Если коэффициент *X* равен, скажем, значению X_0, уровень значимости этого значения определяется так: $a = P(X > X_0)$ при условии, что в генеральной совокупности $X = 0$.

Достоинством коэффициентов *X* является то, что в отличие от коэффициента $\%^2$ либо производных от него само значение X_a и X_b имеет непосредственный смысл — это улучшение вероятности правильного предсказания. Иначе говоря, если для некоторой таблицы *j ~ 0,2, это означает, что мы можем предсказывать модальное значение переменной, располагающейся по колонкам, зная совместное распределение двух переменных на 20% точнее по сравнению с ситуацией, когда мы не знаем этого распределения.

Однако это значение весьма условно. Действительно, коэффициенты *X* являются статистическими мерами и потому точное полученное значение коэффициента бессмысленно. Ведь мы можем повто-

рить опрос для другой выборки (с соблюдением той же процедуры ее построения) и тем не менее почти наверняка получим другое значение коэффициента *X,* поскольку будут опрашиваться другие респонденты. Следовательно, гораздо важнее получить не точечное значение коэффициентов *X,* а доверительный интервал.

При вычислении коэффициентов *X* командой *Crosstabs* наряду с точечными значениями вычисляются также и величины стандартных ошибок. Стандартные ошибки позволяют построить доверительные интервалы с задаваемыми уровнями значимости. В табл. 2.6 приведен фрагмент таблицы, выдаваемой командой *Crosstabs,* в которой вычислены коэффициенты *X,* соответствующие стандартные ошибки и уровни значимости для проверки гипотезы о равенстве нулю этих коэффициентов для данных табл. 2.1.

Таблица 2.6. Значения коэффициентов *X* и связанных с ними статистических показателей для данных табл. 2.1

X	**Value**	**Asymp. Std. Error"**	**Approx. T[4]**	**Approx. Sig.**
Symmetric	0,017	0,018	0,942	0,346
«Как бы вы оценили в настоящее время материальное положение вашей семьи?» Dependent	0,032	0,033	0,947	0,344
«Как бы вы оценили в целом политическую обстановку в России?» Dependent	0,001	0,010	0,103	0,918

Таблица 2.6 содержит одновременно все три коэффициента *X* — *X* (строка *Symmetric), X_a* (следующая за Symmetric строка) и X_b (последняя строка). В колонке *Value* расположены значения соответствующих коэффициентов, а в последней колонке *{Approx. Sig.)* -— уровни значимости для проверки гипотез о равенстве нулю этих коэффициентов. Находящееся в этой колонке значение имеет смысл сравнить с заранее выбранным уровнем значимости, снова обозначим его **а** (в статистике чаще всего используется ***а*** = 0,05). Если наше значение больше а, гипотезу следует принять, если меньше — отвергнуть.

Из табл. 2.6 видно, что при уровне значимости **а** = 0,05 следует принять гипотезу о равенстве нулю всех трех коэффициентов — A_s , *X* , X_b (так как все значения в колонке *Approx. Sig.* превосходят 0,05).

. Таблица 2.6 содержит еще две колонки, одна из которых—*Asymp. Std. Error* (асимптотическая стандартная ошибка) дает нам информацию, необходимую для построения доверительных интервалов. Напомним, что 68%-й доверительный интервал дает *X*± (одна стандартная ошибка); 95%-й доверительный интервал — *X*± (две стандартные ошибки).

Графа *Approx. T* — вспомогательная, и обозначает отношение значения коэффициента *Хк* величине его стандартной ошибки. Отметим, что данный показатель *Т* широко распространен в разных разделах статистики. Фактически он показывает, как соотносится измеренное значение с ошибкой измерения и характеризует то, насколько мы можем доверять полученным коэффициентам. Поясним смысл этого показателя на примере. Предположим, что мы взвешиваем спичку на обычных бытовых весах и получаем, что спичка весит 2 г. Однако точность наших весов составляет ± 10 г. В данном примере показатель *Т* = 0,2 и едва ли мы будем всерьез относиться к полученному значению веса спички. Таким образом, чем больше значение *Т,* тем выше качество полученного измерения.

Из приведенных формул коэффициентов *X* следует, что у них есть очень существенный недостаток — в том случае, когда все модальные частоты лежат в одной колонке либо в одной строке таблицы, соответствующие коэффициенты всегда обращаются в нуль. Таким образом, равенство нулю коэффициентов X_M*Щ* X_t— это необходимое, но не достаточное условие для независимости переменных, образующих таблицу.

Последнее свойство весьма неудобно. Действительно, хотелось бы иметь коэффициенты, которые обладают естественным свойством —

равенство нулю всегда говорит о независимости. Этим качеством обладают коэффициенты, также основанные на прогнозе, но в которых прогнозируется не модальная частота, а весь спектр частот. Это коэффициенты т (тау) Гудмена — Краскэла.

Общая идея оценки качества прогноза для коэффициентов т записывается выражением (2.4), так же как и для коэффициентов *X,* однако сам прогноз строится иначе. Рассмотрим это подробнее.

Из табл. 2.2 следует, что переменная q12 «Как бы вы оценили в целом политическую обстановку в России?» имеет следующее одномерное распределение (табл. 2.7). Случайным образом отберем 2407 респондентов и, базируясь на приведенном одномерном распределении, попытаемся угадать ответы каждого на вопрос q12. Возьмем 43 респондента и скажем, что они отметили в этом вопросе градацию «1». Поскольку вероятность выбора первой градации составляет 1,8%, количество людей, у которых мы правильно угадаем первый вариант ответа, составляет 43 х 0,018 = 0,774. Аналогичным образом поступим со всеми 2407 респондентами (последняя колонка табл. 2.6).

Таблица 2.7. Одномерное частотное распределение переменной q12 и результаты предсказания этого распределения

1	*N*	%	Количество правильно предсказанных ответов
Благополучная, спокойная	43	1,8	0,774
Напряженная	692	28,7	198,604
Критическая, взрывоопасная	1429	59,4	848,826
Затрудняюсь ответить	243	10,1	24,543
N	2407	100	1072,747

Таблица 2.7 показывает, что из 2407 отобранных респондентов, используя предлагаемую модель предсказания, мы сумели правильно предсказать выбор у 1072,747 респондентов. Таким образом, общее качество такого прогноза, который базируется только на знании одномерного распределения переменной q12, составляет 1072,747 / 2407 = 44,57%.

Далее строим модель предсказания, базируясь уже на данных таблицы двумерного распределения переменных qlO и q12, т.е. будем использовать для предсказания значений q12 значения переменной qlO. В табл. 2.8 приведен расчет такого предсказания для первой строки таблицы совместного распределения.

Таблица 2.8. Таблица расчетов коэффициента пропорционального предсказания

Как бы вы оценили в настоящее время материальное положение вашей семьи?		Как бы вы оценили в целом политическую обстановку в России?				Всего
		благополучная, спокойная	напряженная	критическая, взрывоопасная	затрудняюсь ответить	
Хорошее, очень хорошее	*N*	12	48	47	17	124
	%	9,7	38,7	37,9	13,7	100,0
	Количество респондентов с правильным прогнозом	1,164	18,576	17,813	2,329	39,882

2.4

Коэффициенты связи для порядковых данных

В предыдущих рассуждениях о таблицах сопряженности и коэффициентах связи не делалось никаких ограничений либо допущений в отношении уровня измерения тех переменных, которые образуют таблицу. Не использовалась и информация о порядке следования градаций в переменных. Очевидно, что если мы поменяем местами града-

ции переменных, это никоим образом не скажется на значении коэффициентов x^2, Крамера, *Хкх*.

Это является естественным для переменных, измеренных на номинальном уровне. Действительно, номера, которые присваиваются градациям в таких переменных, имеют абсолютно условный смысл. Так, совершенно не имеет значения, присвоен ли в вопросе «Ваш пол» мужчинам код 1, 2 или 28. Главное, чтобы код, присвоенный мужчинам, отличался от кода, присвоенного женщинам.

По этой причине то, что коэффициенты связи никак не реагируют на наш произвол в присвоении определенным градациям тех или иных числовых кодов, является вполне правильным для случая, когда исходные данные получены по номинальным шкалам.

Однако эти рассуждения становятся неверными, когда речь заходит о переменных, измеренных на порядковом уровне. Для такого рода переменных порядок расположения градаций уже существен, поскольку он фиксирует степень выраженности измеряемого свойства. Измерение взаимосвязи в таблицах, построенных с использованием порядковых переменных, вполне возможно и нередко делается с использованием коэффициентов $\%^2$, Крамера, *X* и т. Но эти коэффициенты не используют данные о порядке следования градаций и, следовательно, лишают нас возможности использовать всю содержащуюся в переменных информацию. Для того чтобы устранить этот недостаток, наряду с перечисленными коэффициентами, для порядковых переменных используют и другие меры связи — *коэффициенты ранговой корреляции*.

Для демонстрации принципов работы коэффициентов ранговой корреляции рассмотрим пример (табл. 2.9). Таблица должна ответить на вопрос о том, насколько взаимосвязаны оценка человеком своего материального положения и оценка удовлетворенности жизнью в целом.

Коэффициенты $\%^2$ и Крамера, вычисленные для этой таблицы, показывают, что с большой вероятностью можно утверждать о наличии взаимосвязи между двумя рассматриваемыми показателями, поскольку значимость обоих коэффициентов весьма высока (а > 0,001)-

Однако эти коэффициенты не дают ответа на важный вопрос: возрастает или падает удовлетворенность жизнью в целом с ростом удовлетворенности материальным положением? На интуитивном уровне представляется, что удовлетворенность жизнью должна возрастать с ростом удовлетворенности материальным положением, но коэффициенты не дают возможности это зафиксировать либо хотя бы проверить направление взаимосвязи.

Таблица 2.9. Таблица сопряженности с использованием порядковых переменных

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	15	5	342
Отчасти устраивает, отчасти нет	0	31	649	201	14	895
По большей части не устраивает	1	3	200	340	55	599
Совершенно не устраивает	1	1	49	185	118	354
Всего	10	109	1291	749	195	2354

В настоящее время социологи используют коэффициенты ранговой корреляции — р Спирмена, т Кендэла, у Гудмена — Краскэла. Рассмотрим правила вычисления коэффициента у Гудмена — Краскэла как самого простого и часто используемого при анализе социологических данных.

На первом шаге вычисления коэффициента у фиксируют количество респондентов, у которых значение первой переменной не меньше значений второй переменной. Например, в табл. 2.9 у пяти респондентов значения обоих переменных равны 1, у 35 респондентов — Равны 2 и т.д.

Таблица 2.10. Схема определения показателя *S* для вычисления коэффициента у

Шаг!

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	**35**	**284**	**15**	5	342
Отчасти устраивает, отчасти нет		**31**	**649**	**201**	**14**	895
По большей части не устраивает	1	3	**200**	**340**	**55**	599
Совершенно не устраивает	1	1	**49**	**185**	**118**	354
Всего	10	109	1291	749	195	2354

Шаг 2

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	**35**	284	15	5	342
Отчасти устраивает, отчасти нет		31	**649**	**201**	**14**	895
По большей части не устраивает	1	3	**200**	**340**	**55**	599
Совершенно не устраивает	1	1	**49**	**185**	**118**	354
Всего	10	109	1291	749	195	2354

Ц1аг3

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	15	5	342
Отчасти устраивает, отчасти нет		31	**649**	201	14	895
По большей части не устраивает	1	3	200	**340**	**55**	599
Совершенно не устраивает	1	1	49	**185**	**118**	354
Всего	10	109	1291	749	195	2354

Шаг 4

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	15	5	342
Отчасти устраивает, отчасти нет		31	649	201	14	895
По большей части не устраивает	1	3	200	**340**	55	599
Совершенно не устраивает	1	1	49	185	**118**	354
Всего	10	109	1291	749	195	2354

В табл. 2.10 представлена схема вычисления показателя *S*— количества пар, в которых значение первой переменной не меньше значений второй переменной:

$$S = 5 \times (35 + 31 + 3 + 1 + 284 + 649 + 200 + 49 + 15 + 201 + 340 + 185 + 5 + 14 + 55 + 118) + 35 \times (649 + 200 + 49 + 201 + 340 + 185 + 14 + 55 + 118) + 649 \times (340 + 185 + 55 + 118) + 340 \times 118 = 567\,432.$$

Таблица 2.11. Схема определения показателя D для вычисления коэффициента γ

Шаг!

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	**35**	**284**	**15**	5	342
Отчасти устраивает, отчасти нет		**31**	**649**	**201**	14	895
По большей части не устраивает	1	3	**200**	**340**	55	599
Совершенно не устраивает	1	1	**49**	**185**	118	354
Всего	10	109	1291	749	195	2354

Шаг 2

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	**15**	5	342
Отчасти устраивает, отчасти нет		**31**	**649**	201	14	895
По большей части не устраивает	1		**200**	340	55	599
Совершенно не устраивает	1	1	**49**	185	118	354
Всего	10	109	1291	749	195	2354

Шаг3

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	15	5	342
Отчасти устраивает, отчасти нет		31	**649**	201	14	895
По большей части не устраивает	1	3	200	340	55	599
Совершенно не устраивает	1	1	49	185	118	354
Всего	10	109	1291	749	195	2354

Шаг 4

В какой мере вас устраивает жизнь, которую вы ведете?	Как бы вы оценили в настоящее время материальное положение вашей семьи?					Всего
	очень хорошее	хорошее	среднее	плохое	очень плохое	
Вполне устраивает	5	39	109	8	3	164
По большей части устраивает	3	35	284	15	5	342
Отчасти устраивает, отчасти нет		31	649	201	14	895
По большей части не устраивает	1	3	200	340	55	599
Совершенно не устраивает	1	1	49	185	118	354
Всего	10	109	1291	749	195	2354

В табл. 2.11 представлена схема вычисления показателя D — количества пар, в которых значение первой переменной не меньше значений второй переменной.

$$D = 3 \times (3 + 1 + 1 + 35 + 31 + 3 + 1 + 284 + 649 + 200 + 49 + 15 + 201 + 340 + 185) + 15 \times (1 + 1 + 31 + 3 + 1 + 649 + 200 + 49) + 649 \times (1 + 1 + 3 + 1) + 3 \times 1 = 23\,916.$$

Имея значения *SKD,* можно непосредственно рассчитать коэффициент γ по формуле

$$\gamma = \frac{S - D}{S + D}. \qquad (2.7)$$

Для табл. 2.9 значение γ равно 0,763. О чем говорит такое значение коэффициента, и, более того, как вообще интерпретируются ранговые коэффициенты связи?

В целом ранговые коэффициенты связи характеризуют ситуацию, когда, сопоставляя двух случайно отобранных респондентов, у которых измеряются две порядковые переменные *A* и *B,* мы можем сказать, что если у первого респондента значение переменной *A* больше, чем у второго респондента, то у него будет больше и значение по переменной *B.* Количество пар респондентов, у которых это правило выполняется, и есть построенный показатель *S.* Количество пар респондентов, для которых действует обратное правило, т.е. таких пар, у которых переменная *A* у первого респондента имеет значение больше, чем у второго, а переменная *B* — меньше, фиксируется показателем *D.* Таким образом, коэффициент γ фиксирует то, каких пар больше.

Из формулы (2.7) следует, что коэффициент γ может изменяться в интервале от -1 до +1. Коэффициент равен +1 в случае, когда показатель *D* равен нулю, т.е. в ситуации, когда для всех респондентов верно, что если переменная $A = i$, а переменная $B = j$, всегда $i > j$. Соответственно γ равна -1, когда в той же ситуации переменных *A* и *B* всегда $i < j$.

Что означает ситуация, когда одна пара переменных, скажем, A_1 и A, имеет более высокое (по абсолютной величине) значение коэффициента γ, чем пара переменных 5, и 5_2? Это означает, что для переменных A и A_2 вероятность правильного порядка значений переменных выше, чем для переменных 5, и B_2. Под правильным порядком мы понимаем порядок, при котором если $A = i$, а $B = j$, то всегда $i > j$,

или $i < j$. Вообще, коэффициент γ имеет прямую вероятностную интерпретацию — это разность между вероятностями правильного и неправильного порядка для пары случайно извлеченных из выборки наблюдений[4]. Именно так следует понимать силу связи, которая фиксируется ранговыми коэффициентами корреляции.

Как на практике определить, насколько велико полученное значение коэффициента γ, можно ли сказать, что если в одном исследовании коэффициент γ = 0,5, а в другом — γ = 0,6, то во втором исследовании имеет место более тесная связь между анализируемыми показателями? Поскольку для коэффициента γ известно теоретическое распределение, то пакет SPSS одновременно со значением коэффициента вычисляет также и значение стандартной ошибки. Благодаря этому возможно построение доверительного интервала для коэффициента γ. В табл. 2.12 приведены результаты, которые выводит команда *Crosstabs* при запросе на вычисление коэффициента γ для данных, приведенных в табл. 2.9.

Таблица 2.12. Результаты вычисления коэффициента ранговой корреляции γ для данных табл. 2.8

	Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Gamma	0,763	0,015	37,143	0,000

Основываясь на данных табл. 2.12, можно сказать, что с вероятностью 95% значение коэффициента γ для генеральной совокупности будет находиться в интервале (0,763 ± 0,03). С помощью числа в колонке *Approx. Sig.* (приблизительная значимость) можно оценить справедливость гипотезы *H:* «Величина коэффициента ранговой корреляции γ для анализируемых переменных в генеральной совокупности равна нулю». В табл. 2.12 мы получили, что Approx. Sig = 0,000. Это означает,

[4] См.: Аптон Г. Анализ таблиц сопряженности. М.: Финансы и статистика, 1982. С 37.

что для соответствующего уровня значимости а имеет место неравенств: а < 0,001. Гипотезу H_0 следует отвергнуть, поскольку эта величина намного меньше общепринятого для отвержения гипотезы уровня значимости 0,05.

Если необходимо решить задачу сравнения коэффициентов *y,* вычисленных для двух разных социальных совокупностей, необходимо:

• определить доверительные интервалы для обоих коэффициентов;

• посмотреть, пересекаются ли эти доверительные интервалы. Если они не пересекаются, то мы, с соответствующей доверительной вероятностью, можем утверждать, что эти коэффициенты различны.

Отличие ранговых коэффициентов корреляции от коэффициентов связи, основанных на $\%^2$ либо на модели предсказания, состоит в том, что фиксируют не только наличие либо отсутствие связи, но и, в случае наличия связи, ее направление. Это, несомненно, является достоинством данных коэффициентов, но в определенных случаях может являться и их недостатком. Дело в том, что ранговые коэффициенты корреляции фиксируют только однонаправленность, монотонность формы зависимости (см. рис. 2.6). Например, для всех изображенных на рис. 2.6 зависимостей имеем значение коэффициента *y,* равное +1 или - 1 , несмотря на то что сами формы зависимости существенно разные.

Что произойдет, если зависимость между переменными не имеет однонаправленной связи, как, например, зависимости, изображенные на рис. 2.7? Оказывается, что в ситуации такого рода форм зависимостей ранговые коэффициенты связи оказываются неэффективными. Действительно, если может оказаться, что для части респондентов, например тех, кто имеет малые значения переменной *x* (рис. 2.7, график 1), значение рангового коэффициента связи будет отрицательное, а для тех респондентов, которые имеют большие значения переменной *x,* значение рангового коэффициента будет положительное, то общее значение рангового коэффициента может оказаться равным нулю. И это при том, что, как показывает график, связь между переменными явно есть.

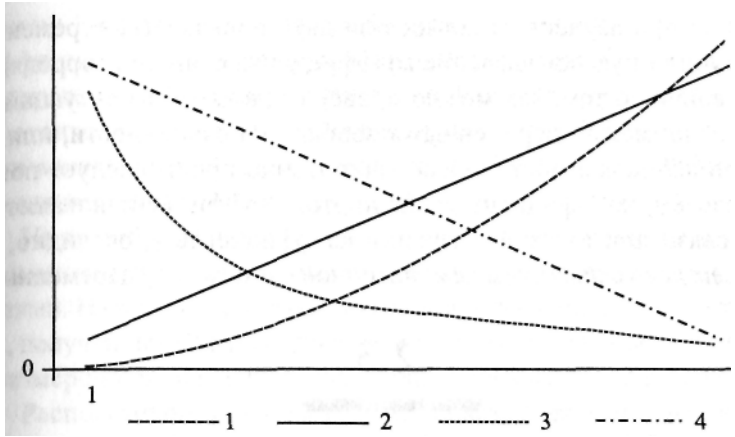


Рис. 2.6. Примеры монотонных зависимостей между переменными

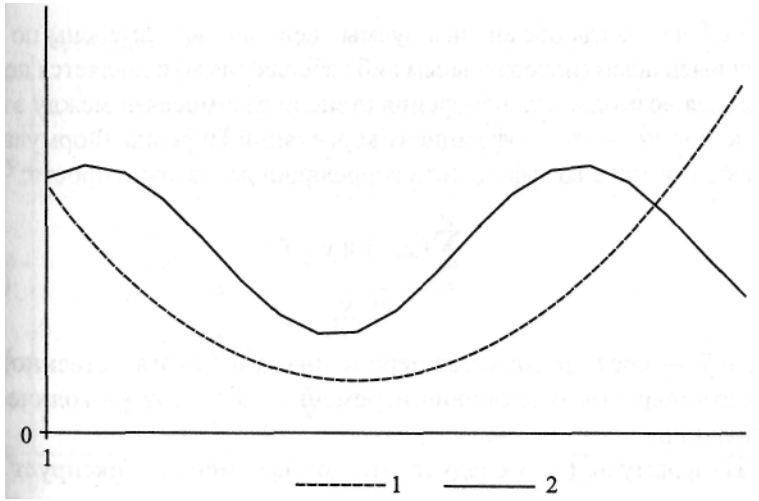


*Рис. 2.*7. Примеры немонотонных зависимостей между переменными

Таким образом, тот факт, что значение рангового коэффициента корреляции равно нулю, говорит не об отсутствии связи, а лишь об отсутствии монотонной связи.

Если при изучении взаимосвязи двух порядковых переменных мы получили нулевое значение коэффициента ранговой корреляции, встает вопрос о том, как можно проверить, с какой из ситуаций мы имеем дело: между переменными вообще нет зависимости, или нет монотонной зависимости? Ответ достаточно прост: следует посчитать, скажем, коэффициент $\%^2$. Если этот коэффициент покажет наличие связи при нулевом значении коэффициента *y,* очевидно, что мы имеем дело с наличием немонотонной связи между переменными.

2.5

Коэффициент корреляции Пирсона

В том случае, когда обе анализируемые переменные измерены по метрическим шкалам (интервальным либо абсолютным) появляется дополнительная возможность измерения степени взаимосвязи между этими переменными —- это коэффициент корреляции Пирсона. Формула для вычисления этого коэффициента корреляции достаточно проста:

$$r = \frac{\sum_{i=1}^{N}(x-\bar{x})(y-\bar{y})}{NS_xS_y}, \qquad (2.8)$$

где x и y — средние значения переменных x и y соответственно; S_x и S — стандартные отклонения переменных x и y; N — количество наблюдений.

Из формулы (2.8) следует, что коэффициент *г* фиксирует степень того, насколько переменные x и y одновременно отклоняются от средних значений. Таким образом, в отличие от ранговых коэффициентов корреляции, которые замеряют монотонный характер связи между переменными, коэффициент корреляции Пирсона учитывает более узкий характер монотонности — линейность. Когда между переменными есть строгая линейная зависимость, значение коэффициента корреляции Пирсона будет равно +1 в случае положительной

связи и -1 в случае отрицательной связи. Так, для графика 2 рис. 2.6 коэффициент корреляции равен +1, а для графика 4——1. В ситуации, когда связь не соответствует линейной, коэффициент корреляции отличается от единицы даже в случае жесткой функциональной связи между переменными. Для графика 1 рис. 2.6 коэффициент корреляции равен 0,975, а для графика 4——0,833.

На практике, когда анализируется зависимость между социологическими переменными, мы имеем дело не с функциональными зависимостями. На рис. 2.8 показана диаграмма рассеяния для реальных данных, полученных в ходе социологического изучения зависимости между размером семьи и величиной среднедушевого дохода семьи.

Расположение точек на рис. 2.8 показывает, что едва ли существует какая-то строгая функция, которая позволит построить кривую, проходящую через все точки реальных данных.

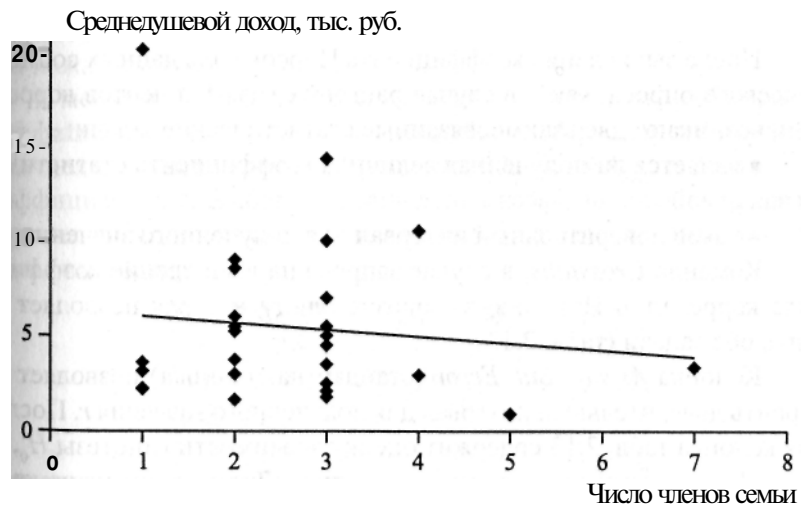


Рис. 2.8. Диаграмма рассеяния для 30 респондентов по переменным «Среднедушевой доход» и «Число членов семьи»

Коэффициент корреляции Пирсона отражает определенную прямую, которая в некотором смысле наилучшим образом фиксирует за-

висимость между двумя переменными[5]. Для данных, представленных на рис. 2.8, коэффициент корреляции Пирсона равен -0,11. Отрицательное значение коэффициента означает, что с ростом размера семьи среднедушевой доход уменьшается. Величина коэффициента 0,11 показывает степень того, насколько реальные данные близки к линейной зависимости (прямой).

Когда мы рассматриваем совместное поведение двух метрических переменных, то целью социологического анализа является установление взаимосвязи, зависимости между этими переменными. При использовании для решения этой задачи коэффициента корреляции Пирсона следует помнить, что нулевое значение этого коэффициента, строго говоря, свидетельствует только об отсутствии *линейной зависимости*. Это, в свою очередь, может свидетельствовать и об отсутствии вообще какой-либо зависимости, и о том, что зависимость есть, но она носит нелинейный характер. Установить с помощью данного коэффициента, с какой из этих ситуаций мы имеем дело в конкретном случае, нельзя.

После вычисления коэффициента Пирсона для данных социологического опроса, как и в случае ранговых коэффициентов корреляции, возникают две взаимосвязанные статистические задачи:

• является ли полученная величина коэффициента статистически значимой;

• каков доверительный интервал для полученного значения.

Команда *Crosstabs,* в случае запроса на вычисление коэффициента корреляции Пирсона, выводит таблицу, которая позволяет решить обе задачи (табл. 2.13).

Колонка *Asymp. Std. Error* (стандартная ошибка) позволяет построить доверительный интервал для полученного значения *г.* Последняя колонка табл. 2.13 содержит оценку значимости гипотезы $\#_0$, которая формулируется следующим образом: «Для двух анализируемых переменных коэффициент корреляции Пирсона равен нулю». В нашем примере (табл. 2.12), если снова взять уровень значимости 5%, гипотезу следует принять.

[5] Прямая в данном случае вычисляется по методу наименьших квадратов — см. главу 4 «Модели регрессионного анализа».

Таблица 2.13. Формат выдачи результатов вычисления коэффициента корреляции Пирсона командой Crosstabs

	Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Pearson's R	-0,109	0,173	-0,580	0,567
N of Valid Cases	30			

2.6

Вычисление коэффициентов связи в команде Crosstabs

Главное меню команды *Crosstabs* в нижней части имеет клавишу Statistics... (см. рис. 2.2). На рис. 2.9 показано меню, которое вызывается нажатием этой клавиши.

В меню Statistics можно выбрать любое количество необходимых коэффициентов связи. Отметим, что выбор коэффициента А, приведет к вычислению всех трех коэффициентов X, X_b, и Я. , а также двух коэффициентов т. Выбор же вычисления коэффициента корреляции Пирсона приводит также и к вычислению рангового коэффициента корреляции Спирмена.

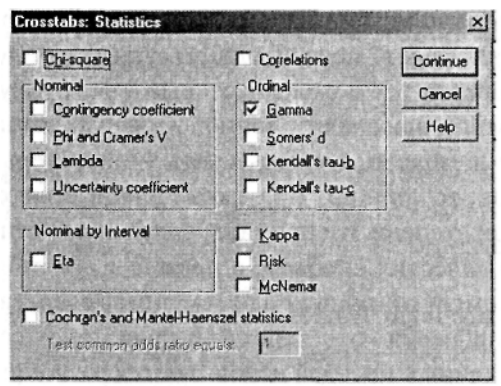


Рис. 2.9. Меню Statistics команды Crosstabs

3
глава

АНАЛИЗ ВЗАИМОСВЯЗЕЙ КАЧЕСТВЕННЫХ И КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

Рассмотренные ранее методы анализа одномерных распределений и таблиц сопряженности были направлены на построение описательных, а не объяснительных моделей изучаемых явлений. Действительно, даже анализ таблиц сопряженности, вычисление коэффициентов связи в этих таблицах подразумевают фиксацию статистических взаимосвязей переменных, а на основе этих выявленных взаимосвязей социолог начинает конструировать объяснительные модели, привлекая социологические теории, свое знание социальной реальности, т.е. ту информацию, которая в анализе данных отсутствует. Как уже отмечалось, ни один математико-статистический метод не может построить объяснительной модели, однако существует достаточно много методов проверки тех моделей, которые конструирует социолог.

Прежде всего, при анализе социологических данных нас интересуют причинные модели, в которых некий показатель выступает как следствие каких-то причин. При таком анализе интересует то, насколько, в какой степени эти причины определяют данное следствие. Целый ряд технических проблем, и прежде всего различия в уровне измерения переменных-причин и переменных-следствий, приводит к тому, что для проверки корректности выдвигаемых социологами однотипных причинных моделей используются различные математико-статистические методы. В данной главе мы рассмотрим методы, позволяющие строить причинные модели в ситуации, когда переменная-следствие измерена по метрической шкале, а переменные-причины — по неметрическим шкалам (порядковым или номинальным).

3.1

Визуализация различий средних значений

Достаточно распространенная задача, с которой сталкивается социолог еще на этапе описания собранных данных, это демонстрация средних значений каких-то количественных показателей в социальных, демографических или каких-то иных группах. Например, необходимо сопоставить величину средней заработной платы в группах респондентов, опрошенных в разных типах населенных пунктов, либо сравнить средний возраст людей, проголосовавших за разных кандидатов на выборах, и т.п.

Данный тип задач напоминает задачи описательного анализа с помощью одномерных частотных распределений, однако в рассматриваемом случае нам требуется получить средние значения количественного показателя не во всей выборке, а отдельно по нескольким группам. Так же как и при анализе одномерных распределений, результатом решения означенной задачи являются либо статистические характеристики, либо графические формы представления данных.

Построение статистических таблиц в рамках пакета программ SPSS выполняется с помощью специальной команды *Means* в рамках блока команд *Compare Means* (рис. 3.1).

В главном меню команды *Means* (рис. 3.2) видно, что необходимо задать два типа переменных. Первый тип переменных называется Dependent List (зависимые переменные) — это переменные, средние значения которых необходимо вычислять. На рис. 3.2 это переменная ЧЧ—размер заработка за последний месяц. Второй тип переменных — Independent List (независимые переменные) — это те переменные, которые определяют разделение всей совокупности опрошенных на оп-

ределенные группы. На рис. 3.2 эту функцию выполняет переменная adm — тип населенного пункта.

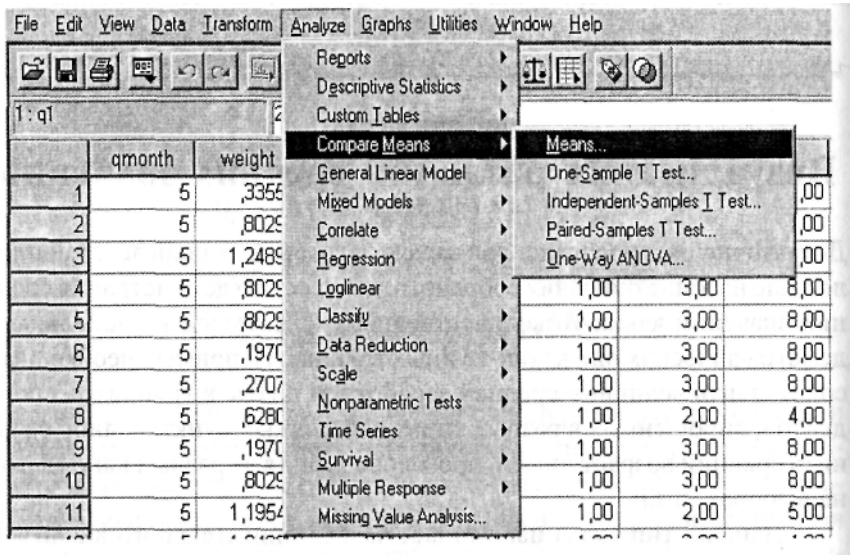


***Рис.** 3.1.* Меню обращения к команде Means

Результаты выполнения команды, изображенной на рис. 3.2, приведены в табл. 3.1.

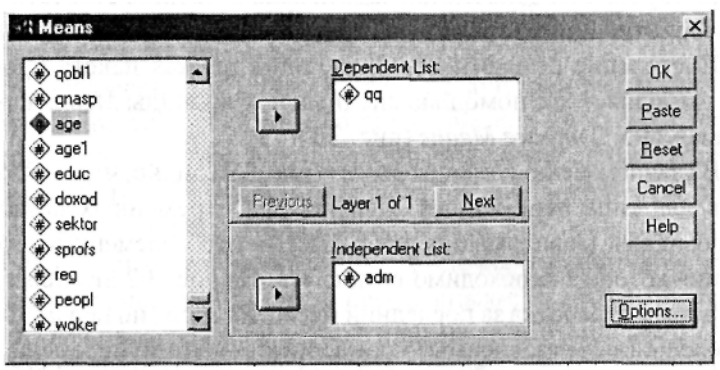


Рис. 3.2. Главное меню команды Means

Таблица 3.1. Результаты выполнения команды Means, приведенной на рис. 3.2

Тип **населенного пункта**	**Mean**	*N*	**Std. Deviation**
Москва, Санкт-Петербург	10823,60	118	13627,817
Большие города	6908,04	328	5066,637
Малые города	6375,22	418	6811,442
Села	5033,98	245	5007,508
Total	6708,22	1109	7252,782

Как показывает табл. 3.1, команда *Means,* если специально не указываются дополнительные параметры, вместе со средними значениями вычисляет еще две статистические характеристики: *N*— количество респондентов в каждой из выделенных групп и *Std. Deviation* (стандартное отклонение) анализируемого показателя в каждой из этих групп.

Команда *Means* наряду с этими параметрами может вычислять и другие, дополнительные характеристики. Выбор характеристик осуществляется нажатием клавиши Options... (в правом нижнем углу главного меню команды *Means,* см. рис. 3.2). На рис. 3.3 приводится меню, которое вызывается нажатием этой клавиши.

Графическое представление средних значений количественной переменной в нескольких группах, задаваемых какой-то неколичественной переменной, осуществляется в рамках блока команд *Graph*. Данный блок команд позволяет строить разные типы графиков. Рассмотрим меню построения столбиковых диаграмм (рис. 3.4, 3.5) для той же пары переменных, которые рассмотрены в примере для команды *Means.*

На рис. 3.6 приведена столбиковая диаграмма, определенная характеристиками, заданными в меню рис. 3.5.

Все рассмотренные возможности представления данных с помощью команд *Means* и *Graph* по характеру решаемых задач идентичны команде *Frequencies* и являются, по сути, командами блока описательной статистики. Интересно в этой связи, что в более ранних версиях пакета SPSS команда *Means* находилась именно в этом блоке команд.

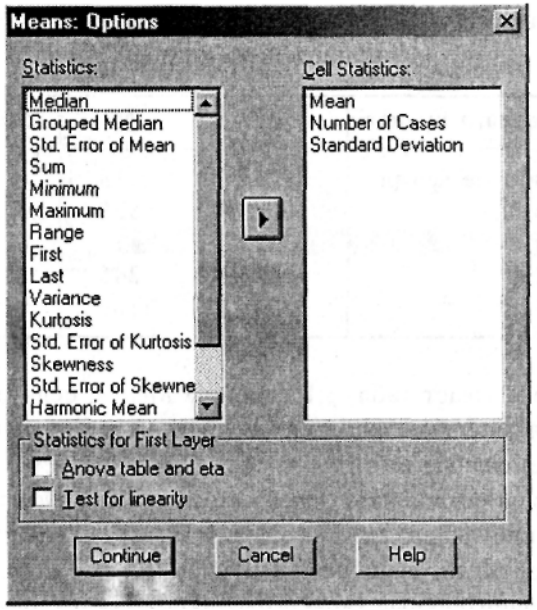


Рис. 3.3. Меню Options команды Means

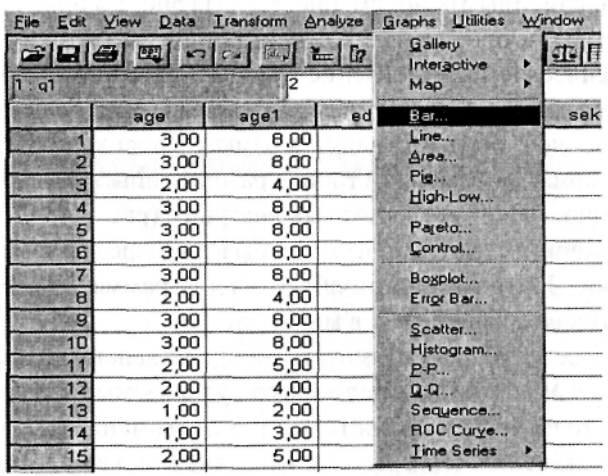


Рис. 3.4. Вызов команды построения столбиковых диаграмм из блока команд Graph

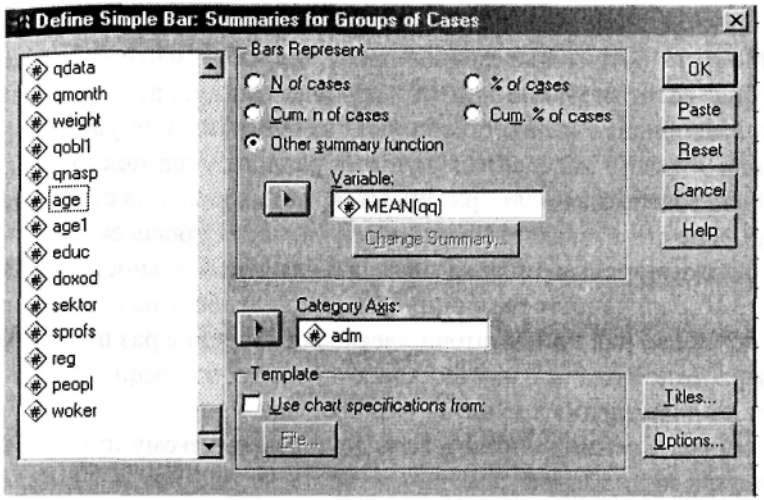
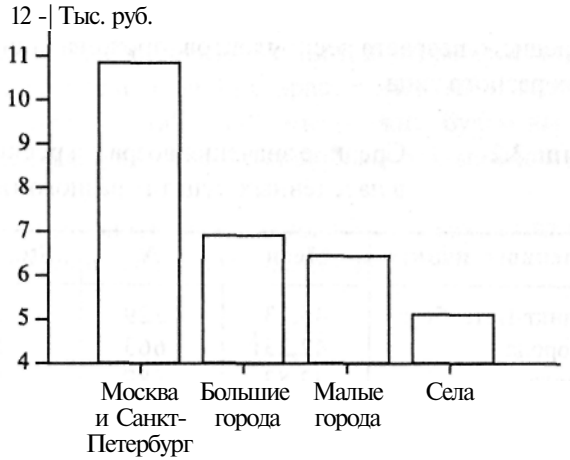


Рис. 3.5. Меню команды построения столбиковых диаграмм



?ис. 5.6. Столбиковая диаграмма значений средней заработной платы в населенных пунктах различного типа

Рассмотренные методы работы со средними значениями показателей, измеренных по метрическим шкалам, не доказывают того, что

эти средние значения в группах, задаваемых неколичественной переменной, статистически различаются между собой, и что, следовательно, у нас есть основания считать неколичественную переменную причиной изменения количественной. Пример сравнения уровней заработной платы в населенных пунктах разного типа показывает, что средняя заработная плата различается весьма и весьма существенно, что, вроде бы, избавляет нас от необходимости проведения дальнейшего статистического анализа. Визуальная убедительность (особенно рис. 3.6) полученных различий сама подталкивает нас к мысли, что! мы *доказали,* нашли достаточно веское объяснение различий в уровне заработной платы в нашей стране — это то, что респонденты проживают в населенных пунктах разного типа.

Однако важно понимать, что, двигаясь по такому пути работы с данными, мы, на самом деле, ничего не доказали. На этом пути мы и не могли ничего доказать, поскольку этот путь — не путь доказательства.

Рассмотрим пример, который не имеет столь очевидного решения. В табл. 3.2 представлены результаты команды *Means* при оценке различий среднего возраста респондентов, проживающих в населенных пунктах разного типа.

Таблица 3.2. Средние значения возраста респондентов в населенных пунктах разного типа

Тип населенного пункта	Mean	*N*	Std. Deviation
Москва, Санкт-Петербург	45,13	229	17,906
Большие города	42,23	663	17,077
Малые города	43,83	887	18,279
Села	45,65	628	18,356
Total	43,99	2407	17,977

Данные табл. 3.2 показывают, что средний возраст респондентов, проживающих в населенных пунктах разного типа, различается, хотя различия и не очень велики. Можем ли мы на основании данных табл. 3.2 утверждать, что средний возраст жителей населенных пунктов разного типа действительно различается, или эти различия носят случайный, статистически незначимый характер? Средства команды *Means* (равно как и команды графического представления данных) не позволяют нам ответить на этот вопрос.

В блоке команд Compare means представлены две команды, которые решают задачу математического доказательства наличия либо отсутствия различий средних значений. Это команды *T-Test* и *One-Way ANOVA.*

3.2

Команда T-Test

Команда *T-Test* (или тест Стьюдента) решает задачу *доказательства* наличия различий средних значений количественной переменной в усеченном виде, а именно в случае, когда имеются лишь две сравниваемые группы. Таким образом, если мы хотим ответить на вопрос о том, различается ли средний возраст у жителей населенных пунктов разного типа (см. табл. 3.2), мы должны будем выполнить эту команду несколько раз, попарно сравнивая разные типы населенных пунктов.

Есть три разновидности команды *T-Test,* каждая из которых соответствует разным исследовательским задачам.

3.2.1

Команда T-Test для сравнения двух независимых выборок

Пусть мы имеем две группы респондентов, для каждой из которых измерены средние значения некоторой количественной переменной. Для социологических исследований важное допущение о том, что эти

две группы (две выборки) являются независимыми, почти всегда выполняется. Действительно, если мы сравниваем выборки в двух типах населенных пунктов либо выборки мужчин и женщин и т.п., мы знаем, что сбор данных в этих группах выполняется независимо. Другими словами то, как отвечали женщины, никак не влияло на ответы мужчин и т.п.

Для статистической модели проверки гипотезы о равенстве средних значений в двух сравниваемых группах с помощью *T-Test* требуются еще допущения о дисперсии анализируемого количественного показателя в этих группах. Практически возможны две ситуации: дисперсии a и σ_2 анализируемой переменной x в двух группах одинаковы либо различны. Эти две ситуации приводят к тому, что для решения поставленной задачи применяются два разных статистических критерия.

Вызов команды сравнения средних значений количественной переменной в двух независимых выборках осуществляется путем перехода к соответствующему меню (рис. 3.7).

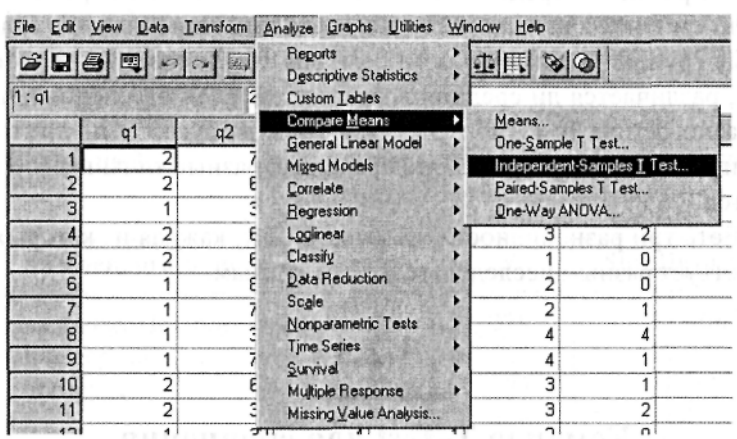


Рис. 3.7. Местоположение команды T-Test для сравнения средних значений в двух независимых выборках

На рис. 3.8 показано меню команды *T-Test* для двух независимых выборок. В окне *Test Variable (s)* указываются имена переменных, сред-

ние значения которых будут сравниваться. В примере это переменная q2 — возраст респондента.

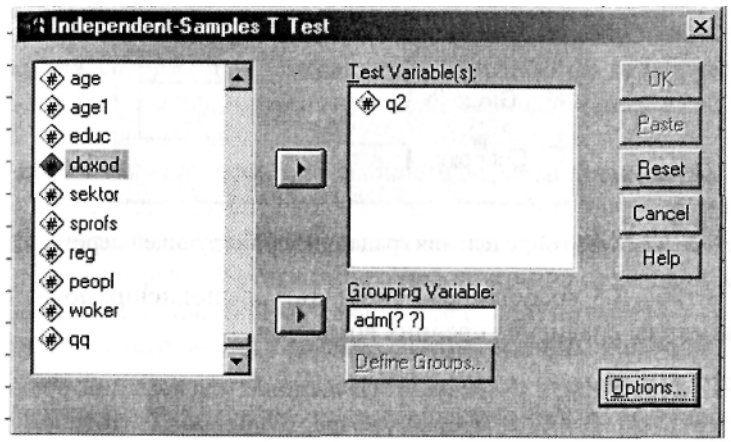


Рис. 3.8. Меню команды T-Test для двух независимых выборок

В окне *Grouping Variable* необходимо указать имя той переменной, которая будет определять разбиение респондентов на группы. В примере это переменная adm — тип населенного пункта. Обратите внимание, что в скобках после имени переменной adm проставлены знаки «??». Эти знаки автоматически проставляет программная система, напоминая, что кроме имени переменной необходимо указать номера градаций этой переменной. Дело в том, что переменная, которая выполняет разбиение респондентов на группы, может иметь более двух градаций (как, кстати, и есть в анализируемом случае — переменная adm имеет 4 градации — см. табл. 3.2). Поскольку команда *T-Test* может сравнивать только две группы респондентов, мы, нажав клавишу *Define Groups,* переходим к меню задания требуемых значений (рис. 3.9).

В предлагаемых окнах меню определения градаций необходимо задать числовые значения переменной. В случае решения задачи сравнения средних возрастов респондентов, проживающих в столицах и больших городах (коды переменной adm равны 1 и 2), необходимо Указать 1 и 2 соответственно.

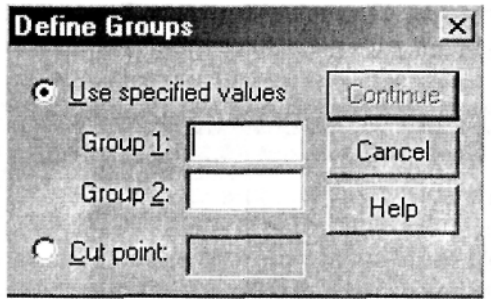


Рис. 3.9. Меню определения градаций группирующей переменной

В табл. 3.3 представлены результаты выполнения подготовленной команды сравнения средних возрастов.

Таблица 3.3. Результаты выполнения команды сравнения средних возрастов респондентов, проживающих в столицах и больших городах

Group Statistics

	Административный статус	N	**Mean**	**Std. Deviation**	Std. Error Mean
Возраст	Москва, Санкт-Петербург	229	45,13	17,906	1,183
	Большие города	663	42,23	17,077	0,663

Independent Samples Test

		Levene's Test for Equality of Variances		T-test for Equality of Means				
		F	**Sig.**	*t*	df	Sig. (2-**tailed)**	Mean Difference	Std. Error Difference
Возраст	Equal variances assumed	1,979	0,160	2,187	890	0,029	2,90	1,325
	Equal variances not assumed			2,138	381,531	0,033	2,90	1,356

Таблица 3.3 показывает, что команда *T-Test* в качестве результата выводит две таблицы. Первая — *Group Statistics* — содержит общую описательную информацию о поведении анализируемой переменной в двух отобранных группах респондентов. Отметим, что информация, содержащаяся в этой таблице, полностью дублирует данные по двум типам населенных пунктов, уже полученные нами с помощью команды *Means* (см. табл. 3.2).

