ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

Кристофер Доугерти

УНИВЕРСИТЕТСКИЙ УЧЕБНИК

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

INTRODUCTION TO ECONOMETRICS

Christopher Dougherty

New York Oxford
OXFORD UNIVERSITY PRESS
1992

Университетский учебник

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

Кристофер Доугерти

Перевод с английского

Рекомендовано Министерством общего и профессионального образования Российской Федерации в качестве учебника для студентов экономических специальностей высших учебных заведений

Экономический факультет
МГУ им. М.В. Ломоносова

МОСКВА ИНФРА·М

Москва 1999

УДК (075.8)330.115
ББК 22.172
Д 71

This translation of Introduction to Econometrics by Christopher Dougherty originally published in English in 1992 is published by arrangement with Oxford University Press

Настоящий перевод издания «Введение в эконометрику» Кристофера Доугерти, опубликованного на английском языке в 1992 году, выходит в свет по согласованию с издательским домом «Oxford University Press»

Перевод и подготовка книги к изданию осуществлены в рамках программы «ТЕМПУС» при участии и содействии Европейского Сообщества и экономического факультета Московского Государственного Университета им. М.В. Ломоносова

Перевод с английского — *Е. Н. Лукаш, О. Ю. Шибалкин, О. О. Замков*

Научный редактор — *О. О. Замков*

Рецензенты — *декан ф-та «Статистика», зав. кафедрой математической статистики и экономики МЭСИ, д-р экон. наук, проф. В. С. Мхитарян; зам. директора ЦЭМИ РАН, д-р физ.-мат. наук, проф. С. А. Айвазян*

Доугерти К.

Д 71 **Введение в эконометрику:** Пер. с англ. — М.: ИНФРА-М, 1999. — XIV, 402 с.

ISBN 0-19-504346-4 (англ.)
ISBN 5-86225-458-7 (русск.)

Книга Кристофера Доугерти — один из самых популярных на Западе вводных учебников эконометрики для студентов-экономистов. Курс эконометрики занимает важное место в современных программах экономических вузов во всем мире наряду с такими предметами, как микроэкономика, макроэкономика, финансовый анализ. Эконометрические методы необходимо знать и ученому, и преподавателю, и практику. Без них нельзя построить сколько-нибудь надежного прогноза, а значит, под вопросом и успех в банковском деле, финансах, бизнесе.

Популярный учебник по эконометрике издается в России впервые. Актуальность его появления на российском книжном рынке связана с острым дефицитом книг по эконометрике. Книгу отличает доступность изложения и вместе с тем высокий научный уровень освещения основных современных идей и методов эконометрики.

Книга может быть рекомендована в качестве базового учебника для студентов экономических специальностей, изучающих курс эконометрики. Ее можно также рекомендовать для самостоятельного ознакомления с этой дисциплиной. Работа может оказаться весьма полезной и при решении широкого круга прикладных проблем, с которыми читатель сталкивается в практической работе.

ISBN 0-19-504346-4 (англ.)
ISBN 5-86225-458-7 (русск.) ББК 22.172

ОТ НАУЧНОГО РЕДАКТОРА ПЕРЕВОДА

Книга Кристофера Доугерти «Введение в эконометрику» стала сейчас одним из наиболее популярных (если не самым популярным) вводным учебником эконометрики для широкого круга студентов-экономистов. Популярность этой книги связана также с получившим известность, тщательно проработанным, очень тонко и современно построенным курсом, который автор читает в Лондонской школе экономики (LSE). Этот курс слушают в течение года все студенты-экономисты LSE, он включен также и в программу знаменитой трехнедельной Летней школы. Курс К. Доугерти — один из тех в этой школе, где, несмотря на довольно высокую стоимость обучения, велик конкурс со стороны студентов, преподавателей, молодых исследователей, банковских работников из стран всего мира.