Вторая таблица, вычисляемая командой *T-Test,* выполняет проверку статистической гипотезы о равенстве средних значений возраста в двух выделенных группах респондентов. Как уже указывалось, *T-Test* реализует разные статистические критерии в ситуациях, когда дисперсии количественной переменной (в нашем примере — возраста) в двух рассматриваемых группах различны либо одинаковы. Таблица *Group Statistics* (табл. 3.3) показывает, что выборочная дисперсия возраста у респондентов, проживающих в столицах, равна 5, = = 17,906, а у респондентов, проживающих в крупных городах, S_2= 17,077. Эти значения достаточно близки, однако можем ли мы с какой-либо уверенностью утверждать, что генеральные дисперсии равны? Первая часть таблицы *Independent Samples Test* выполняет проверку статистической гипотезы о равенстве генеральных дисперсий, т.е. проверяется статистическая гипотеза #$_0$: а, = $а_2$. Этот статистический тест называется тестом Левина проверки равенства дисперсий (Levene's Test for Equality of Variances). Таблица 3.3 показывает, что F-статистика этого теста равна 1,979, а значимость этой статистики (Sig.) — 0,16. Как мы можем интерпретировать полученное значение, гипотезу нужно принять или отвергнуть? Конечно, все зависит от принятого уровня значимости, но величина 0,16 превосходит наиболее употребительные в статистике уровни значимости, значит, гипотезу нужно принять. Таким образом, в нашем случае необходимо использовать статистику для ситуации равных дисперсий.

Вторая часть таблицы *Independent Samples Test* направлена на решение исходной задачи — проверку равенства средних (T-Test for Equality of Means). Здесь непосредственно проверяется статистическая гипотеза Я$_0$: ц, = Ц$_2$,где ц, и ц $_2$ — генеральные средние в соответствующих группах. Таблица *Independent Samples Test* включает две

строки, каждая из которых соответствует одной из ситуации — ра-Я венства дисперсий (Equal variances assumed) или различия дисперсий! (Equal variances not assumed). Поскольку мы выяснили, что в рассмат-И риваемом примере мы имеем дело скорее с ситуацией равенства дис-И персий в двух группах, необходимо ориентироваться на значения I Ьстатистики, приведенное в первой строке.

В анализируемом примере Sig = 0,029 (см. табл. 3.3). Это мень-И ше чем 0,05, но больше чем, например, 0,01. Снова мы оказываемся в I ситуации, когда все сильно зависит от того, какой уровень значимое- • ти мы считаем достаточным. Если остановиться на а = 0,05, гипотезу | следует отвергнуть (но при этом мы должны помнить, что ситуация достаточно неоднозначная). Таким образом, делаем вывод, что сред-ний возраст жителей столиц и крупных городов различен.

Полученный результат важен. Во-первых, он показывает, что| визуальная схожесть двух чисел не может служить доказательством их равенства. Во-вторых, и это более важно, T-Test дает нам инстру-мент статистической проверки, *доказательства* наличия (или отсут-ствия) статистической взаимосвязи двух переменных.

3.2.2

Команда T-Test для одной выборки

В ходе анализа социологических данных нередко возникает ситуация, когда необходимо сравнить среднее значение какой-то количествен-ной переменной с некоторым фиксированным значением. Например, в ходе исследования образа жизни было выяснено, что в среднем рес-понденты тратят на просмотр телепередач около двух часов. Из мате-риалов предыдущих исследований известно, что год назад респон-денты тратили на этот вид деятельности приблизительно 1,8 часа J Можем ли мы, опираясь на эту информацию, *утверждать,* что за прошедший год люди стали больше времени проводить у телевизора или обнаруженная разница носит случайный, статистически не зна чимый характер? Другая исследовательская ситуация определяется

необходимостью оценки репрезентативности проведенного опроса по количественным показателям. Если, скажем, мы провели всероссий-ский опрос, для оценки репрезентативности по параметру «возраст» требуется сопоставить данные опроса с материалами, представляе-мыми органами государственной статистики.

Общим в двух рассмотренных примерах является то, что мы долж-ны оценить значимость различий между данными опроса и некоторы-ми «внешними» цифрами. Переводя эту задачу на язык математичес-кой статистики, можно сказать, что требуется провести проверку гипо-тезы H_0: $\backslash x = c$, где $ц$ — среднее значение количественной переменной; c — константа. Альтернативной гипотезой выступает H: (Л Φ c.

В меню рис. 3.7 необходимо выбрать команду *One-Sample T-Test.* На рис. 3.10 приводится меню этой команды. В данном примере про-веряется, совпадает или нет среднее значение возраста (переменная q2) в проведенном всероссийском репрезентативном опросе с данны-ми о среднем возрасте взрослого населения, имеющимися в материа-лах Госкомстата России[1].

В окне *Test Value* рис. 3.10 указывается то значение, с которым будет сопоставляться среднее значение тестируемой переменной.

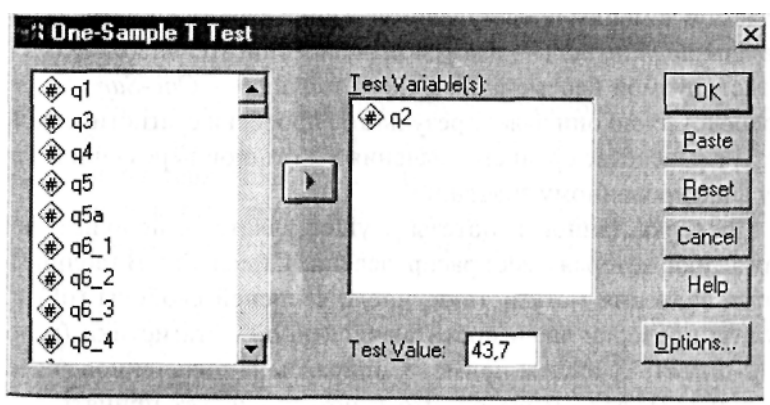


Рис. 3.10. Меню команды T-Test для одной выборки

[1] См.: Российский статистический ежегодник. 2002. М., 2002. С. 87.

В табл. 3.4 приводятся результаты выполнения команды, меню которой показано на рис. 3.10.

Таблица 3.4. Результаты выполнения команды сравнения среднего возраста опрошенных с данными Госкомстата России

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Возраст	2407	43,99	17,977	0,366

One-Sample Test

	Test Value = 43,7			
	t	df	Sig. (2-tailed)	Mean Difference
Возраст	**0,784**	**2406**	**0,433**	0,29

Так же как и результаты выполнения команды сравнения средних в независимых выборках (см. табл. 3.3), результаты работы обсуждаемой команды представляются в виде двух таблиц. В первой — *{One-Sample Statistics)* — даются значения описательных характеристик тестируемой переменной. Вторая таблица — *One-Sample Test* — непосредственно описывает результаты проверки статистической гипотезы о равенстве среднего значения выбранной переменной заданному фиксированному значению.

Проверка данной гипотезы осуществляется с использованием /-статистики, которая имеет распределение Стьюдента. В таблице приводятся значения /-статистики, число степеней свободы (df) и результаты проверки значимости вычисленной /-статистики (колонка *Sig. (2-tailed))*. В нашем примере, опираясь на полученное значение (Sig. = 0,433), делаем вывод, что гипотезу следует принять. Таким образом, мы можем сказать, что полученные данные о среднем возрасте близки к официальной статистике.

3.2.3

Команда T-Test для парных данных

Еще одна исследовательская задача, которая достаточно часто возникает при анализе социологических данных, это ситуация сравнения средних значений двух отдельных переменных на предмет выяснения вопроса о том, среднее значение какой из них больше, а какой — меньше. Например, в ходе панельного социологического исследований «Российский мониторинг экономики и здоровья» (RLMS) одним из направлений изучения является анализ экономических источников жизни россиян. В рамках этого направления в ходе опроса у респондентов спрашивали, какое количество овощей, выращенных на собственных приусадебных или дачных участках, они употребили в течение последнего года, а какую часть овощей продали. В этой связи интересно выяснить, есть ли разница между размерами доходов семей от продажи овощей.

Очевидно, что в данном случае мы снова имеем дело с задачей проверки различия средних значений, но уже в отношении двух отдельных переменных, измеренных в рамках одной выборки. Эту ситуацию иногда называют *сравнением парных данных.* С точки зрения математической статистики в данном случае проверяется статистическая гипотеза H_0: ц, = ц$_2$, где Ц„ Ц$_2$ — средние значения переменных *х* и x_2, против альтернативы *Н: ц$_1$ Ф* ц$_2$.

Решение данной задачи в рамках пакета SPSS осуществляется с помощью команды *Paired Samples T-Test* (Т-тест для парных выборок) (рис. 3.11).

В меню показано, что требуется сравнение средних значений переменных ed8.4b и ed8.5d — переменных, фиксирующих количество свеклы и моркови, которые собрали семьи на своих приусадебных участках[2]. Результаты работы команды приведены в табл. 3.5.

[2] Для примера взяты данные исследования RLMS 9-й волны (октябрь-ноябрь 2001г.).

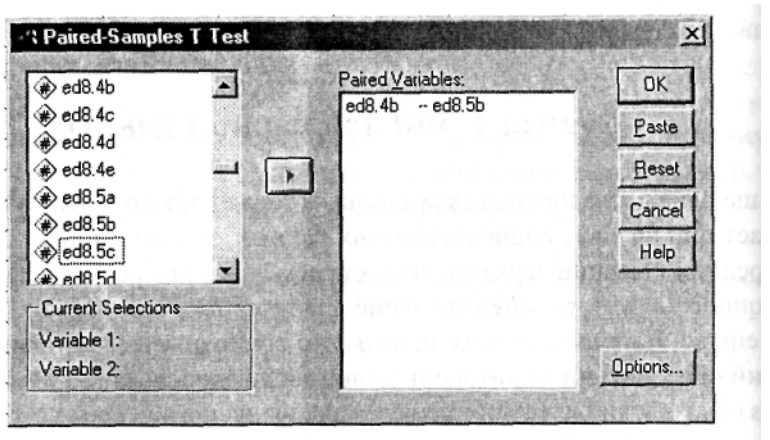


Рис. i.ii. Меню команды сравнения средних значений в парных переменных

Таблица 3.5. Результаты выполнения команды сравнения средних значений двух переменных

Paired Samples Statistics

		Mean	*N*	Std. Deviation	Std. Error Mean
Pair 1	Свекла	96,5359	1841	2335,66058	54,43558
	Морковь	73,1684	1841	1166,70392	27,19154

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Свекла & Морковь	1841	0,996	0,000

Paired Samples Test

		Paired Differences			*t*	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	Свекла — Морковь	23,3675	1178,5864	27,46847	0,851	1840	0,395

Результаты работы команды *T-Test* на парных данных представлены в виде трех таблиц. Первая содержит показатели описательной статистики для пары анализируемых переменных, вторая — коэффициент корреляции Пирсона между этими переменными с оценкой уровня значимости этого коэффициента. Таблица *Paired Samples Test* собственно и посвящена представлению результатов проверки статистической гипотезы H_0. Для проверки этой гипотезы, как и в предыдущих командах *T-Test,* используется /-статистика. В таблице содержатся значения /-статистики, которая имеет распределение Стьюдента, число степеней свободы (df) и оценка значимости полученного значения /-статистики (Sig. 2-tailed). В рамках рассматриваемого примера получаем, что гипотезу следует принять. Это означает, что, несмотря на, как кажется, весьма существенную разницу между средним количеством собираемых свеклы и моркови, мы не имеем статистических оснований утверждать наличие различий в объемах выращивания россиянами этих овощей.

Почему в данном случае довольно большие различия в объемах выращивания рассматриваемых овощей (96,5 кг и 73,2 кг), тем не менее, не дают статистических оснований для констатации статистической разницы? Ответ довольно прост: обе переменные имеют очень высокие значения показателя вариации — стандартного отклонения.

"2 1

Однофакторный дисперсионный анализ

Рассмотренные возможности применения разных модификаций T-Test (теста Стьюдента) показывали и существенные ограничения этого Метода. Например, приведенные в табл. 3.1 результаты работы команды *Means* свидетельствуют о том, что в данном случае число градаций в качественной переменной больше двух. Т-тест позволяет сопоставить только две градации. Как быть в данной ситуации? Иными слова-

ми, как проверить статистическую гипотезу *Щ* $1, = Ц$_2$ = ... = ц где! Цр |Д$_{,2}$, ..., ц$_л$ — средние значения анализируемых переменных в! *п* независимых выборках? Данная задача решается с помощью методов дисперсионного анализа.

С точки зрения построения социологической модели вопрос можно сформулировать следующим образом: оказывает ли значимое влияние на значение некоторой количественной переменной интересующая нас переменная, которая измерена на номинальном или порядковом уровне? В терминах метода дисперсионного анализа та переменная, которая, как мы считаем, должна оказывать влияние на конечный результат, называется *фактором*. Например, если для данных табл. 3.1 мы начнем строить модель объяснения различий в заработных платах респондентов тем, что респонденты проживают в населенных пунктах разного типа, переменная «Тип населенного пункта» будет выступать фактором.

Конкретную реализацию, значение фактора (например, определенный тип населенного пункта) называют *уровнем фактора*. Значение измеряемого признака (в нашем примере — величину заработной платы) называют *откликом*.

Само название дисперсионного анализа происходит из того, что метод проверки статистической гипотезы #$_0$ о равенстве средних значений в нескольких непересекающихся группах респондентов основан на сопоставлении двух оценок дисперсии анализируемой количественной переменной (о чем речь пойдет ниже).

В рамках пакета SPSS программа, реализующая метод однофакторного дисперсионного анализа, называется One-Way ANOVA и расположена в блоке команд *Compare means* (см. рис. 3.1). Название One-Way отражает тот факт, что эта программа выполняет метод *однофакторного дисперсионного анализа,* т.е. анализируется влияние только *одной* качественной переменной (фактора) на количественную переменную. Слово «ANOVA» — аббревиатура и расшифровывается как ANAlisys Of Variance, или дисперсионный анализ.

Главное меню команды *One-Way ANOVA* показано на рис. 3.12.

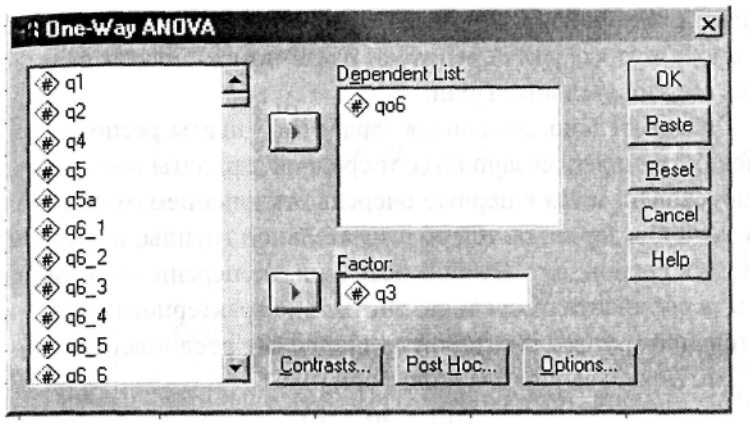


Рис. 3.12. Главное меню команды One-Way ANOVA

На рис. 3.12 рассматривается проверка модели о равенстве средних значений заработной платы (переменная qo6) у респондентов с разным уровнем образования (образование — переменная q3).

Для понимания результатов, которые вычисляет и представляет пользователю команда *One-Way ANOVA,* необходимо рассмотреть принципы работы метода однофакторного дисперсионного анализа. Модель факторного анализа достаточно прозрачна и основана на двух допущениях. Во-первых, анализируя модель влияния образования на заработную плату респондентов, мы предполагаем, что респонденты, имеющие разный уровень образования, имеют разную зарплату. Другими словами, если у нас есть переменная (фактор) «образование» с несколькими уровнями, скажем, *п* уровнями, средние значения заработной платы (отклика) ц,у = 1, ... , *п* при этих уровнях фактора не равны между собой. Важно понимать, что мы не требуем, чтобы *все \х.* были различны. Главное, чтобы не все они были одинаковы.

Во-вторых, мы понимаем, что на заработную плату респондента влияет не только его образование, и, следовательно, зарплаты респондентов, принадлежащих к одной образовательной группе, также могут весьма существенно различаться. Таким образом, зарплата складывается из двух факторов: из средней зарплаты той образовательной

группы, к которой принадлежит респондент, и из каких-то других факторов, которые уже определяют различие в зарплатах внутри каждой из образовательных групп.

Следовательно, дисперсия заработной платы респондента, или степень отличия этой зарплаты от средней зарплаты всех респондентов обусловливается в первую очередь отклонением от общего среднего значения зарплаты той образовательной группы, к которой принадлежит респондент. Второй источник дисперсии — отличия зарплаты респондента от среднего значения, характерного для той образовательной группы, к которой принадлежит респондент.

Это можно записать в виде формулы:

$$S^2 = S_b^2 + S_w^2, \tag{3.1}$$

где S^2 — общая (total) дисперсия; S_b — *межгрупповая* (between) дисперсия, т.е. дисперсия, характеризующая различия средних значений зарплат разных образовательных групп; S^2 — *внутригрупповая* (within) дисперсия, т.е. дисперсия зарплат внутри каждой образовательной группы.

Зарплату отдельного г'-го респондента, принадлежащего ку'-й образовательной группе, обозначенную как *х..,* можно представить как

$$x_{ij} = \mu + (x_{ij} - \mu_j) - (\mu_j - \mu),$$

где |LL— средняя зарплата ву'-й группе, к которой принадлежит респондент; ц — средняя зарплата во всей совокупности. Выражение *(х.* - ц.) в формуле (3.2) говорит о внутригрупповом различии, выражение (ц - \i) — о межгрупповом.

Из формулы (3.2) достаточно просто получить и формулу *разложения дисперсии* на межгрупповую и внутригрупповую. При анализе модели дисперсионного анализа обычно оперируют не полной формулой дисперсии (1.1), а лишь ее числителем, который называют суммой квадратов (Sum of Squares). Суммы квадратов для межгрупповой и внугригрупповой составляющих представлены в формулах (3.3) и (3.4). Общая сумма квадратов — это числитель из формулы дисперсии (1.1):

$$SS_w = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_{ij} - \mu_j)^2; \tag{3.3}$$

$$SS_b = \sum_{j=1}^{n} n_j(\mu_j - \mu)^2; \tag{3.4}$$

$$SS_T = \sum_{j=1}^{n}\sum_{i=1}^{n_j}(x_{ij} - \mu)^2, \tag{3.5}$$

где л. — объем *j-й* группы.

Каждая из сумм квадратов обладает неким числом степеней свободы и будучи на него поделенной (нормированной) представляет собой, по сути дела, дисперсию и в терминологии дисперсионного анализа называется средним квадратом (Mean Square).

Все эти характеристики необходимы для вычисления F-статистики, которая служит инструментом для проверки исходной гипотезы *Н*: ц. = |0. = ... = \i_n. F-статистика — это отношение межгруппового и внутригруппового нормированных средних квадратов. При этом F-статистика имеет F-распределение, что дает нам инструмент для проверки уровня значимости и соответственно для принятия или непринятия гипотезы H_0.

Результаты выполнения команды *One-Way ANOVA,* заданной в меню рис. 3.12, представлены в табл. 3.6.

В табл. 3.6 приводится межгрупповая (Between Groups), внутригрупповая (Within Groups) и общая (Total) суммы квадратов. Далее следуют числа степеней свободы для этих трех сумм (df), средние суммы квадратов и значение F-статистики. Наконец, колонка *Sig.* содержит значимость полученного значения F-статистики.

Полученный результат говорит нам, что вероятность справедливости гипотезы $Я_0$ крайне мала. Другими словами, у нас имеются убедительные причины отвергнуть H_0 и согласиться с альтернативной гипотезой, т.е. с предположением о том, что не все образовательные группы имеют одинаковую среднюю зарплату.

Таблица 3.6. Результаты выполнения команды *One- Way ANOVA* по проверке модели различия средней зарплаты в различных образовательных группах

ANOVA

	Sum of Squares	df	**Mean Square**	*F*	Sig.
Between Groups	3637891345,3	6	606315224,2	5,666	0,000
Within Groups	114704562245,6	1072	107000524,4		
Total	118342453590,9	1078			

3.4

Методы множественных сравнений

Полученный результат, указывающий, что средние зарплаты у респондентов, принадлежащих к разным образовательным группам, различаются, не выглядит окончательным результатом социологического анализа данной проблемы. Это скорее промежуточный, или даже предварительный результат, который подразумевает дальнейшее раскрытие того, в какой из образовательных групп зарплаты больше, в какой меньше, а в каких, быть может, зарплаты одинаковы. Основная процедура дисперсионного анализа не дает возможности ответить на эти вопросы, однако в команде *One-WayANOVA* есть дополнительные возможности, которые направлены на решение этих задач с помощью методов *множественных сравнений*.

Суть методов множественных сравнений состоит в определении различий — совпадений средних значений количественной переменной во всех возможных парах групп, определяемых градациями переменной — фактора. Иными словами, если мы проводим множественные сравнения различий в уровнях заработной платы в образовательных группах, метод множественных сравнений построит все возможные пары уровней образования и сравнит среднюю зарплату в этих парах. Вызов данного метода выполняется нажатием клавиши Post Hoc..., расположенной в нижней части главного меню команды *One-Way ANOVA* (см. рис. 3.12).

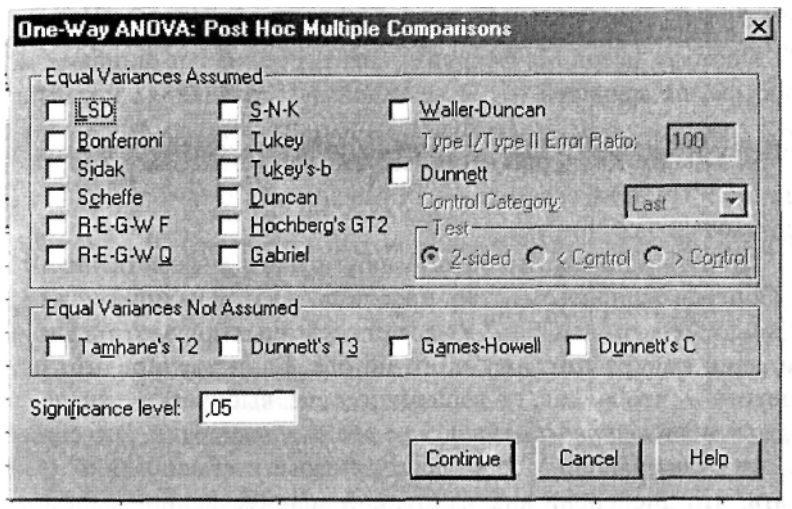


Рис. 3.13. Меню вызова методов множественных сравнений

Как видно из рис. 3.13, программная система SPSS предлагает на выбор 20 методов множественных сравнений. При этом все методы разбиты на две группы: те, в которых предполагается равенство дисперсий количественной переменной во всех группах, задаваемых фактором (Equal Variance Assumed), и те, в которых это равенство не предполагается (Equal Variance Not Assumed). Многообразие предлагаемых методов множественных сравнений определяется разнообразием математических моделей оценки статистических различий средних значений в совокупности всех получаемых пар.

Например, согласно методу Бонферрони, в случае множественных сравнений назначается более строгий уровень значимости. Он определяется следующим образом: задается уровень значимости для

множественных сравнений α_m и в качестве попарного уровня значимости берется $\alpha = (1 / k)\alpha_m$, где k — число сравнений. Пусть A_i — событие, состоящее в том, что мы в i-м сравнении выявили существенное отличие средних, когда средние совпадают, тогда, в соответствии с заданным уровнем значимости, $P\{A_i\} < \alpha$.

$$P\{A_1 + A_2 + ... + A_k\} \le P\{A_1\} + P\{A_2\} + ... + P\{A_k\} < k\alpha = \alpha_m.$$

Поэтому метод Бонферрони гарантирует нас от ошибки с вероятностью, не меньшей α_m. В независимых сравнениях неравенство $P\{A_1 + A_2 + ... + A_k\} < k\alpha$ будет выполняться почти точно, так как $1 - (1 - \alpha)^k \approx k\alpha$. Критерий несколько жестче, чем необходимо, так как средние в группах связаны — их взвешенная сумма равна общему среднему.

При необходимости использования подхода множественных сравнений перед социологом встает проблема — какой из множества предлагаемых методов выбрать? У нас нет серьезных аргументов для рекомендации выбора того или иного метода. Единственное, что можно отметить — что лучше, по всей видимости, выбирать метод из группы, в которой *не предполагается* равенство дисперсий. Это связано с тем, что, как правило, у социолога нет серьезных оснований предполагать, что дисперсии анализируемой количественной переменной будут одинаковы во всех группах, определяемых фактором. Отметим, что, как показывает практика, все предлагаемые методы дают весьма близкие результаты.

Продемонстрируем результаты, которые получаются с помощью метода множественных сравнений при решении задачи сравнения средних уровней заработной платы в разных образовательных группах с помощью метода Тамхена (Tamhen's) (табл. 3.7).

В табл. 3.7 приведено сопоставление всех пар уровней образования на предмет различия уровней заработной платы. Колонка *Mean Difference (I-J)* (разница средних значений в группах 7, J) содержит величины разницы заработной платы в парах сравниваемых групп. Колонка *Std. Error* дает значения стандартной ошибки среднего для рассматриваемой разности, а колонка *Sig.* показывает, с какой вероятностью мы можем принять гипотезу H_0 о равенстве средних в сравниваемых группах.

Таблица 3.7. Результаты выполнения метода множественных сравнений

Образование (I)	Образование (J)	Mean Difference (I-J)	Std. Error	Sig.
Общее начальное или неполное среднее	Общее полное среднее	1196,72	1208,23	1,000
	Профессионально-техническое с неполным средним образованием	-1400,7	2282,92	1,000
	Профессионально-техническое с неполным средним образованием	62,39	1245,70	1,000
	Среднее специальное	378,56	1220,94	1,000
	Неполное высшее	-3990,3	2773,76	0,972
	Высшее	-3320,6	1456,24	0,407
Общее полное среднее	Общее начальное или неполное среднее	-1196,7	1208,23	1,000
	Профессионально-техническое с неполным средним образованием	-2597,4	1999,94	0,991
	Профессионально-техническое с полным средним образованием	-1134,3	582,972	0,679
	Среднее специальное	-818,16	527,998	0,935
	Неполное высшее	-5187,0	2545,94	0,647
	Высшее	-4517,4	953,265	0,000
Профессионально-техническое с неполным средним образованием	Общее начальное или неполное среднее	1400,74	2282,92	1,000
	Общее полное среднее	2597,46	1999,94	0,991
	Профессионально-техническое с полным средним образованием	1779,30	2007,65	1,000
	Среднее специальное	-2589,5	3199,02	1,000
	Неполное высшее	-1919,9	2158,84	1,000
	Высшее			

Продолжение табл. 3.7

Образование (I)	**Образование (J)**	**Mean Difference (I-J)**	**Std. Error**	Sig.
Профессионально-техническое с полным средним образованием	Общее начальное или неполное среднее	-62,39	1245,70	1,000
	Общее полное среднее	1134,33	582,972	0,679
	Профессионально-техническое с неполным средним образованием	-1463,1	2022,80	1,000
	Среднее специальное	316,17	608,883	1,000
	Неполное высшее	-4052,6	2563,94	0,934
	Высшее	-3383,0	1000,33	0,016
Среднее специальное	Общее начальное или неполное среднее	-378,56	1220,94	1,000
	Общее полное среднее	818,16	527,998	0,935
	Профессионально-техническое с неполным средним образованием	-1779,3	2007,65	1,000
	Профессионально-техническое с полным средним образованием	-316,17	608,883	1,000
	Неполное высшее	-4368,8	2552,00	0,876
	Высшее	-3699,2	969,328	0,003
Неполное высшее	Общее начальное или неполное среднее	3990,30	2773,76	0,972
	Общее полное среднее	5187,02	2545,97	0,647
	Профессионально-техническое с неполным средним образованием	2589,56	3199,02	1,000
	Профессионально-техническое с полным средним образованием	4052,69	2563,93	0,934
	Среднее специальное	4368,86	2552,05	0,876
	Высшее	669,61	2672,56	1,000

Окончание табл. 3.7

Образование (I)	**Образование (J)**	**Mean Difference (I-J)**	**Std. Error**	Sig.
Высшее	Общее начальное или неполное среднее	3320,69	1456,22	0,407
	Общее полное среднее	4517,41	953,265	0,000
	Профессионально-техническое с неполным средним образованием	1919,94	2158,85	1,000
	Профессионально-техническое с полным средним образованием	3383,08	1000,35	0,016
	Среднее специальное	3699,25	969,328	0,003
	Неполное высшее	-669,61	2672,56	1,000

3.5

Дисперсионный анализ Краскэла — Уоллиса

До сих пор мы рассматривали ситуацию, в которой сравнивались (и анализировались результаты сравнения) средние значения переменной, измеренной по метрической шкале либо в двух группах *(T-Test),* либо в *n* группах, задаваемых уровнями фактора *(ANOVA).*

Однако эти подходы имеют два существенных недостатка. Во-первых, в основе используемых статистических моделей лежит допущение о том, что в анализируемых выборках (одной или нескольких) рассматриваемые параметры имеют *нормальное* распределение. Например, T-Test для оценки различия средних значений какого-то показателя в двух независимых выборках основан на предположении, что значения этого показателя в данных выборках имеют нормальное

распределение. В определенных случаях такое допущение кажется' вполне естественным. Скажем, если мы пытаемся сравнить средний! рост мужчин и женщин, предположение о том, что данный показа-! тель в этих группах распределен нормально, не выглядит странным Вместе с тем во многих случаях предположение о нормальности обос- і новать довольно трудно, а подчас можно точно сказать, что распреде*! ление резко отличается от нормального.

Вторым недостатком является то, что данные методы предназначены для фиксации различий в значениях переменных, измеренных по количественным (интервальным либо абсолютным) шкалам, а пе-м ременные этого типа в данных социологических исследований встречаются достаточно редко. Материалы анкетных опросов преимущественно состоят из переменных, измеренных по порядковым или номинальным шкалам. Существует подход, позволяющий применять метод дисперсионного анализа для ситуации, когда переменная измерена по порядковой (ранговой) шкале, который называется дисперсионным анализом Краскэла — Уоллиса.

При работе с ранговыми переменными учитывается лишь упорядоченность значений. Суть ранговых (порядковых) шкал состоит в том, что в данных кодируется некоторая числовая информация, но используются только ранги. В ряде методов при вычислении критериев по имеющимся числовым значениям исследуемой переменной объектам приписываются ранги. Для вычисления рангов объекты упорядочиваются от минимального значения переменной к максимальному, и порядковые номера объектов считаются рангами. Если для некоторой последовательности объектов числовые значения переменной повторяются, этим объектам приписывается средний ранг по этой последовательности. Об объектах, ранги которых совпадают, говорят, что они имеют связанные ранги. Наличие связанных рангов в выдаче по ранговым тестам обозначается словом «ties» (связи). Обычно выводится число связей и статистика критерия, скорректированная для связей.

В качестве примера рассмотрим упорядоченную информацию об успеваемости семи студентов.

I

Средний балл	3,0	3,1	4,0	4,2	4,2	4,5	5,0
Ранг	1	2	3	4,5	4,5	6	7

Первые три объекта имеют ранги 1, 2, 3; следующая пара — 4 5 = (4 + 5) / 2, последняя пара — 6 и 7. Если предположить, что первые три студента в этой последовательности — юноши, а остальные — девушки, можно ввести понятие *среднего ранга* у студентов разного пола. Это будут просто средние суммы рангов у студентов разного пола. Соответственно у юношей-студентов средний ранг будет равен 2, у девушек — 4,5.

В основе метода дисперсионного анализа Краскэла — Уоллиса лежит однофакторный дисперсионный анализ, в котором вместо значений переменных используется ранг объекта по исследуемой переменной, проводится сравнение средних произвольного числа групп. Нормированный межгрупповой разброс в условиях гипотезы равенства средних рангов в группах имеет распределение, близкое к распределению $\%^2$.

Метод дисперсионного анализа Краскэла — Уоллиса в пакете программ SPSS выполняется через блок команд *Nonparametric Tests,* в котором выбирается команда *K Independent Samples* (рис. 3.14).

Меню самой команды сравнения средних рангов в группах, определяемых переменной-фактором (в терминах данного метода анализа эта переменная называется группирующей переменной — *Grouping Variable),* представлено на рис. 3.15.

В меню (см. рис. 3.15) проверяется модель влияния пола (переменная q1) на настроение человека (переменная q9). Результаты анализа данной модели записаны в табл. 3.8 и представлены в виде двух таблиц. В первой — *Ranks* приводятся данные о количестве респондентов, принадлежащих каждой из градаций группирующей переменной и средний ранг (Mean Rank) анализируемой переменной в каждой из этих групп.

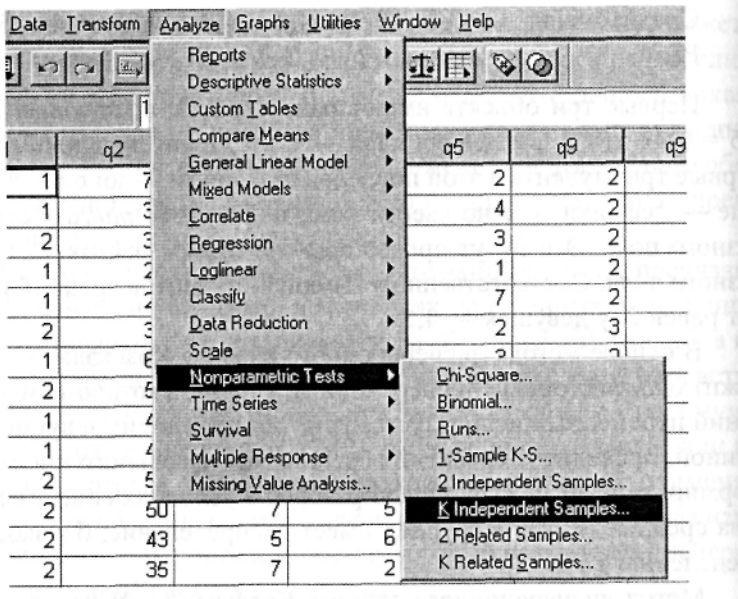


***Рис.** 3.14.* Меню перехода к команде K Independent Samples дисперсионного анализа Краскэла — Уоллиса

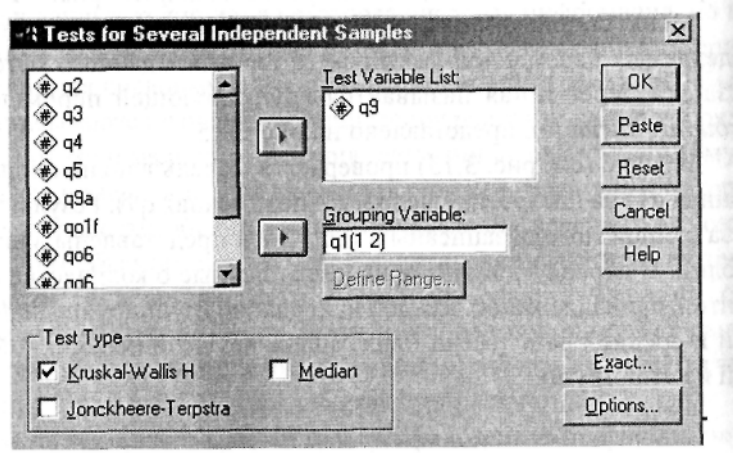


***Рис.** 3.15.* Главное меню команды дисперсионного анализа Краскэла — Уоллиса

Вторая таблица — *Test Statistics* содержит результаты проверки статистической гипотезы о том, что средние ранги в каждой из сравниваемых групп равны между собой. Очевидно, что эта гипотеза эквивалентна гипотезе об отсутствии влияния группирующей переменной на анализируемую переменную. Строка *Asymp. Sig.* представляет значимость тестируемой гипотезы H_0.

Таблица 3.8. Результаты проверки модели влияния пола на настроение

Ranks

	Пол	*N*	**Mean Rank**
Что вы могли бы сказать о своем настроении в последние дни?	Мужской Женский	444 560	471,95 526,72
Total		1004	

Test Statistics

	Что вы могли бы сказать о своем настроении в последние дни?
Chi-Square df Asymp. Sig.	11,638 **1** 0,001

Какой результат демонстрирует табл. 3.8, с точки зрения проверяемой социологической модели? Прежде всего, вычисленная значимость показывает, что у нас имеются все основания отвергнуть исходную гипотезу об отсутствии влияния пола на настроение на уровне значимости **a** = 0,001, т.е. говорить о наличии связи между полом и настроением.

Можем ли мы что-нибудь сказать о характере выявленной зависимости? Иными словами, у кого, в среднем выше настроение — у Мужчин или женщин? Поскольку анализируемая переменная «Настроение» имеет определенный порядок, то таблица *Ranks* табл. 3.8 показывает, что средний ранг ответов на этот вопрос у мужчин ниже, чем

у женщин. Для корректного ответа на поставленный вопрос необхо-) димо взглянуть на формулировку самого вопроса анкеты[3].

q9 Что Вы могли бы сказать о своем настроении в последние дни?

1. Прекрасное настроение.
2. Нормальное, ровное состояние.
3. Испытываю напряжение, раздражение.
4. Испытываю страх, тоску.

Приведенный вопрос показывает, что меньший ранг по шкале переменной соответствует более хорошему настроению. Следовательно, можно сказать, что в среднем настроение у мужчин значительно лучше, чем у женщин.

Необходимо также отметить, что дисперсионный анализ Краскэла — Уоллиса, равно как и в целом методы дисперсионного анализа, J решает только задачу фиксации наличия связи (точнее — отсутствия независимости) между количественной и неколичественной переменными. Мы не получаем информации о форме этой связи, однако, в] некоторых случаях, имеем информацию о ее направлении.

[3] Используется вопрос из исследования ВЦИОМ // Мониторинг общественного мнения: экономические и социальные перемены. 2002. № 6. С. 74.

4 глава

МОДЕЛИ РЕГРЕССИОННОГО АНАЛИЗА

Начнем с примера. Выяснение причин хорошей или плохой успеваемости студентов является, несомненно, сложной задачей. Социологические теории, да и просто здравый смысл подсказывают нам, что среди факторов, влияющих на успеваемость, должны присутствовать:

- уровень подготовки студента;
- активность посещения занятий;
- активность самостоятельной работы;
- способности студента.

Очевидно, что этот список неполон и может быть расширен за счет других характеристик, однако ограничимся пока только этими.

Представим схему влияния различных показателей на успеваемость в виде рисунка (рис. 4.1).

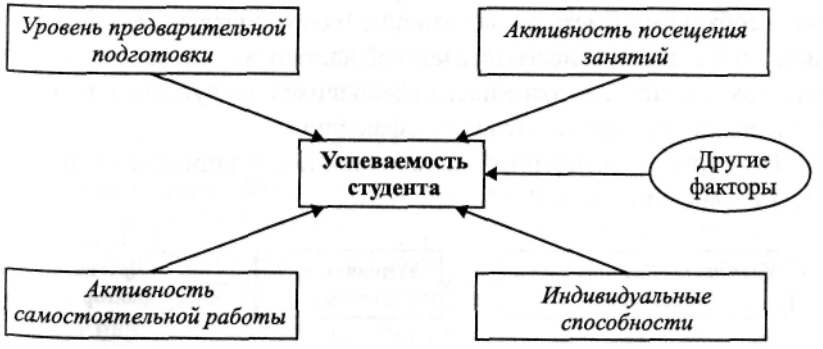


Рис. 4.1. Модель «Успеваемость студента»

не

Рисунок 4.1 можно рассматривать как *модель* успеваемости, или как некоторую схему, которая позволяет систематизировать наши взгляды на изучаемое явление. Анализируя эмпирические данные, можно попытаться проверить, насколько наша модель соответствует тем реальным процессам, которые управляют успеваемостью и данные о которых можно собрать с помощью социологических методов.

Пока, однако, в нашем распоряжении есть только инструменты проверки парных взаимосвязей между переменными — коэффициенты сопряженности и корреляции. При этом сами коэффициенты фактически фиксируют не то, насколько *сильно* взаимосвязаны два показателя между собой, а то, насколько *тесно* они взаимосвязаны.

Теснота взаимосвязи является, несомненно, важной характеристикой, но на практике интереснее сила связи. Так, мы знаем, что если солить еду, она становится солонее. Другими словами, эти характеристики взаимосвязаны, и, по всей видимости, достаточно тесно. Однако крайне важно знать и то, насколько становится солонее блюдо при добавлении определенного количества соли. Зависит это и от характеристик соли, и от особенностей используемых продуктов, и от специфики процесса приготовления, но, согласитесь, без этого знания вкусного блюда не приготовишь.

В модели, представленной на рис. 4.1, для нас принципиально важно не только наличие обозначенных стрелок. Чтобы модель давала нам полезную информацию, которую можно использовать на практике, необходимо иметь представление о силе соответствующих связей, т.е. понимать, какие из показателей влияют на успеваемость сильнее, а какие слабее, а также насколько велико совокупное влияние на успеваемость четырех выделенных факторов.

Решение поставленной задачи начнем с упрощения модели рис. 4.1 к модели рис. 4.2.

Рис. 4.2. Упрощенная модель «Успеваемость студента»

Отличие модели рис. 4.2 от модели рис. 4.1 состоит в том, что мы фокусируемся только на одной причине успеваемости студента — уровне предварительной подготовки, а все остальные факторы включили в «Другие факторы». Социологический смысл данной модели представляется вполне естественным: успеваемость студента зависит от уровня его предварительной подготовки. Разумеется, успеваемость определяется не только этим. Имеется еще множество других факторов, влияющих на успеваемость. Смысл построения модели математической зависимости состоит в выяснении того, каким образом на успеваемость влияет именно уровень предварительной подготовки, каково направление и сила этого влияния.

4.1

Общее описание регрессионной модели

Если о направлении воздействия можно сделать, как представляется, вполне обоснованное предположение: «чем выше уровень предварительной подготовки, тем выше успеваемость», то сформулировать предположения о силе такого воздействия довольно сложно. Попытаемся с помощью анализа данных, содержащих сведения об успеваемости студентов и уровне их предварительной подготовки, найти точные ответы на поставленные вопросы.

Формально предложенную модель зависимости можно записать в виде следующей математической зависимости:

$$y=f(x) + u, \qquad (4.1)$$

где y — показатель «Успеваемость студента»; x — показатель «Уровень предварительной подготовки»;/— функция, описывающая силу и форму влияния *хна у; и* — все остальные факторы, влияющие на^. Задачей построения модели (4.1) становится подбор функции/ которая будет наилучшим образом описывать зависимость *хпу.* Рассмотрим решение этой задачи на примере.

В нашем распоряжении есть данные, в которых в качестве показателя «Уровень предварительной подготовки» выступает суммарный балл, полученный студентом на вступительных экзаменах в вуз, в качестве показателя «Успеваемость» — суммарный балл студента за 1-й семестр обучения в вузе (табл. 4.**1**)[1].

Таблица 4.1. Оценки студентов при поступлении в вуз и по итогам 1 -го семестра обучения

№ студента	Суммарный балл на вступительных экзаменах	Суммарный балл по итогам 1-го семестра обучения
1	32	117,4
2	26	106,7
3	27	120,0
4	27	97,3
5	26	108,0
6	25	124,0
7	25	121,4
8	28	106,7
9	29	105,3
10	27	96,0
11	26	94,7
12	26	89,4
13	25	113,4
14	26	113,3
15	?4	93,3
16	25	118,7
17	25	88,0
18	28	100,0
19	14	78,7
20	18	102,7

[1] Были взяты оценки абитуриентов на вступительных экзаменах в 2002 г. на факультет социологии ГУ ВШЭ. Вступительные испытания проводились по четырем дисциплинам: математика, обществознание, иностранный язык, русский язык. Оценки по первым трем дисциплинам выставлялись по 10-балльной системе, по русскому языку— ПО 5-балльной системе.

Коэффициент корреляции Пирсона между двумя анализируемыми показателями составляет 0,43 и значим на уровне $a = 0{,}06$. Следовательно, у нас есть неплохие основания заключить, что модель, приведенная на рис. 4.2, отражает реально существующие закономерности. Представим данные табл. 4.1 в виде диаграммы рассеяния (рис. 4.3).

Рисунок 4.3 показывает, что есть определенная зависимость между x и y — с ростом значений показателя «Уровень предварительной подготовки» наблюдается тенденция возрастания показателя «Успеваемость». Какова форма этой зависимости, или каков вид функции f в выражении (4.1)? Начнем поиск этой функции с самого простого и удобного класса функций — с линейных функций.

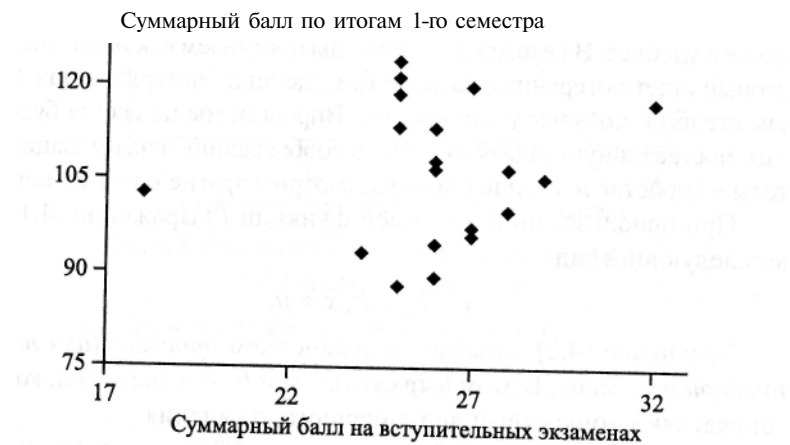


Рис. *4.3.* Оценки студентов при поступлении в вуз и за 1-й семестр обучения

Все оценки за обучение в ГУ ВШЭ выставляются по 10-балльной системе, независимо от формы контроля (как за экзамены, так и за зачеты). При вычислении суммарного балла за семестр оценка по каждому предмету учитывается с определенным весом, который отражает объем часов по данному предмету. Так, если на предмет отводится, скажем, 50 часов, вес его оценки — 1, а если 100 часов, то вес оценки уже 2. Максимально возможная сумма баллов, которые мог набрать студент I курса в 1-м семестре 2002/03 учебного года, —146,7.

Почему именно с линейных? Ведь диаграмма (см. рис. 4.3) показывает нам лишь то, что это должна быть какая-то возрастающая функция, а в этом качестве могут выступать и показательная функция, и логарифм, да и вообще бесконечное число самых разных функций. Причем также видно, что какую бы функцию мы ни взяли, она не будет точно проходить через все точки.

Однако этого и не требуется. Ведь в выражении (4.1) значения y описываются не как/fx), а как сумма/(!x) и *и*. Таким образом, можно сказать, что несовпадения положения точек с графиком некоторой функции/объясняются наличием именно добавки *и*.

Данные соображения, к сожалению, не объясняют, почему мы решили рассматривать именно линейные функции. Объяснение этому лежит совсем в другой плоскости — на самом деле линейные функции проще и удобнее. В некотором смысле мы поступаем как герой анекдота, который ищет потерянные часы не там, где он их потерял, а под фонарным столбом, поскольку там светлее. Впрочем, мы не всегда будем решать поставленную задачу исходя из соображений максимизации простоты и удобства и в конце главы рассмотрим другие виды функций.

При использовании линейной функции/выражение (4.1) примет следующий вид:

$$y = b_g + b_1 x + u. \qquad (4.2)$$

Уравнение (4.2) называется уравнением *простой (или парной) линейной регрессии.* В этом выражении $Ь_0$ и $Ъ_x$ — константы, которы и определяют конкретный вид линейного уравнения.

Представим, как будет выглядеть рис. 4.3, если на нем изобра зить линейную функцию (4.2) (рис. 4.4).

Из каких соображений мы исходили, строя прямую на рис. 4.4. Иными словами, как мы определили параметры $Ь_0$ и b_v которые и дал нам именно такую прямую? Логика вычисления параметров прямо достаточно проста. Прямая должна лежать максимально близко ко все' точкам графика, т.е. сумма расстояний от всех точек до искомой пря мой была бы наименьшей. Подробнее это показано на рис. 4.5.

Оставим для наглядности на графике четыре точки, а остальны сделаем невидимыми. Стрелки E1, E2, E3, E4 — это расстояния д регрессионной прямой соответственно для точек 1, 2, 3, 4. Один и

способов вычисления параметров b_Q и $Ъ_x$ регрессионного уравнения состоит в минимизации суммы (4.3). Иначе говоря, мы стараемся сделать минимальной не сумму расстояний от точек до прямой, а сумму квадратов расстояний:

$$S = E1^2 + E2^2 + E3^2 + E4^2. \qquad (4\text{-}3)$$

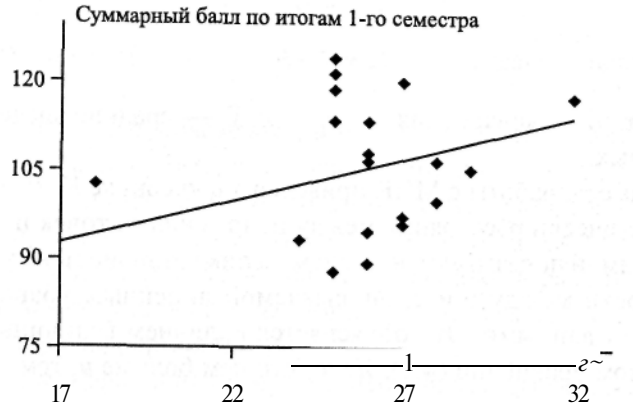


Рис. 4.4. Оценки студентов при поступлении в вуз и за 1-й семестр обучения

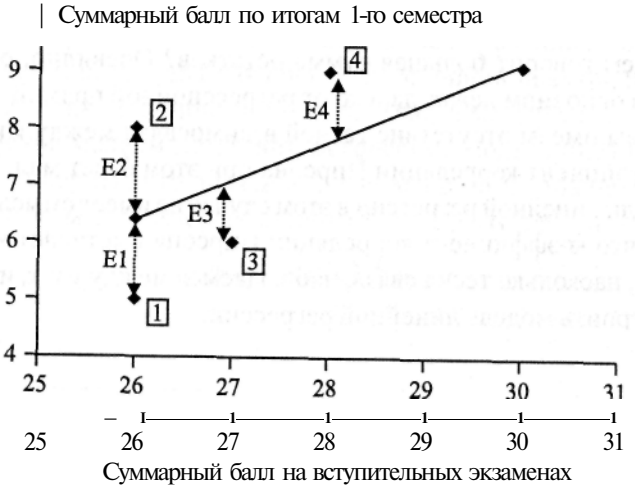


Рис. 4.5. Оценки студентов при поступлении в вуз и за 1-й семестр обучения. Пример с четырьмя наблюдениями

Метод решения задачи вычисления параметров регрессии путем минимизации выражения (4.3) называется *методом наименьших квадратов* (МНК). Оказывается, что S минимальна при следующих значениях b_1 и b_0:

$$b_1 = \frac{cov(x, y)}{D_x}; \quad (4.4)$$

$$b_0 = \bar{y} - b_1\bar{x}, \quad (4.5)$$

где *cov (x, y)* — ковариация x и y; $\bar{x}$ и $\bar{y}$ — средние значения этих переменных.

Примеры работы с МНК приведены в учебнике К. Доугерти[2].

Фактически расстояния между положениями точек и регрессионной прямой показывают, насколько велико отличие между моделью зависимости между y и x, описываемой линейным уравнением, и реальными данными. Это объясняется наличием величины u в регрессионном уравнении (4.2). Ясно, что чем больше и, тем хуже описывает линейная функция реальные данные.

Степень расхождения реальных данных*у-ков* и*у-ков*, вычисленных с помощью найденной функции, *(и),* в регрессионном анализе называются *остатками.* На рис. 4.5 расстояния Е1, Е2, Е3 и Е4 и есть остатки.

О чем говорит большая сумма остатков? Очевидно, о том, что данные в основном лежат далеко от регрессионной прямой. Следовательно, мы имеем отсутствие тесной взаимосвязи между *у их.* Ясно, что коэффициент корреляции Пирсона при этом будет мал. Построение модели линейной регрессии в этом случае не имеет смысла. Можно сказать, что коэффициент корреляции Пирсона выступает индикатором того, насколько тесна связь, наблюдаемая между y и x, и имеет ли смысл строить модель линейной регрессии.

[2] См.: Доугерти К. Введение в эконометрику. М.: ИНФРА-М, 1999. С. 58—60.

Интерпретация коэффициентов регрессии. Используя команду *Regression* пакета SPSS[3], вычислим значения коэффициентов регрессии для данных, представленных в табл. 4.1. Получаем значения: $b_0 = 68{,}4$; $b_1 = 1{,}4$. Итак, модель линейной регрессии будет выглядеть следующим образом:

$$y = 68{,}4 + 1{,}4x, \quad (4.6)$$

где y — успеваемость студента; x — уровень предварительной подготовки.

Коэффициент b_0 показывает, в какой точке регрессионная прямая пересечет ось y. Интерпретировать этот показатель достаточно просто: какую успеваемость по итогам 1-го семестра будут иметь студенты, которые набрали на вступительных экзаменах 0 баллов. Они будут иметь успеваемость 68,4 балла. Очевидно, в рамках данного примера такая ситуация бессмысленна, однако во многих случаях b_0 несет полезную информацию.

Смысл коэффициента 6, интереснее. Он показывает, на сколько баллов возрастает средняя успеваемость студента в 1-м семестре при увеличении на единицу балла на вступительных экзаменах в вуз. Таким образом, мы видим, что увеличение суммарной оценки на вступительных экзаменах на 1 балл дает улучшение успеваемости студента в 1-м семестре на 1,4 балла. На самом деле коэффициент b_x есть не что иное, как тангенс угла наклона регрессионной прямой, и, следовательно, именно он демонстрирует *силу* связи между *у их.*

Качество модели линейной регрессии. Модель (4.2) дает нам основание говорить, что значение y для каждого из анализируемых случаев, т.е. *у.,* мы можем рассматривать как сумму двух компонент:

$$y_i = (b_0 + b_1x_i) + u_i. \quad \textbf{(4.7)}$$

[3] См.: Бююль А., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб.: ООО «ДиаСофтЮП», 2001. С. 271—272.

Для удобства обозначим слагаемое в скобках как *y*.. Тогда выражение (4.7) может быть записано как:

$$y_i = \hat{y}_i + u_i. \quad (4.8)$$

При этом первое слагаемое представляет собой ту часть значения *y* для г'-го случая, которая объясняется линейным влиянием *x*. Что же касается *и.,* то это — результат воздействия всех остальных факторов на *y* для г'-го случая. Другими словами, первое слагаемое — закономерная, объясняемая линейной моделью часть значения *y,* а второе — часть, объясняемая всеми другими, подчас случайными и мало понятными причинами.

Понятно, что регрессионная модель хороша, если большая часть изменений^ объясняется изменением закономерной составляющей *y*. Это соображение подталкивает к определению показателя, который может выступать как характеристика качества регрессионной модели. Традиционно таким показателем принято считать отношение дисперсии *y* к дисперсии^. Обозначают этот показатель как R^2:

$$R^2 = \frac{D(\hat{y})}{D(y)}. \quad (4.9)$$

Показатель R^2 называется *коэффициентом детерминации*. Очевидно, что R^2 всегда положителен и равен единице в ситуации, когда 1 полностью описывает *y,* или когда остатки *и* отсутствуют. Введем в табл. 4.1 колонку *y*, значения которой вычислим по модели (4.6).

Можно показать также, что $R^2 = r^2(y, y)$.

Исходя из дисперсий, приведенных в табл. 4.2, можем рассчитать показатель качества — коэффициент детерминации для модели (4.6).

$$R^2 = \frac{28,6}{157,2} = 0,18.$$

Таким образом, можно констатировать, что регрессионная модель (4.6) объясняет 18% дисперсии>>. Иными словами, успеваемость студентов в 1-м семестре обучения в вузе на 18% объясняется исходным уровнем подготовки студентов.

Таблица 4.2. Оценки студентов при поступлении в вуз и по итогам 1-го семестра обучения

№ студента	**Суммарный балл на вступительных экзаменах (Л:)**	**Суммарный балл по итогам 1-го семестра обучения** *(y)*	**Значения** *y,* **предсказываемые регрессионной моделью (4.4)** *(y)*
1	32	117,4	114,1
2	26	106,7	105,5
3	27	120,0	107,0
4	27	97,3	107,0
5	26	108,0	105,5
6	25	124,0	104,1
7	25	121,4	104,1
8	28	106,7	108,4
9	29	105,3	109,8
10	27	96,0	107,0
11	26	94,7	105,5
12	26	89,4	105,5
13	25	113,4	104,1
14	26	113,3	105,5
15	24	93,3	102,7
16	25	118,7	104,1
17	25	88,0	104,1
18	28	100,0	108,4
19	14	78,7	88,4
20	18	102,7	94,1
Дисперсии	14,1	157,2	28,6

4.2

Особенности использования регрессионных моделей при анализе данных выборочных исследований

В нашем примере поиска зависимости успеваемости студентов от уровня предварительной подготовки мы опирались на данные об оценках 20 студентов и, соответственно, получили результаты, справедливые именно для этих 20 студентов. Поскольку заключительный вывод предыдущего параграфа справедлив только для этих 20 человек, то, строго говоря, ценность этого вывода не велика. Действительно, то, что у некоторых 20 студентов успеваемость на первых этапах обучения в вузе на 18% зависит от уровня предварительной подготовки, является лишь любопытным фактом из жизни этих 20 студентов, и не более того.

Иная ситуация возникает, когда мы говорим, что изучаемые 20 студентов являются *случайной выборкой* из всей совокупности студентов I курса факультета социологии ГУ ВШЭ 2002 г. В этом случае можно утверждать, что результаты, полученные для 20 студентов, с определенной точностью могут быть перенесены и на всю генеральную совокупность, т.е. на всех студентов I курса факультета социологии ГУ ВШЭ 2002 г.

Такое обобщение результатов называют *генерализацией,* а само исследование 20 студентов становится выборочным исследованием[4]. Очевидно, что для прямого утверждения: «Для студентов I курса факультета социологии ГУ ВШЭ 2002 г. успеваемость и уровень предварительной подготовки связаны соотношением (4.6)», у нас нет оснований. Действительно, мы ведь получили этот результат только для

[4] Подробнее об основаниях применения выборочного метода в социологии см Батыгин Г.С. Лекции по методологии социологических исследований. М.: Аспект-Пресс 1995. С. 145—189.

выборки в 20 человек, а для всего I курса, который насчитывает более 100 человек, зависимость может быть иной.

Поскольку сведения об успеваемости и оценках на вступительных экзаменах для всех студентов доступны, не составляет труда повторить вычисления на массиве всего I курса. Однако информация обо всех элементах генеральной совокупности бывает доступна далеко не всегда. Более того, в абсолютном большинстве случаев получение таких сведений либо сопряжено с большими затратами ресурсов (времени, денег), либо вообще невозможно[5]. Именно по этой причине и используют выборочные, а не сплошные исследования.

Мы тоже можем поставить вопрос: как на основании результатов выборочного изучения успеваемости 20 студентов можно делать выводы о характеристиках всей генеральной совокупности, т.е. обо всех студентах I курса факультета социологии 2002 г.? Как могут измениться результаты, верные для 20 человек, когда мы будем переносить их на весь поток I курса?

Если задуматься о направлении этих возможных изменений, то можно предположить, что скорее всего регрессионная прямая, описывающая зависимость успеваемости от уровня предварительной подготовки, будет не той, которую мы получили для 20 студентов (уравнение (4.6)), а какой-то другой. По всей видимости, изменится и показатель качества R^2, описывающий степень приближения прямой к реальным точкам.

Что означает изменение регрессионной прямой? Это означает изменение коэффициентов b_0 и b_y. От чего может зависеть степень такого изменения? Прежде всего, от величины корреляции между x и y. Действительно, если в нашей выборке из 20 студентов мы получили, что корреляция высока, и, следовательно, реальные точки лежат достаточно плотно вокруг регрессионной прямой, то естественно предположить, что и во всей генеральной совокупности картина аналогичная. И при этом сама «истинная» прямая будет близка к той, которая получена по данным выборки.

[5] Например, если мы хотим изучить особенности поведения комаров, то проведение сплошного исследования этих насекомых едва ли возможно даже в ситуации неограниченных ресурсов.

Если в выборке есть немало точек, достаточно далеко отстоящих от прямой, вполне вероятно, что при переходе от выборки к генеральной совокупности число таких точек увеличится. Следовательно, велика вероятность того, что регрессионная прямая существенно изменит свое положение. Таким образом, принципиально важным фактором, влияющим на возможное изменение параметров b_0 и 6_1 при переходе от выборки к генеральной совокупности, является разброс значений *и,* т.е. дисперсия остатков. При этом понятно, что чем больше эта дисперсия, тем сильнее могут измениться b_Q и $Ъ_x$ при генерализации.

Другим фактором, влияющим на устойчивость параметров регрессии, является дисперсия *x.* Действительно, из выражения (4.2) следует, что изменения *y* в определенной степени обусловлены изменениями *x.* Следовательно, чем меньше возможные изменения *x, тещ* вероятнее, что изменения *y* будут происходить из-за влияния *и*

Эти рассуждения вполне логично сочетаются с известными фор мулами для определения стандартных ошибок коэффициентов $Ь_0$ и $Ъ$

$$\text{с.о.}b_0 = \sqrt{\frac{D_u}{n}\left(1+\frac{\bar{x}^2}{D_x}\right)}; \qquad (4.10)$$

$$\text{с.о.}b_1 = \sqrt{\frac{D_u}{nD_x}}, \qquad (4.11)$$

где с.о.6_0— стандартная ошибка коэффициента o_0; с.о.о, — стандартная ошибка коэффициента b_x; D — дисперсия остатка; D_x — дисперсия *x;* $\bar{x}$ — среднее значение *x; n* — объем выборки.

В качестве примера проведем вычисление стандартных ошибок для регрессионной модели (4.6), данные по которой представлены в табл. 4.2. Результаты вычислений сведены в табл. 4.3.

В результате получаем: с.о.6_0= 18,3; со.о, = 0,71.

[6] Подробнее с выводом формул для оценки точности коэффициентов регрессии см.: Доугерти К. Введение в эконометрику. С. 83—85.

Таблица 4.3. Оценки студентов при поступлении в вуз и по итогам 1 -го семестра обучения

№ студента	Суммарный балл на вступительных экзаменах *(x)*	Суммарный балл по итогам 1-го семестра обучения *(y)*	Значения y, предсказываемые регрессионной моделью **(4.4)** (у)	Значения остатков (и)
1	**32**	117,4	114,1	3,3
2	26	106,7	105,5	1,2
3	27	120,0	107,0	13,0
4	27	97,3	107,0	-9,7
5	26	108,0	105,5	2,5
6	25	124,0	104,1	19,9
7	25	121,4	104,1	17,3
8	28	106,7	108,4	-1,7
9	29	105,3	109,8	-4,5
10	27	96,0	107,0	-11,0
11	26	94,7	105,5	-10,8
12	26	89,4	105,5	-16,1
13	25	113,4	104,1	9,3
14	26	113,3	105,5	7,8
15	24	93,3	102,7	-9,4
16	25	118,7	104,1	14,6
17	25	88,0	104,1	-16,1
18	28	100,0	108,4	-8,4
19	14	78,7	88,4	-9,7
20	18	102,7	94,1	8,6
Средние	25,4	104,75	104,75	0,0
Квадрат среднего	645,2			
Дисперсии	14,1	157,2	28,6	128,6

Что дают нам вычисленные значения стандартных ошибок для $b_Q nb_1'$? Они дают оценку точности для этих коэффициентов при переносе результатов модели (4.6) с выборки на генеральную совокуп-

ность[7]. Говорить о том, что зависимость между успеваемостью и уровнем предварительной подготовки студентов описывается уравнением (4.6), мы не имеем права, не указав, с каким уровнем точности можно переносить результаты выборки на генеральную совокупность.

По этой причине, анализируя зависимости типа (4.6), следует отдавать себе отчет в том, что наличие характеристик точности (стандартных ошибок) в этом уравнении принципиально важно. Символически выразим это в виде (4.12).

$$y = 68{,}4 + 1{,}4x. \quad (18{,}3)\ (0{,}71) \qquad (4.12)$$

Вычисленные стандартные ошибки коэффициентов b_0 и $Ъ_x$ дают возможность с определенной, задаваемой нами вероятностью определить *доверительные интервалы* для характеристик регрессионной прямой в генеральной совокупности. Из начального курса математик ческой статистики известно, что величина доверительного интервала параметра A определяется по формуле

$$A^{выб} - \Delta \le A^{ген} \le A^{выб} + \Delta, \qquad (4.13)$$

где Д — А = z x *с.о.А; г* — квантиль нормального распределения; *с.о.А* — стандартная ошибка параметра Л.

Из таблиц нормального распределения следует, что с вероятностью 0,95 величина z равна 1,96[8]. Выражения (4.12) и (4.13) дают нам основания определить, что с вероятность 0,95 значения коэффициентов o_0 и $Ь_1$ для модели (4.2) в генеральной совокупности будут иметь вид

$$b_0 = 68{,}4 \pm 1{,}96 \times 18{,}3;$$
$$b_1 = 1{,}4 \pm 1{,}9 \times 0{,}71.$$

[7] Подробнее об оценке характеристик генеральной совокупности по данным выборки см.: Гмурман Е.В. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998. С. 219—220.

[8] Квантили нормального распределения см., например: Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. С. 965. ТаблицаШ.3.

Таким образом, регрессионное уравнение для генеральной совокупности с вероятностью 95% будет иметь коэффициент $Ь_0$, лежащий в интервале (32,5; 104,3), а коэффициент b_v в интервале (0,01; 2,79).

Вычисленные доверительные интервалы для коэффициентов регрессионной модели достаточно велики. Оказывается, что с вероятностью 95% модель зависимости между уровнем исходной подготовки и успеваемостью студента может иметь разный вид. На рис. 4.6 показана построенная по нашим данным линия регрессии (сплошная линия) и, пунктиром, две из бесконечного числа прямых, которые возможны в границах уравнения (4.12) не для выборки, а для генеральной совокупности.

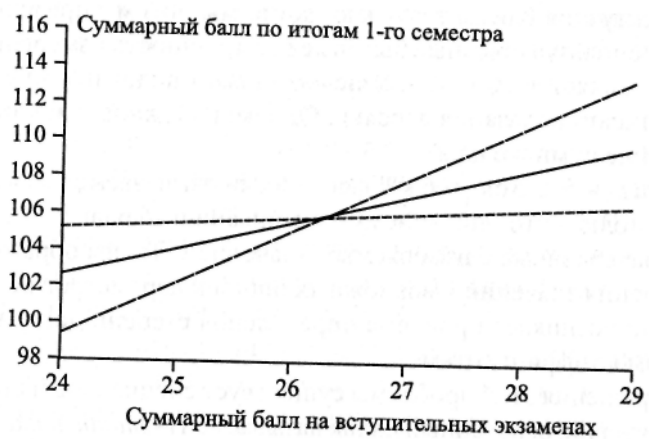


Рис. 4.6. Возможные формы зависимости оценок студентов при поступлении в вуз и за 1-й семестр обучения

Как можно уменьшить такую неопределенность? В формулах для стандартной ошибки регрессионных коэффициентов (4.10), (4.11) есть ᵒЩе один параметр, который мы пока не обсуждали, — *n*.

Действительно, из соображений здравого смысла следует, что Увеличение объема выборки должно приводить к получению более точных оценок параметров регрессии. Формулы (4.10) и (4.11) показывает, что значения стандартных ошибок обратно пропорциональны кор-

ню квадратному из объема выборки. Этот факт достаточно неприятен, поскольку, например, для двукратного увеличения точности оценок параметров регрессии мы должны увеличить объем выборки в 4 раза.

Проверка статистических гипотез о параметрах регрессии. Итак, мы научились вычислять значения коэффициентов для линейной регрессионной модели, умеем оценивать возможную погрешность, которая возникает при генерализации. Однако может возникнуть одна существенная проблема. Не исключено, что неточность определения одного из параметров регрессии (или даже обоих) больше, либо, по крайней мере, близка к значению самих параметров. Например, оценивая параметр b_i для какого-то регрессионного уравнения, мы можем вычислить, что его значение равно пяти, а его стандартная ошибка равна шести.

Эта ситуация близка к случаю, когда мы хотим определить вес муравья, используя обычные бытовые весы, точность измерения у которых составляет ± 10 г. Мы, конечно, можем положить муравья на весы и записать показания стрелки. Однако надежность такого измерения крайне сомнительна.

Когда $b_i = 5$, а с.о.о, = 6 мы фактически оказываемся в ситуации, близкой к только что описанному взвешиванию муравья, — возможная ошибка сравнима с измеряемым значением. Таким образом, после вычисления значений и возможных ошибок параметров регрессии перед нами возникает проблема определения степени доверия к вычисленным коэффициентам.

Для решения этой проблемы существует специальный статистический критерий, основанный на так называемой *t-статистике*. Смысл /-статистики достаточно прозрачен. Она показывает, во сколько раз вычисленное значение параметра больше его стандартной ошибки:

$$t = \frac{b}{\text{с.о.}b}. \qquad (4.14)$$

Понятно, если значение t велико, скорее всего, вычисленному значению Ъ можно доверять. Но давайте разберемся: что такое «t велико» и что такое «можно доверять». Говоря формально, речь пойдет о проверке статистической гипотезы $H_{\text{и}} \setminus Ь_x^m 0$.

Поскольку вычисленная по формуле (4.14) /-статистика является случайной величиной, то для определения того, какое значение вели-

$_{к\,0;}$ а какое нет, необходимо знать закон распределения этой случайной величины. Известно, что /-статистика имеет *t-распределение* (при условии, что в генеральной совокупности 6, = 0), критические точки которого приведены в статистических таблицах и в учебниках *tio* статистике[9]. Используя эти таблицы, мы можем определить, с ка^ой вероятностью можно доверять конкретному значению /-статистики.

Например, рассчитаем значение /-статистики для параметра Ъ из уравнения (4.12).

$$t = \frac{1,4}{0,71} = 2.$$

В таблице /-распределения находим, что для 18 степеней свободы[10] при равенстве нулю генерального значения $о_р$ с вероятностью 0 90 /-статистика должна быть меньше 1,73, с вероятностью 0,95 Меньше 2,1. В нашем случае /-статистика больше этой критической величины и, следовательно, H_0 может быть отвергнута на уровне значимости а = 0,1, но не на уровне 0,95. Это означает, что вряд ли мы имеем основания утверждать, что вычисленному значению 6, можно доверять.

Итак, фактически /-статистика в форме записи (4.14) служит инструментом проверки статистической гипотезы о равенстве нулю параметра Ъ. Почему мы проверяем статистическую гипотезу о равенстве Ъ именно нулю? Тому есть две причины.

Во-первых, в отношении $Ъ_x$ проверка на равенство именно нулю существенна. Действительно, ведь если о, равно нулю, это значит что регрессионная прямая идет параллельно оси абсцисс и, следовательно, y не зависит от x. Таким образом, если мы не можем с высокой вероятностью отвергнуть статистическую гипотезу о равенстве о, нулю, значит, мы не можем принять гипотезу о связи y и x.

Такой подход, однако, не объясняет, почему должны проверять гипотезу о равенстве нулю коэффициента b . Ведь для этого коэффициен-

[9] См., например: Гмурман В.Е. Теория вероятностей и математическая статистика. С. 464. Приложение 3; Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. С. 972. Таблица Ш.6.

[10] Число степеней свободы в ситуации простой регрессии, когда мы оцениваем Два параметра, определяется как n - 2, где n — число наблюдений.

та нуль является вполне приемлемым значением, ничем не отличающимся от любого другого. Здесь вступает в силу второе, сугубо утилитарное соображение. Дело в том, что все компьютерные пакеты программ статистического анализа при вычислении коэффициентов регрессии проверяют статистическую гипотезу об их равенстве именно нулю.

В заключение приведем пример вычисления коэффициентов регрессии командой *Regression* пакета программ SPSS (табл. 4.4).

Таблица 4.4. Пример вычисления коэффициентов регрессии командой Regression пакета программ SPSS

Coefficients

	Unstandardized Coefficients		**Standardized Coefficients**	***t***	Sig.
	B	**Std. Error**	**Beta**		
(Constant)	68,413	18,332		3,732	0,002
x	1,428	0,713	0,427	2,002	0,069

В качестве массива данных для расчетов использованы данны табл. 4.1. При этом в качестве зависимой переменной (y) выступае суммарный балл студентов по итогам 1-го семестра, а в качестве неза висимой переменной *(x)* — суммарный балл на вступительных экза менах. Разберем те характеристики, которые вычисляет SPSS и кото рые приводятся в результирующей таблице.

В первой колонке таблицы указано, какие именно коэффициенть располагаются в соответствующих строках. Коэффициент, который нас обозначался как b_0, в таблице SPSS называется (Constant), в следую щей строке указывается имя переменной, для которой вычисляется рег рессионный коэффициент в данной строке (в нашем случае — *x)*.

Следующие 2 колонки, объединенные заголовком «Unstandardize Coefficients», содержат значения регрессионных коэффициентов (ко лонка *B)* и значения стандартных ошибок для них (колонка *Std. Error)*

Смысл колонки *Standardized Coefficients* будет рассмотрен подробнее при обсуждении модели множественной регрессии. Колонка / содержит значения /-статистики для каждого из коэффициентов. И наконец, колонка *Sig.* — уровень значимости /-статистики. Наличие данной колонки избавляет от поиска в статистических таблицах уровня значимости для полученных значений /-статистики. Данная колонка содержит вероятность, с которой вычисленный для произвольной выборки регрессионный коэффициент будет больше или равен найденному нами значению (при условии равенства нулю коэффициента в генеральной совокупности). Это тот уровень значимости, на котором может быть отвергнута H_0. В нашем примере коэффициент b_0 значим на уровне 0,002, коэффициент 6, незначим (0,069 > 0,05).

Представленные в табл. 4.1 данные, для которых вычислили регрессионную модель зависимости успеваемости от уровня предварительной подготовки, представляют собой случайную выборку из 20 человек из общей совокупности студентов I курса факультета социологии. Поскольку в нашем распоряжении есть данные о всей генеральной совокупности, можно проверить, насколько правильно мы оценили тенденции в этой совокупности по данным выборки.

Оказалось, что уравнение для генеральной совокупности[11] следующее:

$$y = 79{,}3 + 0{,}95x. \qquad (4.15)$$

Уравнение (4.15) существенно отличается от той зависимости, которую мы получили для выборки (4.6), однако значения коэффициентов регрессии лежат в вычисленных нами доверительных интервалах. Обратим внимание, что значение коэффициента детерминации у модели (4.15) для всей выборки оказалось равным $R^2 = 0{,}05$, значимо на уровне и = 0,05, но существенно ниже значения, полученного на выборке.

" Отметим, что вычисления проводились на совокупности 84 студентов I курса факультета социологии ГУ ВШЭ 2002/03 учебного года. Мы не могли включить в совокупность студентов, которые при поступлении в университет имели медаль за окончание школы, получили отличную оценку на профилирующем экзамене и, соответственно, "е сдавали остальные экзамены. Таким образом, у них не было суммарного балла на вступительных экзаменах.

4.3

Ограничения модели регрессии

Изложенные методы вычисления и оценки качества модели регрес сии в целом, равно как и параметров регрессии в частности, справед ливы не всегда. Вполне возможно, что поведение исходных данны не позволит использовать стандартный регрессионный подход. Прин ципиально важно, что те ограничения, которые предъявляет к дан ным статистическая модель регрессионного анализа, одновременн оказываются требованиями и к содержательным социологически моделям, которые строятся на основе моделей регрессионных.

Нормальность распределения остатков. Построение довери тельных интервалов при оценке коэффициентов регрессии происхо дит в предположении, что возможные значения этих коэффициенте подчиняются закону *нормального распределения*. Выражение (4.13 базируется на этом допущении.

В свою очередь, данное предположение напрямую основано н предположении о нормальном распределении остатков *и*. А почем" собственно, такое предположение должно выполняться, бывают л случаи его невыполнения, и что это значит?

На рис. 4.7 представлена функция плотности нормального распре деления. Глядя на этот рисунок, кажется, что требование к нормальности распределения остатков является вполне логичным. С определенным упрощением можно считать, что это требование означает: маленьких остатков должно быть много, а больших остатков — мало. Другими словами, основная масса точек должна лежать близко к регрессионной прямой, и чем дальше от прямой, тем точек должно быть меньше, и лишь небольшое число точек может лежать далеко от прямой.

Из этого рассуждения не следует, однако, что это должно быть именно нормальное распределение. Здесь вступает в силу другое со» ображение, затрагивающее сущность остатков. Из нашей модели (см. рис. 4.2) следует, что остатки — это результат действия большо го числа разнообразных факторов («Другие факторы»), которые воз-

действуют на показатель «Успеваемость», кроме показателя «Уровень предварительной подготовки». Можно предположить, что ни один из большого количества «Других факторов» не влияет на успеваемость в большей степени, чем другие.

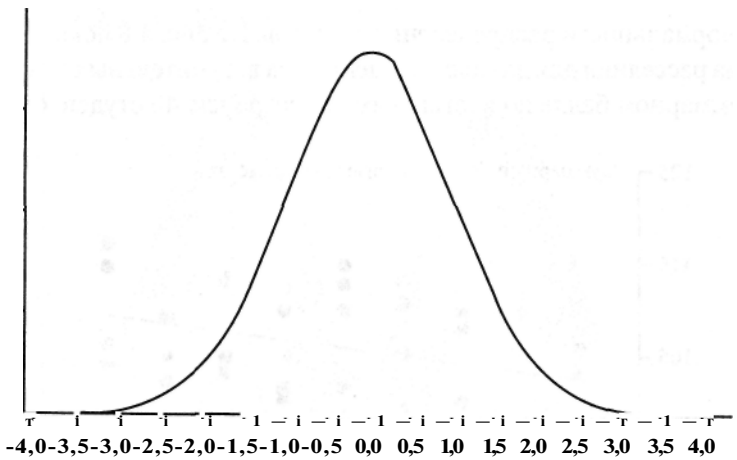


*Рис. 4.*7. Функция плотности нормально распределенной случайной величины (со средним = 0 и дисперсией = 1)

В этой ситуации вступает в силу одна из центральных теорем теории вероятностей — *центральная предельная теорема*. Она утверждает, что «если случайная величина является общим результатом взаимодействия большого числа других случайных величин, ни одна из которых не является доминирующей, то он будет иметь приблизительно нормальное распределение»[12]. Исходя из этой теоремы предположение о нормальности распределения остатков выглядит вполне естественным.

Что произойдет, если условие нормальности распределения остатков будет нарушено? Прежде всего, это значит, что мы не сможем пользоваться формулами определения доверительных интервалов для

[г] **См.: Доугерти К. Введение в эконометрику. С. 82.**

коэффициентов регрессии. А раз так, то у нас нет возможности переносить результаты, полученные на выборке, на характеристики генеральной совокупности. И, следовательно, вычисленная по выборке прямая регрессии будет представлять ценность лишь для этой выборки.

Рассмотрим гипотетический пример, в котором нарушается правило нормальности распределения остатков. На рис. 4.8 показана диаграмма рассеяния для данных об оценках на вступительных экзаменах и о суммарном балле по итогам 1-го семестра для 40 студентов.

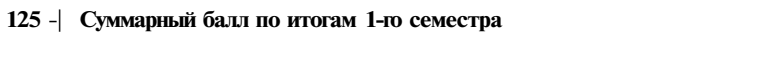


Рис. 4.8. **Гипотетический пример распределения оценок при поступлении в вуз и оценок за 1-й семестр обучения**

Коэффициент линейной корреляции Пирсона для этих данных составляет 0,33 и значим на уровне $a = 0{,}03$. Следовательно, мы можем утверждать, что модель линейной зависимости между переменными имеет место. Параметры линейной регрессии даны в (4.16) (в скобках — значения стандартных ошибок):

$$\begin{aligned} R^2 &= 0{,}11; \\ b_0 &= 83{,}3\ (10{,}6); \\ b_1 &= 0{,}92\ (0{,}42). \end{aligned} \qquad (4.16)$$

Таким образом, представляется, что мы вполне можем анализировать регрессионную модель. Проверим, однако, выполняется ли для данных рис. 4.8 требование нормальности распределения остатков. На рис. 4.9 изображена гистограмма распределения остатков.

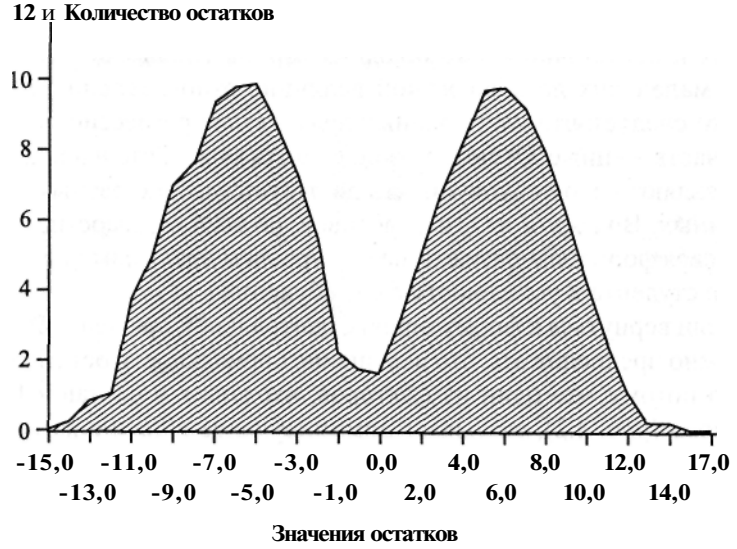


Рис. 4.9. **Гистограмма распределения остатков для регрессионной модели рис. 4.8**

Рисунок 4.9 показывает, что распределение остатков явно отличается от нормального. Многие авторы указывают, что когда идет контроль на нормальность распределения остатков регрессии, нет необходимости требовать жесткого выполнения этого требования[13]. Однако гистограмма (см. рис. 4.9) слишком не похожа на нормальную кривую. Она больше напоминает гистограмму случайной величины, которая является суммой двух нормально распределенных случайных величин с разными средними. Какие выводы можно сделать из такого

[13] **См., например: Тюрин Ю.Н., Макаров А. А.** Статистический анализ данных на компьютере. С. **255.**

распределения остатков? Первый вывод — пользоваться значениями регрессионных коэффициентов и стандартных ошибок (4.16) для определения с фиксированной вероятностью доверительных интервалов для регрессионных коэффициентов, базируясь на формуле (4.13), нельзя.

Второй вывод более содержателен. Гистограмма на рис. 4.9 показывает, что в нашей модели достаточно много больших положительных и достаточно много больших отрицательных остатков. Остатков маленьких по абсолютной величине относительно немного. Из этого следует, что часть данных лежит выше регрессионной прямой, а часть — ниже. Отсюда можно сделать вывод, что наши данные представляют собой совокупность двух существенно разных массивов данных. В каждом из этих массивов, по всей видимости, наблюдается своя форма зависимости между уровнем предварительной подготовки студента и успешностью его обучения в вузе.

Если вернуться к формулировке центральной предельной теоремы, можно предположить, что нарушение нормальности остатков произошло потому, что один из факторов, входящих в состав «Других факторов» (рис. 4.2), оказывает доминирующее влияние на величины остатков и что, следовательно, нормальное распределение может быть нарушено.

Выделим из данных (рис. 4.8) точки, которые лежат выше регрессионной прямой (массив 1), и точки, которые лежат ниже регрессионной прямой (массив 2), и построим регрессии для каждого из этих массивов (рис. 4.10).

Две построенные регрессионные модели имеют показатели качества гораздо более высокие, чем одна модель, общая для всех данных. Если общая модель имела значение $R^2 = 0{,}11$, модель для массива 1 имеет $R^2 = 0{,}61$, а для массива 2 — $R^2 = 0{,}60$. Значительно отличаются и параметры моделей: для массива 1 $Ь_0 = 83{,}1$ (5,5); $Ъ_x = 1{,}14$ (0,22). Для массива 2 $Ь_0 = 65{,}9$ (6,8); 6, = 1,38 (0,26).

Таким образом, контроль на нормальность распределения остатков позволил получить важный результат. Наши данные содержат две разные совокупности респондентов и в каждой из этих совокупностей наблюдаются свои закономерные взаимосвязи между уровнем исходной подготовки и успеваемостью. К сожалению, метод регрессионного анализа не может сказать, что это за две совокупности. Может быть, это юноши и девушки, может быть — студенты из Москвы и из других городов и т.д. Наша задача — это поиск признака, который делит всю совокупность опрошенных на две группы. Важно, что с помощью контроля формальных ограничений метода регрессионного анализа мы вышли на интересный социологический результат.

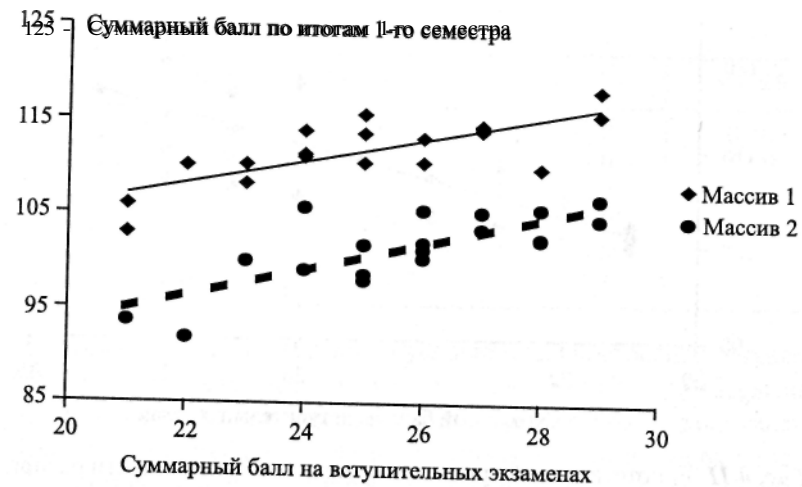


Рис. 4.10. Разбиение данных рис. 4.8 на 2 массива данных и построение регрессионной модели для каждого из массивов

Равная дисперсия распределения остатков (гомоскедастичность). Это ограничение метода достаточно легко понять. На рис. 4.11 показан гипотетический пример распределения данных, который демонстрирует, что с увеличением значения *x* возрастает разброс (дисперсия) точек вокруг регрессионной прямой.

К чему приводит такая картина данных с точки зрения оценок регрессионных коэффициентов? В формулах (4.10) и (4.11) для оценки стандартных ошибок коэффициентов $Ь_0$ и 6, присутствует величина D_u — дисперсия остатков. Для данных, представленных на рис. 4.11, дисперсия остатков составляет 21,7. Однако, если разбить весь массив данных на студентов, получивших на вступительных экзаменах невысокий балл ($x < 25$), и студентов, получивших высокий балл

($x > 25$), окажется, что дисперсия остатков в этих двух массивах существенно разная. Для тех, у кого $x < 25$, дисперсия остатков равна 7,5, а > для тех, у кого $x > 25$, она равна 33,8.

Рис. 4.11. Гипотетический пример с нарушением однородности распределения данных вокруг регрессионной прямой

В табл. 4.5 приведены значения параметров регрессии, рассчитанные для данных в целом и для двух подмножеств данных.

Данные табл. 4.5 показывают, что значения самих регрессионных коэффициентов во всех трех моделях одинаковы. Однако стандартные ошибки регрессионных коэффициентов для тех данных, у которых $x > 25$, гораздо больше, чем те, которые мы получаем, основываясь на данных массива в целом[14]. Следовательно, базируясь на

[14] Обратите внимание, что и для части массива с $x < 25$ стандартные ошибки коэффициентов также больше, чем для массива в целом, хотя и не столь существенно. Это может показаться странным, учитывая тот факт, что дисперсия остатков для этой части массива значительно меньше, чем у массива в целом. Однако, как указывалось при обсуждении формул (4.10) и (4.11), стандартные ошибки зависят не только от дисперсии остатков D_u, но и от дисперсии x и объема выборки n, причем обе величины стоят в знаменателях формул для определения стандартных ошибок. Поскольку дисперсия x для массива $x < 25$ уменьшилась (ведь x меняется в этом подмассиве от 21 До 25, а не от 21 до 29, как во всем массиве), равно как и уменьшилось я, то, несмотря на уменьшение $Z>_u$, значения стандартных ошибок все равно возросли.

общей для всего массива данных линии регрессии, мы рискуем дать ошибочную модель, по крайней мере, для части массива.

Таблица 4.5. Характеристики регрессионных моделей для данных гипотетического примера рис. 4.11 и для двух подмножеств данных

Параметры регрессии	Все данные	$x \le 25$	$x > 25$
R^2	0,56	0,52	0,10
b_0	59,0	59,0	59,0
с.о.b_0	3,97	5,79	18,13
b_1	2,00	2,00	2,00
с.о.b_1	0,16	0,25	0,66
Дисперсия остатков	21,7	7,5	33,8

Таким образом, обязательным условием для построения регрессионной модели является требование одинакового разброса наблюдений вокруг линии регрессии для всех значений x. Это требование называется требованием *гомоскедастичности,* что означает одинаковый разброс.

С социологической точки зрения нарушение гомоскедастичности, т.е. *гетероскедастичность,* фактически означает, что для разных значений x мы должны строить разные регрессионные модели. Действительно, пример (см. рис. 4.11) показывает, что характер зависимости между уровнем предварительной подготовки студента и его успехами в начале обучения в вузе для студентов, набравших на вступительных экзаменах не более 25 баллов, существенно отличается от аналогичной зависимости для студентов, набравших более 25 баллов. В первой группе студентов зависимость между оценками на вступительных экзаменах и оценками в вузе гораздо более тесная, чем для студентов второй группы. Даже простой подсчет коэффициента корреляции Пирсона для этих двух показателей показывает, что в первой группе $r = 0{,}72$, а во второй — $r = 0{,}34$.

Основным выводом, который можно сделать при обнаружении гетероскедастичности, является необходимость разделения массива на несколько относительно гомоскедастичных подмассивов и построение для каждого из них отдельной модели регрессии. Представляется, что при таком подходе и с содержательной точки зрения результаты будут гораздо адекватнее.

Проверка ограничений регрессионной модели. Как уже отмечалось, основной метод контроля нормальности распределения остатков и гомоскедастичности — это анализ остатков. Большинство статистических пакетов анализа данных предоставляют для этого удобные средства. В рамках команды *Regression* пакета программ SPSS последовательность действий будет следующей.

1. В меню команды *Linear Regression,* после задания зависимой и независимой переменных, необходимо выбрать меню, вызываемое клавишей Save (рис. 4.12).

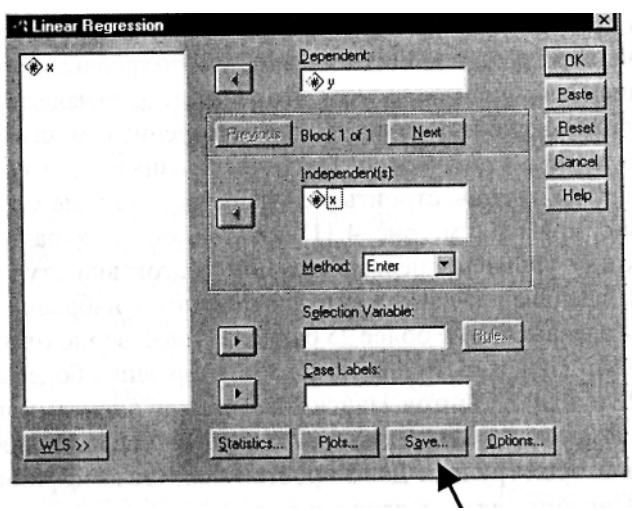


***Рис.** 4.12.* Меню команды Linear Regression пакета программ SPSS

2. В меню Save в разделе «Residuals» (остатки) необходимо поставить галочку против позиции *Unstandardized* (не стандартизованные) (рис. 4.13). Это приведет к созданию в матрице данных SPSS новой переменной со служебным именем res_1. В качестве значений данной переменной будут находиться остатки, вычисленные командой *Regression* для линейной регрессионной модели. На рис. 4.14 приводится фрагмент матрицы данных SPSS для примера рис. 4.11с вновь созданной переменной res_1.

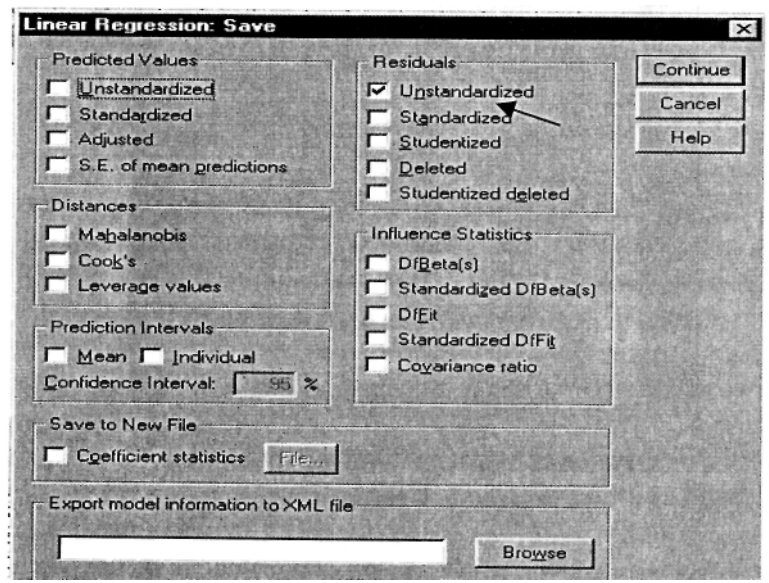
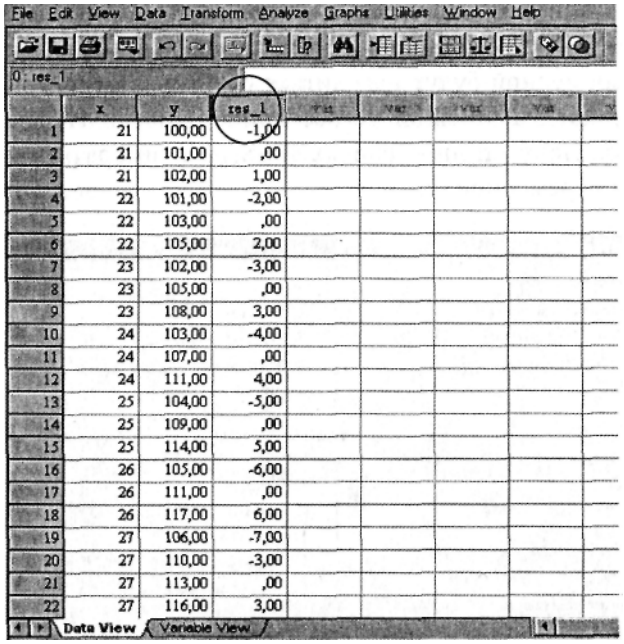


Рис. 4.13. Меню Save команды Regression

3. Полученные таким образом значения переменной res1 можно, с помощью команд меню *Graphs — Histogram,* проверить на нормальность распределения остатков. Методом построения двумерного графика *(Graphs — Line)* можно оценить гомоскедастичность. В последнем случае в качестве переменной по оси *x* следует использовать независимую переменную из регрессионной модели, а в качестве переменной по оси *y* — переменную со значениями остатков.



	x	y	res_1
1	21	100,00	-1,00
2	21	101,00	,00
3	21	102,00	1,00
4	22	101,00	-2,00
5	22	103,00	,00
6	22	105,00	2,00
7	23	102,00	-3,00
8	23	105,00	,00
9	23	108,00	3,00
10	24	103,00	-4,00
11	24	107,00	,00
12	24	111,00	4,00
13	25	104,00	-5,00
14	25	109,00	,00
15	25	114,00	5,00
16	26	105,00	-6,00
17	26	111,00	,00
18	26	117,00	6,00
19	27	106,00	-7,00
20	27	110,00	-3,00
21	27	113,00	,00
22	27	116,00	3,00

Рис. 4.14. Фрагмент матрицы данных SPSS с добавленной переменной res1

4.4

Множественный регрессионный анализ

В начале главы, на рис. 4.1 была представлена модель зависимости успеваемости от четырех различных характеристик: от уровня подготовки студента; активности посещения занятий; активности самостоятельной работы; индивидуальных способностей. В дальнейшем мы упростили эту модель, сосредоточив свое внимание на анализе воздействия только одного фактора — уровня предварительной подготовки студента, а остальные показатели, равно как другие, не зафиксированные на рис. 4.1, мы объединили в группу «Другие факторы» и рассматривали их скорее как мешающие построить упрощенную модель успеваемости (см. рис. 4.2).

Благодаря линейной регрессионной модели мы выяснили, что уровень предварительной подготовки студентов на 18% определяет их успеваемость на первых этапах обучения в вузе, построили модель линейной регрессии, которая описывает указанную зависимость (4.6). Попытаемся теперь вернуться к рис. 4.1, снова упростив эту модель, но сделав ее все-таки сложнее, чем модель на рис. 4.2 (рис. 4.15).

Рис. 4.15. Упрощенная модель «Успеваемость студента»

К сожалению, в нашем распоряжении нет данных, в которых систематически фиксировалась бы степень активности самостоятельной работы студентов. Ограничимся активностью посещения ими обязательных занятий. Будем рассматривать пример, данные которого приведены в табл. 4.6.

Мы могли бы повторить весь путь построения модели простой линейной регрессии, изучив зависимость успеваемости от активности посещения занятий. Однако модель рис. 4.15 подразумевает исследование влияния на успеваемость одновременно двух показателей: активности посещения занятий и уровня предварительной подготовки. Для построения математической модели одновременного влияния нескольких факторов (независимых переменных, предикторов) на зависимую переменную используют усложненный вариант простой линейной регрессии — модель *множественной линейной регрессии*.