В современных программах подготовки экономистов курс эконометрики уверенно занял одно из ключевых мест. Не зная достаточно хорошо этого предмета, не владея его инструментарием, невозможно ни проверить представляемые в учебниках, книгах и статьях эмпирические зависимости, ни получить новые такие зависимости, а значит — и выдвинуть новые теории. Без эконометрических методов нельзя построить сколь-нибудь надежного прогноза, а значит — под вопросом и успех в банковском деле, финансах, бизнесе. Поэтому курс эконометрики входит в «ядро» учебных программ современного экономического вуза наряду с такими предметами, как микроэкономика, макроэкономика, финансовый анализ. Он, кроме того, должен быть тесно связан с перечисленными курсами, давая не абстрактно-формальные, а прикладные знания.

К сожалению, в России современный прикладной курс эконометрики пока еще не стал обязательным элементом программы в каждом экономическом вузе. Этот предмет изучается в основном студентами, специализирующимися в применении математических и статистических методов в экономике, изучается достаточно глубоко, но формально, нередко в отрыве от современных экономических теорий и их приложений. Для преподавания прикладного курса эконометрики студентам массовых экономических специальностей обычно не хватает ни квалифицированных педагогов, ни учебников, ни места в учебных программах. В то же время отечественные ученые внесли весьма значительный вклад в развитие эконометрических методов, особенно на начальном этапе. Такие имена, как А. А. Марков, А. М. Ляпунов, П. Л. Чебышев, Е. Е. Слуцкий, обязательно много раз упоминаются в любом курсе прикладной статистики и эконометрики. Правда, в течение ряда десятилетий развитие этих направлений не поощрялось, поскольку эконометрические методы способны были

выявить те или иные нежелательные для властей тенденции экономического развития. В настоящее время, когда перед высшим экономическим образованием в России стоит (и решается) задача выхода на мировой уровень, одним из ключевых моментов является развитие преподавания эконометрики.

Нельзя сказать, что у нас не издавались ранее книги по эконометрике, как отечественные, так и переводные. В 1960–1970-х и в начале 1980-х годов были изданы на русском языке известные книги Г. Тинтнера, Э. Маленво, Т. Андерсона, Дж. Джонстона, М. Дж. Кендалла и А. Стьюарта, Э. Кейна и др. В 1980-е годы публикация переводных изданий по эконометрике практически прекратилась, но проблему с литературой частично решали отечественные книги (Е. М. Четыркина и И. Л. Калихмана, С. А. Айвазяна, И. С. Енюкова и Л. Д. Мешалкина и др.). С конца 1980-х годов книги по эконометрике на русском языке почти не публиковались, и в настоящий момент остро стоит проблема с литературой любого уровня сложности. Прежние издания, во-первых, стали библиографической редкостью, а во-вторых, в них не отражены идеи и результаты последних десятилетий, когда эконометрика и ее приложения бурно развивались, а на Западе появилось новое поколение литературы в данной области.

Книга К. Доугерти поможет закрыть наиболее значительный, базовый пробел: отсутствие понятного, доступного современного учебника. Такого учебника не было ранее и в англоязычной литературе, о чем автор упоминает в предисловии к английскому изданию. Хотя книга очень проста в изложении, в ней с достаточной строгостью затрагиваются практически все основные современные базовые идеи и методы эконометрики, на которых строятся и научные исследования, и гораздо более продвинутые учебные курсы. Поэтому книга может широко использоваться как в преподавании курса эконометрики для студентов экономических специальностей, так и для самостоятельного ознакомления с этой дисциплиной. Книга может оказаться очень полезной и при выборочном ознакомлении с прикладными проблемами, с которыми читатель может столкнуться в практической работе.

Книга К. Доугерти издается в России в рамках проекта ТЕМПУС, осуществляемого совместно экономическим факультетом МГУ и тремя европейскими университетами: Лондонской школой экономики (Англия), университетом Сорбонна (Франция), университетом Тилбург (Нидерланды). Целью данного проекта является модернизация университетского экономического образования и сближение его с мировыми стандартами. Помимо данной книги проектом предусмотрено издание еще целого ряда учебников, таких, как: «Принципы корпоративных финансов» Р. Брейли и С. Майерса, «Международная экономика. Теория и политика» П. Кругмана и М. Обстфельда, «Экономическое развитие» М. Тодаро, «Структура отраслевых рынков» Ф. Шерера и Д. Росса, «Экономика организаций» К. Менара.