Общий вид модели множественной линейной регрессии — это естественное развитие уравнения (4.2) для простой линейной регрессии:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n + u. \qquad (4.17)$$

Таблица 4.6. Оценки студентов при поступлении в вуз и по итогам 1 -го семестра обучения

№ студента	Суммарный балл на вступительных экзаменах (pe,)	Суммарный балл по итогам 1-го семестра обучения *(y)*	Процент занятий, пропущенных студентом *(x,)*
1	32	117,4	1
2	26	106,7	3
3	27	120,0	1
4	27	97,3	12
5	26	108,0	15
6	25	124,0	3
7	25	121,4	10
8	28	106,7	12
9	29	105,3	18
10	27	96,0	10
11	26	94,7	12
12	26	89,4	20
13	25	113,4	5
14	26	113,3	7
15	24	93,3	10
16	25	118,7	12
17	25	88,0	15
18	28	100,0	11
19	14	78,7	15
20	18	102,7	5

Так же, как и в модели простого регрессионного анализа, принимая зависимость *y* от нескольких *x* в форме (4.17), мы делаем очень сильное допущение о линейной форме этой зависимости. Как правило, для такого допущения у нас нет сколько-нибудь серьезных социологических оснований. Использование модели именно линейного регрессионного анализа основано, прежде всего, на хорошей разработанности этого метода. Для обоснования применимости данной модели к конкретным социологическим данным необходимо провести

отдельное исследование, о чем мы будем говорить, обсуждая нелинейные регрессионные модели.

Приступая к построению множественной регрессионной модели, прежде всего необходимо ответить на вопрос: существует ли вообще хоть какая-то зависимость между *y* и предикторами? Быть может, никакой зависимости нет и наши усилия по построению модели заведомо обречены на неудачу?

Как и в ситуации простой регрессионной модели, индикатором наличия зависимости выступает коэффициент корреляции Пирсона. При выборе независимых переменных для модели (4.17) целесообразно вычислить корреляции между *y* и предикторами.

Коэффициенты корреляции для данных табл. 4.6 составляют: $r_{xl} = 0{,}43$; $r_{л} = -0{,}62$, они высоко значимы и, следовательно, построение модели множественной регрессии для этих данных имеет смысл.

Точно так же, как и в модели простой регрессии, для вычисления значений регрессионных коэффициентов 6_0, b_v b_2,..., b_n в множественной регрессии используется метод наименьших квадратов. И так же, как в ситуации простой регрессии, важнейшей задачей является оценка точности регрессионных коэффициентов. Формула для оценки стандартной ошибки коэффициента регрессии $Ъ_x$ для случая двух независимых переменных приведена ниже (4.18). Формула для оценки стандартной ошибки $Ъ_г$ будет такой же, только индекс x_x заменен на $x_г$ Эта формула отличается от формулы стандартной ошибки для простой линейной регрессии (4.11) наличием второго сомножителя.

$$\text{с.о.}b_1 = \sqrt{\frac{D_u}{nD_{x1}}}\sqrt{\frac{1}{1-r^2_{x1,\ x2}}}, \qquad (4.18)$$

где с.о.Oj — стандартная ошибка коэффициента $Ъ_A$; D — дисперсия остатка; D_{xl} — дисперсия x_x; n — объем выборки; $r_{Лх2}$ — квадрат коэффициента корреляции Пирсона для переменных x_f и *•'.

Таким образом, при вычислении стандартной ошибки для регрессионных коэффициентов, наряду с дисперсией остатков и дисперсией независимой переменной, у нас появляется еще один источник ошибки — корреляция между независимыми переменными. При этом

из формулы (4.18) следует, что чем больше значение этого коэффициента (чем теснее связаны независимые переменные между собой), тем больше будет величина стандартной ошибки.

Точно так же, как и для случая простой регрессии, вычисляются *i* значения ^-статистики (формула (4.14)), которая, с одной стороны, показывает, во сколько раз значение регрессионного коэффициента больше его стандартной ошибки, с другой стороны, служит для оценки вероятности того, что соответствующий регрессионный коэффициент равен нулю.

Как и в случае простой регрессии, нам необходим инструмент общей оценки качества построенной множественной регрессионной модели. Напомним, что в простой регрессии эту функцию выполнял 1 коэффициент детерминации Л^2 (4.9), который показывает, какую часть от общей дисперсии *y* объясняют независимые переменные. Ничто не мешает нам и в множественной регрессионной модели также использовать R^2 для оценки качества этой модели.

Дополним табл. 4.6 колонкой *y* и вычислим значения R^2 для этой модели (табл. 4.7).

В обсужденном примере мы получили достаточно большое зна-] чение коэффициента R^2 и можем, вроде бы, утверждать, что уровень ! исходной подготовки студента и активность посещения занятий в зна- | чительной степени определяют его успехи в учебе. А если бы R^2 оказался равен 0,2, либо вообще 0,05? В этом случае наша радость по поводу качества построенной модели была бы гораздо скромнее. Более того, вполне может возникнуть и более серьезный вопрос: а может быть, полученное значение вообще статистическая случайность | и связи между анализируемыми показателями на самом деле нет?

Если аналогичные сомнения возникают у нас в отношении значений регрессионных коэффициентов, то, как уже отмечалось, мы можем вычислить стандартные ошибки и, используя Г-статистику, проверить, можем ли мы отвергнуть гипотезу о равенстве нулю генерального значения соответствующего коэффициента. А есть ли такого рода инструменты для R^2*1* Можем ли каким-то образом вычислить доверительный интервал для полученного значения R^2*?*

Таблица 4.7. Оценки студентов при поступлении в вуз и по итогам 1-го семестра обучения

№ студента	Суммарный балл на вступительных экзаменах *(x,)*	Суммарный балл по итогам 1-го семестра обучения *(y)*	Процент занятий, пропущенных студентом (д:,)	Значения *y*, вычисляемые линейной регрессионной моделью *(y)*
1	32	117,4	**1**	124,2
2	**26**	106,7	**3**	114,4
3	**27**	120,0	**1**	118,2
4	27	97,3	12	103,8
5	**26**	108,0	**15**	98,6
6	**25**	124,0	**3**	113,2
7	**25**	121,4	10	104,0
8	28	106,7	12	105,0
9	29	105,3	18	98,3
10	27	96,0	10	106,4
11	26	94,7	12	102,6
12	26	89,4	20	92,0
13	**25**	113,4	**5**	110,6
14	**26**	113,3	**7**	109,2
15	24	93,3	10	102,8
16	**25**	118,7	12	101,4
17	**25**	88,0	**15**	97,4
18	**28**	100,0	11	106,3
19	14	78,7	**15**	84,3
20	18	102,7	**5**	102,2
Дисперсии		157,2		80,2

$$R^2 = \frac{80,2}{157,2} = 0,51.$$

Ответ, к сожалению, отрицательный. У нас нет таблицы критических значений R^2, и по этой причине мы не можем пойти по пути, который используем для оценки значимости регрессионных коэффи-

циентов. Метод, который применяется для вычисления уровня значимости K^1, более громоздкий. Рассмотрим его подробнее.

В модели регрессионного анализа мы предполагаем, что каждое значение зависимой переменной складывается из того значения, которое предсказывается моделью, — *y,* и некоторой ошибки (остатка) — *u.*

$$y = \hat{y} + u. \tag{4.19}$$

В этом случае дисперсия у может быть представлена в виде суммы:

$$D_y = D_{\hat{y}} + D_u. \tag{4.20}$$

Исходя из определения дисперсии перепишем последнее выражение

$$\sum \frac{(y_i - \bar{y})^2}{n} = \sum \frac{(\bar{y}_i - \bar{y})^2}{n} + \sum \frac{(u_i - \bar{u})^2}{n}. \tag{4.21}$$

Умножив обе части уравнения на *n* и вспомнив, что $\bar{u} = 0$, мы получаем выражение:

$$\sum (y_i - \bar{y})^2 = \sum (\bar{y}_i - \bar{y})^2 + \sum u_i^2. \tag{4.22}$$

Левая часть уравнения (4.22) представляет собой *общую сумму квадратов* отклонений *y* от его средней. В литературе это выражение принято обозначать знаком *TSS* (Total Sum of Squares). Первое слагаемое в правой части (4.22) является той частью суммы квадратов отклонений от средней, которая объясняется *регрессионной моделью* и обозначается как *ESS* (Explained Sum of Squares). Наконец, последний член в уравнении (4.22) есть не что иное, как просто *сумма квадратов остатков RSS* (Residuals Sum of Squares)[15]. Таким образом, уравнение (4.22) можно представить в виде:

$$TSS = ESS + RSS. \tag{4.23}$$

[15] Обратите внимание, что в этих обозначениях коэффициент детерминации (4.9) можно переписать как $R^2 = \frac{ESS}{TSS}$.

Для оценки значимости коэффициента детерминации R^2 используется *F-статистика,* которая вычисляется как отношение средних квадратов по формуле:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}}, \tag{4.24}$$

$$ti - k - X$$

где *n* — число наблюдений; *к* — число независимых переменных.

Таким образом, F-статистика представляет собой отношение объясненной суммы квадратов (в расчете на одну переменную) к необъясненной сумме квадратов (в расчете на одну степень свободы). Таблицы критических значений F-статистики приведены во многих учебниках и, следовательно, мы легко можем установить уровень значимости коэффициента детерминации для конкретного случая, что и дает возможность оценки достоверности коэффициента R^2.

К сожалению, данный метод оценки коэффициента детерминации не дает возможности построения доверительного интервала для R^2. Следовательно, получив некоторое значение R^2 по результатам анализа данных в выборке, мы не сможем оценить значение этого коэффициента в генеральной совокупности.

При выполнении команды регрессионного анализа большинство статистических пакетов проводят оценку значимости R^2 через разложение дисперсии по схеме (4.23) и рассчитывают значение F-статистики. Команда *Regression* пакета SPSS выводит эту информацию в таблице, называемой ANOVA. В табл. 4.8 и 4.9 приводятся результаты выполнения команды *Regression* пакета SPSS для данных табл. 4.5.

Во второй колонке *Sum of Squares* табл. 4.8 находятся суммы квадратов из формулы (4.23): в первой строке — *ESS,* во второй строке — *RSS,* в последней строке — *TSS.* В колонке *Mean Square* находятся те же суммы квадратов, но уже деленные на числа степеней свободы (см. знаменатель формулы (4.24)). В следующей колонке — значение ^-статистики, и, наконец, в последней колонке *Sig.* — тот уровень значимости, на котором мы можем отвергнуть гипотезу о равенстве нулю R^2. Таким образом, табл. 4.7 показывает, что мы можем отвергнуть

гипотезу об отсутствии влияния предикторов на .у на уровне значимости **а** = 0,002 < 0,05. Иными словами, с вероятностью *P* = 0,998 мы можем заключить, что суммарный балл на вступительных экзаменах и процент пропущенных занятий влияют на успеваемость студента.

Таблица 4.8. Результаты разложения дисперсии при выполнении регрессионного анализа данных табл. 4.5

ANOVA

	Sum of Squares	df	Mean **Square**	*F*	Sig.
Regression	1524,514	2	762,257	8,860	0,002
Residual	1462,576	17	86,034		
Total	2987,090	19			

Таблица 4.9. Коэффициенты регрессии при выполнении регрессионного анализа данных табл. 4.5

Coefficients

	Unstandardized Coefficients		**Standardized Coefficients**	*t*	Sig.
	B	**Std. Error**	**Beta**		
(Constant)	87,354	15,638		5,586	0,000
Суммарный балл на вступительных экзаменах	1,193	0,572	0,357	2,086	0,052
Процент пропущенных занятий	-1,316	,390	-0,577	-3,376	0,004

В табл. 4.9 мы получили различные показатели, касающиеся рег рессионных коэффициентов. Смысл и значение этих показателе' идентичны смыслу показателей, вычисляемых в случае простой ли нейной регрессии (см. табл. 4.4 и комментарии к ней). Однако ест одна колонка, значение которой мы пока не обсуждали. Это колонк

Standardized Coefficients, содержащая стандартизованные коэффициенты регрессии.

Необходимость в стандартизованных коэффициентах регрессии продемонстрируем на примере. Изучается влияние на частоту покупки определенного товара двух факторов: величины дохода *(x,)* и возраста покупателя (x_2). В результате проведенного регрессионного анализа было получено следующее уравнение:

$$y = 0{,}3 + 0{,}01x_1 + 0{,}15\,x_2, \quad (0{,}05)\ (0{,}001)\ \ (0{,}01) \qquad (4.25)$$

где/ — частота покупки товара; JC, — доход; x_2 — возраст.

Влияние обеих переменных на *у* высоко значимо: ^-статистика для о, и $Ъ_2$ равна 10 и 15 соответственно, что явно превышает разумные критические значения. При этом, поскольку коэффициент при переменной x_2 в 15 раз выше, чем коэффициент при x_v кажется, что на частоту покупки возраст влияет гораздо сильнее, чем доход.

В этом рассуждении, однако, не учтен один важный факт. А именно то, что интервал изменения возраста составляет менее 40 единиц (в данном случае — лет), поскольку в исследовании опрашивались респонденты от 25 до 60 лет. Интервал изменения дохода составляет несколько тысяч единиц (рублей). А именно, масштаб изменениях, в сотни раз больше, чем масштаб изменения x_2. Таким образом, суммарное воздействие дохода может оказаться гораздо существеннее, чем суммарное влияние возраста.

Данная ситуация вполне типична при построении регрессионных моделей для анализа социологических данных. Поскольку размерности используемых переменных могут быть очень разные, оказывается, что регрессионные коэффициенты *Ь.* часто не дают нам возможности сказать, какая же из переменных сильнее влияет на^.

Для решения задачи сопоставления влияния независимых переменных на *у* используют стандартизованную форму регрессионного уравнения. При этом подходе все переменные в уравнении регрессии стандартизуют, т.е. вместо *у* и всех предикторов используют их стандартизованные значения:

$$Z_{yi} = \frac{y_i - \bar{y}}{\sigma_y};$$
$$Z_{xi} = \frac{x_i - \bar{x}}{\sigma_x}. \qquad (4.26)$$

Как изменится регрессионное уравнение, если вместо *у их мы* будем использовать соответственно Z и *Z1* Во-первых, поскольку в результате преобразования (4.26) не изменятся коэффициенты корреляции между всеми переменными, показатель качества регрессионной модели R^2 не изменится. Во-вторых, если вспомнить, что коэффициент 6_0 вычисляется по формуле

$$b_0 = \bar{y} - b_1\bar{x}, \qquad (4.27)$$

становится ясно, что при такой замене b_0 в регрессионном уравнении станет равным нулю. В результате стандартизованная форма регрессионного уравнения будет выглядеть следующим образом:

$$Z_y = \beta_1 Z_{x1} + \beta_2 Z_{x2} + ... + \beta_n Z_{xn}. \qquad (4.28)$$

Что дает нам такая измененная форма уравнения регрессии? Для построения нашей модели, вообще говоря, ничего. Поскольку в отличие от использовавшихся в основном уравнении предикторов все Z_{xl} в уравнении (4.28) имеют одинаковый масштаб измерений, то коэффициенты р. в этом уравнении сравнимы между собой. Таким образом, сопоставляя эти коэффициенты между собой, мы можем понять, какая из переменных оказывает на^ более сильное влияние.

Таким образом, глядя на коэффициенты Beta колонки 4 табл. 4.9 видно, что активность посещения занятий влияет на успеваемость студента в 1,6 раз сильнее, чем уровень его предварительной подготовки.

Подчеркнем, что стандартизованные коэффициенты регрессии не заменяют нестандартизованных. У них другой смысл и назначение. Если нестандартизованные коэффициенты показывают, на сколько меняется *у* при изменении соответствующего *х* на единицу, стандартизованные коэффициенты позволяют сопоставить между собой общую степень воздействия каждого из *х на у.*

Ограничения модели множественного регрессионного анализа. Как и при построении модели простой линейной регрессии, для корректного вычисления стандартных ошибок регрессионных коэффициентов в модели множественной регрессии необходимо выполнять требования нормального распределения остатков регрессии и гомоскедастичности. Наряду с этими ограничениями у модели множественной регрессии есть и свое специфическое ограничение, которое называется требованием *отсутствия мулыпиколлинеарности.*

Из формулы (4.18) вычисления стандартной ошибки коэффициентов регрессии следует, что наличие высокой корреляции между какой-то парой независимых переменных приводит к резкому увеличению значений стандартных ошибок у соответствующих регрессионных коэффициентов. Рассмотрим пример, поясняющий суть данной проблемы.

Не вызывает сомнения, что на покупательское поведение человека значительно влияет размер его дохода. При этом можно предположить, что для товаров, на которые распространяется модель ситуативной покупки, более существенно влияние показателя личного дохода, а для товаров длительного пользования — среднедушевой доход. Предположим, что при изучении моделей потребления некоторого товара мы хотим узнать, какой из двух показателей оказывает более существенное влияние.

В табл. 4.10 приведены гипотетические данные по анализируемым показателям.

Таблица 4.10. Матрица, содержащая модельные данные по трем выбранным показателям

№ респондента	Количество покупок товара за последнее время	Среднедушевой доход респондента, руб.	Личный доход респондента, руб.
1	2	1000	1000
2	3	5500	4000
3	3	7000	5000
4	2	2000	2000
5	1	10000	7000
^ 6	4	5000	5000

Окончание табл. 4.10

№ респондента	Количество покупок товара за последнее время	Среднедушевой доход респондента, руб.	Личный доход респондента, руб.
7	**7**	6000	6000
8	**8**	3000	2000
9	**3**	2000	2000
10	**5**	12000	6000
11	**6**	10000	6000
12	10	10000	10000
13	**1**	3000	3000
14	**3**	4000	4000
15	**4**	6700	6700
16	**7**	15000	7500
17	**2**	4000	5000
18	**9**	9000	6500
19	**9**	7000	7000
20	**3**	3000	4000

На первом шаге анализа построим две модели простой регрессии, для того чтобы понять, как влияет на частоту покупки каждый из рассматриваемых показателей. В табл. 4.11 представлены результаты построения этих моделей.

Таблица 4.11. Параметры моделей простой линейной регрессии при двух различных независимых переменных

Описание модели	Параметры модели			
	b0	b1	*R2*	Значимость
Независимая переменная — среднедушевой доход респондента	2,33 **(1,13)**	0,0004 (0,0002)	0,23	0,03
Независимая переменная — личный доход респондента	1,06 (1,33)	0,0007 (0,0002)	0,32	0,01

Данные табл. 4.11 показывают, что обе модели высоко значимы. Что же покажет регрессионная модель с одновременным участием двух означенных переменных в качестве независимых?

Результаты построения этой модели весьма неожиданны. Значение R^2 этой модели составило 0,32 при значимости 0,04. Это первая неожиданность — значимость одновременного воздействия на у двух переменных меньше, чем отдельно любой модели.

Вторую неожиданность дает таблица регрессионных коэффициентов (табл. 4.12).

Таблица 4.12. Регрессионные коэффициенты

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
	B	**Std. Error**	**Beta**		
(Constant)	1,078	1,371		0,786	0,443
Среднедушевой доход респондента	0,00004	0,0003	0,046	0,133	0,896
Личный доход респондента	0,0007	0,0004	0,529	1,512	0,149

Из табл. 4.12 следует, что *обе* переменных оказывают слабо значимое влияние на *y*. Это уже совершенно непонятно, поскольку R^2 Достаточно высоко значим, т.е. совокупное влияние двух переменных существенно.

Объяснение этим парадоксам легко найти, если подсчитать коэффициент корреляции Пирсона двух независимых переменных. Он составляет 0,82 и, следовательно, в данном примере мы столкнулись со случаем нарушения ограничения мультиколлинеарности. Оказывается, что в ситуации сильной корреляции независимых переменных Доверять оценкам коэффициентов регрессии нельзя. Следовательно, мы не можем решить задачу выявления более сильно влияющих факторов с использованием метода множественной регрессии.

Визуальный контроль диаграммы рассеяния часто показывает, что даже когда большинство точек лежит более или менее близко к регрессионной прямой, есть, как правило, небольшое число точек, у которых расстояние с прямой весьма велико. На рис. 4.16 показана диаграмма рассеяния для гипотетического массива данных по 20 наблюдениям. Регрессионная модель достаточно хорошо описывает данные: $R^2 = 0{,}44$; $\alpha = 0{,}001$.

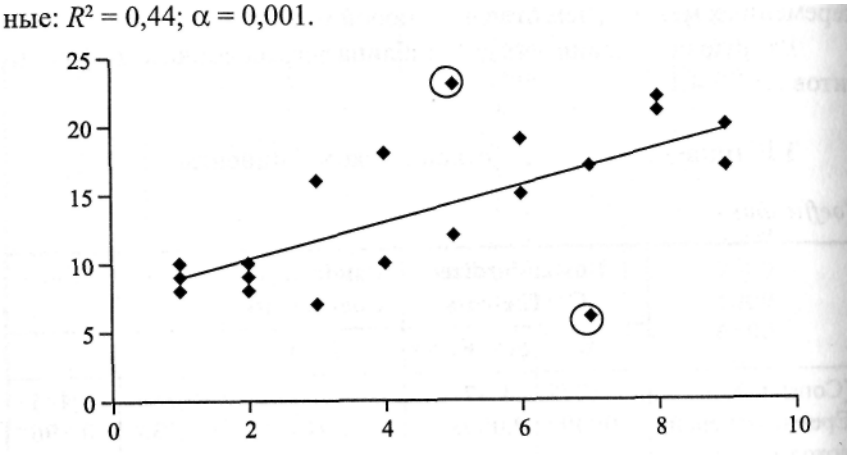


Рис. 4.16. Диаграмма рассеяния для гипотетического примера

Однако на диаграмме можно увидеть две точки, которые располагаются достаточно далеко от прямой (на рис. 4.16 они обведены кругами). С социологической точки зрения наличие такого рода точек достаточно примечательно. Оказывается, что есть два наблюдения, которые, по всей видимости, плохо вписываются в ту тенденцию, которая существует для 18 остальных наблюдений. Такого рода точки, резко выпадающие из общей тенденции и соответственно далеко отстоящие от регрессионной прямой, в регрессионном анализе принято называть *выбросами*.

Наличие выбросов — весьма негативный факт, как с математической, так и с содержательной точки зрения. С математической точки зрения выбросы ухудшают нормальность распределения остатков и увеличивают их дисперсию, что влечет увеличение стандартных ошибок регрессионных коэффициентов и уменьшение коэффициента

детерминации. С социологической точки зрения все еще хуже. Возникает подозрение, что наши данные неоднородны. В них есть часть наблюдений, для которых характерен один вид зависимости *y* от x, и другая часть, у которых эта зависимость иная. Мы же строим для всех данных одну, единую модель, которая в результате не будет описывать ни одну из этих частей данных. В некотором смысле все это напоминает вычисление средней температуры по больнице, в которой у половины больных температура 42°, а у половины — 32°. В среднем температура составляет 37°, и, опираясь на эту цифру, можно сказать, что больные, в основном, близки к выздоровлению.

Следует отметить, что появление выбросов при построении регрессионных моделей для социологических данных — явление весьма распространенное. Одной из причин их появления бывают ошибки ввода данных. Например, при вводе данных в компьютер для показателя дохода оператор совершил ошибку и вместо «10 000 рублей» ввел «1000 рублей». При построении регрессионной модели эта анкета, скорее всего, окажется выбросом. Таким образом, анализ выбросов может служить эффективным инструментом контроля данных.

Второй причиной появления выбросов при анализе социологических данных является попадание в выборку специфических совокупностей респондентов, которые по некоторым параметрам резко отличаются от остальной выборки. Например, при сборе данных по всероссийской выборке в массив вполне могут попасть работники нефтедобычи из Тюменской области. Поскольку средние зарплаты у данной категории респондентов значительно выше, чем в среднем по стране, то при построении регрессионной модели они могут оказаться выбросами. Очевидно, в такой ситуации строить общую модель нецелесообразно. Следует разделить массив на достаточно однородные группы и построить модели для каждой из них. Таким образом, анализ выбросов может помочь выделить специфические группы респондентов из общего массива данных.

К чему приведет удаление выбросов из данных примера на Рис. 4.16? Во-первых, к резкому улучшению качества модели регрессии. Коэффициент детерминации вырос с 0,44 до 0,72. Стандартные ошибки регрессионных коэффициентов уменьшились в 1,5 раза.

Во-вторых, изменились сами значения регрессионных коэффициентов, т.е. изменилось содержание регрессионной модели. По нашему мнению, модель с удаленными выбросами адекватнее отражает исследуемые закономерности.

Важным вопросом, который необходимо решить при анализе выбросов, является следующий: в какой момент определенное наблюдение следует считать выбросом? Две точки, обозначенные на рис. 4.16 как выбросы, для наглядности изображены действительно далеко отстоящими от прямой. А если бы они располагались чуть-чуть ближе к прямой, они все равно являлись бы выбросами, или уже нет? Где та граница, которая отделяет выбросы от «нормальных» данных?

Однозначного ответа на этот вопрос нет. В каждом конкретном случае ответ приходится искать исходя, прежде всего, из решаемой социологической задачи.

Определяя какое-то наблюдение как выброс, мы исходим из величины остатка. Остаток—это расстояние между реальным значением *y,* которое есть у данного респондента, и значением *y*, которое предсказывает респонденту модель. Исходя из того, что такое в нашей задаче *y,* мы и задаем границу, определяющую выброс. Например, при построении модели влияния уровня предварительной подготовки на успеваемость студента в качестве *y* у нас выступал средний балл, полученный студентом в 1-м семестре (см. табл. 4.1, рис. 4.3)J Как выбросы определим наблюдения, для которых остаток превышает 15 по абсолютной величине. Почему мы выбрали «15» в качестве границы? Исходя из здравого смысла — кажется, что те респонденты, у которых предсказанное значение среднего балла за 1-й семестр отличается от реального на 15 и более, плохо вписываются в построенную модель и этих респондентов из модели лучше удалить.

А можно ли было взять в качестве порогового значения 10? Ведь отклонение на 10 тоже достаточно сильное. При определении порогового значения для выбросов необходимо обратить внимание на дисперсию остатков. Для данных табл. 4.1 и регрессионной модели (4.6) стандартное отклонение а остатков составляет 11,3. Отсюда следует, что если мы будем использовать 10 в качестве границы (величин меньше а), то, в силу требования нормальности распределения остат

1

ков, в выбросы у нас попадет более — случаев, что весьма нежелательно. Следовательно, при определении границы выбросов важным фактором выступает разброс остатков.

Команда *Linear Regression* пакета SPSS предлагает в качестве выбросов считать случаи, когда значение остатка выходит за границу трех стандартных отклонений остатков (используется правило 3а). На рис. 4.17 приведено меню Statistics команды *Linear Regression,* в котором обведена часть, фиксирующая диагностику выбросов. Заметьте, что предлагается выбросами считать значения остатков, выходящие за 3а. Однако, в том окне, где на рис. 4.17 стоит цифра «3», можно указать и любое другое целое число.

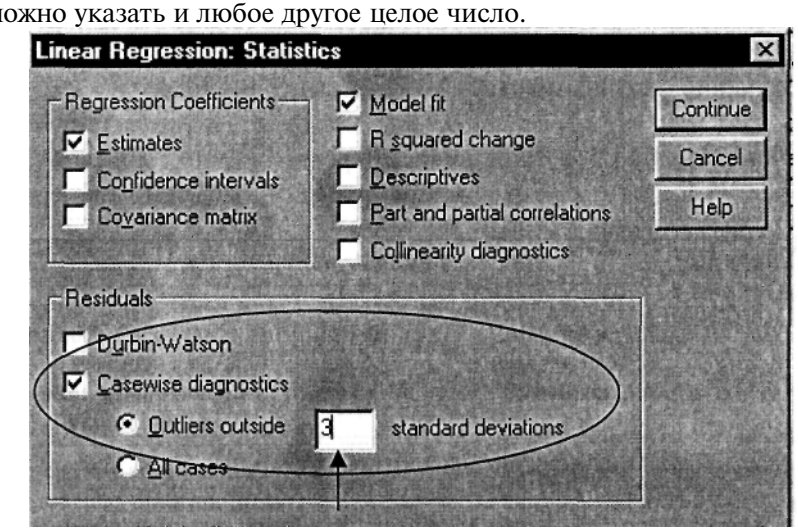


Рис. 4.17. Меню Statistics команды Linear Regression

При выборе параметров, обозначенных в меню, команда регрессии напечатает номера тех наблюдений, в которых значения остатков выходят за границы трех стандартных отклонений.

Наряду с теми ограничениями метода линейного регрессионного анализа, о которых мы говорили (нормальность распределения ос-

татков; гомоскедастичность; отсутствие мультиколлинеарности), есть еще одно очень серьезное ограничение — уровень измерения переменных, используемых в модели. Все рассуждения, статистические характеристики и меры связи, которые использовались при построении модели регрессии, применимы только к показателям, измеренным на интервальном или абсолютном уровнях[16]. В отношении социологических данных это очень неприятно, поскольку большинство переменных, с которыми работают социологи, измерены на порядковом или номинальном уровнях.

Если не преодолеть ограничение на уровень измерения переменных, окажется, что область применения регрессионных моделей в социологии весьма ограничена. Оказывается, что преодолеть это ограничение можно, причем несколькими путями.

В табл. 4.13 приведены модификации регрессионного подхода для ситуаций с разным уровнем измерения переменных.

Таблица 4.13. Разновидности регрессионных моделей в зависимости от уровня измерения переменных

Уровень измерения y	Уровень измерения x		
	Интервальный или абсолютный для всех x	Порядковый для всех x	Для некоторых x интервальный или абсолютный, для некоторых — порядковый, либо номинальный
Интервальный или абсолютный	Классическая регрессионная модель	Классическая регрессионная модель с использованием фиктивных переменных	Классическая регрессионная модель с использованием фиктивных переменных

[16] Об измерении социологических показателей см.: ТолстоваЮ.Н. Измерение в социологии: Курс лекций. М: ИНФРА-М, 1998.

Окончание табл. 4.13

Уровень измерения y	Уровень измерения x		
	Интервальный или абсолютный для всех x	Порядковый для всех x	Для некоторых x интервальный или абсолютный, для некоторых — порядковый либо номинальный
Порядковый	Множественная логистическая регрессия	Порядковая регрессия	Множественная логистическая регрессия с использованием фиктивных переменных
Номинальный с несколькими значениями	Множественная логистическая регрессия	Множественная логистическая регрессия с использованием фиктивных переменных	Множественная логистическая регрессия с использованием фиктивных переменных
Номинальный с двумя значениями	Бинарная логистическая регрессия	Бинарная логистическая регрессия с использованием фиктивных переменных	Бинарная логистическая регрессия с использованием фиктивных переменных

Несмотря на кажущуюся сложность и объемность табл. 4.13 (и соответственно многообразие регрессионных моделей), в ней легко разобраться, если учесть, что во всех моделях, наряду с классической идеей регрессии, присутствуют еще два новых подхода. Во-первых, это идея фиктивных переменных и, во-вторых, идея логитов.

4.5

Регрессионная модель с использованием фиктивных переменных

Включение в регрессионные модели переменных, измеренных на порядковом и номинальном уровнях, во многих случаях является абсолютно необходимым. Например, когда мы строили модель зависимости успеваемости от уровня предварительной подготовки, вполне логичным казалось предположение о том, что эта зависимость может быть разной для юношей и для девушек. Проверить это предположение можно, построив для этих двух групп студентов две отдельные модели и сравнив полученные результаты. Есть, однако, более эффективный, и, как будет видно далее, более общий метод — введение в регрессионную модель *фиктивных переменных*[11].

Для иллюстрации дополним табл. 4.1 данными о половой принадлежности студента (табл. 4.14).

Таблица 4.14. Оценки студентов при поступлении в вуз и по итогам 1 -го семестра обучения

№ студента	Суммарный балл на вступительных экзаменах	Суммарный балл по итогам 1-го семестра обучения	Пол (0 - - мужской; 1 - - женский)
1	**32**	117,4	**1**
2	**26**	106,7	**1**
3	**27**	120,0	**1**
4	**27**	97,3	**1**

[17] Наряду с термином «фиктивные переменные» в русскоязычной литературе для таких переменных используются также термины «индексные переменные», «псевдопеременные». В англоязычной литературе всегда используется только один термин — Dummy variables.

Окончание табл. 4.14

№ студента	Суммарный балл на вступительных экзаменах	Суммарный балл по итогам 1-го семестра обучения	Пол (0 — мужской; 1 — женский)
5	**26**	108,0	
6	**25**	124,0	*
7	**25**	121,4	0
8	**28**	106,7	1
9	**29**	105,3	
10	**27**	96,0	0
11	**26**	94,7	
12	**26**	89,4	**1**
13	**25**	113,4	**1**
14	**26**	113,3	**1**
15	**24**	93,3	0
16	**25**	118,7	
17	**25**	88,0	
18	**28**	100,0	¹
19	**14**	78,7	0
20	**18**	102,7	1

Определим средние значения двух рассматриваемых оценок для юношей и для девушек (табл. 4.15).

Таблица 4.15. Средние оценки, полученные на вступительных экзаменах и по итогам 1 -го семестра юношами и девушками

Пол	Средняя сумма баллов на вступительных экзаменах	Средний суммарный балл по итогам 1-го семестра	*N*
Женский	**26,2**	106,6	**16**
Мужской	**22,5**	97,3	**4**
Всего	**25,5**	104,7	**20**

Данные табл. 4.15 показывают, что как на вступительных экзаменах, так и по итогам 1 -го семестра оценки девушек несколько выше, чем оценки юношей. Таким образом, если мы будем строить регрессионные модели зависимости успеваемости от уровня предварительной подготовки, то, скорее всего, это будут две прямые. Одна из них (данные по девушкам) расположена несколько выше другой (данные по юношам). Поэтому следует строить две модели, а не одну. Можно *m* тем не менее, свести это к одной модели? Оказывается, можно.

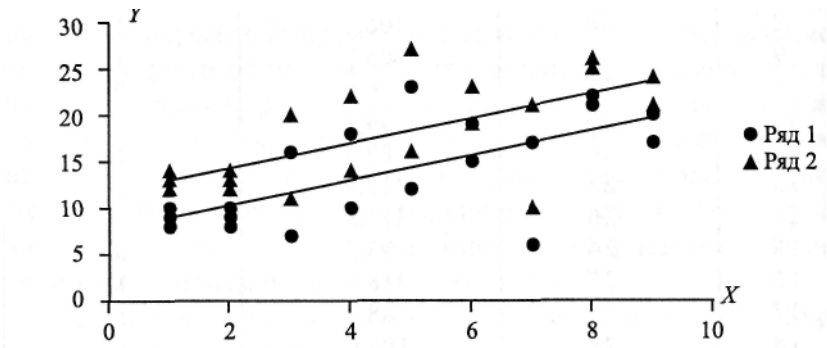


Рис. 4.18. Гипотетическая модель для двух регрессионных моделей

Предположим, что мы имеем две регрессионные модели, аналогичные тем, которые изображены на рис. 4.18 и которые записываются в виде уравнений:

$$y = 7{,}6 + 1{,}3x \quad \text{(ряд 1)};$$
$$y = 11{,}6 + 1{,}3x \quad \text{(ряд 2)}. \qquad (4.29)$$

Очевидно, что коэффициенты b_j в этих уравнениях должны быть одинаковы, поскольку прямые на рис. 4.18 идут параллельно. Два ряда данных, представленных в уравнениях (4.29), можно представить в виде одного уравнения:

$$y = 7{,}6 + 4D + 1{,}3x. \qquad (4.30)$$

В уравнении (4.30) D — переменная, которая принимает значение нуль, если это данные из ряда 1, и единица, если данные принадлежат ряду 2.

На уравнение (4.30) можно взглянуть как на модель множественной регрессии с двумя независимыми переменными x и D. При таком подходе две прямые на рис. 4.18 становятся одним графиком для регрессионной модели (4.30). Принципиально важно, что в данном примере переменная D — фактически номинальная переменная, которая делит всю совокупность на две части — ряд 1 и ряд 2.

Таким образом, модели множественной регрессии типа (4.30), в которые входит дихотомическая переменная, могут описывать зависимости, куда в качестве одного из x входит переменная, измеренная на номинальном уровне. Если вернуться к данным табл. 4.14, можем построить модель одновременного влияния на успеваемость и уровня предварительной подготовки и пола студентов.

Проведя вычисления для данных всей генеральной совокупности, получаем следующее регрессионное уравнение:

$$y = 81{,}7 + 10{,}9x_1 + 0{,}53x_2,$$
$$(11{,}2) \quad (3{,}2) \quad (0{,}46) \qquad (4.31)$$
$$\alpha = 0{,}001 \; \alpha = 0{,}001 \; \alpha = 0{,}26$$

где x_x — фиктивная переменная «Пол студента»; x_2 — переменная «Суммарный балл на вступительных экзаменах»; $R^2 = 0{,}16$.

Введение в модель, объясняющую успеваемость, переменной «Пол студента» принципиально меняет не только вид модели (сопоставьте модель (4.12) с моделью (4.31)), но и ее содержательную интерпретацию. Модель (4.12) показывает, что успеваемость на 18% объясняется уровнем предварительной подготовки студентов. Модель (4.31) говорит нам, что уровень предварительной подготовки студентов значимого влияния на успеваемость не оказывает, а вот пол влияет на успеваемость, и притом существенно.

Для оценки достоверности полученного результата необходимо проверить выполнение для модели (4.31) ограничений метода множественного регрессионного анализа.

1. *Ограничение мультиколлинеарности.* Коэффициент корреляции Пирсона между переменными x^ и x_2 составляет 0,27. Этот коэффициент хотя и значим с а = 0,02, однако, очевидно, мультиколлинеарности между независимыми переменными нет.

2. Нормальность распределения остатков. На рис. 4.19 показана гистограмма распределения остатков для модели (4.31) для данных, фрагмент которых представлен в табл. 4.14. Пунктиром обозначена кривая нормального распределения. Хотя представленная гистограмма и не совпадает с нормальным распределением, представляется, что общий вид этой гистограммы позволяет использовать описанные подходы к оценке значимости коэффициентов регрессии.

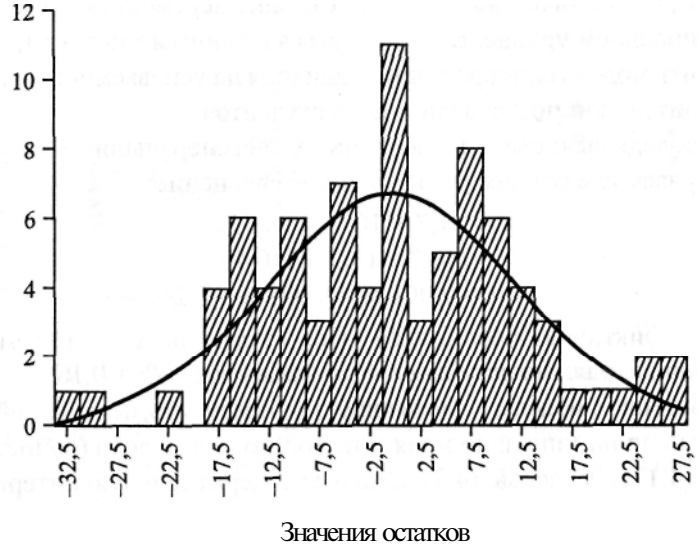


Рис. 4.19. Гистограмма распределения остатков модели (4.31)

3. *Гомоскедастичность.* Мы рассматривали проблему равенства дисперсий остатков при разных значениях *x* в ситуации простой линейной регрессии. В случае модели множественной регрессии, вместо проверки равенства дисперсий остатков по всем независимым переменным, можно проверить это равенство при различных значениях *y*. Таблица 4.16 содержит данные для проверки гомоскедастичности и показывает, что, хотя и нет точного равенства дисперсии при разных значениях *y* (чего на практике практически никогда не бывает), тем

_н_е менее, поскольку нет резких отклонений, можно констатировать, _ч_хо существенных нарушений гомоскедастичности нет.

Таблица 4.16. Проверка равенства дисперсии остатков для модели (4.31)

Значения *y*	**Дисперсия остатков**	***N***
Менее 87	**50,5**	**8**
87—90	**44,5**	**8**
90—94	65,9	**8**
94—96	**54,5**	**8**
96—100	**53,2**	**7**
100—106	**65,1**	**9**
106—110	**51,6**	**7**
110—114	**53,5**	10
114—120	**44,6**	**9**
Более120	**50,7**	10

Таким образом, проверка выполнения ограничений регрессионного метода свидетельствует о том, что для модели (4.31) они выполняются.

Рассмотренные способы применения дихотомических переменных в модели регрессии открывают перед нами возможности введения в регрессию переменных, измеренных на номинальном и порядковом уровне. Для реализации этих возможностей существует процедура создания из номинальных или порядковых переменных нескольких дихотомических переменных, которые называются фиктивными переменными.

Процедура эта состоит в следующем. Предположим, что мы имеем переменную с четырьмя градациями, измеренную на номинальным уровне. Пусть это будет вопрос о семейном положении. Очевидно, что с социологической точки зрения семейное положение является характеристикой, оказывающей существенное влияние на многие поведенческие, мотивационные, ценностные аспекты жизни индивида. В этой связи включение семейного положения в число независимых переменных весьма желательно для построения многих социологических моделей.

Разделим переменную «Семейное положение» на фиктивные переменные следующим образом.

Семейное положение

1. Холост (не замужем)——> *Q*: 1. Холост (не замужем)
   0. Иное семейное положение
2. Женат (замужем)——> *Q2:* 1. Женат (замужем)
   0. Иное семейное положение
3. Разведен(а)——> *Q3:* 1. Разведен(а)
   0. Иное семейное положение
4. Вдовец (вдова)——> *QA:* 1. Вдовец (вдова)
   0. Иное семейное положение

Одну переменную «Семейное положение» мы преобразовали в четыре дихотомические переменные, которые в совокупности эквивалентны одной исходной переменной. Эквивалентны в том смысле, что вся информация, которая содержится в ответе респондента на исходный вопрос, без потерь может быть извлечена из значений новых четырех переменных. Более того, на самом деле для восстановления информации исходного вопроса достаточно любых трех из четырех созданных переменных. Действительно, если мы оставим лишь первые три переменные — *Q\, Q2, Q3,* кажется, что мы можем потерять ответы тех респондентов, которые отметят семейное положение «4». Однако при таком ответе респондента значения переменных *Q\,Q2u Qb* будут равны нулю. Таким образом, значение нулю данных трех переменных означает, что переменная *QA* будет равна единице. Если хотя бы одна из переменных *Ql,Q2* или *Q3* равна единице, это означает, что переменная *QA* равна нулю.

Общее правило, которое следует из рассмотренного примера, — всю информацию, которая содержится в переменной с ./Vградациями, можно сохранить, используя *N*- 1 дихотомическую переменную.

Что мы выигрываем, заменяя одну исходную переменную несколькими дихотомическими? Выигрываем мы многое: у нас появляется возможность включения переменной «Семейное положение», измеренной на номинальном уровне, в регрессионную модель. Правда,

придется включать не одну переменную, а несколько дихотомических, но, самое главное, что мы можем изучать степень воздействия на *y* не только количественных показателей, но любых социологических переменных.

Интерпретация коэффициентов регрессии при фиктивных переменных. Смысл коэффициентов регрессии при фиктивных переменных принципиально отличается от коэффициентов при обычных количественных переменных. Напомним, что нестандартизованный коэффициент Ъ. показывает, на сколько единиц изменяется значение *y* при изменении *x.* на одну единицу. Для понимания смысла регрессионных коэффициентов при фиктивных переменных вернемся к примеру, в котором мы создали три фиктивных переменных для переменной «Семейное положение». Если выполнить процедуру построения модели множественной регрессии с использованием этих переменных, модель будет выглядеть следующим образом:

$$y = b_0 + b_1 Q_1 + b_2 Q_2 + b_3 Q_3. \qquad (4.32)$$

Что показывает в этой модели коэффициент 6_0? Обратите внимание, что в ситуации, когда исходная переменная «Семейное положение» имеет значение «4», т.е. когда респондент отметил в вопросе позицию «Вдовец (вдова)», то переменные *Q\, Ql* и *Q3* будут равны нулю. Таким образом, уравнение (4.32) для таких респондентов превращается в выражение $y - b_0$. Отсюда и смысл коэффициента $Ь_0$ — это среднее значение *y* для группы респондентов, для которой не создано фиктивной переменной.

Чему будет равно среднее значение *y* для тех респондентов, которые на вопрос о семейном положении отметили позицию «1» — холосты (не замужем)? Для этих респондентов фиктивная переменная *Q* будет равна единице, а остальные — нулю. Таким образом, уравнение (4.32) приобретает следующий вид: $y = 6_0 + Ъ,$. Это выражение показывает, что среднее значение *y* для респондентов, имеющих семейное положение «1», на $Ъ_x$ отличается от среднего значения *y* у респондентов, имеющих семейное положение «4». Отсюда вытекает общая закономерность, объясняющая смысл регрессионных коэффициентов при фиктивных переменных.

Коэффициент *Ь.* при фиктивной переменной *х.* показывает, на сколько среднее значение *у* в группе респондентов, для которых значение фиктивной переменной *х.* равно единице, отличается от среднего значения *у* в группе респондентов, для которых не создано фиктивной переменной. Все коэффициенты *Ь.* при фиктивных переменных показывают величину различия с одной группой респондентов. Таким образом, группа, для которой не создано фиктивной переменной, выступает эталонной, с которой и сопоставляются все остальные группы. Такую группу обычно называют *контрольной группой*.

Если вернуться к примеру с созданием фиктивных переменных для показателя «Семейное положение», возьмем в качестве *у* величину заработка респондента и построим регрессионную модель с фиктивными переменными[18]:

$$y = \underset{\substack{(373,6)\\ \alpha = 0,001}}{1805,8} + \underset{\substack{(407,1)\\ \alpha = 0,03}}{915,3\,Q1} + \underset{\substack{(385,3)\\ \alpha = 0,18}}{515,6\,Q2} + \underset{\substack{(428,8)\\ \alpha = 0,69}}{172,9\,Q3}. \qquad (4.33)$$

Из модели (4.30) видно, что средний заработок респондентов с семейным положением «4» (вдовцов) составляет 1805,8 руб. Средний заработок холостяков (семейное положение «1») выше заработка вдовцов на 915,3 руб. Средний заработок группы женатых (замужних) респондентов выше заработка вдовцов на 515,6 руб. Заработок разведенных респондентов выше заработка вдовцов в среднем на 172,9 руб.

О чем говорят значения доверительных интервалов и уровни значимости для регрессионных коэффициентов при фиктивных переменных? Например, для коэффициента при *Q* мы можем утверждать, что с вероятностью 95% разность между средней зарплатой вдовцов и холостяков лежит в интервале (101,1 - 1729,5) руб. Уровень значимости показывает, с какой вероятностью мы можем утверждать, что средний размер зарплаты у соответствующей группы не отличается от средней зарплаты респондентов контрольной группы. Для уравнения (4.33) это означает, например, с вероятностью 0,69, что гипотеза о том, что средний уровень зарплаты у холостяков не отличается от зарплаты группы вдовцов (контрольной группы)[19] может быть отвергнута лишь на уровне значимости а = 0,69. Значит, мы должны ее принять.

[18] Данные для расчета этой модели взяты из всероссийского опроса, проведе ного ВЦИОМ в мае 2001 г. в рамках исследования «Мониторинг социальных и эко мических перемен».

О выборе контрольной группы. Удобная и социологически прозрачная интерпретация результатов регрессионного анализа с использованием фиктивных переменных зависит от выбора контрольной группы. Обсуждая значение каждого из регрессионных коэффициентов, мы говорим, что они показывают, насколько среднее значение *у* в этой группе больше (или меньше) среднего значения^ в контрольной группе. Чтобы такое сопоставление между двумя группами было интересным, смысл контрольной группы должен быть понятен. Например, если в качестве контрольной группы возьмем респондентов, которые затруднились с ответом на вопрос, то сама эта группа, в большинстве случаев, будет крайне неоднородной и противоречивой. Действительно, группа затруднившихся с ответом обычно включает и тех, кто поленился ответить, и тех, кто после мучительных размышлений так и не смог выбрать один из предложенных вариантов, и тех, кто просто ничего не знает по теме вопроса, и, наверное, еще какие-то группы респондентов.

Таким образом, если мы будем говорить, что «в анализируемой группе среднее значение *у* больше, чем в группе затруднившихся ответить», то социологического смысла в этом будет немного. Эталон для сопоставления должен представлять социологически понятную группу респондентов. Тогда и само сравнение будет представлять интерес.

Вторым требованием к выбору контрольной группы является ее объем. Что произойдет, если в качестве контрольной группы мы выберем очень маленькую группу? Например, если контрольная группа будет составлять, скажем, 3% всей выборки. В этом случае соответствующая фиктивная переменная в 3% всех случаев будет иметь значение «1» и в 97% случаев — «0». Если объем выборки при этом будет составлять 500 респондентов, дисперсия этой фиктивной переменной будет 0,006.

[19] Для того чтобы полученный результат приобрел статус достоверного, мы должны проверить, в какой степени для модели (4.33) выполняются ограничения регрессионной модели.

Вернувшись к формуле определения стандартной ошибки для ко-; эффициентов множественной регрессии (4.18), увидим, что в знаменателе этой формулы находится D_x — дисперсия x. Ясно, что при такой низкой дисперсии x показатель стандартной ошибки будет большим.

Рассмотрим пример, который показывает влияние размера выбираемой контрольной группы на получаемые результаты[20]. В качестве y возьмем величину заработной платы респондента: каким был размер вашего заработка, доходов от основной работы, полученных в прошлом месяце (после вычета налогов). В качестве переменной, влияющей на размер доходов, используем самооценку респондентом социального статуса: к какому слою в обществе вы бы, скорее всего, себя отнесли:

1. К низшему слою.
2. К рабочим.
3. К низшей части среднего слоя.
4. К средней части среднего слоя.
5. К высшей части среднего слоя.
6. К высшему слою.
7. Затрудняюсь ответить.

На первом шаге удалим из массива данных респондентов, затруднившихся с ответом. Из оставшихся шести градаций вопроса необходимо определить группу, которая будет взята в качестве контрольной.

Для дальнейшего сравнения в качестве контрольной группы целесообразно взять одну из полярных групп — первую или последнюю Однако табл. 4.17 показывает, что последняя группа крайне мала. Если с содержательной точки зрения эта совокупность достаточно однородна и социологически понятна, в ситуации малой по объему контрольной группы нет ничего страшного. Рекомендация отнесения к контрольной группе достаточно больших совокупностей респондентов является не столько требованием, сколько пожеланием. Однако в такой рекомендации еще один резон, который проявится в обсуждаемой далее ситуации создания нескольких совокупностей фиктивных переменных.

[20] Данные для расчета этой модели взяты из всероссийского опроса, проведенного ВЦИОМ в мае 2001 г. в рамках исследования «Мониторинг социальных и экономических перемен».

Таблица 4.17. Результат расчета командой Frequencies пакета SPSS ответов на вопрос анкеты: «К какому слою в обществе вы бы, скорее всего, себя отнесли?»

		Frequency	Percent	Valid Percent
Valid	К низшему слою	260	10,8	11,7
	К рабочим	869	36,1	39,0
	К низшей части среднего слоя	356	14,8	16,0
	К средней части среднего слоя	658	27,3	29,5
	К высшей части среднего слоя	77	3,2	3,4
	К высшему слою	10	0,4	0,5
	Total	2231	92,7	100,0
Missing	Затрудняюсь ответить	176	7,3	
Total		2407	100,0	

Показатель качества для модели составляете^2 = 0,033 с **а** = 0,001.

В интерпретации результатов, представленных в табл. 4.18, есть еще одна специфика — из регрессионных коэффициентов следует, что, с одной стороны, разница в зарплате между представителями контрольной группы и респондентами, отнесшими себя к другим группам среднего слоя, достаточно велика, а с другой — /-статистика показывает, что эта разница незначима (почти все ***а*** > 0,05).

Из этого факта можно сделать два вывода. Во-первых, величина зарплаты лиц, относящих себя к высшему классу, приблизительно равна зарплате тех, кто относит себя к другим социальным слоям. Этот вывод напрямую следует из табл. 4.18 регрессионных коэффициентов. Второй вывод естественным образом вытекает из первого: зарплаты людей, относящих себя к разным социальным слоям, равны между собой. Действительно, если, с одной стороны, зарплата относящих себя к высшему слою не отличается от зарплаты относящих себя к высшей части среднего слоя, а с другой стороны, зарплата относящих себя к высшему слою не отличается от зарплаты относя-

щих себя к рабочим, то, кажется, что можно заключить, что зарплата относящих себя к рабочим не отличается от зарплаты относящих себя к верхней части среднего слоя. Иными словами, если $A = B$ и $A = C$, то можно заключить, что $B = C$. Это свойство в математике называют *транзитивностью*.

Таблица 4.18. Результат расчета командой Regression пакета SPSS параметров регрессии для случая контрольной группы «Принадлежность к высшему слою»

Coefficients

	Unstandardized Coefficients		**Standardized Coefficients**	*t*	Sig.
	B	**Std. Error**	**Beta**		
(Constant)	3112,1	1139,4		2,7	0,01
К низшему слою	-1825,8	1194,4	-0,16	-1,5	0,13
К рабочим	-980,5	1145,0	-0,20	-0,9	0,39
К низшей части среднего слоя	-1007,8	1155,1	**-0,15**	-0,9	0,38
К средней части среднего слоя	-320,6	1147,9	**-0,06**	-0,3	0,78
К высшей части среднего слоя	443,9	1207,9	0,03	0,4	0,71

Оказывается, что в отношении коэффициентов свойство транзитивности не соблюдается. Это легко продемонстрировать, если построить регрессионную модель для тех же переменных, но в качестве контрольной группы взять, скажем, респондентов, относящих себя к нижнему слою. Регрессионные коэффициенты этой модели приведены в табл. 4.19.

Показатель качества для модели составляет $R^2 = 0{,}033$ с $\mathbf{a} = 0{,}001$.

Вначале отметим, что показатель качества в двух обсуждаемых моделях одинаков, что неизбежно, поскольку количество информации в совокупностях фиктивных переменных в обоих случаях одинаково. Далее, модель (см. табл. 4.19) подтверждает одну часть первого вывода — зарплата тех, кто относит себя к высшему слою, слабо отличается от зарплаты тех, кто относит себя к низшему слою. Что же касается второго вывода, то данные табл. 4.19 его опровергают. Действительно, зарплату в контрольной группе мы можем считать отличающейся от зарплат всех групп, кроме группы респондентов, относящих себя к высшему слою (все а, кроме последнего, меньше 0,05).

Таблица 4.19. Результат расчета командой Regression пакета SPSS параметров регрессии для случая контрольной группы «Принадлежность к низшему слою»

Coefficients

	Unstandardized **Coefficients**		**Standardized Coefficients**	*t*	Sig.
	B	**Std. Error**	**Beta**		
(Constant)	1286,3	358,6		3,6	0,000
К рабочим	845,3	376,0	0,17	2,2	0,025
К низшей части среднего слоя	818,0	405,8	0,12	2,0	0,044
К средней части среднего слоя	1505,2	384,8	0,28	3,9	0,000
К высшей части среднего слоя	2269,7	537,9	0,17	4,2	0,000
К высшему слою	1825,8	1194,5	0,05	1,5	0,127

Следовательно, вывод о равенстве зарплат во всех группах не подтверждается. Подробнее вопрос о причинах нарушения транзитивности и следствиях этого обсуждается в разд. 3.4 «Метод множественных сравнений».

Продемонстрированный метод использования номинальных либо порядковых переменных в регрессионной модели открывает большие перспективы для включения в число независимых переменных широ-

кого списка самых разных показателей. Есть, однако, определенная специфика использования нескольких переменных в таких моделях.

Несколько групп фиктивных переменных. Расширим список переменных, влияющих на заработную плату из примера модели (4.33), включением в этот список переменной «Образование респондента».

Включение в модель (4.33) двух групп фиктивных переменных дает нам модель (4.34).

$$y = 2527{,}7 + 853{,}6\,Q1 + 484{,}6\,Q2 + 179{,}8\,Q3 - 594{,}4\,Q4 - 925{,}7\,Q5 - 819{,}4\,Q6. \quad (4.34)$$

(408,9) (404,2) (382,9) (426,2) (310,7) (220,1) (209,2)

$\alpha = 0{,}001$ $\alpha = 0{,}04$ $\alpha = 0{,}21$ $\alpha = 0{,}67$ $\alpha = 0{,}06$ $\alpha = 0{,}001$ $\alpha = 0{,}001$

Какой смысл имеет коэффициент b_0 в этой модели? Напомним, что в модели (4.33) коэффициент b_0 был равен среднему значению y в контрольной группе. В модели (4.34) мы имеем две группы фиктивных переменных и соответственно две контрольные группы. Соответственно, в модели (4.34) контрольной группой будет пересечение двух контрольных групп. Иными словами, контрольная группа — это вдовцы (вдовы) с незаконченным или полным высшим образованием и со средней зарплатой 2527,7 руб.

Какой смысл у коэффициента о.? Он показывает, как отличается среднее значение у в г'-й группе от среднего значения у в объединении контрольных групп, либо от контрольной группы, образованной для соответствующей группы фиктивных переменных.

Взаимодействие переменных. Попытаемся теперь определить степень влияния на зарплату одновременно семейного положения и образования респондента. Две независимые переменные, каждая из которых имеет четыре градации, в совокупности дают нам 16 возможных сочетаний значений. Для каждого из этих сочетаний потребуется создание своей фиктивной переменной, кроме одного сочетания, которое будет выбрано контрольной группой. В табл. 4.20 сочетание (4,4) было выбрано контрольной группой, и соответственно переменная *Q44* в таблице отсутствует (для демонстрации эта клетка в таблице заштрихована).

Таблица **4.20.** Список фиктивных переменных для включения в регрессионную модель двух номинальных переменных —- «Образование» и «Семейное положение»

Семейное положение	**Образование**			
	общее начальное или неполное среднее	**общее полное среднее**	**среднее специальное**	**незаконченное высшее, высшее**
Холост (не замужем)	**211**	**612**	613	**614**
Женат (замужем)	**621**	622	623	624
Разведен(а)	031	632	633	634
Вдовец (вдова)	**Q41**	642	643	

Использование фиктивных переменных для угла **наклона.** В рассмотренных примерах одновременного включения в модель количественных и номинальных (порядковых) переменных, последние преобразовывались в наборы фиктивных переменных так, что получаемые регрессионные прямые шли параллельно друг другу (см. модель (4.29) рис. 4.18). В примере (4.29) это означает, что зависимость успеваемости от уровня предварительной подготовки у юношей и девушек одинакова, только у юношей исходный уровень подготовки ниже. Аналогичный подход фактически заложен в изложенном выше подходе использования фиктивных переменных.

С точки зрения содержательных социологических моделей предположение о параллельности регрессионных прямых для разных социальных групп в большинстве случаев выглядит надуманным. Возможно ли в рамках регрессионного подхода преодолеть это ограничение? Можно, причем с использованием тех же фиктивных переменных.

При введении фиктивных переменных для изменения угла наклона регрессионная модель будет выглядеть следующим образом:

$$y = b_0 + (b_1 + b_2 Q1)x_1 + b_3 Q1, \quad (4.35)$$

где Л:, — количественная переменная; $Q\backslash$ — фиктивная переменная. Выражение (4.35) можно переписать в следующем виде:

$$y = b_0 + b_1 x_1 + b_2 Q1 x_1 + b_3 Q1. \quad (4.36)$$

Поскольку переменная $Q\backslash$ является фиктивной, уравнение (4.36) представляет собой два уравнения. Одно для ситуации $Q\backslash = 0$, а другое — 61 = 1:

$$\begin{array}{ll} y = b_0 + b_1 x_1 + b_2 x_1 + b_3 & (Q1 = 1); \\ y = b_0 + b_1 x_1 & (Q1 = 0). \end{array} \quad (4.37)$$

Построение модели (4.36) дает регрессионную модель с разными углами наклона для двух разных уровней $Q\backslash$.

4.6

Логистическая регрессия

Фактическим ограничением регрессионного анализа является то, что зависимая переменная должна быть либо количественной, т.е. иметь интервальный или абсолютный уровень измерения, либо дихотомической. Последнее очень важно, поскольку во многих случаях мы хотим изучить влияние разных факторов на электоральное поведение, на потребительское поведение и др. Когда зависимая переменная^ -— дихотомическая («проголосует за определенную партию — не проголосует», «купит определенный товар — не купит»), используется метод *логистической регрессии*.

Непосредственно включить в регрессионную модель дихотомический *y* нельзя. Однако это можно сделать, если вместо *y* использовать некоторую производную от *y* функцию — *логит*.

Отношение шансов и логит. Отношение вероятности того, что событие произойдет, к вероятности того, что оно не произойдет $P/(1 - P)$, называется отношением шансов (или отношением предпочтения). С этим отношением связана модель логистической регрессии, получаемая за счет непосредственного задания зависимой переменной в виде $Z = \mathrm{Ln}(P/(1-P))$, где $P = P\{Y = 1|X^1, \ldots, X^p\}$. Переменная Z называется *логитом*. Модель логистической регрессии определяется уравнением регрессии

$$Z = B_0 + B_1 X^1 + \ldots + B_p X^p. \quad (4.38)$$

Что мы получим, с точки зрения знания о зависимости *y,* от совокупности $\{x_v \ldots, x_p\}$ если будем знать зависимость ZOT ЭТИХ переменных? Ведь на самом деле нас интересует именно *y,* а не какой-то там логит. Рассмотрим пример.

Пусть мы анализируем детерминанты электорального поведения респондентов. Введем для тех, кто проголосовал за партию *X,* значение $y = 1$, а для тех, кто проголосовал за другую партию, $y = 0$. Если по результатам исследования мы получили, что логит голосования за партию Хдля мужчин равен -0,847, а для женщин -1,386, это значит, что отношение предпочтения для мужчин равно $e^{(-0{,}847)} = 0{,}43$, а для женщин — $e^{(-1{,}386)} = 0{,}25$. Иными словами, среди мужчин за партию *X* проголосовали 43% опрошенных, а среди женщин — 25%. Следовательно, зная логит, мы получаем прямую информацию о поведении^.

Правая часть уравнения (4.38) повторяет обычную запись модели множественной регрессии в (4.17).

Отношение шансов может быть записано в виде

$$P/(1-P) = e^{B_0 + B_1 X^1 + B_2 X^2 + \ldots + B_p X^p} = e^{B_0} e^{B_1 X^1} \ldots e^{B_p X^p} = e^{B_0} (e^{B_1})^{X^1} \ldots (e^{B_p})^{X^p}. \quad (4.39)$$

Отсюда получается, что если модель верна, при независимых X^1, -У изменение X^* на единицу вызывает изменение отношения шансов в e^{a}"раз.

Логистическая регрессия решает задачу построения модели прогноза вероятности события (y = 1} в зависимости от переменных •*„ ..., $л_д$. Иначе эта связь может быть выражена в виде зависимости $P\{y = 1|X\} = f(X)$.

Логистическая регрессия выражает эту связь в виде формулы

$$P\{Y=1|X^1,...,X^p\}=\frac{e^z}{1+e^z},\ \text{где}\ Z=B_0+B_1X^1+...+B_pX^p. \qquad (4.40)$$

Название «логистическая регрессия» происходит от логистического распределения, имеющего функцию распределения $F(x)=\frac{e^{(x-a)/k}}{1+e^{(x-a)/k}}$. Таким образом, модель, представленная этим видом регрессии, по сути, является функцией распределения этого закона, в которой в качестве аргумента используется линейная комбинация независимых переменных.

Решение уравнения с использованием логита. Механизм решения уравнения (4.40) можно представить следующим образом.

1. Получаются агрегированные данные по переменными, в которых для каждой группы, характеризуемой значениями $X_j=(X_j^1,...,X_j^p)$, подсчитывается доля объектов, соответствующих событию {*Y*= 1}. Эта доля является оценкой вероятности $\hat{P}_j=P\{Y=1|X_j^1,...,X_j^p\}$. В соответствии с этим для каждой группы получается значение логита Z.

2. На агрегированных данных оцениваются коэффициенты уравн $Z=B_0+B_1X^1+...+B_pX^p$. сожалению, дисперсия Z здесь зависит от значений *X,* поэтому для логита применяется специальная техника оценки коэффициентов — метод взвешенной регрессии.

Еще одна особенность состоит в том, что в реальных данных очень часто группы по *X* оказываются однородными по *Y,* поэтому оценки *Pj* становятся равными нулю или единице. Таким образом, оценка логита для них не определена:
для этих значений $Z=\mathrm{Ln}(0/(1-0))=-\infty, Z=\mathrm{Ln}(1/(1-1))=\infty)$

В настоящее время для оценки коэффициентов используется метод максимального правдоподобия, лишенный этого недостатка. Тем не менее проблема, хотя и не в таком остром виде остается: если оценки вероятности для многих групп оказываются равными нулю или единице, оценки коэффициентов регрессии имеют слишком большую дисперсию. Поэтому, имея в качестве независимых переменных такие признаки, как, например, душевой доход в сочетании с возрастом, их следует укрупнить по интервалам, приписав объектам средние значения интервалов.

Неколичественные данные. Если в обычной линейной регрессии для работы с неколичественными переменными нам приходилось подготавливать специальные фиктивные переменные, в реализации логистической регрессии в SPSS это может делаться автоматически. Для этого в диалоговом окне специально предусмотрены средства, сообщающие пакету, что ту или иную переменную следует считать категориальной. При этом, чтобы не получить линейно зависимых переменных, максимальный код значения рассматриваемой переменной (или минимальный, в зависимости от задания процедуры) не перекодируется в дихотомическую (индексную) переменную. Впрочем, средства преобразования данных позволяют не учитывать любой код значения. Существуют и другие способы перекодирования категориальных (неколичественных) переменных в несколько дихотомических, но мы будем пользоваться только указанным, как более естественным.

Взаимодействие переменных. В процедуре логистической регрессии в SPSS предусмотрены средства для автоматического включения в уравнение переменных взаимодействий. В диалоговом окне в списке исходных переменных для этого следует выделить имена переменных, взаимодействия которых предполагается рассмотреть, затем переправить выделенные имена в окно независимых переменных кнопкой с текстом *>a* x *Ь>.*

На рис. 4.20 показано меню вызова команды логистической регрессии в пакете SPSS.

Главное меню команды логистической регрессии представлено на рис. 4.21. В данном меню тестируется модель анализа влияния на то, потребляет ли респондент спиртные напитки (переменная ет80) следующих показателей:

- курит ли респондент (переменная ет71);
- величина заработка (переменная ejlO);
- пол (переменная eh5);

- наличие подчиненных на работе (переменная ej6);
- переменная взаимодействия пол — доход (переменная qql)[21].

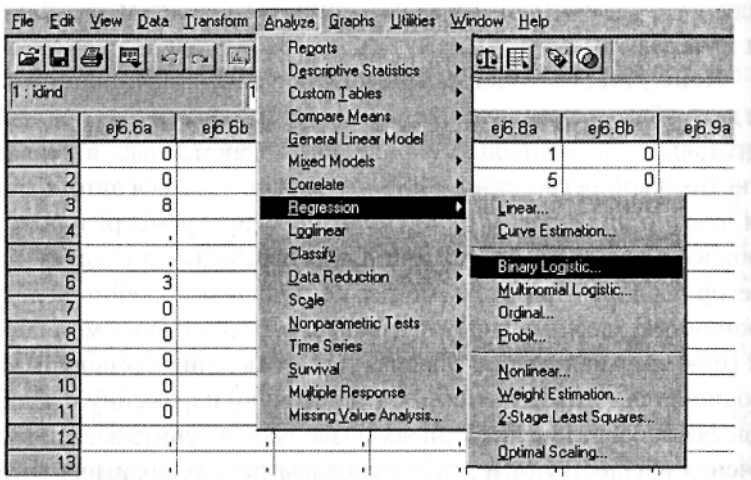


Рис. 4.20, Меню вызова команды логистической регрессии

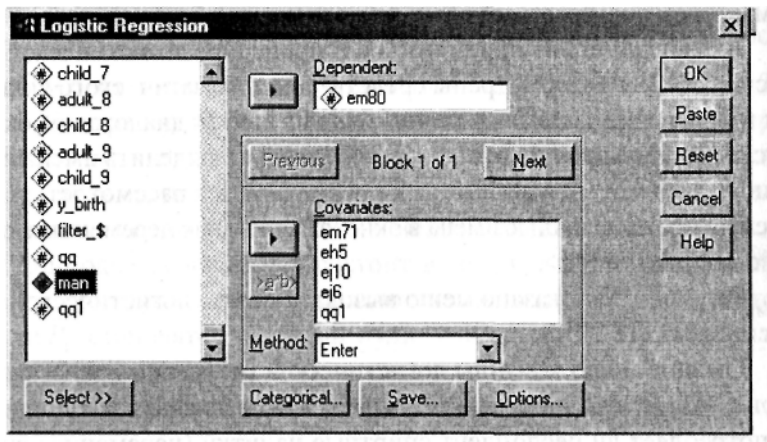


Рис. 4.21. Главное меню команды логистической регрессии

[21] Данные взяты из исследования проведенного в октябре-ноябре 2001 г. Описание РМЭЗ см.: Сваффорд М.С., Косолапое М.С., Козырева П.М. Российский мониторинг экономического положения и здоровья россиян (РМЭЗ): измерение благосостояния россиян в 90-е годы // Мир России. 1999. № 3. С. 153—172.

Результаты работы команды, показанной на рис. 4.21, приведены в табл. 4.21.

Таблица 4.21. Результаты выполнения команды логистической регрессии, представленной на рис. 4.21

Model Summary

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
4398,137	0,056	0,077

Variables in the Equation

	B	S.E.	**Wald**	df	Sig.	**Exp(B)**
EM71	0,692	0,089	60,725		0,000	1,999
EJ10	0,096	0,066	2,131		0,144	1,101
EH5	0,763	0,163	21,978		0,000	2,145
EJ6	0,053	0,037	2,032		0,154	1,055
QQl	-0,119	0,041	8,264		0,004	0,888
Constant	-2,971	0,319	86,804		0,000	0,051

Classification Table

		Predicted		Percentage Correct
		В течение последних 30 дней вы употребляли алкогольные напитки?		
Observed		да	нет	
В течение последних 30 дней вы употребляли алкогольные напитки?	да	2262	175	92,8
	нет	1031	166	13,9
Overall Percentage				66,8

Следует обратить внимание, что зависимая переменная здесь должна быть дихотомической, и ее максимальный код считается ко-

дом события, вероятность которого прогнозируется. Поскольку переменная ет80 закодирована: 1 — употреблял, 2 — не употреблял, то будет прогнозироваться вероятность неупотребления алкоголя.

Результаты работы команды логистической регрессии представлены в виде нескольких таблиц. Первая из них содержит общую оценку качества построенной модели *(Model Summary)*.

• -2 Log Likelihood — удвоенный логарифм функции правдоподобия со знаком минус;

• Cox & Snell R^A2 и Nagelkerke R^A2 — псевдокоэффициенты детерминации, полученные на основе отношения функций правдоподобия моделей лишь с константой и всеми коэффициентами.

Эти коэффициенты следует использовать при сравнении очень похожих моделей на аналогичных данных, что практически нереально, поэтому мы не будем на них останавливаться.

Следующая таблица — *Classification Table,* или таблица правильного предсказания. По ней мы видим, что для 2262 человек, потребляющих алкоголь, а также для 166 респондентов, не потребляющих алкоголь, модель правильно предсказывает этот факт. Таким образом, для 2428 респондентов модель правильно предсказывает потребление — непотребление алкоголя. Это число составляет 66,8% общего числа анализируемых респондентов и может рассматриваться как еще одна характеристика качества построенной модели.

При этом *Classification Table* показывает не только общее качество предсказания модели, но и качество предсказания отдельных градаций зависимой переменной. Так, из таблицы видно, что модель правильно предсказывает потребление алкоголя в 92,8% случаев, а непотребление алкоголя лишь в 13,9% случаев.

На основе модели логистической регрессии можно строить предсказание, произойдет или не произойдет событие $\{Y = 1\}$. Правило предсказания, по умолчанию заложенное в процедуру *Logistic Regression,* устроено по следующему принципу: если $P = P\{Y = 1\ Xj,..., Xj\} > 0{,}5$, считаем, что событие произойдет; $Pj = P\{Y = 1\ Xj,..., Xj\} < 0{,}5$, считаем, что событие не произойдет. Это правило оптимально с точки зрения минимизации числа ошибок, но очень грубо с точки зрения исследования связи. Зачастую вероятность со-

бытия $P\{Y = 1\}$ мала (значительно меньше 0,5) или велика (значительно больше 0,5), поэтому получается, что все имеющиеся в данных сочетания Хпред сказывают событие или все предсказывают противоположное событие.

Классификационная таблица показывает, насколько правильно наша модель предсказывает, потребляет ли респондент алкоголь на основе предложенных независимых переменных.

Коэффициенты регрессии. Основная информация, как и должно быть в ситуации регрессионной модели, содержится в таблице коэффициентов регрессии. Прежде всего следует обратить внимание на значимость коэффициентов регрессии. Наблюдаемая значимость вычисляется на основе статистики Вальда, которая связана с методом максимального правдоподобия и может быть использована при оценках различных параметров.

Универсальность статистики Вальда позволяет оценить значимость не только отдельных переменных, но и в целом значимость категориальных переменных, несмотря на то что они дезагрегированы на индексные переменные. Статистика Вальда имеет распределение $\%^2$. Число степеней свободы равно единице, если проверяется гипотеза о равенстве нулю коэффициента при обычной или индексной переменной и, для категориальной переменной, равно числу ее значений без единицы (числу соответствующих индексных переменных). Квадратный корень из статистики Вальда приближенно равен отношению величины коэффициента к его стандартной ошибке — так же выражается f-статистика в обычной линейной модели регрессии.

В таблице коэффициентов (см. табл. 4.21) почти все переменные значимы на уровне 5%. Закрыв глаза на возможное взаимодействие между независимыми переменными (коллинеарность), можно считать, что курение и принадлежность к мужскому полу повышают вероятность употребления алкоголя.

Кроме того, в табл. 4.21 представлены значения экспонент коэффициентов, e^{β}.

Согласно модели и полученным значениям коэффициентов, при фиксированных прочих переменных, принадлежность к мужскому полу увеличивает отношение шансов употребления и неупотребления ал-

коголя в 2,15 раза, курения — в 2 раза, а прибавка к зарплате 100 руб. — на 10%, правда, такая прибавка мужчине одновременно уменьшает это отношение на 11%. Быть начальником — значит увеличить отношение шансов на 5%.

О статистике Вальда. Недостаток статистики Вальда состоит в том, что при малом числе наблюдений она может давать заниженные оценки наблюдаемой значимости коэффициентов. Для получения более точной информации о значимости переменных можно воспользоваться пошаговой регрессией, метод FORWARD LR (LR — likelihood ratio — отношение правдоподобия), тогда для каждой переменной будет выдана значимость включения/исключения, полученная на основе отношения функций правдоподобия модели. Поскольку основная выдача построена на статистике Вальда, первые выводы удобнее делать на ее основе, а потом уже уточнять результаты, если это необходимо.

Сохранение переменных. Программа позволяет сохранить множество показателей, среди которых наиболее полезным является, по всей видимости, предсказанная вероятность. Вызов возможности сохранения характеристик, вычисляемых командой логистической регрессии, осуществляется с помощью клавиши Save... в главном меню этой команды (см. рис. 4.21).

5 глава

ИССЛЕДОВАНИЕ СТРУКТУРЫ ДАННЫХ

Собирая данные, исследователь руководствуется определенными гипотезами. Полученная в ходе исследования информация относится к избранным предмету и теме исследования, но нередко она представляет собой сырой материал, в котором нужно изучить структуру показателей, характеризующих объекты, а также выявить однородные группы объектов. Информацию лучше представить в геометрическом пространстве, лаконично отразить ее особенности в классификации объектов и переменных. Такая работа создает предпосылки к выявлению типологий объектов и формулированию «социального пространства», в котором обозначены расстояния между объектами наблюдения, позволяет наглядно представить свойства объектов.

5.1

Факторный анализ

Социологический смысл модели факторного анализа состоит в том, что измеряемые эмпирические показатели, переменные считаются следствием других, глубинных, скрытых от непосредственного измерения характеристик — латентных переменных. Например, если мы фиксируем степень доверия респондента к различным государствен-

ным институтам, то вполне логично предположить, что нет отдельных «доверий» к Государственной Думе, Совету Федерации, Счетной палате и т.п. Скорее у респондента есть общее отношение к институтам центральной власти, которое и определяет, как респондент отвечает на отдельные вопросы по доверию к каждому отдельному институту.

При этом важно, что это общее, единое отношение к государственным институтам, формируя отношение к каждому из них, не определяет отношения к отдельному институту на 100%. Таким образом, ответ респондента на вопрос о том, насколько он доверяет какому-то конкретному государственному институту, находится под влиянием двух составляющих: общего фактора отношения к государственным институтам и отдельного отношения именно к данному конкретному институту.

Схематично модель факторного анализа можно представить следующим образом (рис. 5.1).

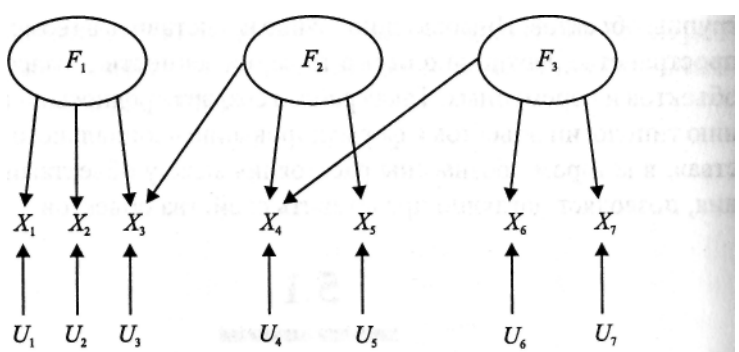


Рис. 5.1. Условное представление модели факторного анализа: *F, F,* F_3 — *общие факторы,* каждый из которых влияет на определенную совокупность переменных; $X_1, X_2, \ldots, X_7$ — переменные, формируемые на основании ответов опрашиваемых; *U, Щ ..., Щ —уникальные факторы,* каждый из которых влияет только на одну переменную

Уравнение факторного анализа имеет вид

$$X_i = \sum_{k=1} a_{ik} F_k + U_i, \qquad (5.1)$$

где a_{ik} — факторные нагрузки.

Обычно (хотя и не всегда) предполагается, что *X.* стандартизованы, а факторы $F_1 F_2, \ldots, F_m$ независимы и не связаны со специфическими факторами *U* (хотя существуют модели, выполненные в других предположениях). Предполагается также, что факторы *F.* стандартизованы.

В этих условиях факторные нагрузки $a._к$ совпадают с коэффициентами корреляции между общими факторами и переменными *X.*. Дисперсия *X.* раскладывается на сумму квадратов факторных нагрузок и дисперсию специфического фактора:

$$S_{x_i}^2 = H_i^2 + S_{u_i}^2, \qquad (5.2)$$

где

$$H_i^2 = \sum_k a_{ik}^2. \qquad (5.3)$$

Величина Я,^{2} называется общностью, S,^{2} — специфичностью. Другими словами, общность — это часть дисперсии переменных, объясненная общими факторами, специфичность — часть не объясненной общими факторами дисперсии.

В соответствии с постановкой задачи необходимо искать такие факторы, при которых суммарная общность максимальна, а специфичность — минимальна.

Метод главных компонент. Один из наиболее распространенных методов поиска факторов, метод главных компонент, заключается в последовательном поиске факторов. Вначале определяется первый фактор, который объясняет наибольшую часть дисперсии, затем независимый от него второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т.д. Математическая реализация метода главных компонент достаточно сложна, поэтому для пояснения идеи метода используем условное изображение (рис. 5.2).

Смысл данной схемы в следующем. Для построения первого фактора берется прямая, проходящая через начало координат и облако

рассеяния данных. Объектам можно сопоставить расстояния от их проекций на эту прямую до центра координат, причем для одной из половин прямой (по отношению к нулевой точке) можно взять эти расстояния с отрицательным знаком. Такое построение представляет собой новую переменную, которую назовем осью. При построении фактора находится такая ось, чтобы дисперсия переменных вокруг оси была минимальна. (Заметим, что в определенном смысле эта первая ось строится по той же модели, что регрессионная прямая в регрессионном анализе.) Это означает, что эта ось объясняет максимум дисперсии переменных. Найденная ось после нормировки используется в качестве первого фактора. Если облако данных вытянуто в виде эллипсоида, фактор совпадет с направлением, в котором вытянуты объекты, и по нему (по проекциям) с наибольшей точностью можно предсказать значения исходных переменных.

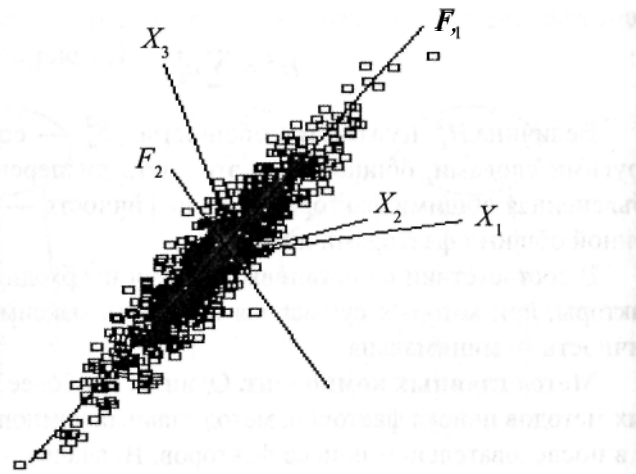


Рис. 5.2. Условное представление модели главных компонент:
X_3 *X2 X3* — наблюдаемые переменные; F_1 F_2 — факторы

Для поиска второго фактора строится ось, перпендикулярная первому фактору, также объясняющая наибольшую часть дисперсии, не объясненной первой осью. После нормировки ось становится вторым фактором. Если данные представляют собой плоский эллипсоид в трехмерном пространстве, два фактора позволяют точно описать эти данные.

Максимально возможное число главных компонент равно количеству переменных. Иными словами, если мы хотим на 100% описать значения *m* переменных, то для этого потребуется столько же, т.е. *m* главных компонент. Сколько главных компонент необходимо построить для *оптимального* представления рассматриваемых исходных переменных?

Обозначим через $X_к$ объясненную главной компонентой F_k часть суммарной дисперсии совокупности исходных переменных. По умолчанию, в пакете SPSS предусмотрено продолжать строить факторы, пока $K_к > 1$. Напомним, что переменные стандартизованы, и поэтому нет смысла строить очередной фактор, если он объясняет часть дисперсии, меньшую, чем приходящуюся непосредственно на одну переменную. При этом следует учесть, что $\lambda_1 > \lambda_2 > \lambda_3$..

Заметим, что значения λ_k — это собственные значения корреляционной матрицы переменных *X.,* поэтому в выдаче пакета SPSS они будут помечены текстом Eigen Value (собственные значения)[1].

Интерпретация факторов. Как же можно понять смысл того, что скрыто в найденных факторах? Основной информацией, которую использует для этого исследователь, являются факторные нагрузки. Для интерпретации необходимо приписать каждому фактору какой-то термин, понятие. Этот термин появляется на основе анализа корреляций фактора с исходными переменными. Например, если при анализе успеваемости школьников фактор имеет высокую положительную корреляцию с оценкой по алгебре, геометрии и большую отрицательную корреляцию с оценками по рисованию, можно предположить, что этот фактор характеризует точное мышление.

Не всегда такая интерпретация возможна. Для повышения интерпретируемости факторов добиваются большей контрастности матрицы факторных нагрузок. Такое улучшение результата называется

[1] Подробнее см.: Ростовцев П.С., Ковалева Г.Д. Анализ социологических данных с применением статистического пакета SPSS.

методом *вращения факторов*. Его суть состоит в следующем. Если вращать координатные оси, образуемые факторами, мы не потеряем в точности представления данных через новые оси, и при этом факторы не будут упорядочены по величине объясненной ими дисперсии, зато появляется возможность получить более контрастные факторные нагрузки. Вращение состоит в получении новых факторов — в виде специального вида линейной комбинации имеющихся факторов:

$$\hat{F}_i = \sum_{k=1}^{m} b_{ik} F_k. \qquad (5.4)$$

Чтобы не вводить новые обозначения, факторы и факторные нагрузки, полученные вращением, будем обозначать теми же символами, что и до вращения. Для достижения хорошей интерпретируемости существует достаточно много методов, которые состоят в оптимизации подходящей функции от факторных нагрузок. Рассмотрим реализуемый пакетом SPSS метод варимакс. Этот метод состоит в максимизации дисперсии квадратов факторных нагрузок для переменных:

$$\sum_i \left[\sum_k a_{ik}^4 / m - \left(\sum_k a_{ik}^2 / m \right)^2 \right] \to \max.$$

Чем сильнее разойдутся квадраты факторных нагрузок к концам отрезка [0,1], тем больше будет значение целевой функции вращения, тем четче интерпретация факторов.

Следует иметь в виду, что интерпретация полученных факторов в значительной степени связана с представлениями исследователя о характере изучаемого явления. По сути дела в процесс интерпретации включается большой объем информации, которая не связана с анализом собранных данных. В результате глубинное понимание смысла получаемых факторов может быть отнесено скорее к методам качественного, а не количественного исследования.

Индивидуальные значения факторов. Математический аппарат, используемый в факторном анализе, в действительности позволяет не вычислять непосредственно главные оси. И факторные нагрузки до и после вращения факторов, и общности вычисляются за счет операций с корреляционной матрицей. Поэтому оценка значений факторов для объектов является одной из проблем факторного анализа.

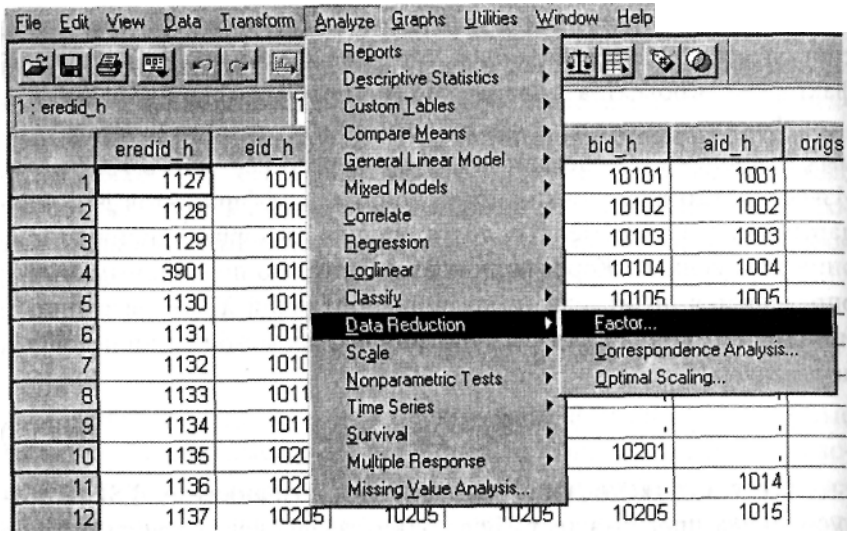
Факторы, имеющие свойства полученных с помощью метода главных компонент, определяются на основе регрессионного уравнения. Известно, что для оценки регрессионных коэффициентов для стандартизованных переменных достаточно знать корреляционную матрицу переменных. Корреляционная матрица по переменным X_i и F_k определяется по модели и матрице корреляций X_i. Регрессионным методом находятся факторы в виде линейных комбинаций исходных переменных:

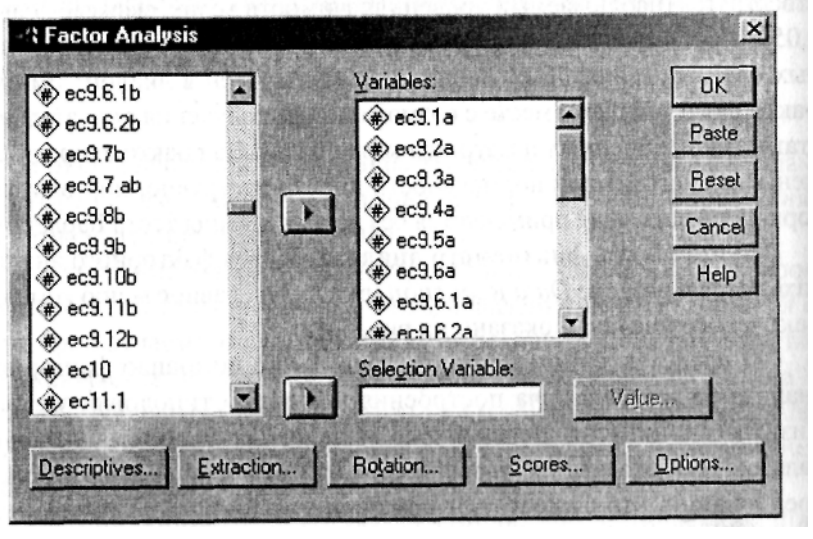
$$F_k = \sum_i c_{ki} X_i. \qquad (5.5)$$

Статистические гипотезы в факторном анализе. В SPSS предусмотрена проверка теста Барлетта о сферичности распределения данных. В предположении многомерной нормальности распределения проверяется, не диагональна ли матрица корреляций. Если гипотеза не отвергается (наблюдаемый уровень значимости велик, скажем, больше 0,05) — нет смысла в факторном анализе, поскольку направления главных осей случайны. Тест Барлетта предусмотрен в диалоговом окне факторного анализа, вместе с возможностью получения описательных статистик переменных и матрицы корреляций. На практике предположение о многомерной нормальности проверить трудно, поэтому факторный анализ чаще применяется без использования теста Барлетта.

Выполнение факторного анализа. Метод факторного анализа находится в *разделе Data Reduction* (рис. 5.3). Главное меню команды факторного анализа показано на рис. 5.4.

В представленном меню (см. рис. 5.4) с помощью факторного анализа решается задача построения некоторой типологии условий жизни респондентов. Включенные в анализ переменные фиксируют наличие или отсутствие у респондентов предметов. Данная модель предполагает, что существуют некоторые глубинные факторы, которые проявляются в указанных переменных.





	Initial	Extraction
У вас есть:		
холодильник	1,000	0,591
отдельная морозильная камера	1,000	0,467
стиральная машина	1,000	0,584
черно-белый телевизор	1,000	0,748
цветной телевизор	1,000	0,754
видеомагнитофон или видеоплеер	1,000	0,547
фен	1,000	0,481
компьютер	1,000	0,374
легковой автомобиль	1,000	0,416
грузовой автомобиль	1,000	0,469
мотоцикл, мотороллер, моторная лодка	1,000	0,363
трактор или мини-трактор	1,000	0,531
садовый домик	1,000	0,568
дача или другой дом	1,000	0,303
другая квартира или часть квартиры	1,000	0,186

Compo-nent	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumula-tive, %	Total	% of Variance	Cumula-tive, %
1	2,776	18,504	18,504	2,776	18,504	18,504
2	1,273	8,489	26,992	1,273	8,489	26,992
3	1,208	8,052	35,045	1,208	8,052	35,045
4	1,105	7,364	42,409	1,105	7,364	42,409
5	1,018	6,789	49,197	1,018	6,789	49,197
6	0,988	6,584	55,782			

Окончание табл. 5.1

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative, %	Total	% of Variance	Cumulative, %
7	0,935	6,233	62,015			
8	0,913	6,089	68,103			
9	0,883	5,885	73,989			
10	0,820	5,468	79,457			
11	0,808	5,388	84,845			
12	0,739	4,926	89,771			
13	0,654	4,357	94,129			
14	0,507	3,383	97,511			
15	0,373	2,489	100,000			

Component Matrix

	Component				
	1	2	3	4	5
У вас есть:					
холодильник	0,464	-0,0076	-0,410	0,406	0,206
отдельная морозильная камера	0,283	0,081	0,229	0,218	-0,529
стиральная машина	0,540	0,069	-0,305	0,404	0,177
черно-белый телевизор	-0,435	0,301	0,344	0,566	0,170
цветной телевизор	0,739	-0,156	-0,394	-0,161	-0,043
видеомагнитофон или видеоплеер	0,684	-0,035	0,205	-0,177	0,071
фен	0,602	-0,081	0,275	-0,119	0,149
компьютер	0,340	-0,150	0,475	-0,075	0,072
легковой автомобиль	0,545	0,166	0,272	0,049	-0,121
грузовой автомобиль	0,123	0,592	-0,056	-0,309	-0,072
мотоцикл, мотороллер, моторная лодка	0,088	0,533	-0,169	0,115	0,171
трактор или мини-трактор	0,127	0,671	-0,019	-0,234	-0,099
садовый домик	0,243	-0,021	0,021	0,351	-0,620
дача или другой дом	0,298	0,06	0,314	0,206	0,263
другая квартира или часть квартиры	0,155	0,030	0,270	-0,009	0,297

Результаты факторного анализа выводятся в виде трех таблиц. Первая — таблица *Communalities* демонстрирует, какую часть дисперсии каждой из включенных в анализ переменных объясняет предлагаемая факторная модель. Таблица показывает, что, скажем, переменная, фиксирующая наличие у респондента телевизора, объясняется моделью приблизительно на 75%. В то же время переменная, фиксирующая наличие у респондента другой квартиры, объясняется лишь на 18,6%. По всей видимости, эту переменную следовало бы исключить из анализа, поскольку она плохо объясняется построенной моделью.

Таблица *Total Variance Explained* содержит информацию о дисперсии, объясненной моделью. Из таблицы видно, что первая главная компонента объясняет 18,5% общей дисперсии, вторая — 8,5% и т.д. В представленной модели было отобрано пять главных компонент (факторов), которые в совокупности объясняют 49,2% общей дисперсии.

Таблица *Component Matrix* называется *матрицей факторных нагрузок* и служит для интерпретации полученных факторов. В рассматриваемом примере первый фактор имеет высокие корреляции с наличием у респондента следующих предметов: цветной телевизор, стиральная машина, видеомагнитофон, фен, легковой автомобиль. Другими словами, можно сказать, что этот фактор — характеристика современной, городской, достаточно обеспеченной семьи. А вот второй фактор будет выражен скорее у сельской семьи, поскольку имеет высокие факторные нагрузки с переменными: наличие грузового автомобиля, мотоцикла, трактора.

Поскольку более удобную матрицу факторных нагрузок дают методы вращения факторов, рассмотрим ту же факторную матрицу, но уже после вращения (табл. 5.2). Само вращение факторной матрицы можно выполнить, используя клавишу Rotation..., расположенную в главном меню команды факторного анализа (см. рис. 5.4).

В отличие от матрицы факторных нагрузок до вращения, матрица после вращения заметно удобнее — в ней почти все факторные нагрузки либо большие, либо маленькие, и, следовательно, такая матрица проще для интерпретации.

Таблица 5.2. Матрица факторных нагрузок после вращения

Rotated Component Matrix

	Component				
	1	2	3	4	5
У вас есть:					
холодильник	0,046	0,102	0,759	-0,035	0,032
отдельная морозильная камера	0,130	0,0012	-0,012	0,053	0,668
стиральная машина	0,155	0,081	0,735	0,040	0,105
черно-белый телевизор	0,0017	-0,864	0,010	0,010	0,020
цветной телевизор	0,169	0,706	0,460	0,046	0,111
видеомагнитофон или видеоплеер	0,596	0,383	0,144	0,096	0,119
фен	0,622	0,274	0,121	0,014	0,061
компьютер	0,575	0,069	-0,122	-0,115	0,101
легковой автомобиль	0,473	0,112	0,129	0,199	0,351
грузовой автомобиль	0,010	0,102	-0,068	0,674	0,0041
мотоцикл, мотороллер, моторная лодка	-0,023	-0,176	0,298	0,485	-0,081
трактор или мини-трактор	0,023	0,020	-0,043	0,723	0,067
садовый домик	-0,077	0,031	0,123	-0,065	0,736
дача или другой дом	0,485	-0,179	0,188	-0,0028	0,0095
другая квартира или часть квартиры	0,391	-0,095	0,020	0,015	-0,154

Проблема определения числа факторов. Как уже отмечалось, полное описание дисперсии исходных признаков возможно только в ситуации, когда число факторов равно числу исходных признаков. Основная направленность факторного анализа — это именно сокращение числа показателей, и, следовательно, мы идем на то, что полученные факторы не будут на 100% объяснять исходную информацию, и то, сколько же именно процентов будет объяснено, зависит от того, какое число факторов будет получено. Матрица объясненной дисперсии (см. табл. 5.1) показывает, что если взять три фактора, они объяснят примерно 35% исходной информации, а 8 факторов — уже около 68% информации. Какой процент является приемлемым, на каком числе

факторов остановиться? Точного ответа на этот вопрос нет, однако есть несколько подходов, дающих основания для решения этой проблемы.

Первый подход — формально-статистический. Есть определенные математические основания, говорящие, что целесообразно отбирать столько факторов, сколько существует собственных чисел корреляционной матрицы, больше единицы. Данный критерий называется критерием Кайзера. Таблица объясненной дисперсии (см. табл. 5.1) показывает, что в нашем примере таких чисел пять и потому в данной модели было отобрано именно пять факторов. Отметим, что критерий Кайзера по отбору числа факторов в команде факторного анализа SPSS используется по умолчанию.