Хотелось бы верить, что эта книга — лишь первая ласточка в новой волне переводных и отечественных публикаций по эконометрике, с помощью которых эта дисциплина займет подобающее ей место в экономическом образовании, научных исследованиях и практической работе в России.

О. О. Замков

ПРЕДИСЛОВИЕ

Эта книга предназначена для студентов, изучающих годовой курс эконометрики. Она отражает определенную потребность, вызванную последними изменениями в программах эконометрической подготовки студентов и не учтенную в ранее написанных учебниках. За последнее десятилетие преподавание эконометрики на университетском уровне достигло своего «совершеннолетия». Курсы эконометрики, предлагавшиеся обычно как курсы на выбор в магистерских программах по экономике, сейчас все в большей мере становятся обязательными. Это обусловлено несколькими факторами. Возможно, важнейшим из них является растущее признание того, что определенное понимание методов эмпирических исследований является не просто желательной, но весьма существенной частью базовой подготовки экономиста и что ограничивающиеся прикладной статистикой курсы неадекватны этой задаче. Без сомнения, это привело к тому, что курсы эконометрики для аспирантов стали намного более продвинутыми, вследствие чего недостаточное знакомство с эконометрикой стало препятствием для поступления в аспирантуру ведущих университетов. Сыграл свою роль и «фактор предложения». Волна, поднявшая эконометрику на столь высокий уровень в экономическом образовании, идет вслед за другой волной, поднявшей значение математики и статистики. Без предшествующего улучшения подготовки в области количественных методов анализа выдвижение эконометрики в «ядро» программ экономических вузов было бы невозможным. Данный сдвиг был также связан с увеличением числа квалифицированных преподавателей эконометрики.

Вследствие происшедших изменений слушатели курсов эконометрики в большей степени различаются по своим возможностям, чем ранее. Это уже не только меньшинство, избравшее сложный путь математической специализации. Типичный студент сейчас — это обычный студент-магистр экономического профиля, изучивший базовые, но не продвинутые курсы математического анализа и статистики. «Демократизация» эконометрики создала необходимость подготовки более широкого спектра учебников, чем прежде, в особенности для новичков. Студенты, изучившие продвинутые курсы математики, уже несколько лет пользуются рядом завершенных учебников, отдельные из которых выдержали уже два или три издания. Меньше повезло новичкам, и этот текст адресован главным образом им.

Цель этой книги — обеспечить базу для изучения годового курса, позволяющего затем студентам продолжить изучение предмета в аспирантуре. Важнейшей задачей было сократить до минимума математические требования к читателю. Почти у каждого есть свой предел математической сложности изложения, которую он готов принять. Если этот предел превышен, читатель тратит большую часть усилий на техни-

ческую сторону вопроса вместо его сущности. Появляется усталость, страдает понимание, и путь исследования становится поденной работой или даже хуже того.

К счастью, математическое бремя намного облегчается, если коэффициенты регрессии выразить через выборочные ковариации и дисперсии. Для преподавателей и наиболее одаренных студентов степень математизированности может быть не столь важной, но это не так для той аудитории, которой адресована эта книга. Многие из этих людей, по-видимому, чувствовали себя не совсем уверенно в предшествующих математических курсах. Принятый здесь подход делает возможным расставание с устрашающими обозначениями типа Σ, становящимися непреодолимым препятствием для многих студентов. Это означает, что для знакомства с соответствующими понятиями понадобятся определенные затраты времени (глава 1), но эти понятия легко усваиваются и потраченные усилия затем многократно окупаются. Это не пустые рассуждения. Когда несколько лет назад я внес эти изменения в свой курс, где до этого пользовался обозначениями типа Σ, то обнаружил у студентов значительное улучшение восприятия, особенно при рассмотрении свойств регрессионных оценок.