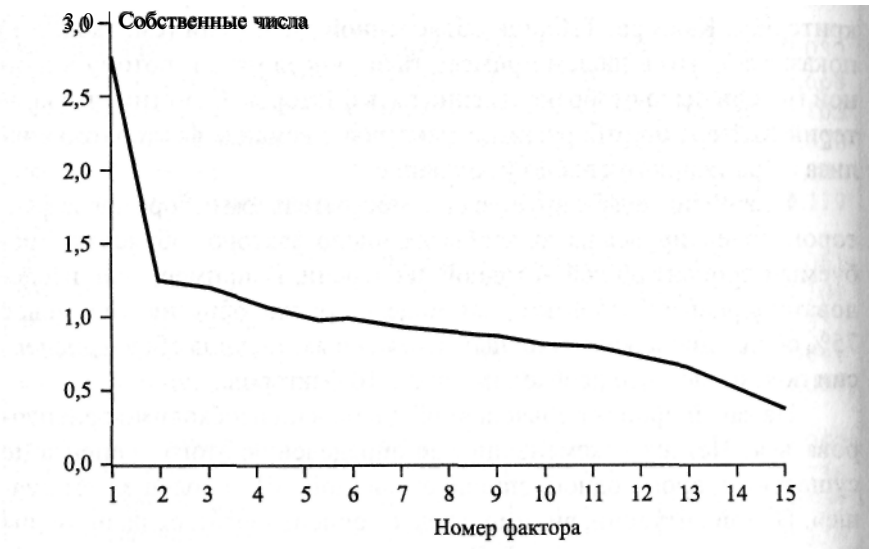
Второй подход базируется на самостоятельном отборе числа факторов, ориентируясь на то, чтобы это число факторов объясняло требуемый процент общей исходной дисперсии. Например, если исследователь решает, что факторная модель должна объяснять не менее 75% общей дисперсии исходных переменных, таблица общей дисперсии показывает, что необходимо взять 10 факторов.

На какой процент объясненной дисперсии необходимо ориентироваться? Четких рекомендаций по определению этого процента не существует, кроме одной, вполне очевидной: «Чем больше, тем лучше». В этой ситуации, видимо, следует ориентироваться на примеры предыдущих исследователей. В социологии, как правило, встречаются факторные модели, в которых объясняется 60—75% дисперсии, хотя можно привести примеры и с большими, и с меньшими процентами.

Есть еще один подход, который базируется на методе так называемой каменной осыпи. Суть метода в следующем. Строится график, в котором по оси абсцисс откладываются номера факторов, а по оси ординат — значения собственных чисел для каждого из факторов. Пример такого графика для модели табл. 5.1 показан на рис. 5.5. Как говорилось вначале, все собственные числа в методе главных компонент вычисляются в порядке убывания, поэтому график будет представлять собой понижающуюся кривую.

Далее на этом графике определяют точки, в которых происходит более или менее резкое понижение. В нашем примере (см. рис. 5.5) можно сказать, что действительно резких понижений нет. Хоть сколь-

нибудь резкое понижение происходит от 9-го к 10-му фактору. Рекомендация метода «каменной осыпи» состоит в том, что надо отобрать столько факторов, сколько точек на графике расположено до момента такого рода резкого понижения, т.е. в нашем примере лучше брать 9 факторов, а не 10.



Номер фактора

Рис. 5.5. График каменной осыпи для модели табл. 5.1

Важно понимать, что ни один из изложенных подходов к определению числа факторов не дает нам доказательных оснований по отбору числа факторов. У исследователя остается большая свобода в решении этого вопроса. Основным критерием является максимальное удобство в построении наиболее правдоподобной модели, что, естественно, ни в каком смысле не может считаться строгим основанием.

Определение числа факторов происходит в меню Extraction, вызов которого осуществляется нажатием соответствующей клавиши в главном меню команды факторного анализа (см. рис. 5.1). В меню Extraction (рис. 5.6) также находится окно, выбрав которое, можно получить график «каменной осыпи» (окно *Scree plot).*

В части меню Extract мы определяем, как будет проводиться выбор числа факторов через значения собственных чисел *(Eigenvalues over* — собственные числа больше, чем…), или через непосредственное указание требуемого числа факторов *(Number of factors).* В любом случае мы должны указать точное значение либо собственных чисел, либо числа факторов, что будет основанием для отбора числа факторов в модели.

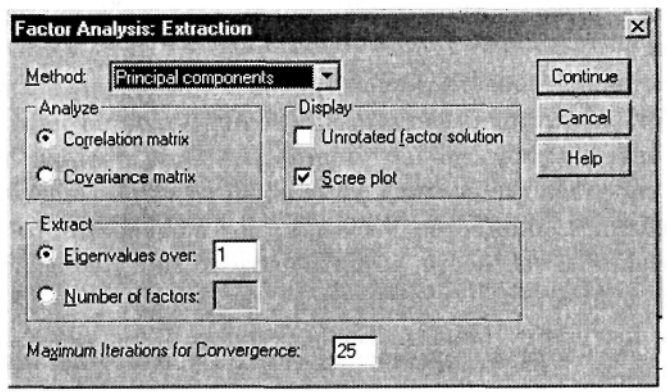


Рис. 5.6. Меню Extraction команды факторного анализа

Уровень измерения переменных, используемых в факторном анализе. Поскольку исходной информацией для метода факторного анализа является матрица коэффициентов корреляции Пирсона, это автоматически диктует нам возможность использования в данном методе переменных, измеренных только по количественным (интервальным либо абсолютным) шкалам, либо дихотомических переменных.

5.2

Кластерный анализ

Если процедура факторного анализа сжимает матрицу признаков в матрицу с меньшим числом переменных, кластерный анализ дает нам

группы единиц анализа, иначе — выполняет классификацию объектов. Иными словами, если в факторном анализе мы группируем столбцы матрицы данных, в кластерном анализе группируются строки. Синонимами термина «кластерный анализ» являются «автоматическая классификация объектов без учителя» и «таксономия».

Если данные понимать как точки в признаковом пространстве, задача кластерного анализа формулируется как выделение «сгущений точек», разбиение совокупности на однородные подмножества объектов.

При проведении кластерного анализа обычно определяют различные типы расстояний на множестве объектов; алгоритмы кластерного анализа формулируют в терминах этих расстояний. Мер близости и способов вычисления расстояний между объектами существует великое множество, их выбирают в зависимости от цели исследования. В частности, евклидово расстояние лучше использовать для количественных переменных, расстояние $\%^2$ — для исследования частотных таблиц, имеются также меры для бинарных переменных.

5.2.1

Иерархический кластерный анализ

Процедура иерархического кластерного анализа в SPSS предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют переменные.

Этот метод реализует иерархический агломеративный алгоритм. Его смысл заключается в следующем. Перед началом кластеризации все объекты считаются отдельными кластерами, которые в ходе алгоритма объединяются. Вначале берется N объектов и между ними попарно вычисляются расстояния. Далее выбирается пара объектов, ко-

торые расположены наиболее близко друг от друга, и эти объекты объединяются в один кластер. В результате количество кластеров становится равным N - \. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования определяют способы вычисления расстояния между объектами и определения близости между кластерами.

Для определения расстояния между парой кластеров могут использоваться разные подходы. В SPSS предусмотрены следующие методы, определяемые на основе расстояний между объектами.

- Среднее расстояние между кластерами (Between-groups linkage).
- Среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage).
- Расстояние между ближайшими соседями — ближайшими объектами кластеров (Nearest neighbor).
- Расстояние между самыми далекими соседями (Furthest neighbor).
- Расстояние между центрами кластеров (Centroid clustering), или центроидный метод. Недостатком этого метода является то, что центр объединенного кластера вычисляется как среднее центров объединяемых кластеров, без учета их объема.
- Метод медиан — тот же центроидный метод, но центр объединенного кластера вычисляется как среднее всех объектов (Median clustering).
- Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

Расстояния и меры близости между объектами. У нас нет возможности сделать полный обзор всех коэффициентов, поэтому остановимся лишь на некоторых.

Пусть имеются два объекта $X = (X_x, \ldots, X)_m$ и $Y = №', \ldots, У)_m$. Используя эту запись, определим основные виды расстояний в процедуре кластерного анализа.

• Евклидово расстояние (Euclidian distance) —

$$d(X,Y)=\sqrt{\sum_{i=1}^{m}(X_i-Y_i)^2}.$$

• Квадрат евклидова расстояния (Squared Euclidian distance) -

$$d(X,Y)=\sum_{i=1}^{m}(X_i-Y_i)^2.$$

Евклидово расстояние и его квадрат разумно применять для ана лиза количественных данных.

• Мера близости — коэффициент корреляции

$S(X,Y)=(\sum_{i=1}^{m}Z_{X_i}Z_{Y_i})/(m-1)$, где Z_{X_i} и Z_{Y_i} — компоненты стандар тизованных векторов X и Y. Эту меру целесообразно использоват для выявления кластеров переменных, а не объектов.

Стандартизация. Непосредственное использование переменны в анализе может привести к тому, что классификацию будут опреде лять переменные, имеющие наибольший разброс значений. Поэтом применяются следующие виды стандартизации.

• Z-стандартизация *(Z-Scores)*. Из значений переменных вычита ется их среднее, и эти значения делятся на стандартное отклонение.

• Разброс от -1 до 1. Линейным преобразованием переменны добиваются разброса значений от -1 до 1.

• Разброс от 0 до 1. Линейным преобразованием переменны добиваются разброса значений от 0 до 1.

• Максимум 1. Значения переменных делятся на их максимум.

• Среднее 1. Значения переменных делятся на их среднее.

• Стандартное отклонение 1. Значения переменных делятся н стандартное отклонение.

• Возможны преобразования самих расстояний, в частности, можно расстояния заменить их абсолютными значениями, это актуально для коэффициентов корреляции. Можно также все расстояния преобразовать так, чтобы они изменялись от 0 до 1.

Выполнение иерархического кластерного анализа. На рис. 5.7 показано меню вызова команды иерархического кластерного анализа. Главное меню команды иерархического кластерного анализа представлено на рис. 5.8.

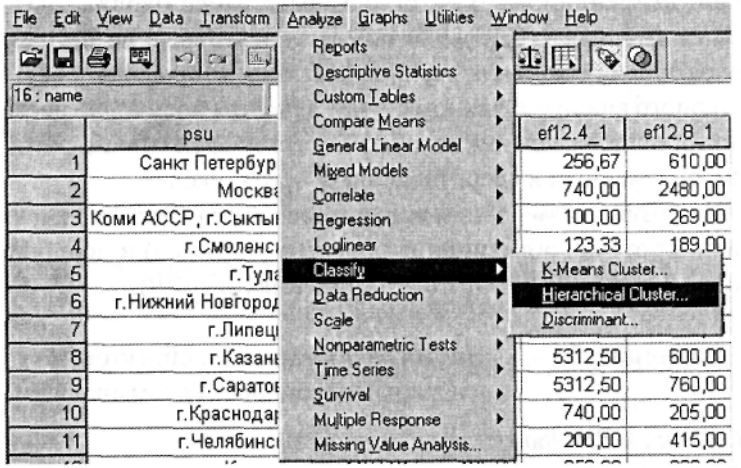


Рис. 5.7. Вызов команды иерархического кластерного анализа

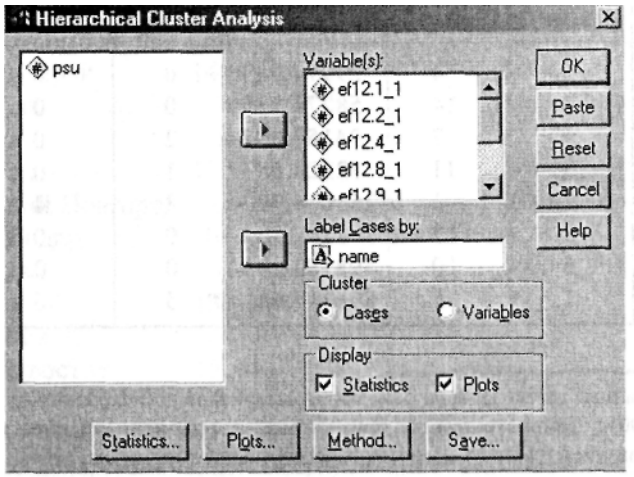


Рис. *5.8.* Главное меню команды иерархического кластерного анализа

Приведенный пример (см. рис. 5.8) решает задачу классификации единиц анализа, в качестве которых выступают несколько городов России. В каждом из городов респондентам задавали вопросы о размерах доходов их семей, полученных из различных источников: пенсий, стипендий, алиментов, возврата ранее одолженных денег, продажи имущества[2]. Далее были рассчитаны средние значения этих доходов среди респондентов, проживающих в городах опроса. Целью кластерного анализа в данном случае является получение нескольких групп городов, население которых достаточно сходно по размеру доходов, полученных из перечисленных источников.

По результатам работы иерархического кластерного анализа составили протокол объединения объектов (табл. 5.3) и дендрограмму, демонстрирующую ход этого объединения (рис. 5.9).

Таблица 5.3. Протокол объединения объектов в иерархическом кластерном анализе

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	29974,948	0	0	4
2	1	14	58278,238	0	0	3
3	1	7	94158,037	2	0	5
4	3	11	172229,687	1	0	5
5	1	3	263034,790	3	4	8
6	5	12	302187,863	0	0	8
7	6	10	498182,113	0	0	10
8	1	5	896117,681	5	6	10

[2] Данные вычислены на основании материалов исследования РМЭЗ, октябрь-ноябрь 2001 г. Демонстрируемый пример имеет формат иллюстрации и не может служить основанием для социологических рассуждений по вопросам структуры доходов в рассматриваемых населенных пунктах, поскольку исследование не содержит данных, репрезентирующих население данных городов.

Окончание табл. 5.3

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
9	9	15	915664,247	0	0	13
10	1	6	1204070,792	8	7	12
11	2	13	1792091,613	0	0	12
12	1	2	3643327,865	10	11	14
13	8	9	9363162,158	0	9	14
14	1	8	29635213,961	12	13	0

* * * HIERARCHICAL CLUSTER ANALYSIS * * *
Dendrogramm using Average Linkage (Between Groups)

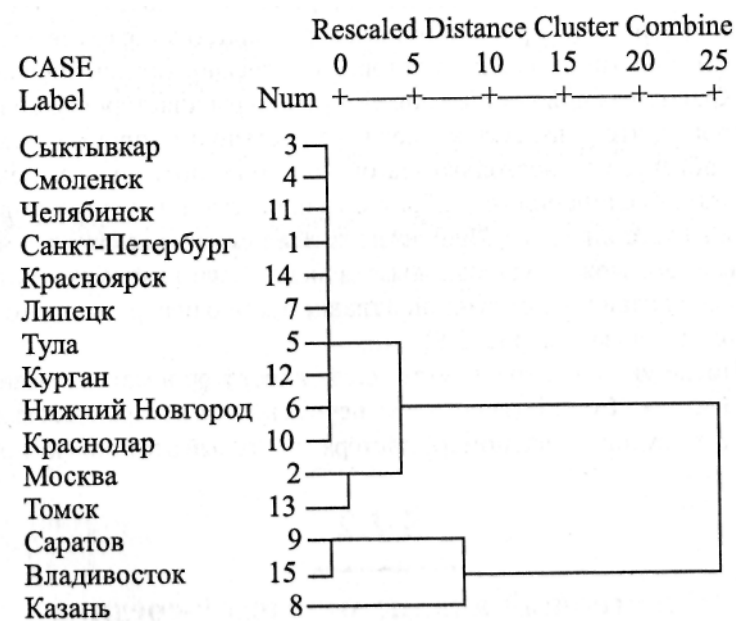


Рис. *5.9.* Дендрограмма, демонстрирующая объединение объектов в иерархическом кластерном анализе

Из табл. 5.3 видно, что, например, на первом шаге произошло объединение 3-го и 4-го объектов, поскольку между этими объектами было наименьшее расстояние (колонка *Coefficients)*. В колонке *Next Stage* (следующий этап) указывается, что в следующий раз тот кластер, который получен на первом шаге, будет задействован в объединении на четвертом шаге. Таким образом, когда на четвертом шаге указано, что одним из объединяемых объектов является объект номер 3, надо иметь в виду, что это не сам 3-й объект, а уже то, что получилось в результате объединения 3-го и 4-го объектов на первом шаге.

Процесс агрегирования данных может быть представлен графически деревом объединения кластеров *(Dendrogramm)*. Дендрограмма наглядно демонстрирует, что, например, объект «Казань» располагается достаточно далеко от других объектов и был объединен с парой объектов «Саратов — Владивосток» только на предпоследнем шаге.

На практике интерпретация кластеров требует достаточно серьезной работы, изучения разнообразных характеристик объектов для точного описания типов объектов, которые составляют тот или иной класс.

Крайне важной составляющей процедуры кластерного анализа является то, что у нас есть возможность остановить процесс объединения объектов за несколько шагов до конца, поскольку конечный результат объединения всех объектов в один кластер не представляет практического интереса. И если мы хотим получить, скажем, четыре кластера, это можно указать, вызвав меню Save нажатием соответствующей клавиши, показанной в главном меню иерархического кластерного анализа (см. рис. 5.8).

После указания требуемого числа кластеров в матрице данных автоматически будет создана новая переменная, в которой для каждого объекта будет указан номер кластера, в который этот объект попал.

5.2.2

Кластерный анализ методом Л-средних

Процедура иерархического кластерного анализа эффективна для малого числа объектов. Ее преимущество в том, что каждый объект можно, образно говоря, пощупать руками. Но эта процедура не годится для массивов большого объема из-за трудоемкости агломеративного алгоритма и слишком большого размера и практической бессмысленности дендрограмм.

В такой ситуации наиболее приемлем алгоритм, носящий название метода «^-средних». Он реализуется в пакете командой меню *k-means*.

Алгоритм заключается в следующем: выбирается заданное число *к* точек и на первом шаге эти точки рассматриваются как «центры» кластеров. Каждому кластеру соответствует один центр. Объекты распределяются по кластерам по принципу: каждый объект относится к кластеру с ближайшим к этому объекту центром. Таким образом, все объекты распределились по *к* кластерам.

Затем заново вычисляют центры этих кластеров, которыми после этого момента считаются покоординатные средние кластеров. После этого опять перераспределяют объекты. Вычисление центров и перераспределение объектов происходит до тех пор, пока центры не стабилизируются.

Рис. 5.10 демонстрирует главное меню команды *k-means*.

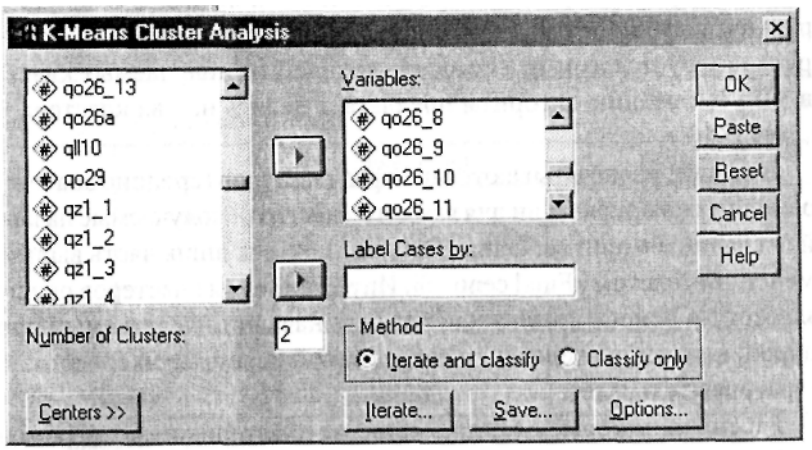


Рис. 5.10. Главное меню команды *k-means*

Часто переменные, используемые в кластеризации, имеют разный диапазон изменений, например рост и вес, килограммы и граммы. В этих условиях основное влияние на кластеризацию окажут переменные, имеющие большую дисперсию. Поэтому перед кластеризацией полезно стандартизовать переменные. К сожалению, в данной команде кластерного анализа средства стандартизации не предусмотрены, в отличие от процедуры иерархического кластерного анализа.

Часть переменных может иметь неопределенные значения, расстояния до центров рассчитывают по определенным значениям. Для использования такой возможности в меню Options следует выбрать параметр обработки пропущенных данных Pairwise.

Говоря о допустимом уровне измерения для переменных при кластеризации, необходимо помнить, что команда использует только евклидово расстояние. Следовательно, корректные результаты при применении данного метода можно ожидать только на основе метрических переменных.

Ключевым вопросом, который необходимо решить при подготовке к кластерному анализу, является вопрос о количестве получаемых кластеров. В силу специфики алгоритма метода *k-means,* в отличие от иерархического кластерного анализа, в данном случае в обязательном порядке требуется изначально задать количество получаемых кластеров. (По умолчанию алгоритм предлагает делить на два кластера — см. рис. 5.10.)

В выдаче распечатываются центры кластеров (средние значения переменных кластеризации для каждого кластера), получаемые на каждой итерации алгоритма. Однако для нас полезна лишь часть выдачи, помеченная текстом «Final centres». Интерпретация кластеров осуществляется на основе сравнения средних значений, выдаваемых процедурой, а также исследования сохраненной переменной средствами статистического пакета.

Рассмотрим пример, когда в качестве кластеризуемых переменных берутся переменные, фиксирующие наличие в семьях респондентов разных предметов длительного пользования[3]. Возьмем 4 кластера. Такая классификация может грубо, но наглядно показать различие семей по благосостоянию.

Таблица 5.4. Результаты работы команды кластерного анализа *k-means*

Final Cluster Centers

	Cluster			
	1	2	3	4
В семье есть:				
цветной телевизор	1	0	1	1
фотоаппарат	0,5	0,1	**ОД**	0,9
миксер	0,1	0,0	**ОД**	1,0
электродрель	0,8	0,0	0,0	0,6
отдельный морозильник	0,1	0,0	0,0	0,2
микроволновая печь	0,0	0,0	0,0	0,2
видеомагнитофон	0,6	0,0	0,1	0,8
видеокамера	0,0	0,0	0,0	ОД
пылесос	0,9	0,2	0,5	0,9
домашний компьютер	0,0	0,0	0,0	0,2
автомобиль	0,4	0,0	0,1	0,6
проигрыватель компакт-дисков	0,1	0,0	0,0	**0,3**

Number of Cases in each Cluster

		Unweighted	Weighted
Cluster	1	426	426
	2	398	398
	3	1073	1073
	4	510	510
Valid		2407	2407
Missing		0,000	0,000

[3] Данные исследования «Мониторинг экономических и социальных перемен». Проведено ВЦИОМ в мае 2001 г. по всероссийской репрезентативной выборке.

С помощью табл. 5.4 имеем следующую интерпретацию полученных кластеров. Поскольку кодировка используемых вопросов «1 — есть; 0 — нет», то мы можем сказать, что у 50% респондентов, попавших в кластер 1, есть фотоаппарат, у 40% — автомобиль и т.д.

Кластер 1 —респонденты из достаточно зажиточных семей, имеющие дома большинство из предлагаемых предметов длительного пользования.

Кластер 2 — респонденты из наиболее бедных семей, у которых нет практически ничего из предметов длительного пользования.

Кластер 3 — респонденты из семей более зажиточных, чем в кластере 2, но обладающие лишь небольшим набором предметов.

Кластер 4 — респонденты из наиболее зажиточных семей, имеющие большинство из предлагаемых предметов длительного пользования.

Имеется масса возможностей изучить и сравнить полученные классы, используя сохраненную в виде переменной классификацию. Например, можно посмотреть, какая доля респондентов проживает в городах, а какая — в селах, каков средний доход респондентов в каждом из кластеров и т.п.

Принципиальным вопросом для понимания содержания полученных кластеров — групп респондентов является то, насколько действительно эти группы однородны. В меню Save команды *k-means* можно сохранять не только переменную, фиксирующую номер кластера, к которому отнесен респондент, но и переменную, измеряющую расстояние каждого респондента от центра «его» кластера. В табл. 5.5 представлены средние расстояния для разбиения, рассмотренного в табл. 5.4.

Таблица 5.5. Средние значения расстояний от центра для четырех кластеров табл. 5.4

Cluster Number of Case	Mean	*N*	Std. Deviation
1	1,1369504	426	0,24006403
2	0,4002193	398	0,39150538
3	0,8222608	1073	0,32131648
4	1,2606004	510	0,28033532
Total	0,9010882	2407	0,42375184

Данные табл. 5.5 показывают, что кластер 2 наиболее однородный, а кластеры 1 и 4 однородны, но в меньшей степени. По всей видимости, целесообразно провести другую кластеризацию, увеличив число кластеров. Это должно привести к разбиению кластеров 1 и 4 на более однородные группы.

Многомерное шкалирование

Многомерное шкалирование заключается в построении переменных на основе имеющихся расстояний между объектами. В частности, если даны расстояния между городами, программа многомерного шкалирования должна восстановить систему координат (с точностью до поворота и единицы длины) и приписать координаты каждому городу, так чтобы карта и изображение городов в этой системе координат зрительно совпали. Близость может определяться не только расстоянием в километрах, но и другими показателями, такими, как размеры миграционных потоков между городами, интенсивность телефонных звонков, а также расстояниями в многомерном признаковом пространстве. В последнем случае задача построения искомой системы координат близка к задаче, решаемой факторным анализом, — сжатию данных, описанию их небольшим числом переменных. Нередко важно наглядное представление свойств объектов: полезно придать координаты переменным, расположить в геометрическом пространстве переменные. С технической точки зрения это всего лишь транспонирование матрицы данных. Для определенности мы будем говорить о создании геометрического пространства для объектов, специально оговаривая случаи анализа множества их свойств. В социальных исследованиях методом многомерного шкалирования создают зрительный образ «социального пространства» объектов наблюдения или свойств. Для такого образа наиболее приемлемо создание двумерного пространства.

Основная идея метода состоит в приписывании каждому объекту значений координат, так чтобы матрица евклидовых расстояний между объектами в этих координатах оказалась близка к матрице расстояний между объектами, определенной из каких-либо соображений ранее.

Метод весьма трудоемок и рассчитан на анализ данных, имеющих небольшое число объектов.

Евклидово пространство. Пусть мы определили *r* шкал $X^r,\ldots,$ X^r. Расстояние между парой объектов *i* и/ определяется по формуле

$$d_{ij} = \sqrt{\sum_{k=1}^{r}(X_i^k - X_j^k)^2}.$$

Для однозначности задания шкал предполагается, что $\sum_{i}^{n} X_i^k = 0$ и $\sum_{i}^{n}\sum_{k}^{r}(X_i^k)^2 = nr$, где *n* — число объектов. Кроме того, по аналогии с методом главных компонент, первой шкалой обычно называется шкала с наибольшей дисперсией, вторая — имеет вторую наибольшую дисперсию и т.д.

Идея многомерного шкалирования. В многомерном шкалировании выделяются два направления: метрическое и неметрическое. Первая из предложенных моделей — модель метрического многомерного шкалирования — имеет вид

$$L\{S\} = D^2 + E, \qquad (5.6)$$

где *L {S}* — линейное преобразование исходной матрицы расстояний; D^2 — матрица квадратов расстояний, полученная на основе созданных шкал; £ — матрица отклонений модели от исходных данных.

Линейное преобразование дает матрицу преобразованных расстояний *T = L{S}.* Цель многомерного метрического шкалирования — поиск оптимальных шкал с помощью линейного преобразования матрицы исходных расстояний, минимизирующих ошибку *E.*

Шепард и Краскэл совершили существенный прорыв, разработав метод неметрического шкалирования. Суть этого метода состоит

в нелинейном преобразовании расстояний. Модель неметрического шкалирования имеет вид

$$M\{S\} = D^2 + E, \qquad (5.7)$$

где *M{S)* — монотонное преобразование исходной матрицы расстояний.

Монотонное преобразование дает матрицу преобразованных расстояний *T = M{S}.*

Качество подгонки модели. Для измерения качества подгонки модели был предложен показатель

$$S\text{-}stress = \left(\frac{\|E\|}{\|T\|}\right)^{1/2} \qquad (5.8)$$

где норма матрицы ММ означает сумму квадратов элементов матрицы. Слово «stress» в английском языке имеет множество значений, одно из них — нагрузка. Этот показатель изменяется от 0 до 1. Равенство нулю означает точную подгонку модели, единице — полную ее бессмысленность.

Кроме того, оценить качество модели можно с помощью показателя stress index Краскэла, который получается с использованием матрицы не квадратов расстояний, а расстояний. Заметим, что алгоритм оптимизирует S-stress, а не stress index.

Еще один показатель качества модели, RSQ, представляет квадрат коэффициента корреляции между матрицами T и *D.* Таким образом, так же как в регрессионном анализе, RSQ может быть интерпретирован как доля дисперсии преобразованных расстояний *T,* объясненная матрицей расстояний *D.*

Вызов процедуры многомерного шкалирования. На рис. 5.11 и 5.12 показаны пути вызова метода многомерного шкалирования и главное меню этой команды.

По умолчанию в процедуре проводится неметрическое шкалирование, кнопкой *Model* можно переключиться на метрическое шкалирование.

Исходная матрица расстояний. По умолчанию в процедуре предполагается, что исходная матрица расстояний вводится из файла SPSS. Но у исследователя подготовленная матрица расстояний быва-

ет весьма редко. Поэтому чаще используется возможность вычисления расстояний на основе имеющихся данных, которая реализуется в диалоговом окне команды в разделе *Distances* включением пункта *Create distances from data*. Здесь предусмотрен такой же широкий набор мер близости и расстояний, как и в иерархическом кластерном анализе. Их можно выбрать, воспользовавшись кнопкой *Measure* **в** разделе *Distances,* при этом можно определить, что визуализируется, матрица расстояний между объектами или переменными.

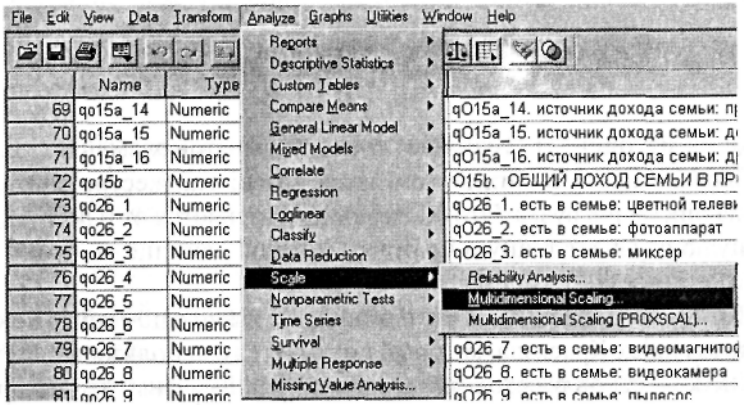


Рис. 5.11, Путь вызова команды многомерного шкалирования

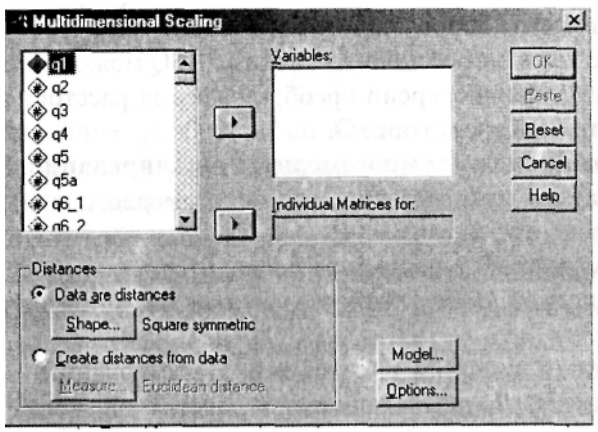


Рис. 5.12. Главное меню команды многомерного шкалирования

Пример построения шкал. В качестве примера исследуем данные по средней обеспеченности семей дорогостоящими предметами быта, электроникой, средствами транспорта и дачами (всего 9 предметов) в 38 территориальных общностях (данные RLMS, 1996). В результате применения процедуры шкалирования территориальные общности должны расположиться в двумерном геометрическом пространстве, построенном исходя из расстояний по 9 переменным.

Для этого получим файл, в котором объектами будут территориальные общности, а переменными — обеспеченность семей указанными предметами. Значения переменных — доли семей, обладающих ими. Исходными данными здесь являются ответы на вопрос: имеете ли вы холодильник; имеете ли вы морозильник; имеете ли вы стиральную машину и т.д. (1 — да, 2 — нет, 9 — нет ответа) в файле анкет семьи.

Интерпретация результатов многомерного шкалирования. Для интерпретации можно изучить связь полученных шкал с имеющимися данными, в частности с исходными переменными, по которым строилась матрица расстояний.

В нашем примере таблица ранговых корреляций с исходными переменными свидетельствует о том, что первое измерение (Diml) характеризует уровень благосостояния жителей территориальных образований в целом, второе измерение связано с приверженностью их садоводству (табл. 5.6).

Наглядную картину дает непосредственное размещение объектов (территориальных общностей) на поле рассеяния в построенном геометрическом пространстве (рис. 5.13). На графике видно, что шкала Diml имеет больший разброс, чем шкала Dim2, а значит, объясняет большую часть разброса расстояний объектов. Зримо подтверждается интерпретация первой шкалы 1: по разным полюсам Diml стоят Ханты-Мансийский автономный округ — весьма богатый регион и Пензенская область, Кабардино-Балкария — беднейшие части России.

Поскольку мы не обладаем информацией о развитии садоводства, для проверки интерпретации второй шкалы полезно рассмотреть диаграмму рассеяния Dim2 и доли семей, имеющих садовые домики (рис. 5.14). Рисунок показывает, что указанная выше интерпретация небезосновательна.

Таблица 5.6. Коэффициенты ранговой корреляции Спирмена, построенных шкал с обеспеченностью предметами быта

		Холодильник	Стиральная машина	Черно-белый телевизор	Цветной телевизор	Видеомагнитофон	Фен	Легковой автомобиль	Садовый домик	Дача или другой дом
Diml		0,844	0,265	-0,820	0,950	0,773	0,929	0,426	0,226	0,305
	Sig.	0,000	0,108	0,000	0,000	0,000	0,000	0,008	0,408	0,659
Dim2		-0,112	-0,156	-0,145	0,113	0,402	0,240	0,262	+0,687	0,532
	Sig-	0,502	0,350	0,385	0,501	0,212	0,148	0,112	0,000	0,004

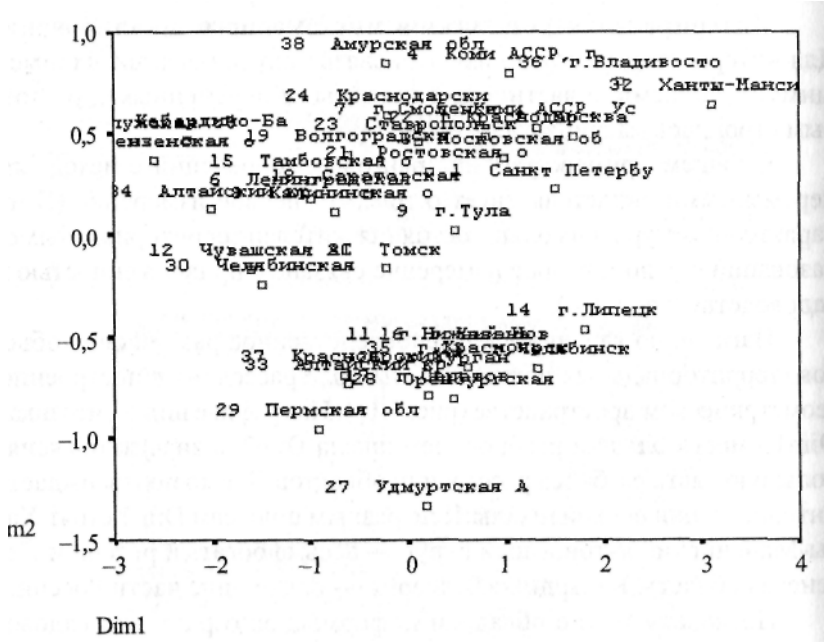


Рис. 5.13. Представление объектов в сконструированном геометрическом пространстве

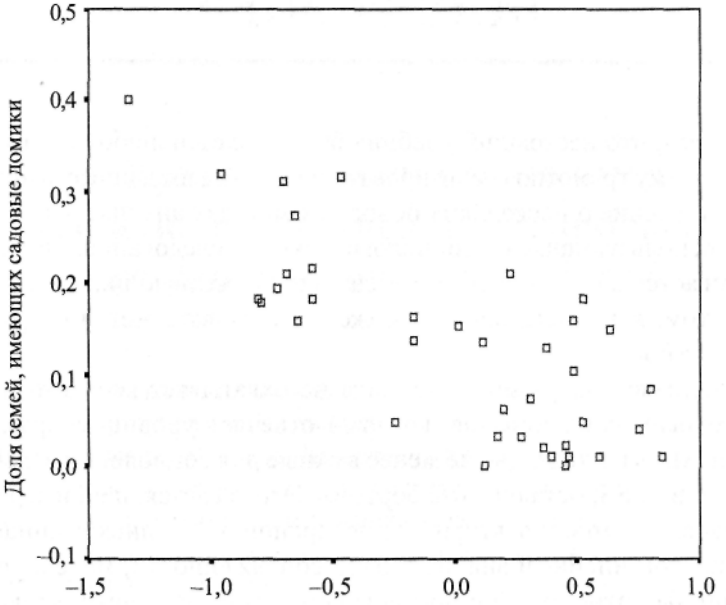


Рис. 5.14. Поле рассеяния второй шкалы, порожденной процедурой многомерного шкалирования, и доли семей, имеющих садовые домики

ПОСЛЕСЛОВИЕ

Надеемся, что настоящий учебник будет полезен любому социологу, желающему грамотно анализировать собранные им данные. Книга дает представление о нескольких основных методах анализа данных, широко использующихся в социологических исследованиях. В ней рассказывается о сути каждого метода, о соответствующих ограничениях, о том, как исследователь может реализовать метод с помощью пакета SPSS.

Однако содержание учебника не охватывает весь тот арсенал социологических методов, который отвечает уровню современной науки. Многие подходы, не менее важные для социолога и задействованные в SPSS, остались «за бортом». Это касается, например, таких методов, как многофакторный дисперсионный и дискриминантный анализ, логлинейный анализ таблиц сопряженности, моделирование причинных отношений с помощью систем структурных уравнений, методы анализа временных рядов, совместный анализ, анализ соответствий, методы поиска взаимодействий (например, система алгоритмов answer trees), технологии нейронных сетей.

Остались без рассмотрения многие подходы к анализу данных, не включенные в названный пакет: некоторые алгоритмы кластерного анализа, латентно-структурный анализ.

То, что ряд методов, важных для социолога, не описан в учебнике, естественно: книга соответствует курсу лекций, рассчитанному на три модуля (примерно 1,5 семестра). За такой срок вряд ли можно качественно преподнести студентам материал большего объема. Выражаем надежду, что дело, начатое А.О. Крыштановским, будет продолжено, и мы в не очень отдаленном будущем увидим книги, посвященные названным и другим методам, не затронутым в предлагаемом учебнике, но важным для социологии, книги, в которых наряду с теоретическим описанием методов будут присутствовать конкретные рекомендации по их реализации на компьютере.

ПРИЛОЖЕНИЯ

А.О. Крыштановский и его вклад в развитие отечественной социологии и высшего социологического образования

Летом 2005 г. ушел из жизни один из основателей и первый декан факультета социологии Государственного университета — Высшей школы экономики (ГУ ВШЭ) Александр Олегович Крыштановский (11 февраля 1955 — 2 августа 2005). Будучи специалистом по анализу социологических данных и методике социологических исследований, он являлся также одним из создателей и первым заведующим кафедрой методов сбора и анализа социологической информации (организована в 1999 г., одновременно с факультетом социологии).

Талантливый специалист и организатор, принципиальный и в то же время добрый и отзывчивый человек, один из самых любимых студентами профессор факультета, Александр Олегович оставил о себе добрую, светлую память в душах всех знавших его людей. Нам, сотрудникам руководимого им факультета, хотелось бы, чтобы эта память сохранялась как можно дольше. И в первую очередь, в сознании молодежи. Именно молодому поколению в первую очередь адресуется эта книга.

А.О. Крыштановский родился 11 февраля 1955 г. в Саратове. В 1978 г. окончил Московский институт электронного машиностроения, факультет прикладной математики.

Первая часть трудового пути А.О. Крыштановского была тесно связана с Институтом социологии АН СССР (РАН). Организованный в 1969 г., институт долгое время был основной организацией страны, которая знакомила советских (российских) социологов с новейшими достижениями в области методики социологических исследований (и в первую очередь с математическими методами, эффективно использующимися в социологии), как западными, так и отечественными. И человеку, желающему успешно работать в социологии, было чему здесь научиться.

Вхождение А.О. Крыштановского в социологию началось именно с учебы — с поступления в заочную аспирантуру Института социологии АН СССР. Успешно ее окончив, Александр Олегович защитил кандидатскую диссертацию на тему «Теоретические и методические проблемы построения банка социологических данных», стал кандидатом философских наук. Вероятно, у читателя может вызвать интерес тот факт, что специалист по анализу социологических данных стал кандидатом *философских* наук. Дело в том, что современная система присуждения степеней родилась далеко не сразу. Постепенно социология институциализировалась, «вычленяясь» из философии. Жизнь обгоняла бюрократические рамки. Социология превращалась в науку, для которой использование математического аппарата становилось необходимым, естественным шагом. А «крышей» для этой науки продолжали служить философские кафедры, отделения и т.д. И одной из тех «капель», которые все-таки «точили камень», явилась целая серия осуществленных в Институте социологии защит кандидатских диссертаций, посвященных развитию методов анализа социологических данных. Среди них была и диссертация Александра Олеговича.

Учиться он продолжал всю жизнь и всегда старался следить за новыми достижениями мировой науки в области методического обеспечения социологических исследований. Так, в 1991 г. он прошел стажировку по методам социологических исследований и анализа данных в университете Сарри и Манчестера (Великобритания) по программе Британского Совета. В 1992 г. — стажировался по методам анализа данных CITY University (Лондон, Великобритания).

Эти стажировки Александр Олегович проходил, будучи сотрудником Института социологии Академии наук СССР, где работал с 1981 по 1993 г. (сначала научным сотрудником, потом — старшим научным сотрудником), с момента поступления в академическую аспирантуру (с перерывом в 1988—1990 гг. был ведущим научным сотрудником НИИкниги). Александр Олегович был одним из самых авторитетных работников Отдела методики института. Блестящее знание пакетов программ, добросовестное отношение к делу привлекало к нему многих сотрудников института. К нему обращались за консультациями, помощью в обработке данных. Он внимательно отвечал на все вопросы, всегда старался помочь коллегам. Александр Олегович был одним из идейных вдохновителей создания в институте Всесоюзного банка социологических данных, предложил много идей по его организации и принципам использования. В 1984—1989 гг. возглавлял группу по созданию такого банка. За разработки в этой области был награжден серебряной медалью ВДНХ СССР.

Уже в стенах института ярко проявился педагогический талант А.О. Крыштановского: умение понятным языком донести до сознания слушателя-гуманитария сложные математические построения.

Александра Олеговича как ведущего специалиста профильного академического института в 1985 г. пригласили читать лекции по анализу социологических данных в МГУ, где создавалось подразделение, готовящее профессиональных социологов: отделение социологии на философском факультете. Будущим социологам он читал лекции до 1988 г.

Александр Олегович активнейшим образом участвовал в обучении этих специалистов методам анализа данных. Он читал лекции по основам информатики и вычислительной техники и для аспирантов Института социологии.

Полностью талант А.О. Крыштановского как педагога раскрылся сразу после ухода из Института социологии РАН. Однако связь с институтом не прерывалась. Так, в 1998 г., когда при Российской академии наук был создан ГУ ГН — Государственный университет гуманитарных наук, Александр Олегович вернулся в стены института в качестве преподавателя факультета социологии — обучал магистров-социологов. Но это продолжалось недолго — в 1999 г. его целиком захватила педагогическая работа в ГУ ВШЭ.

Коротко о том, чем Александр Олегович занимался, уйдя из ИСАН, но еще «не дойдя» до ГУ ВШЭ. В 1993 г. некоторое время он заведовал отделом обработки данных в Институте маркетинговых исследований «ГФК-Москва», работал заведующим информационным отделом Междисциплинарного академического центра социальных наук «Интерцентр», который совместно с ВЦИОМ участвовал в проведении мониторингов, осуществляя работу по накоплению, документированию, агрегированию и хранению материалов исследований.

С 1995 г. А.О. Крыштановский работал заместителем декана факультета социологии Московской школы социальных и экономических наук. Одновременно, будучи профессором этого факультета, активно занимался педагогической деятельностью.

Но еще ярче педагогические и организаторские способности Александра Олеговича проявились именно в ГУ ВШЭ, где он работал с 1999 г.

Первый набор на отделение социологии ГУ ВШЭ был осуществлен в 1996 г. В 1999 г., с приходом А.О. Крыштановского, создается факультет социологии. Под руководством А.О. Крыштановского факультет стал одним из ведущих в России центров подготовки высококвалифицированных специалистов-социологов.

В 2005 г. студентами факультета социологии ГУ ВШЭ стали около 100 абитуриентов из разных регионов России. Всего на факультете социологии в настоящее время обучается более 500 студентов.

На факультете ведется активная академическая деятельность, преподаватели факультета участвуют в научных исследованиях. Проводится большая работа по вовлечению студентов в научную деятельность. В 1999 г. при активной поддержке Александра Олеговича по инициативе студентов на факультете был организован студенческий клуб «Город», объединяющий учащихся разных курсов. Этот клуб ведет большую исследовательскую работу, в том числе — ежегодный мониторинг студенческой и преподавательской жизни в ГУ ВШЭ.

В 2003 г. факультетом совместно с Московской высшей школой социальных и экономических наук открыто обучение по двухгодичной магистерской программе «Комплексный социальный анализ». Выпускники получают диплом магистра социологии Манчестерского университета (Master of Arts in Sociology) и диплом магистра Государственного университета — Высшей школы экономики.

С сентября 2005 г. в магистратуре ГУ ВШЭ действует еще одна программа — «Прикладные методы анализа рынков», подготовленная под руководством А.О. Крыштановского, который принимал активное участие в приеме вступительных экзаменов. Но учиться магистранты начали уже без него...

Огромную работу провел Александр Олегович и по организации кафедры методов сбора и анализа социологической информации, преподаватели которой в настоящее время ведут более 20 дисциплин как на факультете социологии, так и на других факультетах ГУ ВШЭ. К работе на кафедре привлечены ведущие специалисты Москвы.

Отметим, что в организационно-педагогической деятельности А.О. Крыштановского большую роль сыграл его предыдущий опыт человека, умеющего практически анализировать социологические данные, хорошо знающего все подводные камни реальной работы социолога-аналитика. В частности, организовывая работу кафедры, он активно задействовал свое знание деятельности маркетинговых компаний.

А.О. Крыштановский проводил регулярные встречи маркетологов, практических специалистов в области исследования рынка со студентами. Эти встречи позволили наладить прямое общение между студентами и профессионалами в области социологических исследований.

С 2001 г. Крыштановский по совместительству руководил департаментом учебных проектов консалтинговой группы «Русинфомар». И здесь он

активнейшим образом выступал за повышение качества работы отечественных маркетологов, широкое использование современных социологических методов, пытался наладить соответствующую систему обучения. По инициативе А.О. Крыштановского на факультете была организована базовая кафедра на основе Института маркетинговых исследований «ГФК-Русь».

Конечно, не только в маркетинге Александр Олегович видел ту большую нишу, в которой могут найти применение своим силам выпускники факультета. Не в меньшей мере он думал о необходимости обучения студентов грамотному изучению общественного мнения. По инициативе Крыштановского к преподаванию на кафедре были привлечены и ведущие специалисты ФОМ и ВЦИОМ. Более того, при факультете социологии ГУ ВШЭ на основе Фонда «Общественное мнение» была организована еще одна базовая кафедра. Соответствующий учебный предмет был определен как «опросная фабрика ФОМ» и в качестве цели кафедры фигурировало обеспечение теоретического и практического преподавания этого предмета. К сожалению, фактически работать базовая кафедра начала уже без Александра Олеговича.

Параллельно А.О. Крыштановский активно занимался и преподавательской работой. Он читал ряд авторских курсов, так или иначе связанных с анализом социологических данных: «Анализ социологических данных», «Современные методы анализа социологических данных», «Методы анализа временной динамики социальных явлений», «Анализ временных рядов», «Методы выборочных исследований в социологии».

Нельзя не сказать хотя бы коротко о научной деятельности А.О. Крыштановского.

Направление его научных интересов определялось в первую очередь тем, что он любил и умел анализировать реальные социологические данные. Любые теоретические аспекты использования разных методов он оценивал с помощью обработки данных реальных исследований. Его интересовало, как наиболее оптимально тот или иной подход можно реализовать на компьютере. Александр Олегович разработал много предложений по организации данных на компьютере, использованию анализа данных в сравнительных и вторичных исследованиях, по организации выборки (в частности, он занимался проблемой взвешивания выборки). Интересные результаты были им получены в области регрессионного анализа и комплексного использования факторного и кластерного анализа.

Будучи на «ты» с самым передовым для того или иного момента времени программным обеспечением, Александр Олегович писал о том, каким образом его надо использовать для решения насущных задач социолога.

Конечно, именно подобная сугубо практическая направленность работ Крыщ. тановского парадоксальным образом привела к тому, что некоторые его научные результаты сегодня можно считать имеющими только историческое значение: вместе с развитием техники ушли и некоторые теоретические проблемы, связанные с привязкой анализа данных к определенным технологиям. Тем не менее в каждой его работе, даже если она посвящена анализу каких-либо аспектов использования уже устаревшей технологии, можно найти нечто, полезное и в наше время. В этом читатель может убедиться, проанализировав работы, список которых дан в конце настоящей книги.

Ниже представлены статьи Александра Олеговича, посвященные таким аспектам проведения социологических исследований, которые не связаны непосредственно со спецификой вычислительной техники и не перестают быть актуальными и сейчас. Это некоторые проблемы построения выборочной совокупности; ряд вопросов, связанных с корректностью применения в социологии одного из самых популярных статистических методов — регрессионного анализа; методические нюансы комплексного использования методов факторного и кластерного анализа. Одна из статей (об отношении населения России к деятельности президента) служит небольшой иллюстрацией огромной работы Александра Олеговича по анализу данных реальных социологических исследований.

Ремонт **выборки**

А.А. Давыдов, А.О. Крыштановский

Практика проведения социологических исследований показывает, что, как бы тщательно ни был спланирован полевой этап сбора информации, всегда имеют место смещения выборок по социально-демографическим характеристикам, пропущенные ответы в анкетах и некоторые другие моменты, снижающие качество социологической информации.

Для того чтобы свести к минимуму влияние этих нежелательных факторов, в методической литературе [1—3] рекомендуется проводить ремонт выборки. Однако у большинства социологов нет ясного представления о практической реализации этого необходимого этапа социологического исследования. Так, среди 300 исследований, содержащихся во Всесоюзном банке социологических данных ИС АН СССР, лишь в десяти осуществлялся ремонт выборки. Для сравнения отметим, что за рубежом ремонт выборки уже давно стал распространенным методом повышения качества социологической информации.

Причины нашего отставания в применении ремонта выборки очевидны: отсутствие вычислительной техники, специализированного программного обеспечения, методических пособий, недостаточная квалификация исследователей и ряд других факторов. В настоящее время положение меняется в лучшую сторону, поэтому мы считаем разговор о ремонте выборки актуальным.

Что же такое ремонт выборки? В узком смысле — это уравнивание выборочных и генеральных распределений социально-демографических характеристик респондентов. В широком — первичная статистическая обработка данных, включающая коррекцию:

- смещения социально-демографических характеристик респондентов;
- неоднородности массивов данных;
- резко выделяющихся и восстановление пропущенных ответов. Таким образом, цель ремонта выборки — повышение качества уже собранной информации. Добиться этой цели можно, если использовать избыточную информацию, которая содержится в собранных данных. Это важнейшее положение, к сожалению, редко учитывают социологи.

Рассмотрим общие методологические принципы, на которых базируется логика ремонта выборки. Первый — доминирование неформальных процедур принятия решений. Успех при ремонте выборки достигается благодаря

глубокому знанию изучаемой проблемы, процедуры выборочного исследования, ситуации опроса респондентов и т.д., а собственно математические процедуры играют подчиненную роль. Такой подход в корне противоположен мнению, согласно которому ремонт выборки — это недостойные серьезного социолога математические манипуляции, уводящие в сторону от познания социальной реальности.

Второй принцип — оптимизация. Повышение качества собранной информации осуществляется благодаря максимальному уменьшению влияния нежелательных факторов при минимальном искажении, вносимом ремонтом выборки. Поясним это положение примером. Допустим, мы опросили мужчин больше, чем требовалось, и теперь уменьшаем их количество до требуемой квоты. В результате объем выборки может оказаться недостаточным для решения поставленных задач. Значит, сокращать объем можно только до некоторого оптимума. При этом надо помнить, что ремонт не заменяет расчета выборки и качественной работы анкетеров. Он лишь облегчает усилия по ее расчету и реализации.

Рассмотрим процедуры ремонта выборки с того момента, когда данные введены в компьютер и «очищены» от ошибок перфорации.

Коррекция неоднородности сбора данных. Неоднородность сбора данных возникает по двум причинам. Во-первых, на полевой стадии исследования практически всегда сбор информации осуществляют несколько анкетеров, различающихся степенью подготовки, социально-демографическими и личностными характеристиками. Кроме того, анкетеры не всегда проводят опрос в одинаковых условиях. Все это оказывает воздействие на ответы респондентов, причем достаточно сильное [4].

Во-вторых, при проведении почтового опроса сбор информации растягивается иногда на несколько недель, и данные, поступившие в разные периоды полевого этапа, могут отражать временные изменения изучаемого явления или разные группы респондентов.

Перед исследователем возникает вопрос: правомерно ли объединять эти данные в один общий массив, если анализировать их вместе нельзя, поскольку в этом случае мы получим существенные искажения? Можно ли рассматривать несколько совокупностей данных в качестве различных выборок, полученных из одной и той же генеральной совокупности [5]? В случае, когда группы данных оказались неоднородными, перед исследователем возникают проблемы коррекции.

Если различия в массивах данных обусловлены влиянием анкетера ил условиями опроса, следовало бы провести повторный опрос в данно

выборочной совокупности с помощью другого анкетера или анкетеров. Но это в идеале, а на практике — дефицит времени, материальных и людских ресурсов. Кроме того, могут произойти существенные изменения в общественной жизни, и неясно, что мы получим: информацию, которая отражает предыдущий период, или результаты изменений в общественной жизни. Следует также учитывать, что повторный опрос одних и тех же людей ведет к искажениям, обусловленным психологическими особенностями, а если опрашивать других респондентов со схожими социально-демографическими характеристиками, изменения в оценках могут быть обусловлены новым объектом.

В этой ситуации на помощь может прийти одна из основных процедур ремонта выборки — перевзвешивание данных. Суть ее состоит в следующем: с помощью эмпирически найденных весов так скорректировать данные, чтобы влияние смещений снизить до оптимальных пределов. Величина смещения находится либо с помощью экспертного опроса, либо как отклонение средних значений в подвыборках от среднего значения по всему массиву.

При перевзвешивании данных возникает проблема правильного выбора эталона, по отношению к которому будет осуществляться коррекция неоднородности. Эталон можно выбрать исходя из содержательных соображений либо конструировать его как нечто среднее из тех подвыборок, которые имеются в наличии. Одним из таких эталонов может выступать средневзвешенная величина смещения по всему массиву. В табл. 1 представлены этапы подобной коррекции.

Таблица 1. Коррекция неоднородности данных*

Номер анкетера	1	2	3
Исходное количество анкет	80	54	101
Балл смещения	2	1	3
Средневзвешенная величина смещения	(80×2 + 54×1 + 101×3) : 235 ≈ 2,2		
Процедура расчета веса	2,2 : 2 = 1,1	2,2 : 1 = 2,2	2,2 : 3 ≈ 0,73
Коррекция количества анкет	80×1,1 = 88	54×2,2 = 119	101×0,73 = 74

* *Примечание*. Эксперты оценивали смещение с помощью 3-балльной шкалы, где 1 балл — смещение незначительное, а 3 балла—смещение очень значительное.

В результате коррекции объем массива увеличился на 46 анкет (на 20% первоначального объема). Если бы в качестве эталона была выбрана не средневзвешенная величина смещения, как в табл. 1, а минимальное смещение

(1 балл), объем скорректированного массива сократился бы на 108 анкет (46% начального объема).

Таким образом, стремление к максимальному уменьшению смещения привело к сокращению исходного массива почти вдвое. В то же время использование средневзвешенной величины смещения в качестве эталона заставило продублировать каждую анкету второго анкетера. И хотя в методической литературе высказывается мнение, что нельзя одну и ту же анкету включать в машинную обработку более 10—11 раз [2], все же дублирование анкет увеличивает влияние индивидуальных особенностей опрашиваемых, которое в каждом конкретном исследовании различно. Поэтому в одном исследовании правомерно увеличить подмассив вдвое, а в другом — нет. Проблема верхних границ дублирования остается открытой, поэтому знание проблемы и здравый смысл — основные критерии принятия правильного решения. Мы рассмотрели коррекцию смещений, вызванных влиянием анкетера. Аналогично проводится коррекция смещений, обусловленных различием массивов данных во времени.

После того как веса найдены и массив скорректирован[1], с помощью статистических критериев следует еще раз провести проверку на однородность. В случае, если подмассивы снова оказались неоднородными, требуется новое перевзвешивание, но уже с другими весами, и затем проверку надо повторить. Если скорректированные подмассивы опять окажутся неоднородными, их следует обрабатывать отдельно.

Коррекция распределений социально-демографических характеристик респондентов. После сбора информации практически всегда наблюдается смещение социально-демографических характеристик опрошенных, по сравнению с генеральной совокупностью. Прежде чем приступать к коррекции, полезно выявить влияние социально-демографических признаков на ответы респондентов. Этот анализ может быть осуществлен с помощью двумерных таблиц сопряженности или множественного номинального анализа. Например, нами установлено, что социально-демографические признаки слабо связаны с ответами об удовлетворенности работой и жизнью, оценкой темпов перестройки, одобрением деятельности политических лидеров, оценкой внешнеполитических событий и др. Для этих индикаторов перевзвешивание по социально-демографическим характеристикам не нужно.

[1] Процедура перевзвешивания реализована в пакете программ «Социолог», который используется в ИС АН СССР с 1984 г.

Возможны три ситуации. Первая — ответы респондентов не связаны с социально-демографическими характеристиками, в этом случае коррекция не проводится. Вторая ситуация — какая-то социально-демографическая характеристика, например пол, тесно связана со всеми содержательными вопросами, или третья — разные вопросы могут быть связаны с различными характеристиками. В этом случае коррекция проводится по схеме, описанной в [3].

Из табл. 2 следует, что в результате коррекции количество мужчин уменьшилось на 28 человек, и на столько же увеличилось число женщин. Дублирование анкет женщин основывается на базовом принципе выборочного метода [6], согласно которому каждый индивид несет всю информацию, представленную в его социально-демографической группе. В табл. 2 приведен пример, когда на ответы респондентов оказывает влияние только одна характеристика — пол. Практика показывает, что так бывает далеко не всегда. Значительно чаще встречается ситуация, когда на ответы респондентов оказывают влияние две, например возраст и образование, или три и более социально-демографические характеристики. Для этого случая одним из авторов статьи разработаны метод и соответствующие программы для ЭВМ, рассчитывающие веса для нескольких признаков одновременно [7]. Здесь отметим следующее: использование данных программ в ИС АН СССР показало их высокую эксплуатационную надежность, а главное — простоту в обращении.

Таблица 2. Коррекция по одной социально-демографической характеристике

Пол	Выборочная совокупность		Генеральная совокупность, %	Расчет веса	Скорректированная совокупность	
	численность	%			численность	%
Мужчины	**100**	**66,6**	**48**	48:66,6 = 0,72	100x0,72 = 72	48
Женщины	**50**	**33,4**	**52**	52:33,4=1,56	50x1,56 = 78	52

После коррекции выборочных распределений социально-демографических признаков можно приступать к следующему этапу ремонта выборки — коррекции резко выделяющихся и восстановлению пропущенных ответов.

Коррекция резко выделяющихся ответов респондентов. В практике опросов общественного мнения встречаются ответы респондентов, которые сильно отличаются от основной массы ответов. Это может быть обусловлено ошибкой самого респондента или ошибкой регистрации ответа

интервьюером, иногда — особенным мнением респондента или резким изменением условий опроса. Установить истинную причину отклонения практически невозможно. Резко выделяющиеся ответы затрудняют анализ данных, поэтому вполне естественно стремление их как-то найти и скорректировать. Выявлению резко выделяющихся наблюдений посвящено большое количество научных публикаций (например, [5]), поэтому мы не будем подробно останавливаться на этой задаче, и рассмотрим проблему коррекции подобных наблюдений. Самый простой способ — удалить данный ответ или всю анкету из дальнейшего анализа. Эта возможность предусмотрена в пакетах «Социолог», BMDP-79, SPSS и ряде других. Однако когда объем выборки невелик, это обходится слишком дорого, особенно если резко выделяющихся ответов много.

Второй способ — отнесение резко выделяющихся ответов к градации «другое». Этот прием применяется при кодировке открытых вопросов и с успехом может быть использован при коррекции резко выделяющихся ответов, поскольку «отнесение» таких ответов в одну градацию обеспечивает ее наполнение и делает возможным дальнейший анализ.

Третий способ — уменьшение дробности шкалы. Например, Г.И. Саганенко отмечает, что шкалу в пять-семь-девять градаций почти всегда приходится сводить к трем-четырем [6]. Эту задачу можно решить с помощью статистических критериев, например критерия Фишера, который показывает, значимо ли различаются доли ответов респондентов. Наш опыт свидетельствует, что уменьшение дробности шкалы позволяет эффективно бороться с резко выделяющимися ответами.

Коррекция пропущенных ответов. Данный вид смещений возникает чаще всего в открытых вопросах и вопросах табличного типа. Самый простой способ коррекции — исключение из дальнейшего анализа пропущенных ответов или всей анкеты. Если объем выборки большой, это весьма рациональный подход. В условиях выборок малых и средних объемов распространенными способами коррекции являются: отнесение пропущенного ответа к градации «затрудняюсь ответить», замена пропущенного ответа каким-либо средним значением, рассчитанным по имеющимся данным, или значением, вычисленным с помощью регрессии. Названные процедуры реализованы в пакетах «Социолог», BMDP-79, SPSS и ряде других. Выбор того или иного способа коррекции пропущенных ответов в значительной мере зависит от последующего анализа данных. Например, при расчете одномерного частотного распределения пропущенный ответ логично отнести к категории «затрудняюсь ответить». Однако если предполагается факторный анализ, такой подход неприемлем, поскольку эта категория исключается из обработки. При

планировании факторного анализа более естественно заполнение пропусков модальным, медианным или среднеарифметическим значением, вычисленным по всему массиву или в социально-демографической группе того респондента, который не ответил на вопрос. Предполагается, что мода, медиана или среднеарифметическое значение отражают общую тенденцию, а другие ответы, отклоняющиеся от этих значений, обусловлены влиянием личностных особенностей респондентов, различиями в ситуации опроса и другими случайными факторами.

При планировании логлинейного анализа коррекция пропущенных ответов осуществляется другим способом. Напомним, что в логлинейном анализе от частоты в каждой ячейке таблицы сопряженности берется логарифм, и, если там нет наблюдений, то, строго говоря, данный анализ невозможен. Поэтому в статистической литературе рекомендуют перед проведением логлинейного анализа в каждую ячейку таблицы сопряженности добавить некоторое небольшое число, как правило, в интервале от 0,25 до 1,00 [8], что позволяет вычислить логарифм при отсутствии ответа. Доказано, что подобная «добавка» не сказывается на качестве результата [9]. (Данная процедура реализована в программе логлинейного анализа пакета BMDP-79, где в каждую ячейку таблицы сопряженности добавляется число 0,5.)

До сих пор мы рассматривали ситуации, когда пропущен содержательный ответ. А что делать, если отсутствует какая-либо социально-демографическая характеристика? В этом случае можно поступить так: если социально-демографические характеристики не связаны с содержательными ответами, то анкете с пропущенными значениями следует присвоить наиболее часто встречающиеся в выборке социально-демографические характеристики, либо определить их случайным образом или пропорционально (если таких анкет много). Если же связь есть, следует определить, к ответам какой группы (например, мужчин или женщин) ближе ответы в анкете, где графа «пол» не указана, и внести этот признак.

Итак, мы осуществили ремонт выборки, и теперь следует оценить смещения, вносимые самим ремонтом. Для этой цели нужно найти эталон, по отношению к которому будет рассчитываться смещение. Возможны два эталона — внутренний и внешний. Процедура построения внутреннего эталона может быть следующей: из выборочной совокупности формируется небольшая подвыборка, в которой практически отсутствуют смещения по социально-демографическим признакам, резко выделяющиеся и пропущенные ответы. По данной подвыборке рассчитываются процентные распределения, связи и т.д., а затем сравниваются с данными, полученными после ремонта.

Мера отклонения от результатов эталонной подвыборки и будет выступать показателем смещения, вносимого ремонтом. Допустимость смещений легко выявляется с помощью статистических показателей, например, $\%^2$ распределения. Процедура построения внешнего эталона несколько иная. Во время полевого этапа данные собираются по двум выборочным планам. Один — основной, по которому будет осуществляться ремонт выборки и дальнейший анализ, а второй — для создания эталона. Подразумевается, что эталонная подвыборка не участвует в общем анализе, собирается особенно тщательно, а распределения социально-демографических характеристик точно соответствуют распределению в генеральной совокупности.

С нашей точки зрения, предпочтение следует отдавать внешнему эталону, поскольку при построении внутреннего могут возникать сложности, обусловленные ограниченным объемом выборки [6].

В заключение отметим еще одно принципиальное положение. Если данных много, ремонт выборки может осуществляться за счет сокращения выборочной совокупности. Это наиболее рациональный подход к ремонту выборки, поскольку данная стратегия не опирается ни на какие дополнительные допущения. Если объем выборки незначителен, для ее ремонта нужно принимать ряд дополнительных допущений, которые не следуют из собранного материала и истинность которых трудно проверить. Таким образом, возникает дилемма: опрашивать большое количество респондентов, не беспокоясь о качестве, в надежде на «капитальный» ремонт выборки, или опрашивать значительно меньшее количество респондентов, но с высоким качеством, предполагая «косметический» ремонт. Ответ следует искать в размере затрат (материальных, временных и др.), вытекающих из каждого решения, в особенностях изучаемой проблемы, целях и задачах исследования и ряде других факторов.

Литература

1. Джессен Р. Методы статистических обследований. М.: Финансы и статистика, 1985.

2. Петренко Е.С., Ярошенко Т.М. Социально-демографические показатели в социологических исследованиях. М.: Статистика, 1979.

3. Процесс обработки данных анкетных опросов на ЭВМ. М.: ИС АН СССР, 1985.

4. Погосян Г.А. Метод интервью и достоверность социологической информации. Ереван: Изд-во АН Армянской СССР, 1985.

5. Айвазян С.А., Енюков И.С, Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.

6. Саганенко Г.И. Надежность результатов социологического исследования. Л.: Наука, 1983.

7. Крыштановский А.О., Кузнецов А.Г. Перевзвешивание выборки // Комплексный подход к анализу данных в социологии. М.: ИС АН СССР, 1988.

8. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Т. 1. М.: Финансы и статистика, 1982.

9. Аптон Г. Анализ таблиц сопряженности. М.: Финансы и статистика, 1983.

Первая публикация: СОЦИС. 1989. № 5. С. 100—105.

Некоторые вопросы перевзвешивания выборки

А.О. Крыштановский, А.Г. Кузнецов

Построение выборки, репрезентирующей изучаемую социологами генеральную совокупность, представляет, как правило, весьма сложную задачу. Возникающие здесь проблемы и пути их решения для некоторых стратегий построения выборок нашли отражение в литературе [1], [2], [3]. Вместе с тем правильное определение на этапе подготовки исследования тех единиц, которые должны быть включены в выборку, еще не означает, что собранная в итоге информация будет репрезентовать генеральную совокупность.

Связано это с тем, что точно выдержать выработанный план выборки на практике невозможно. Различного рода «недоступные» единицы приводят к необходимости опроса других лиц и, следовательно, искажают структуру выборки. Существуют разные способы замены существующих респондентов [см., например, 4; 5], однако на практике их точное выполнение также сопряжено со значительными сложностями (в частности, существенно возрастает длительность и стоимость этапа сбора информации).

Возможность искажения структуры выборки приводит к необходимости этапа контроля соответствия характеристик реальной выборки характеристикам изучаемой совокупности. Контроль этот осуществляется путем сравнения распределений тех параметров генеральной совокупности, которые имеются в распоряжении исследователя, и распределений тех же параметров, полученных в результате исследования.

На практике нередки случаи, когда несовпадение распределений значительно превышает границы соответствующих доверительных интервалов, т.е. имеются нарушения репрезентативности выборки. Причины этих нарушений должны изучаться, возможности их исправления две: проводить добор единиц, которых оказалось в выборке непропорционально мало, или удалять единицы, которых оказалось непропорционально много; перевзвешивать выборку.

Каждый из подходов имеет свои плюсы и минусы. Первый путь в случае добора недостающих единиц требует дополнительного опроса, а значит, дополнительных сил и времени. В случае удаления лишних единиц теряется часть информации, на сбор которой уже израсходованы определенные средства. С другой стороны, вычисление доверительных интервалов по выборкам, полученным таким образом, не вызывает затруднений. Эти выборки вполне корректны с точки зрения статистики.

Перевзвешивание не подразумевает какого-либо изменения реального числа объектов в выборке. При таком подходе соответствие характеристик выборочной и генеральной совокупностей достигается за счет введения на этапе обработки данных специального параметра — переменной веса. Переменная веса указывает, сколько раз должны учитываться ответы той или иной группы респондентов. Например, если в выборке оказалось мужчин в 2 раза меньше, чем в генеральной совокупности, переменная веса равна 1 для женщин и 2 для мужчин. При обработке данных с использованием такой переменной веса ответ каждого мужчины будет учитываться дважды, т.е. объем этой группы респондентов как бы удваивается, чем достигается соответствие выборочной и генеральной совокупности по этому параметру.

Достоинством процедуры перевзвешивания является то, что она не требует никаких реальных манипуляций с выборкой, а выполняется на этапе обработки информации[1]. Следовательно, перевзвешивание много дешевле изменения реального числа единиц в выборке. Недостаток перевзвешивания — изменение структуры выборки, причем не случайным образом. При перевзвешивании остается открытым вопрос о том, как следует вычислять доверительные интервалы для всех получаемых на основе выборки величин (процентов, средних и т.п.).

Другой вопрос, связанный с использованием переменной веса, это вопрос о ее вычислении. Если необходимо провести перевешивание выборки только по одному параметру, вычисление переменной веса не представляет трудностей. Формула для ее вычисления, позволяющая получить такие значения, которые не изменяют объема выборки, может быть записана следующим образом:

$$K_i = \frac{A_i}{B_i}, \qquad (1)$$

где K_i — первое значение переменной веса; i — номер градации переменной, по которой производится перевзвешивание; A_i — доля i-й градации в распре-

[1] Большинство развитых программных систем обработки социологических данных позволяет выполнять все основные алгоритмы информации (построение многомерных распределений, регрессионный, факторный, дисперсионный и др. виды анализа) с помощью переменной веса.

делении переменной в генеральной совокупности; B — доля г-й градации в распределении переменной в выборочной совокупности.

Рассмотрим пример вычисления переменной веса для данных одного из исследований, проведенных ИС АН СССР (табл. 1).

Таблица 1. Пример вычисления значений переменной веса

Демографическая группа	Выборка, %	Генеральная совокупность, %	Значение переменной веса
Мужчины	37	43	1,162
Женщины	63	57	0,905

Описанная процедура перевзвешивания выборки по одной переменной достаточно известна, хотя на практике за последние два года работы ВЦ ИС АН СССР применялась лишь трижды. Связано это, на наш взгляд, не с тем, что реальные выборки хорошо соответствуют генеральным совокупностям, а с крайне малым распространением репрезентативных исследований[2].

Однако несовпадения генеральной и выборочной совокупностей в реальности случаются не по одному, а по нескольким параметрам одновременно. Обычно в этом случае социологи считают, что достаточно перевзвесить выборку по одному из параметров, и это приведет к ее автоматической корректировке по другим параметрам. Но это верно только, когда есть сильная связь между тем параметром, по которому проводится перевзвешивание, и теми, которые также должны быть откорректированы. Наличие такой связи подчас более чем сомнительно, и поэтому задачу о перевзвешивании выборки необходимо ставить одновременно по нескольким переменным.

В принципе, случай перевзвешивания по нескольким переменным не отличается от перевзвешивания по одной переменной. Если имеется совместное распределение интересующих исследователя переменных в выборочной и генеральной совокупностях, получение значений весовой переменной проводится по формуле (1) [6, с. 127]. Однако на практике социолог, как правило, не имеет совместных распределений интересующих его признаков

[2] Из 200 исследований, представленных во Всесоюзном банке социологических данных ИС АН СССР, не более чем в десяти предпринимались реальные усилия для получения репрезентативных выборок.

Для сравнения: 80% исследований, хранящихся в Роупервском центре (крупнейший архив социологических данных США), репрезентативны.

в генеральной совокупности, что значительно усложняет вычисление значений переменной веса. В этой ситуации задачу можно поставить следующим образом. Имеется совместное распределение нескольких переменных на выборочной совокупности, при этом безусловные распределения некоторых из них (быть может, всех) отличаются от соответствующих распределений в генеральной совокупности. Необходимо так изменить совместное распределение в выборочной совокупности, чтобы безусловные распределения всех анализируемых переменных соответствовали распределениям в генеральной совокупности.

Проиллюстрируем постановку задачи на примере таблицы для случая двух переменных (табл. 2). В алгебраической форме двумерную задачу можно записать в виде системы уравнений (2).

Таблица 2. Пример таблицы сопряженности для двумерного перевзвешивания

Переменная 1		1	2	3	
Переменная 2	1	«и	", 2	«13 m_{n}	«,.т,.
	2	« 2 1	«22 m_{22}	«23 т 2 3	"I."I.
		«.!	«. 2 ".2	". 3	$n_{«}m_{••}$

и — доля выборки, попавшая в клетку (i,j); $n_{\%\cdot}$, л.$_{\#}$ — маргинальные доли в выборочной совокупности; $m_{\%\cdot}$, $m._{щ}$ — маргинальные доли в генеральной совокупности; $m..$ — искомые доли в выборке; тогда значение весовой переменной $k_{ij} = \frac{m_{ij}}{n_{ij}}$

$$\sum_{j=1}^{H} m_{1j} = m_{1\bullet};$$

$$\sum_{j=1}^{H} m_{kj} = m_{k\bullet}; \qquad \textbf{(2)}$$

$$\sum_{i=1}^{K} m_{i1} = m_{\bullet 1};$$

...

где K и H — число градаций в двух анализируемых переменных; $m_{i\bullet}$, $m_{\bullet j}$ ($i = 1 \div H; j = 1$ т X) — задаваемые константы (распределения в генеральной совокупности); т.. — искомые значения.

Система уравнений (2) состоит из K+ //уравнений с K x //переменными. Такая система уравнений имеет бесконечное множество решений. Из всех возможных решений целесообразно выбрать такое, которое дает совместное распределение изучаемых характеристик, минимально отличающееся от полученного в процессе исследования. Следовательно, должна решаться задача минимизации функционала:

$$\sum_i \sum_j (m_{ij} - n_{ij})^2 \qquad (3)$$

при ограничениях (2), а также с учетом естественного требования

$$m_{ij} > 0 \text{ для } i, j. \qquad (4)$$

Получаемые решения m.. легко преобразуются в искомые значения весовой переменной $k_{ij} = \frac{m_{ij}}{n_{ij}}$. Однако следует учесть, что очень большие и очень малые значения k_{ij} крайне нежелательны, поскольку при них доля соответствующей группы респондентов в выборке резко меняется. По этой причине ограничения (4) заменим на более сильные:

$$Cn_{ij} \geq m_{ij} \geq \frac{1}{C} n_{ij}, \qquad (5)$$

где С — заранее заданная константа, определяющая максимально и минимально возможные значения переменной веса k_{ij}.

Таким образом, полученная задача математического программирования: найти такие решениm_{ij}, которые доставляют минимум функционалу (3) при ограничениях (2) и (5).

Существуют разные способы решения такой задачи. Нами использовался следующий способ.

Задача на поиск условного минимума была сведена к задаче на безусловный минимум с помощью метода штрафных функций. Если есть задача поиска условного минимума функционала

$$F(x) \text{ -» min} \qquad (6)$$

при TV ограничениях вида

$$G_i(x) = 0, i = l, \ldots, N, \qquad (7)$$

вполне эквивалентна задача на безусловный минимум следующего функционала

$$<Kx) = P(x) + \text{^K\&(x)^mm}, \qquad (8)$$

где z — четное число; K. — достаточно большие числа (г = 1, ...,7V).

Таким образом, за нарушение ограничений (7) функционал (8) «наказывается» штрафом. Задача минимизации (8) решается с помощью метода прямого поиска [7]. Программа, реализующая этот алгоритм, приводится в работе [8].

В заключение рассмотрим пример применения описанного подхода для перевзвешивания выборки одного из исследований, проведенных в ИС АН СССР, и уже частично рассмотренного в (см. табл. 1).

Как показало сравнение характеристик выборочной и генеральной совокупностей, произошло смещение выборки не только по полу, но и по возрасту. При этом перевзвешивание только по одному параметру (полу) не дает необходимой корректировки по возрасту. Были вычислены значения весовой переменной, которые вместе с исходными данными приведены в табл. 3.

Таблица 3. Пример вычисления переменной веса при взвешивании выборки по полу и возрасту

		Возраст, лет									Выборка, %
		1	2	3	4	5	6	7	8	9	
Пол	1	B=0,2	B=2,8	B=4,9	B=10,9	B=7,4	B=3,9	B=3,6	B=2,5	B=2,7	37
		K=3,0	K=2,2	K=1,8	K=1,0	K=1,0	K=1,0	K=0,6	K=0,6	K=0,6	
	2	B=1,4	B=5,1	B=6,7	B=14,8	B=10,6	B=4,5	B=5,8	B=4,7	B=7,3	63
		K=1,4	K=1,0	K=1,0	K=1,0	K=1,0	K=0,6	K=1,0	K=0,6	K=1,0	
Выборка, %		1,6	7,9	11,6	25,7	18,0	8,4	9,4	7,2	10,0	
Генеральная совокупность, %		3,7	12,6	14.2	23,1	16,8	7,3	7,7	5,1	9,5	

B — доля выборки, попавшая в данную клетку;

— значение переменной веса для данной клетки (приводится с точностью до одного знака после запятой).

При вычислении значений переменной веса для табл. 3 в качестве константы С, определяющей границы изменений этой переменной (см. формулу (5)), была выбрана 3, т.е. переменная веса может изменяться от 1/3 до 3.

Литература

1. Кокрен У. Методы выборочного исследования. М.: Статистика, 1976.

2. Территориальная выборка в социологических исследованиях. М.; Наука, 1980.

3. Саганенко Г.И. Надежность результатов социологического исследования. Л.: Наука, 1980.

4. Петренко Е.С., Ярошенко Т.М. Социально-демографические показатели в социологических исследованиях. М.: Статистика, 1979.

5. Ноэль Э. Массовые опросы: Введение в методику демоскопии. М.: Прогресс, 1978.

6. Рукавишников В.О., Паниотто В.И., Чурилов Н.Н. Опросы населения. М.: Финансы и статистика, 1984.

7. Hook R., Jeeves T.A. Direct Search Solution of Numerical and Statistical Problems // J.ASM. 1961. Vol. 8. N 2. P. 212—229.

8. Агеев М.И., Алик В.П., Марков Ю.И. Библиотека алгоритмов 1516—2006. М.: Советское радио, 1981. (Техническая кибернетика).

Первая публикация: Комплексный подход к анализу данных в социологии: сб. науч. ст. М.: Ин-т социологии АН СССР, 1989. С. 16—24.

Отношение населения России к деятельности президента

А. О. Крыштановский

В газетах и на телевидении регулярно приводятся данные массовых опросов о положении в стране, отношении населения к деятельности того или иного политического лидера, рейтинги политиков и т. п. Однако эти сведения носят фрагментарный характер. Как правило, отсутствует информация о том, где и как проходил опрос. Обычно сообщают лишь название организации, проводившей исследование, и объем выборки. Поэтому отследить динамику оценок невозможно. Требуется серия повторных исследований.

Мониторинг экономических и социальных перемен в России осуществляется Всероссийским центром изучения общественного мнения совместно с Междисциплинарным академическим центром социальных наук (ИНТЕРЦЕНТР). Время опроса— 1993—1994 гг. Исследование[1] посвящено изучению социального контекста экономических преобразований России [1], в том числе политических ориентации населения.

Проанализируем ответы на вопрос: «Как вы думаете, способствует ли выходу из нынешнего кризиса деятельность президента Б. Ельцина»: нет ответа; способствует; не способствует; не оказывает существенного влияния; не знаю, затрудняюсь ответить.

Вопрос допускает несколько толкований. Ведь под кризисом можно понимать: общую, долговременную политическую и экономическую ситуацию в России; ситуацию, которая складывается в данный момент (например, во взаимодействии Верховного Совета РФ с президентской властью весной, летом и осенью 1993 г.); ситуацию конкретного политического и экономического напряжения в определенном регионе (например, шахтерские забастовки).

Поскольку опрос проводился по всей России, а политико-экономическая ситуация в регионах существенно различается, однозначное толкование ответа респондента практически невозможно. Мы исходим из того, что ответ

[1] Объем выборки ежемесячного исследования составляет 4 тыс. человек. Опрос проводился методом анкетирования в присутствии интервьюера. Выборка репрезентирует взрослое население России по основным социально-демографическим параметрам. Отклонение средней не превышает 3 процентных пункта с вероятностью 0,95.

на данный вопрос — некоторый комплексный показатель оценки деятельности президента[2], который включает оценку всех перечисленных выше моментов.

В августе 1993 г. 16% взрослого населения России ответили на него утвердительно. Ниже этого уровня популярность президента не падала. В тот же период максимальное число опрошенных (36%) оценили деятельность Б.Н. Ельцина негативно (рис. 1).

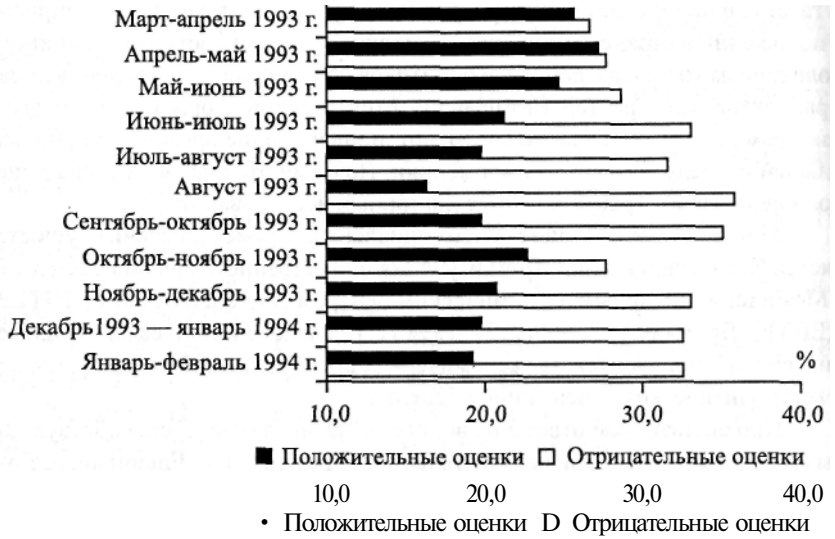


Рис. 1. Оценка деятельности президента Б.Н. Ельцина в зависимости от времени опроса

В табл. 1 приводятся некоторые социально-демографические характеристики респондентов, одобряющих и не одобряющих деятельность президента России. Естественно, возникает вопрос о том, менялся ли качественный состав этих групп респондентов. Существенных изменений социально-демографических характеристик в группах не произошло — почти все различия лежат в рамках доверительного интервала (см. табл. 1). Таким образом, группы поддержки президента приблизительно те же, что и год назад.

[2] Под положительной оценкой («сторонники») подразумевали ответ «деятельность способствует выходу из кризиса», под отрицательной оценкой («противники») — ответ «деятельность не способствует выходу из кризиса».

Таблица 1. Социально-демографические характеристики респондентов, одобряющих и не одобряющих деятельность президента

Группы респондентов	Время опросов				
	Мужчины, %	Средний возраст, лет	Жители города, %	Лица с образованием ниже среднего, %	Лица с высшим образованием, %
В целом					
Сторонники	**57**	*43*	*78*	*22*	*17*
Противники	*49*	*45*	*74*	*26*	*15*
В том числе:					
1993 г.					
Март-апрель					
Сторонники	**50**	**43**	**79**	**19**	**17**
Противники	**50**	**46**	**72**	**27**	**18**
Апрель-май					
Сторонники	**52**	**43**	**82**	**24**	**17**
Противники	**51**	**45**	**69**	**26**	**15**
Май-июнь					
Сторонники	**51**	**43**	**77**	**22**	**16**
Противники	**49**	**45**	**76**	**25**	**15**
Июнь-июль					
Сторонники	**52**	**43**	**78**	**35**	**18**
Противники	**48**	**47**	**72**	**37**	**14**
Июль-август					
Сторонники	**54**	**43**	**77**	**21**	**18**
Противники	**47**	**45**	**75**	**21**	**14**
Август					
Сторонники	**55**	**43**	**77**	**16**	**16**
Противники	**48**	**45**	**72**	**28**	**15**
Сентябрь-октябрь					
Сторонники	**49**	**42**	**80**	**20**	**17**
Противники	**50**	**46**	**73**	**27**	**14**
Октябрь-ноябрь					
Сторонники	**51**	**43**	**76**	**22**	**16**
Противники	**49**	**45**	**75**	**22**	**15**
Ноябрь-декабрь					
Сторонники	**51**	**44**	**73**	**22**	**18**
Противники	**51**	**44**	**74**	**22**	**14**

Окончание таблицы 1

Группы респондентов	Время опросов				
	Мужчины, %	Средний возраст, лет	Жители города, %	Лица с образованием ниже среднего, %	Лица с высшим образованием, %
Декабрь 1993 г.— январь 1994 г.					
Сторонники	49	43	**76**	**21**	**19**
Противники	**46**	44	**76**	**24**	**16**
Январь-февраль 1994 г.					
Сторонники	**52**	**42**	**77**	**23**	**16**
Противники	**48**	44	**77**	**22**	**14**

Важный показатель поддержки деятельности президента — наличие его сторонников в регионах. На рис. 2 представлены данные по регионам России, где проводилось исследование.

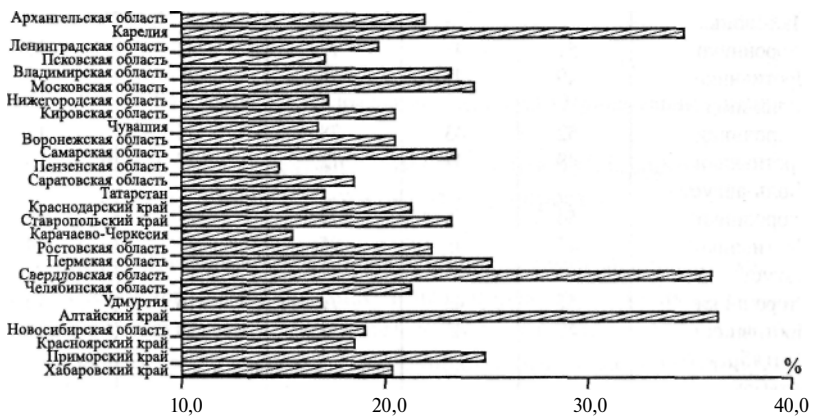


Рис. 2. Доля сторонников президента в регионах (обобщенные результаты опросов в марте 1993 г. — феврале 1994 г.)

Как и следовало ожидать, максимум поддержки (35% опрошенных) президент имеет в Свердловской области. Труднее интерпретировать данные по другим областям. Однако полученные результаты нельзя рассматривать

как точное отражение оценок деятельности президента в разных областях, краях и автономных республиках, поскольку выборка представительна для России в целом, но не репрезентативна для регионов.

Наряду с вопросом об отношении к деятельности президента в анкете был вопрос: как вы думаете, способствует ли выходу из внешнего кризиса деятельность правительства России. Корреляция ответов на два эти вопроса очень высока (табл. 2). Все коэффициенты сопряженности Крамера высоко значимы и демонстрируют сильную взаимосвязь оценок деятельности правительства и президента. При этом характер зависимости отражает сходство оценок: более половины респондентов дают одинаковую оценку правительству и президенту (высокие значения коэффициента сопряженности могут быть и в ситуации устойчиво противоположных оценок).

Таблица 2. Коэффициенты сопряженности Крамера между ответами на вопрос об оценке деятельности президента и правительства России

Время исследования	Значение коэффициента
1993 г.	
Март-апрель	**M!**
Апрель-май	**0,41**
Май-июнь	0,39
Июнь-июль	**0,40**
Июль-август	**0,43**
Август	0,40
Сентябрь-октябрь	**0,42**
Октябрь-ноябрь	0,48
Ноябрь-декабрь	**0,45**
Декабрь 1993 г. — январь 1994 г.	**0,54**
Январь-февраль 1994 г.	**0,37**

В анкетах девяти (из одиннадцати) исследований, наряду с вопросами об оценке деятельности президента и правительства России, были вопросы об оценке деятельности Р.И. Хасбулатова и Верховного Совета Российской Федерации, 15,8% респондентов затруднились ответить на все четыре вопроса. Очевидно, это политически пассивные люди. Разумеется, группа неоднородна, но в соответствии с замечанием П. Бурдье [2] можно предположить, что ее членов отличает прежде всего низкий уровень образования. Действительно, лиц с высшим и незаконченным высшим образованием в группе 7,2%, в то время как в среднем по всему массиву — 14,2, а лиц с образованием ниже среднего — 15,9% (в среднем по массиву — 9,7%).

Средний возраст политически пассивных респондентов практически не отличается от среднего возраста всех опрошенных. Это странно, поскольку для нашей страны характерна достаточно высокая обратная связь между возрастом и уровнем образования. Что касается половой принадлежности, то в группе отмечаются существенные отличия от остального массива: среди политически пассивных 70% женщин (в среднем по массиву — 55%).

К числу политически пассивных можно отнести и тех, кто ответил, что позитивно оценивает деятельность всех рассматриваемых ветвей власти. Учитывая, что поведение последних в течение года кардинально различалось, можно сказать, что полная поддержка свидетельствует либо о неинформированности, либо о равнодушии респондентов. Численность этой группы очень невелика (0,6%). По уровню образования она сходна с первой группой, но женщин там меньше — 36%.

В последнее время получил распространение тезис о росте политической пассивности россиян. На рис. 3 показано, как изменилась численность политически пассивной группы населения по результатам опросов (группы 1 и 2 объединены). За девять месяцев исследования эти изменения незначительны и лежат в границах доверительного интервала[3].

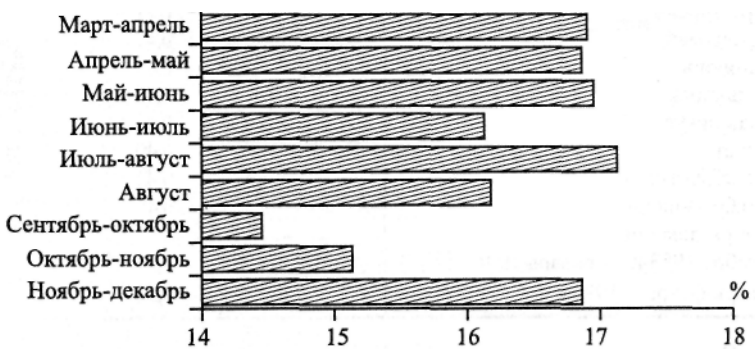


***Рис.** 3.* Численность группы политически пассивных респондентов (март — декабрь 1993 г.)

Продолжительное время деятельность президента Б.Н. Ельцина протекала в условиях политической конфронтации с Верховным Советом РФ и его спикером Р.И. Хасбулатовым. Эти политические лидеры не только выступали оппонентами по самым разным вопросам, но и казались политическими противниками. В этой связи представляется возможной следующая версия: уменьшение поддержки президента в определенные периоды происходило за счет перехода части его сторонников в группу поддержки Хасбулатова, и наоборот.

Разумеется, речь идет только о грубой, в рамках описанного выше подхода, оценке политического поведения.

В табл. 3 приведены данные, позволяющие заключить, что эта версия не подтверждается: снижение поддержки одного из лидеров не компенсируется увеличением поддержки другого.

Таблица 3. Отношение россиян к деятельности Б.Н. Ельцина и Р.И. Хасбулатова, %

Время опроса	Число опрошенных	Деятельность Ельцина		Деятельность Хасбулатова	
		одобряют	не одобряют	одобряют	не одобряют
1993г					
Март-апрель	3988	25,7	**26,6**	4,7	47,4
Апрель-май	3992	**27.3**	**27,8**	4,3	50,6
Май-июнь	3902	24,8	**28,7**	5,3	**47,0**
Июнь-июль	3902	**21,4**	33,1	5,3	49,6
Июль-август	3956	17,5	31,7	**6,7**	**43,9**
Август	3905	**16,1**	35,7	5,9	**46,5**
Сентябрь-октябрь	3962	**19,8**	**35,0**	**6,0**	**50,4**
Октябрь-ноябрь	3984	**22,9**	**27,9**	**2,6**	**54,7**
Ноябрь-декабрь	3941	**20,7**	**32,8**	**2Д**	37,7
Всего	35 532	21,8	**31,0**	**4,8**	47,5

Можно сказать, что систематическое отслеживание политической ситуации в стране, изучение направлений и причин трансформаций политических настроений россиян требует проведения постоянных сравнительных и панельных исследований с использованием сопоставимых методик и репрезентативных выборок. Только такой подход позволит качественно анализировать изменения политических ориентации населения.

Литература

1. Заславская Т.П. Социологический мониторинг экономических и социальных перемен в России // Экономические и социальные перемены: мониторинг общественного мнения. № 1. М.: Аспект Пресс, 1993. С. 3—10.

2. Бурдье П. Социология политики. М.: Socio-Logos, 1993.

Первая публикация: Социологический журнал. 1994. № 3. С. 144—150.

Ограничения метода регрессионного анализа

А. О. Крыштановский

В статье рассматриваются некоторые проблемы, связанные с использованием регрессионного анализа в социологии. Обсуждаются ограничения, обусловленные неравенством дисперсий (гетероскедастичностью) и мультиколлинеарностью в регрессионных моделях. Предлагается несколько подходов к снижению последствий нарушения этих ограничений.

Ключевые слова: регрессионная модель, линия регрессии, коэффициент детерминации, фиктивные переменные, гетероскедастичность, коэффициент Спирмена, мультиколлинеарность.

Построение регрессионных моделей на сегодняшний день, несомненно, является наиболее широко применяемым методом многомерного статистического анализа социологических данных. За последние несколько лет более половины статей, анализирующих эмпирические данные, в таких американских социологических журналах, как American Journal of Sociology и American Sociological Review, основаны на использовании регрессионных моделей.

Достаточно распространены регрессионные методы и среди российских социологов, специалистов, использующих опросные методики. Вместе с тем многие особенности и ограничения регрессионных моделей обычно остаются вне сферы внимания исследователей, что подчас приводит к неточным либо просто ошибочным результатам. В данной статье рассматриваются некоторые особенности использования регрессионных методов при анализе данных массовых опросов.

Проблема недостаточности одного уравнения

Традиционная модель множественного линейного регрессионного анализа подразумевает поиск показателей (обозначаемых *X),* определяющих значение отдельной количественной переменной, обозначаемой *Y.* Структура связи в данной модели предполагается линейной. Иными словами, ищется следующая форма зависимости:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n + U,$$

$$Y=B_0 + BX_{\prime} + B\hat{}_1 + \ldots + вх_n + u, \quad (1)$$

где *U*— так называемый остаточный член, фиксирующий ту часть информации *Y,* которая не объясняется иксами.

Регрессионный анализ показывает, во-первых, качество модели, т.е. степень того, насколько данная совокупность иксов объясняет *Y.* Показатель качества называется коэффициентом детерминации R^2 и показывает, какой процент информации У можно объяснить поведением иксов. Во-вторых, регрессионный анализ вычисляет значения коэффициентов *B.,* т.е. определяет, с какой силой каждый из *X.* влияет на *Y.*

Методологическим недостатком такого подхода является то, что данная зависимость ищется единой для всей совокупности опрошенных респондентов. Иными словами, мы предполагаем, что для всех людей характер зависимости У от иксов единый. В том случае, когда выборочная совокупность достаточно однородна, такого рода допущение имеет под собой определенные основания. Однако, если анализируются, скажем, детерминанты электоральных предпочтений на основе данных всероссийской выборки, допущение об однородности детерминант для чукотского оленевода и для московского профессора выглядит не очень убедительным.

Единая форма уравнения в этой ситуации сильно огрубляет реальную зависимость, качество модели неизбежно оказывается весьма низким, а смысл регрессионных коэффициентов, фиксирующих степень влияния иксов на *Y,* можно приравнять к пресловутому показателю «средней температуры по больнице».

Вполне очевидно, что гораздо разумнее строить отдельные модели для существенно различающихся между собой групп респондентов. Однако доведение такого подхода до логического завершения чревато опасностью полного релятивизма. Действительно, всегда можно найти более или менее убедительные аргументы в пользу того, что по анализируемой проблеме механизмы формирования оценок различны у женщин и мужчин, у горожан и сельских жителей, у инженеров и рабочих и т.д. и т.п. Следовательно, для каждой группы необходимо строить свою модель, что не очень конструктивно, поскольку количество таких моделей ограничивается лишь фантазией социолога по разбиению всей совокупности на отдельные группы.

Оказывается, однако, что есть определенные формальные критерии, позволяющие определять границы групп, для которых действуют одинаковые либо различные механизмы. Рассмотрим такие критерии на примере простейшей задачи.

В качестве зависимой переменной мы взяли придуманный нами в учебных целях индекс зажиточности, измеряемый по количеству предметов длительного пользования, которые есть у респондента в семье[1]. Задачей являлось определение степени влияния возраста на этот индекс. Данные взяты из всероссийского социологического опроса, проведенного ВЦИОМ в ноябре 1999 г. по репрезентативной национальной выборке. Объем выборки — 2388 человек.

Сам индекс зажиточности — это просто сумма ответов респондента по каждой из отмеченных в вопросе 19 позиций. Естественно, что данный индекс фиксирует лишь количественный, но не качественный аспект зажиточности, поскольку и проигрыватель дисков, и автомобиль, и дача входят в этот индекс с одинаковым весом. Однако мы рассматриваем этот индекс исключительно как инструмент для демонстрации метода.

График зависимости индекса зажиточности от возраста приведен на рис. 1. Представленная здесь регрессионная модель выглядит следующим образом:

$$\text{Индекс зажиточности} = 5{,}97 - 0{,}049 \times \text{Возраст}. \qquad (2)$$

Качество полученной модели не очень высоко — коэффициент детерминации равен 0,097, но, с другой стороны, этот коэффициент значим с $P > 0{,}999$, с той же вероятностью значимы регрессионные коэффициенты и, если взглянуть на результаты оптимистически, можно сказать, что возраст почти на 10% определяет степень зажиточности российского населения.

Построенная модель дает нам единый механизм влияния возраста на индекс зажиточности независимо от значения возраста. Другими словами, модель утверждает, что с увеличением возраста на 10 лет респондент в среднем теряет 0,5 вещи, независимо от того, 20 лет респонденту или 70. Однако жизненный опыт и здравый смысл подсказывают, что это, скорее всего, не

[1] Вопрос анкеты выглядел следующим образом: отметьте, пожалуйста, в приведенном ниже списке вещи, которые есть в вашей семье:

1) цветной телевизор;	2) фотоаппарат;
3) радио-часы;	4) миксер;
5) электродрель;	6) стерео-, радиосистема;
7) отдельный морозильник;	8) микроволновая печь;
9) видеомагнитофон;	10) видеокамера;
11) пылесос;	12) домашний компьютер;
13) пианино (фортепиано);	14) новый автомобиль;
15) подержанный автомобиль;	16) дача, дом на садовом участке;
17) дом в деревне;	18) участок, где вы выращиваете овощи, фрукты;
19) проигрыватель компакт-дисков;	20) нет ничего из перечисленного.

так. Действительно, можно предположить, что в 20—40 лет респонденты скорее увеличивают количество вещей, а в пожилом возрасте, в силу сокращения доходов, уже начинают терять. Полученный же средний коэффициент — «минус 0,5 вещи за 10 лет», таким образом, вообще ни к кому не применим, это то усреднение, которое фактически ничего не описывает.

Рис. 1. Взаимосвязь возраста и индекса зажиточности (количества вещей, имеющихся в семье респондента)[2]

Проблема, которая следует из предыдущего рассуждения, следующая: если делить весь жизненный цикл респондента (с точки зрения индекса зажиточности) на два этапа, то где находится это пороговое значение, где кончается один этап и начинается другой? Эту постановку проблемы можно перевести на язык регрессионной модели следующим образом. Если строить для двух совокупностей респондентов две регрессионные модели, то как определить, когда эти две модели отличаются друг от друга и где это отличие максимально?

Для решения этой задачи существует специальный статистический тест, называемый тестом Чоу. Он показывает, является ли значимым улучшение

[2] Из анализа были исключены респонденты младше 20 лет и старше 80. После такого исключения объем выборки составил 2233 респондента.

качества регрессионной модели после разделения выборки [1, с. 282—285]. Для этого используется F-статистика, вычисляемая следующим образом:

Улучшение качества модели / Использование степени свободы

Необъясненная дисперсия / Число остающихся степеней свободы

или

$$(U_P\text{-}U_A\text{-}U_B)/(k \quad + \quad l)$$
$$(U_A+U_B)/(n\text{-}2k\text{-}2)'$$

где U — сумма квадратов остатков для единой модели; U_A — сумма квадратов остатков для первой модели; U_B — сумма квадратов остатков для второй модели; $к = 1$ (в данном примере); n — объем выборки.

Получаемая таким образом F-статистика имеет ^-распределение с $(к+1)и(п\text{-}2хк\text{-}2)$ степенями свободы и позволяет определить статистическую значимость улучшения объясняющей силы модели при переходе от одного уравнения к двум.

Таким образом, возвращаясь к анализируемому примеру, можно разделить возрастную шкалу на два интервала, построить для каждого из этих интервалов свою линию регрессии и с помощью теста Чоу определить, произошло ли улучшение качества модели. Проблема, однако, состоит еще и в том, как выбрать точку разбиения.

Действительно, можно разбить возраст на интервалы «20—25 лет» и «старше 25 лет» и получить, что тест Чоу значим. Можно предложить какое-либо другое разбиение и получить тот же результат (в ходе наших экспериментов с данной моделью оказалось, что почти любое разбиение дает значимые различия по тесту Чоу). Эта закономерность имеет положительную сторону. Она означает, что две отдельных линии регрессии почти всегда лучше описывают реальную ситуацию, чем одна единственная, и это, как представляется, немаловажный результат.

С другой стороны, мы не получаем ответа на вопрос, на какие же все-таки интервалы лучше разбить возрастную шкалу. Решение данной проблемы в свете вышеизложенного выглядит достаточно просто. Необходимо перебрать все возможные, разумные с социологической точки зрения, разбиения и взять то из них, которое дает наибольшее увеличение показателя качества модели, основываясь на F-статистике теста Чоу.

В табл. 1 представлены значения F-статистики для нескольких предпринятых разбиений.

Для всех значений F-статистики в число степеней свободы одинаково — (2,2229). Они (значения F-статистики) значимы на 0,1% уровне и, поскольку число степеней свободы одинаково, мы можем сравнивать значения F-статистики между собой и выбирать максимальное. Как видно из табл. 1, наилучшим разбиением являются интервалы «20—44 года», «45 лет и старше».

На рис. 2 показана модель с двумя линиями регрессии при разбиении шкалы возраста на эти два интервала. Как же выглядят две полученные линии регрессии? Первая из них (для интервала возраста «20—44 года») дает значение коэффициента детерминации $R^2 = 0{,}002$, которое соответствует незначимой (с. $P > 0{,}18$) величине дисперсионного F-отношения. Иными словами, для респондентов из этого возрастного интервала нет значимого влияния возраста на значение индекса зажиточности, и для них этот индекс — просто константа, равная 4,8.

Таблица 1. Значения ^-статистики для различных разбиений шкалы возраста

Точки разбиения возраста на интервалы	Значения .F-статистики
40	20,63
41	21,78
42	22,84
43	23,99
44	26,22
45	24,98
46	25,80
47	24,24
48	25,56
49	25,56
50	25,97
51	25,70
52	25,46
53	25,55
54	24,16
55	24,16

Рис. 2. Взаимосвязь возраста и индекса зажиточности (количества вещей, имеющихся в семье респондента) и модель двух уравнений регрессии

Для второго интервала возраста (45 лет и старше) коэффициент детерминации $K^2 = 0{,}172$, регрессионная зависимость высоко значима и уравнение выглядит следующим образом[3]:

$$\text{Индекс зажиточности} = \underset{(0{,}4)}{8{,}97} \; \underset{(0{,}007)}{-0{,}097} \times \text{Возраст}. \qquad \textbf{(4)}$$

Подводя итог, можно констатировать, что по сравнению с моделью, описывающей зависимость одним уравнением, перейдя к двум отдельным уравнениям, мы получили гораздо более адекватную картину. В интервале до 45 лет возраст не влияет на индекс зажиточности, а начиная с 45 лет значение индекса падает со скоростью приблизительно «минус одна вещь в 10 лет».

Отметим, что использование техники фиктивных *(dummy)* переменных позволяет не только записать полученные нами два уравнения в виде одного, но и сразу проводить оценивание регрессионной модели для двух разделенных по возрасту совокупностей.

[3] В скобках под значениями коэффициентов регрессии приводятся величины стандартных ошибок, позволяющие оценить доверительные интервалы.

Для нашего примера регрессионное уравнение будет выглядеть следующим образом:

$$\text{Индекс зажиточности} = 4{,}8 + 4{,}1\ D - 0{,}1\ D \times \text{Возраст}, \qquad (5)$$

где D — фиктивная переменная, построенная следующим образом: $D = 0$ — для респондентов 20—44 лет, $D = 1$ — для респондентов 45—80 лет.

Очевидно, что уравнение (5) объединяет в рамках одной модели и уравнение (4), и тот факт, что у респондентов до 45 лет индекс зажиточности от возраста не зависит и равен константе — 4,8. Полученное для модели (5) значение коэффициента детерминации $R^2 = 0$, Ц 8.

Представляется, однако, что хотя запись модели в виде единого регрессионного уравнения, несомненно, удобнее, модель (5) имеет принципиальный недостаток. Получение одного значения коэффициента детерминации скрывает от нас тот факт, что для одной части опрошенных вообще нет значимой зависимости индекса зажиточности от возраста, а для другой части эта зависимость есть. При таком подходе гораздо естественнее и полезнее было бы вычисление не единого R^2, а отдельных значений этого коэффициента для двух возрастных совокупностей.

Гетероскедастичность

Еще одну проблему, возникающую при использовании метода регрессионного анализа по отношению к социологическим данным, высвечивает попытка изучить взаимосвязь того же индекса зажиточности с доходом респондента. В этом примере в качестве икса взята переменная «суммарный доход семьи респондента». Данные — тот же массив всероссийского репрезентативного опроса, проведенного ВЦИОМ в ноябре 1999 г.

На рис. 3 демонстрируется зависимость количества вещей в семье респондента от суммарного месячного дохода[4]. Построенная модель, на первый взгляд, достаточно хороша, поскольку коэффициент детерминации $R^2 = 0{,}26$. Само уравнение выглядит следующим образом:

[4] Из анализа исключены респонденты, чей суммарный месячный доход менее 250 руб. и более 9000 руб. Такого рода исключение является стандартной процедурой при обработке данных о доходах/расходах, когда исключаются 1—3% наибольших и наименьших доходов как либо не являющихся достоверными, либо редко встречающихся, т.е. не типичных.

Индекс зажиточности = 2,12 + 0,0009 X Суммарный доход. (6)
(0,08) (0,00003)

Таким образом, с ростом дохода семьи на 1000 руб. количество вещей увеличивается приблизительно на единицу. Рисунок 3 показывает, однако, что при построении регрессионного уравнения нарушается одно из ограничений метода — требование гомоскедастичности, или второе условие Гаусса — Маркова (1, с. 79—82].

Суть этого ограничения проста: разброс точек вокруг линии регрессии должен быть достаточно равномерен по всей протяженности линии икса. График (см. рис. 3) показывает, что это требование нарушено. При небольших значениях X (при невысоких размерах суммарного дохода) отклонения кривой от линии регрессии относительно невелики, но с увеличением дохода возрастают и отклонения.

Рис. 3. Зависимость количества вещей, имеющихся в семье, от суммарного дохода семьи и линия регрессии

Прежде всего отметим, что в тех случаях, когда в качестве зависимых переменных выступают деньги (заработок, суммарный доход и т.п.), то традиционно более эффективным является использование в регрессионном уравнении логарифма от зависимой переменной. Связано это с тем, что воздействие величины прироста (либо уменьшения) дохода на большинство социологических показателей зависит не только от величины прироста, но и от того значения, к которому этот прирост (уменьшение) происходит. Действительно, увеличение дохода на 100 руб. достаточно существенно для семей, имеющих доход в 500 руб. И такое же увеличение (уменьшение) мало заметно для семей с доходом в 10 000 руб.

Переход к логарифму дохода вместо дохода в качестве зависимой переменной в уравнении (6) улучшает качество регрессионной модели — $R^2 = 0,3$, однако принципиально ничего не меняет. Отклонения реальных значений индекса зажиточности от предсказываемых моделью, во-первых, остаются во многих случаях достаточно большими, и, во-вторых, эти отклонения не постоянны по оси X. На рис. 4 представлен график роста дисперсии отклонений реальных значений индекса зажиточности от регрессионной кривой с логарифмом суммарного дохода в качестве независимой переменной.

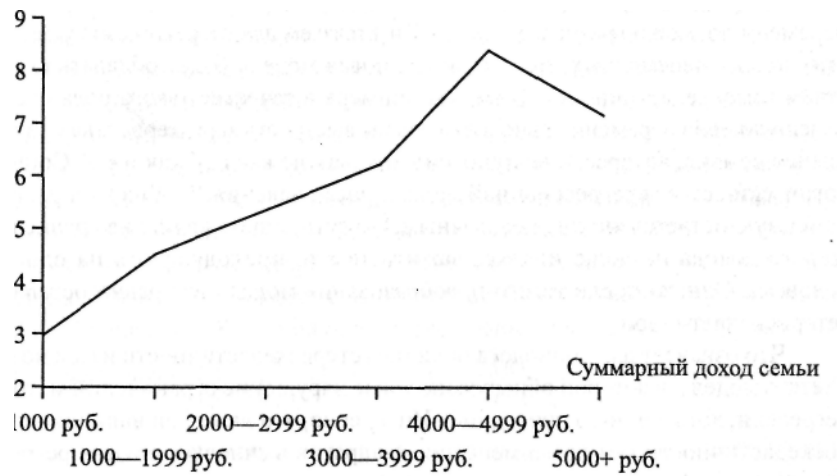


Рис. 4. Дисперсия остатков между значениями индекса зажиточности и регрессионной кривой

Если не ограничиваться визуальной констатацией нарушения требования гомоскедастичности, можно использовать статистические тесты, которые покажут наличие/отсутствие нарушения данного ограничения. Одним из возможных тестов в данной ситуации является тест ранговой корреляции Спирмена.

Суть теста Спирмена для решения поставленной задачи достаточно проста. Ранжируются все значения X (в нашем случае — значения суммарного дохода), затем все значения остатков — отклонений индекса зажиточности от регрессионной кривой и, наконец, выясняется вопрос о наличии взаимосвязи в расположении полученных рангов с помощью следующей статистики, называемой коэффициентом Спирмена [2, с. 48]:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N^2(N-1)}.$$

Полученное для анализируемого примера значение коэффициента Спирмена равно 0,83, значимо для $P > 0{,}999$, и это подтверждает визуально установленный факт гетероскедастичности.

В качестве метода борьбы с гетероскедастичностью рекомендуется искать переменные, которые сильно связаны как с *Y,* так и с *X.* Найдя такую переменную, можно разделить на нее *Xн Yn* затем искать регрессию уже для этих новых переменных. Если повезет, новая модель будет обладать свойством гомоскедастичности. В нашем примере в качестве такого рода «компенсирующей» переменной вполне могла бы выступать характеристика «число членов семьи», которая, очевидно, сильно связана как с *X,* так и с 7. Социологический смысл регрессионной модели после деления *X* и Уна данную переменную остается вполне прозрачным. По сути, ищется влияние среднедушевого дохода на долю индекса зажиточности, приходящуюся на одного человека. Однако после такого преобразования модель все равно осталась гетероскедастичной.

Что означает для социолога наличие гетероскедастичности и можно ли считать модель, в которой обнаружено такое нарушение ограничений метода регрессии, хоть в чем-то пригодной? Интересно, что даже при наличии гетероскедастичности метод наименьших квадратов вычисляет оценки регрессионных коэффициентов несмещенными, т.е. уравнение (6) остается корректным в части значений коэффициентов, хотя неверными становятся значения ошибок коэффициентов. Насколько велики эти ошибки, в литературе нам найти не удалось, кроме замечания, что «...дисперсия оценки коэффициента наклона может быть в 3 раза больше при использовании обычного МНК (метод наименьших квадратов) по сравнению с тем случаем, когда делается поправка на гетероскедастичность» [1, с. 215].

Мультиколлинеарность

Еще одной серьезной проблемой, с которой приходится сталкиваться при построении регрессионных моделей в социологии, является проблема зависимости независимых переменных (т.е. иксов) между собой. Напомним, что хотя классический регрессионный анализ предполагает, что иксы независимы между собой, в любых реальных приложениях оказывается, что так бывает достаточно редко. Действительно, как правило, между иксами есть корреляция, и, подчас, достаточно высокая. Само по себе это является нарушением регрессионной модели и носит название мультиколлинеарности.

При каких значениях взаимосвязи между иксами можно сказать, что мы сталкиваемся с проблемой мультиколлинеарности? В некоторых работах можно встретить рекомендации, указывающие пороговое значение как 0,7 [3]. Однако известно, что «не существует точного граничного значения уровня корреляции переменных, при котором возникает проблема мультиколлинеарности» [4, с. 290]. Рассмотрим конкретный пример, когда данная проблема возникает и какими осложнениями для социолога она чревата.

В качестве зависимой переменной будем рассматривать все тот же индекс зажиточности. Ранее было доказано, что, впрочем, и так очевидно, что на значение этого индекса оказывает существенное влияние суммарный доход семьи. Однако вполне содержательным является и следующий социологический вопрос. Что влияет на индекс зажиточности сильнее — суммарный доход или среднедушевой? Для решения этого вопроса можно построить регрессионную модель, в которой в качестве независимых переменных будут выступать два этих показателя. Для того же массива данных, полученных ВЦИОМ в ноябре 1999 г., получается следующее регрессионное уравнение:

Индекс зажиточности = 2,2 + 0,001 x Суммарный доход - 0,00057 x Среднедушевой доход. *rq*
(0,08x0,00003) (0,00014)

Коэффициент детерминации для модели (7) равен 0,27, все коэффициенты в уравнении значимы с $P > 0{,}999$. Сама модель дает вполне ожидаемый ответ на поставленный вопрос о степени важности двух рассматриваемых показателей для индекса зажиточности. С ростом суммарного дохода (при постоянном среднедушевом) индекс зажиточности растет со скоростью «одна вещь на 1000 руб». Иными словами, при постоянном среднедушевом доходе, т.е. при одновременном увеличении суммы дохода и числа членов семьи, индекс зажиточности возрастает. С другой стороны, при фиксированном суммарном доходе увеличение среднедушевого дохода ведет к умень-

шению индекса зажиточности со скоростью «0,6 вещи на 1000 руб.». Этот факт менее очевиден. Его можно попытаться проинтерпретировать так: при фиксированном суммарном доходе увеличение среднедушевого говорит об уменьшении размера семьи и, соответственно, в меньшей семье будет меньше вещей.

При этом уравнение (7) показывает, что положительное влияние суммарного дохода почти в 2 раза выше, чем отрицательное влияние среднедушевого дохода.

Таким образом, наша модель дала достаточно естественные, с социологической точки зрения, результаты, и можно было бы этим удовлетвориться. Однако, если взглянуть на коэффициент корреляции между суммарным и среднедушевым доходами, то он окажется весьма высоким — 0,77, и, следовательно, мы имеем дело с мультиколлинеарностью модели (7). Чем это грозит?

Основной недостаток регрессионной модели в случае мультиколлинеарности — неустойчивые значения коэффициентов модели. Мы провели численные эксперименты с формированием 100 случайных 50% подвыборок и вычислением для каждой из них моделей типа модели (7). В 38% случаев коэффициент при показателе «среднедушевой доход» давал значения 95% доверительного интервала, отличающиеся от истинного, в качестве которого у нас выступали значения коэффициента для полной выборки. Следовательно, вполне можно допустить, что и сама модель (7) с большой вероятностью даст неверные значения коэффициентов при переносе результатов на генеральную совокупность.

Подводя итог, можно сказать следующее. Современные статистические пакеты сделали техническую сторону обработки данных массовых опросов и социологических исследований весьма простой и доступной. Для того чтобы выполнить факторный, или регрессионный, или какой-либо другой анализ, достаточно несколько раз нажать на соответствующие иконки и, вроде бы, получить готовый результат. Однако на самом деле все значительно сложнее. Обработка данных, относящихся к любой предметной области, требует как знания существа и специфики методов многомерного статистического анализа, так и хорошей подготовки в самой предметной области. Продемонстрированные в статье сложности применения регрессионных моделей в социологии — только небольшой пример тех проблем, которые возникают, если серьезно подходить к анализу данных.

-i.Ubiv-' • ' | V" V i"i-It. • ';>!.(i'K'f'Л'-'-.i"i'"i:,"- •'••••ii'•'"•.•h.T':•>№"•

Литература

1. Доугерти К. Введение в эконометрику. М.: ИНФРА-М, 1999.

2. ГлинскийВ.В., Ионин В.Г. Статистический анализ. М.: Филинъ, 1998.

3. Lewis-Beck M. Applied Regression: An Introdaction. SageUniver. Series Paper on Quantitative Applications in the Social Sciences, 07-022. Beverly Hills, CA: Sage, 1980.

4. Уотшем Т.Дж., Паррамоу К. Количественные методы в финансах. М.: ЮНИТИ, 1999.

Первая публикация: Социология 4М. 2000. № 12. С. 96—112.

«Кластеры на факторах» — об одном распространенном заблуждении

А.О. Крыштановский

Статья посвящена обоснованию некорректности проведения классификации объектов на данных, полученных после факторизации переменных. Эмпирическое обоснование проводится на тестовой матрице данных, сформированной с помощью датчика случайных чисел. В матрице представлены две разные модельные группы респондентов, характеризующиеся 15 переменными.

Ключевые слова: метод факторного анализа, метод главных компонент, кластерный анализ, датчик случайных чисел.

Задача построения классификации единиц исследования весьма распространенна как в социологических, так и в маркетинговых исследованиях. Получение *однородных групп* объектов (респондентов), т.е. таких групп, которые приблизительно одинаково ведут себя в одинаковых ситуациях (одинаково отвечают на вопросы анкеты) — типичная задача сегментирования.

Определенной проблемой при этом является то, что количество параметров, по которым требуется достижение однородности, во многих случаях весьма велико (нередко — несколько десятков). В этой ситуации непосредственная классификация объектов (как правило, с использованием методов кластерного анализа) приводит к плохо интерпретируемым результатам. Действительно, кластерный анализ методом ^-средних (без задания содержательно осмысленных центров кластеров) в качестве исходных точек выбирает максимально далеко отстоящие друг от друга точки, которые на практике часто действительно трудно интерпретируемы. Далее, весь массив разделяется на однородные группы с точки зрения близости к этим «непонятным» объектам. Нет ничего удивительного, что результат становится маловразумительным.

Распространенным подходом в данной ситуации считается двухстадийный метод, когда на первом этапе к исходным признакам применяется факторный анализ с целью получения некоторых латентных показателей (факторов), объединяющих в некоторые группы (факторы) сами признаки. На втором этапе используют кластерный анализ для получения некоторых групп, однородных в смысле средних величин индивидуальных значений построенных факторов.

На первый взгляд такой подход представляется вполне логичным и естественным. Действительно, в данном случае мы проводим кластеризацию

небольшого количества исходных признаков, при котором специфическое поведение даже одного из этих признаков может вызвать сильное смещение результирующей кластеризации, а классифицируем объекты по 3—4 переменным (факторам), каждая из которых при этом имеет более или менее вразумительную интерпретацию.

Данный подход, по нашим наблюдениям, достаточно широко используется в практике как социологических, так и маркетинговых исследований. Такой путь рекомендуется и в достаточно широко распространенной книге А. Бююля и П. Цёфеля[1]. К сожалению, внешняя логичность такого подхода не учитывает базовых положений метода факторного анализа, которые, как нам представляется, приводят к тому, что из такого рода классификаций хоть сколь-нибудь обоснованных выводов получиться не может.

Для иллюстрации высказанных соображений был проведен описанный ниже эксперимент.

Массив данных

С помощью датчика случайных чисел был создан тестовый массив из 500 объектов, содержащий ответы двух групп респондентов на 15 вопросов. Файл синтаксиса SPSS по созданию массива приведен ниже.

```
IF (A = 1) B1 = 10 × NORMAL (1).
IF (A = 2) B1 = 20 + 10 × NORMAL (1).
* A—переменная, определяющая принадлежность объекта к одной из двух групп.
DO REPEAT R = B2 to B15.
IF (A = 1) R = B1 + 20 × NORMAL (1).
IF (A = 2) R = B1 + 20 × NORMAL (1) + 10.
END REPEAT.
```

Средние значения всех переменных в двух группах достаточно сильно различаются между собой (табл. 1), и, следовательно, можно считать, что

[1] Бююль А., Цёфель П. SPSS: искусство обработки информации. М.; СПб.; Киев: DiaSoft, 2002. С. 394—398. В данной главе факторный анализ назван почему-то «факториальным», но это, по всей видимости, редакционные погрешности. Отметим, что такой же подход пропагандируют авторы и в разделе, в котором обсуждается кластерный анализ методом ^-средних (с. 404—409).

эти две группы представляют разные совокупности объектов, что, по логике исследования, должно обнаружиться с помощью методов классификации.

Таблица 1. Статистические характеристики модельных переменных в двух группах

Переменная	Номер группы	Среднее значение	Стандартное отклонение
B1	**1**	-0,88	10,38
	2	19,97	10,03
B2	**1**	-1,64	23,35
	2	31,48	19,86
B3	**1**	-6;93	23,23
	2	28,36	21,95
B4	**1**	-1,32	23,10
	2	29,78	22,32
B5	**1**	-0,90	23,99
	2	33,07	22,66
B6	**1**	-0,49	22,71
	2	31,70	22,15
B7	**1**	-0,12	22,52
	2	30,25	22,44
B8	**1**	-1,14	21,92
	2	31,27	23,26
B9	**1**	-0,75	21,79
	2	31,00	22,94
B10	**1**	0,16	21,77
	2	29,09	21,88
B11	**1**	-2,24	21,96
	2	31,29	23,70
B12	**1**	-0,82	23,16
	2	28,06	22,48
B13	**1**	0,40	24,70
	2	29,43	20,11
B14	**1**	-0,42	21,21
	2	31,95	22,41
B15	**1**	-1,99	22,15
	2	29,62	21,64

Примечание. Средние значения всех переменных различаются с вероятностью P>0,99.

В табл. 2 представлены значения коэффициентов корреляции для созданных 15 переменных.

Таблица 2. Матрица коэффициентов корреляции Пирсона для 15 модельных переменных

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	ВЦ	B12	B13	B14	B15
B1	1	0,688	0,662	0,700	0,689	0,660	0,661	0,672	0,661	0,683	0,690	0,689	0,699	0,678	0,651
B2	0,688	1	0,479	0,478	0,509	0,495	0,495	0,461	0,434	0,521	0,465	0,481	0,496	0,512	0,481
B3	0,662	0,479	1	0,452	0,464	0,500	0,466	0,441	0,428	0,442	0,437	0,499	0,522	0,459	0,433
B4	0,700	0,478	0,452	1	0,507	0,448	0,449	0,491	0,498	0,477	0,528	0,508	0,500	0,460	0,480
B5	0,689	оде	0,464	0,507	1	0,478	0,459	0,453	0,456	0,510	0,486	0,515	0,509	0,475	0,480
B6	0,660	0,495	0,500	0,448	0,478	1	0,484	0,489	0,473	0,461	0,437	0,504	0,441	0,474	0,459
B7	0,661	0,495	0,466	0,449	0,459	0,484	1	0,490	0,457	0,461	0,475	0,474	0,446	0,450	0,467
B8	0,672	0,461	0,441	0,491	0,453	0,489	0,490	1	0,442	*от*	0,486	0,424	0,435	0,479	0,469
B9	0,661	0,434	0,428	0,498	0,456	0,473	0,457	0,442	1	0,421	0,454	0,475	0,477	0,492	0,436
B10	0,683	0,521	0,442	0,477	0,510	0,461	0,461	0,504	0,421	1	0,502	0,449	0,481	0,472	0,440
B11	0,690	0,465	0,437	0,528	0,486	0,437	0,475	0,486	0,454	0,502	1	0,521	0,450	0,505	0,483
B12	0,689	0,481	0,499	0,508	0,515	0,504	0,474	0,424	0,475	0,449	0,521	1	0,463	0,484	0,494
B13	0,699	0,496	0,522	0,500	0,509	0,441	0,446	0,435	0,477	0,481	0,450	0,463	1	0,484	0,478
B14	0,678	0,512	0,459	0,460	0,475	0,474	0,450	0,479	0,492	0,472	0,505	0,484	0,484	1	0,483
B15	0,651	0,481	0,433	0,480	0,480	0,459	0,467	0,469	0,436	0,440	0,483	0,494	0,478	0,483	1

Примечание. Все коэффициенты значимы на уровне больше 0,01.

Факторный анализ

/Для матрицы корреляций (см. табл. 2) использовался факторный анализ с помощью метода главных компонент[2], результаты которого приведены в табл. 3.

Как видно из табл. 3, четыре первых фактора объясняют почти 67% вариации. Такой процент объясненной дисперсии, как правило, считается вполне приемлемым при использовании факторного анализа в социологических и маркетинговых исследованиях. Таблица 3 показывает также, что данный факторный анализ не является особенно удачным, поскольку общности демонстрируют неравномерность в объяснении дисперсии отдельных переменных (особенно для переменных B14, B15 по сравнению с переменной B1), и, по

[2] Строго говоря, метод главных компонент не является методом факторного анализа, однако мы использовали его сознательно, поскольку именно этот метод чаще всего применяется при решении прикладных задач.

всей видимости, было бы целесообразно увеличить число факторов. Однако, учитывая «модельность» примера, мы не будем этого делать, тем более что, по нашим наблюдениям, на значения общностей исследователи внимания чаще всего вообще не обращают.

Таблица 3. Матрица факторных нагрузок, процент объясненной дисперсии, общности для модельных данных

	Факторы				**Общности**
	1	**2**	3	4	
B1	0,959	-0,007	-0,016	-0,020	0,920
B2	0,730	0,054	0,219	-0,218	0,631
B3	0,699	| 0,457 |	0,116	-0,068	0,716
B4	0,728	-0,171	| -0,3011	-0,036	0,652
B5	0,729	0,020	-0,092	-0,260	0,609
B6	0,711	0,224	0,260	0,277	0,699
B7	0,704	-0,007	0,280	0,245	0,634
B8	0,705	|-0,318	0,293	0,164	0,710
B9	0,692	0,071	-0,296	0,340	0,687
B10	0,714	-0,249	0,232	-0,326	0,731
B11	0,723	-0,334	-0,177	0,020	0,665
B12	0,728	0,145	-0,220	0,117	0,614
B13	0,719	0,243	-0,143	|-0,312 |	0,694
B14	0,721	-0,039	-0,042	0,059	0,527
B15	0,704	-0,080	-0,092	0,051	0,513
Процент объясненной дисперсии	**53,8**	4,4	4,3	4,2	

На рис. 1 представлены гистограммы распределения построенных индивидуальных значений факторов. Общий вид распределений напоминает нормальные кривые.

Обратим внимание еще на одну распространенную ошибку в интерпретации результатов факторного анализа. Как правило, исследователи устанавливают определенную «точку отсечения» для значений факторных нагрузок и значения, меньшие этой точки, в интерпретации не участвуют. Рассмотрим, например, матрицу, приведенную в табл. 3. Если установить в качестве «точки отсечения» значение 0,3, для объяснения, скажем, 3-го фактора будут использоваться переменные B9, B10, B13. Далее индивидуальные значения 3-го фактора будут рассматриваться именно как индекс, характеризующий поведение данных трех переменных.

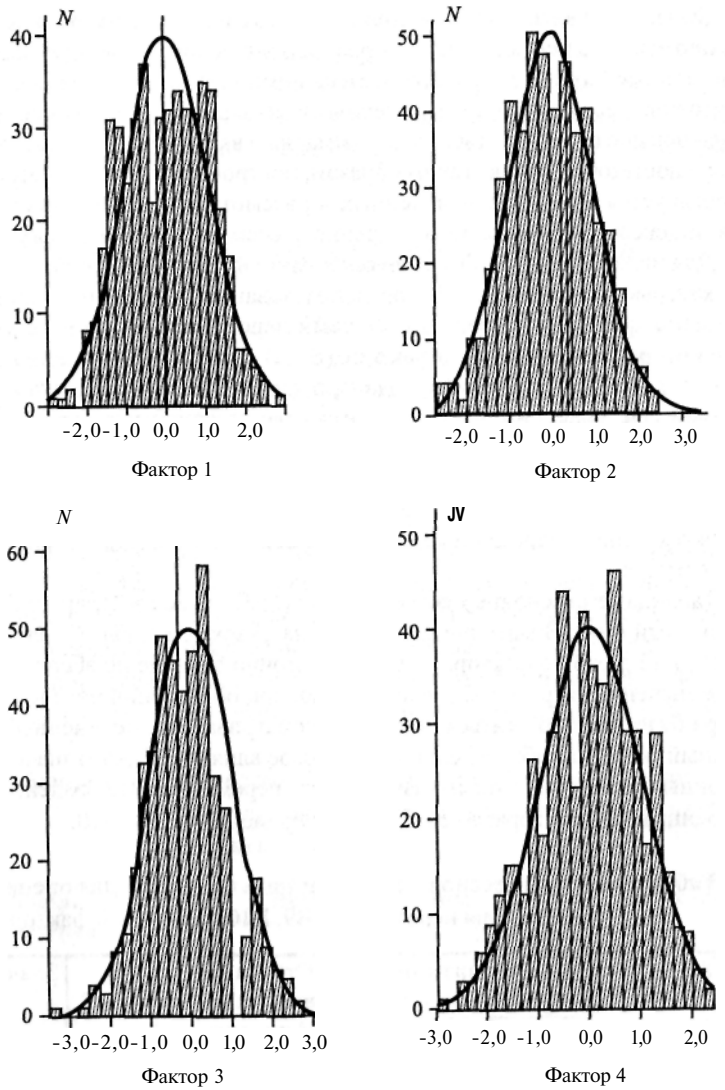


Рис. 1. Гистограммы распределения индивидуальных значений первых четырех факторов для анализа главных компонент табл. 3

Однако при вычислении индивидуальных значений фактора используются не только те переменные, которые легли в основу интерпретации, но и все остальные (хотя и, разумеется, с меньшими весами). Проблема состоит в том, что хотя веса остальных переменных и меньше, однако этих переменных гораздо больше и соответственно их суммарный вклад в полученные значения фактора достаточно велик. Таким образом, построенный фактор, который интерпретируется на основе включенных в рассмотрение переменных, становится индексом, отражающим поведение совсем иных переменных.

Для иллюстрации этой мысли оценим то, на сколько значения 3-го фактора, которые вычисляет SPSS при использовании стандартной процедуры сохранения факторов, определяются теми переменными, которые участвовали в интерпретации данного фактора (B9, B10, B13). Для решения этой задачи построим регрессионную модель, в которой в качестве зависимой переменной выступает 3-й фактор, а в качестве независимых переменных — переменные, которые легли в основу интерпретации фактора (B9, B10, B13).

Коэффициент R^2, определяющий качество такой модели, в нашем примере составил 0,26. Иными словами, лишь 26% поведения 3-го фактора объясняются теми тремя переменными, которые используются для интерпретации этого фактора.

Таблица регрессионных коэффициентов (табл. 4) демонстрирует еще один любопытный факт. В матрице факторных нагрузок (см. табл. 3) переменные B9, B10, B13 для 3-го фактора имеют достаточно близкие по абсолютной величине значения нагрузок и, следовательно, при объяснении поведения этого фактора будет предполагаться, что три рассматриваемые переменные имеют на данный фактор приблизительно одинаковое влияние. Однако значения регрессионных коэффициентов показывают, что переменная B13 воздействует на построенный фактор гораздо слабее, чем переменные B9 и B10.

Таблица 4. Регрессионные коэффициенты модели для оценки влияния переменных B9, BIO, B13 на 3-й фактор

	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	*t*	Значимость
	B	Стандартная ошибка	Beta		
Константа	0,055	0,047		1,178	0,239
B9	-0,015	0,002	-0,418	-9,235	0,000
B10	0,019	0,002	0,495	10,911	0,000
B13	-0,007	0,002	-0,182	-3,882	0,000

Таким образом, традиционная интерпретация поведения 3-го фактора как индекса, отражающего поведение переменных B9, B10 и B13, дает абсолютно неадекватную картину. Во-первых, эти три переменные объясняют лишь 26% поведения фактора. Во-вторых, степень влияния данных переменных на построенный фактор не может быть объяснена на основе матрицы факторных нагрузок.

Иерархический кластерный анализ

В качестве первого метода классификации созданного модельного массива используем иерархический метод кластерного анализа с разбиением на два кластера. Переменными будут выступать индивидуальные значения четырех построенных факторов. В качестве параметров кластеризации выбираются те, которые предлагаются SPSS по умолчанию.

Дендрограмма, построенная программой иерархического кластерного анализа, не позволяет увидеть две группы, которые заданы в модельном массиве. При разбиении массива на два кластера результат получается абсолютно неудовлетворительный — в одном кластере оказывается один объект, а во втором кластере — 499. Даже разбиение на 7 кластеров показывает, что массив разделяется на один большой кластер, два средних и четыре мелких. При этом принадлежность объектов, исходно относящихся к двум заданным группам по построенным кластерам, достаточно произвольна (табл. 5).

Таблица 5. Количество объектов из двух модельных групп, разнесенное по 10 кластерам (кластеризация на четырех факторах)

Номера кластеров	Исходные группы		Всего
	1	2	
1	237	211	448
2	6	18	24
3	2	20	22
4	0	1	1
5	3	0	**3**
6	1	**0**	1
7	1	0	1
Всего	250	250	500

Таблица 5 показывает, что применение иерархического кластерного анализа для выделения двух модельных групп не дает хоть сколь-нибудь приемлемого результата.

Попробуем провести иерархический кластерный анализ, используя в качестве переменных не построенные ранее факторы, а непосредственно 15 исходных переменных. Результат такой кластеризации представлен в табл. 6, которая показывает, что полученную классификацию можно вполне признать удовлетворительной, поскольку лишь 60 (около 12%) объектов были отнесены к неверным группам.

Таблица 6. Количество объектов из двух модельных групп, разнесенное по двум кластерам (кластеризация на исходных переменных)

Номер кластера	Исходная группа		Всего
	1	2	
1	205	**15**	220
2	**45**	235	280
Всего	250	250	500

Кластерный анализ методом А-средних

Представленная в SPSS команда *k-means* (Аг-средних) является гораздо более технологичной по сравнению с программой иерархического кластерного анализа и соответственно используется гораздо чаще. Вначале проведем разбиение модельного массива на два кластера на построенных ранее факторах, не задавая начальные центры кластеров. Соответствие исходных групп построенным объектам представлено в табл. 7.

Таблица 7. Количество объектов из двух модельных групп, разнесенное по двум кластерам (кластеризация на четырех факторах)

Номер кластера	Исходная группа		Всего
	1	2	
1	**174**	**99**	273
2	**76**	**151**	227
Всего	**250**	**250**	500

Результат кластеризации трудно признать удовлетворительным, поскольку почти 35% объектов были классифицированы ошибочно.

Кластеризация с помощью алгоритма ^-средних при использовании в качестве переменных не построенных факторов, а непосредственно исходных показателей, дает гораздо более приемлемые результаты (табл. 8). При таком разбиении менее 9% объектов классифицируются ошибочно.

Таблица 8. Количество объектов из двух модельных групп, разнесенное по двум кластерам (кластеризация на исходных переменных)

Номер кластера	Исходная группа		Всего
	1	2	
1	220	**13**	233
2	**30**	237	267
Всего	250	250	**500**

Обсуждение результатов

Может создаться впечатление, что основной причиной недопустимо низкого качества кластеризации наших модельных данных при использовании индивидуальных значений факторов как переменных является плохая исходная факторная модель. Действительно, матрица факторных нагрузок (см. табл. 3) весьма неудобна для интерпретации. В результате мы имеем факторы, как показано на примере 3-го фактора, с невысокими факторными нагрузками, т.е. слабо связанные с исходными переменными. Когда же выяснилось, что три переменные, выбранные для интерпретации 3-го фактора, объясняют его лишь на 26%, было трудно ожидать хороших результатов от кластеризации на факторах.

Традиционно для улучшения (скорее—упрощения) матрицы факторных нагрузок используют вращение факторной матрицы. В табл. 9 приведена матрица факторных нагрузок после вращения матрицы табл. 3 методом варимакс.

После проведенного вращения ситуация несколько улучшилась. Коэффициент R^2 показывает, что 3-й фактор объясняется переменными В8 и В10 почти на 60%. Однако остаются отмеченные ранее недостатки интерпретации поведения фактора как индекса, отражающего выделенные переменные, основанной на матрице факторных нагрузок. Так, регрессионный коэффициент при переменной В8 в два раза меньше коэффициента при переменной В10, хотя факторные нагрузки у этих переменных отличаются лишь на 20%.

Таблица 9. Матрица факторных нагрузок после вращения варимакс, процент объясненной дисперсии, общности для модельных данных

	Факторы				Общности
	1	**2**	3	4	
B9	0,704				0,687
B4	0,660				0,652
B11	0,618				0,665
B12	0,583				0,614
B1	0,555				0,920
B15	0,503				0,513
B14					0,527
B13		0,697			0,694
B3		0,674			0,716
B5		0,523			0,609
B2		0,503			0,631
B10			0,729		0,731
B8			0,591	0,515	0,710
B6				0,707	0,699
B7				0,637	0,634
Процент объясненной дисперсии	**20,5**	**16,1**	**15,3**	**14,8**	

Примечание. В матрице не приводятся факторные нагрузки меньше 0,5.

Не спасает вращение матрицы факторных нагрузок и при решении задачи кластеризации объектов, основанной на значениях факторов. Применение алгоритма Л-средних к факторам, полученным по результатам ортогонального вращения, дает точно такое же решение, как и разбиение на кластеры, основанное на факторном анализе без вращения (см. табл. 7).

Другим возможным объяснением плохого качества кластеризации на факторах может быть то, что факторная модель (неважно, с вращением или без) объясняет далеко не всю дисперсию исходных признаков (в рассматривавшемся примере — 67%). Соответственно построенные факторы включают лишь исходной информации переменных, и, следовательно, кластеризация получается низкого качества из-за потери значительной части исходной информации. Однако это объяснение является несостоятельным.

Мы повторили эксперимент с модельным массивом данных, выделив не четыре, как ранее, а десять факторов. Очевидно, что такой факторный анализ становится гораздо хуже интерпретируемым, но зато он объясняет более 88% дисперсии исходных переменных. Кажется, что качество кластеризации, основанной на значениях таких десяти факторов, должно быть близким к кластеризации на исходных переменных (см. табл. 8), и, по крайней мере, должно быть лучше, чем качество кластеризации, основанной на четырех факторах (см. табл. 7). На самом деле качество кластеризации на десяти факторах посредством применения метода ^-средних гораздо хуже, чем качество кластеризации на четырех факторах. Количество ошибочно классифицированных объектов при использовании индивидуальных значений десяти факторов составляет 56%, притом что для случая четырех факторов этот показатель был равен 35%.

Причиной выявленных «странностей» является то, что все предлагаемые в традиционных статистических пакетах (SPSS, STATISTICA и др.) методы факторного анализа строят ортогональные факторы[3]. Далее в случае использования в факторном анализе нескольких десятков переменных полученные индивидуальные значения факторов имеют, как правило, распределения, достаточно близкие к нормальному (за исключением случаев тех факторов, которые имеют очень высокие нагрузки для небольшого числа переменных). Таким образом, если взглянуть на полученный массив переменных (факторов), которые подвергаются кластеризации, мы увидим, что это данные из независимых переменных с многомерным нормальным распределением. Ясно, что кластеризация такого массива все равно может быть проведена, поскольку нет таких данных, которые нельзя кластеризовать. Другое дело, что полученный результат будет иметь вполне случайный характер, и его качество будет определяться лишь интерпретационными способностями исследователя. Вообще «замечательность» таких эвристических методов, как факторный и кластерный анализ, состоит в том, что качество получаемых с их помощью результатов верифицируется лишь критерием «правдоподобности», что целиком находится в руках исследователя.

Первая публикация: Социология 4М. 2005. № 21. С. 172—187.

[3] Мы не рассматриваем здесь сюжеты неортогонального вращения факторов.

Учебно-методические и научные труды А.О. Крыштановского

1. Возможности информационно-поисковой системы по обеспечению сравнительного анализа // Проблемы сравнительных социологических исследований. М.: ИСИ АН СССР, 1982.

2. Организация системы сбора, хранения и анализа данных в социологии на ЕС ЭВМ // Применение математических методов и ЭВМ в социологических исследованиях. М.: ИСИ АН СССР, 1982 (в соавторстве с О.В. Лакутиным).

3. Использование архива социологических исследований для проведения вторичного и сравнительного анализа // Методологические и методические аспекты сравнительных исследований. М., 1984.

4. Информационные базы и программные ресурсы ИСИ АН СССР (Методические рекомендации по их использованию). М., 1984 (в соавторстве с В.Г. Андреенковым).

5. О возможностях использования систем обработки информации на ЭВМ специалистами в области общественных наук // Комплексные методы в изучении истории с древнейших времен до наших дней. М.: Ин-т истории СССР АН СССР, 1984.

6. Банк данных ИСИ АН СССР. М., 1985 (в соавторстве с В.Г. Андреенковым).

7. Проблемы накопления и анализа на ЭВМ данных социологических исследований / Отв. ред. В.Н. Иванов, А.А. Стогний. М.: Наука, 1986 (в соавторстве с В.Г. Андреенковым и В.А. Чередниченко).

8. Процесс обработки данных анкетных опросов. М., 1986 (в соавторстве с В.Г. Андреенковым).

9. Банк социологических данных (Информационные ресурсы социологических центров СССР). М., 1987 (в соавторстве с В.Г. Андреенковым).

10. Некоторые вопросы перевзвешивания выборки // Математические методы и модели в социологии. М., 1988 (в соавторстве с А.Г. Кузнецовым).

11. Возможности комплексного применения методов многомерного статистического анализа в современных программных системах // Анализ социологических данных на ЭВМ. М., 1989.

12. Ремонт выборки // Социологические исследования. 1989. № 5 (в соавторстве с А.А. Давыдовым).

13. Банк данных мониторинга // Экономические и социальные перемены: мониторинг общественного мнения. 1993. № 9.

14. Отношение населения России к деятельности президента // Социологический журнал. 1994. № 3.

15. Активность и достижительность в структуре трудовых ценностей российского населения // Социально-трудовые исследования. Вып. IV. М.: ИМЭМО РАН, 1996 (в соавторстве с В. Магуном, Ю. Аржаковой).

16. Методы анализа временных рядов // Экономические и социальные перемены: мониторинг общественного мнения. 2000. № 2.

17. Ограничения метода регрессионного анализа // Социология: методология, методы, математические модели. 2002. № 12.

18. Стратегии адаптации высших учебных заведений: Экономический и социологический аспекты. М.: ГУ ВШЭ, 2002 (в соавторстве с С.Л. Зарецкой, Н.Л. Титовой и др.).

19. «Кластеры на факторах» — об одном распространенном заблуждении // Социология: методология, методы, математические модели. 2005. № 21.

Крыштановский, А. О.

К85 Анализ социологических данных с помощью пакета SPSS [Текст]: учеб. пособие для вузов / А. О. Крыштановский; Гос. ун-т — Высшая школа экономи-ки. — М. : Изд. дом ГУ ВШЭ, 2006. — 281, [3] с. — (Учебники Высшей школы экономики). — Прил.: с. 225—281. — 2000 экз. — ISBN 5-7598-0373-5 (в пер.).

В основе учебного пособия лежит курс лекции по анализу социологических данных, читавшийся автором на протяжении ряда лет на факультете социологии Государственного университета — Высшей школы экономики.

Рассматриваются методы, используемые социологами на практике: построение и анализ одномерных и двумерных частотных таблиц; анализ взаимосвязи качественных и количественных переменных с помощью теста Стьюдента и модели однофакторного дисперсионного анализа; построение моделей регрессии; поиск «латентных переменных» методами факторного анализа, главных компонент, многомерного шкалирования; получение многомерных группировок с помощью кластерного анализа. Подробно описана реализация данных методов с помощью пакета SPSS — одной из самых распространенных в мире систем статистической обработки данных социальных исследований.

Для студентов высших учебных заведений, обучающихся по специальности «Социология», атакже специалистов-социологов.

УДК 303.4:004.9
ББК 60.5в7

Учебное издание

Серия «Учебники Высшей школы экономики»

Крыштановский Александр Олегович

Анализ социологических данных с помощью пакета SPSS

Зав. редакцией *О.А, Шестопалова*
Редактор *Л.И. Кузнецова*
Художественный редактор *А.М. Павлов*
Корректор *Е.Е. Андреева*
Компьютерная верстка и графика: *НЕ. Пузанова*

ЛР № 020832 от 15 октября 1993 г. продлена до 14 октября 2003 г.
Подписано в печать 22.02.2006 г. Формат 60x84 $^{1}/_{16}$. Бумага оф сетная.
Гарнитура Times New Roman. Печать офсетная. Тираж 2000 экз. Усл. печ. л. 16,51.
Уч.-изд. л. 16,3. Заказ № 132. Изд. № 567

ГУ ВШЭ. 125319, Москва, Кочновский проезд, 3
Тел./факс: (495) 772-95-71

Издательство ООО «МАКС Пресс».
105066, г. Москва, Елоховский пр., д. 3. сто 2
Тел. 939-38-90,939-38-91. Тел./факс 939-38-91.