Второй заботой было удаление неэконометрических «камней преткновения», сдерживавших развитие понимания предмета. Многие сложности, возникающие во вводном курсе, не являются техническими по своей природе (и многие технические моменты, например использование фиктивных переменных, вовсе не представляют проблемы). Например, многие студенты затрудняются описать регрессионные результаты словесно, в терминах, понятных неспециалисту. Если рассматривается регрессия в логарифмах, нередко странная заминка возникает даже в том случае, если заранее было показано математически, что коэффициент наклона характеризует эластичность. Проблема, без сомнения, заключается в том, что, изучая эластичность в базовом курсе математики, студенты были столь поглощены математическими вопросами, что у них не оставалось времени как следует разобраться с их практическими приложениями. Для них эластичность оставалась не до конца понятым математическим объектом. Эта книга содержит ряд отступлений в виде вставок, где рассматриваются подобные проблемы.

Ряд отступлений, кроме того, посвящен примерам экономических приложений. Нельзя оставить рассмотрение гипотезы Фридмена о постоянном доходе в качестве отдельных и несвязанных вопросов курсов макроэкономической теории и эконометрики. Поскольку нереалистично ожидать включения эконометрических аспектов в студенческие курсы экономической теории, задача нахождения этой связи неизбежно ложится на эконометристов.

Эксперименты по методу Монте-Карло

Характерной чертой этой книги является широкое использование экспериментов по методу Монте-Карло, наиболее подходящему для проведения эконометристами экспериментов лабораторного типа в контролируемых условиях. Многие студенты, слушающие вводный курс, предпочитают смотреть на такой анализ с двух точек зрения: математических рассуждений и числовой иллюстрации. Можно проследить за математическим исследованием свойств статистической оценки и согласиться с ее логичностью, но вместе с тем и получить пользу, видя подтверждение всего этого числовым примером в форме эксперимента по методу Монте-Карло. Такого рода анализ помогает, так сказать, запе-

чатлеть все это в памяти и почувствовать определенную уверенность. И как и в других дисциплинах, для многих людей развитие познаний в эконометрике напоминает не арифметическую прогрессию (как бы это ни казалось в ретроспективе), а процесс развития и обобщения, в котором технические навыки и интуитивное понимание развиваются во взаимодействии.

Упражнения с функциями спроса

Еще одной особенностью книги является последовательность упражнений с функциями спроса, создающих стержень для практической работы и являющихся существенным компонентом вводного курса. Упражнения с функциями спроса обеспечивают непрерывность и дают возможность наблюдать воздействие теоретических продвижений на спецификацию модели и технику оценивания. Как и эксперименты по методу Монте-Карло, они позволяют лучше запомнить то, что уже было познано аналитически, и обеспечивают ту действенную вовлеченность, которая часто отсутствует при простом представлении результатов оценивания регрессии для комментариев.

Данные для упражнений с функциями спроса приведены в табл. Б.1 и Б.2 приложения Б. Предполагается, что студенты будут разбиты на группы для практической работы по курсу и каждый студент в группе получит задание, связанное со спросом на определенное благо.

Структура книги

Большинство студентов, приступающих к изучению вводного курса эконометрики, уже прослушали один или несколько курсов по статистике и, значит, должны быть хорошо знакомы с фундаментальными понятиями. Факт, однако, состоит в том, что многие люди нуждаются в повторном знакомстве с этим материалом, прежде чем они действительно освоят его. Поскольку нет смысла пытаться изучать вводный курс эконометрики без должного понимания таких категорий, как несмещенность, эффективность и состоятельность, стоит при необходимости начать с рассмотрения такого материала. Этому посвящен обзор в начале книги. В главе 1 описана математическая основа, а оставшиеся главы покрывают обычные для вводного курса темы и разбиты на три части. Первая часть книги (главы 2–5) содержит основы регрессионного анализа. В ее второй части (главы 6–8) рассматриваются некоторые наиболее общие проблемы, возникающие при использовании регрессионного анализа, а в третьей части представлены некоторые дальнейшие продвижения. В заключительной части дается краткая последующая ориентация.

Лондон, декабрь 1990 года *Кристофер Доугерти*

СОДЕРЖАНИЕ

Обзор: Случайные переменные и теория выборок 3

1. Ковариация, дисперсия и корреляция 34

1.1. *Выборочная ковариация* 34
1.2. *Несколько основных правил расчета ковариации* 38
1.3. *Альтернативное выражение для выборочной ковариации* 42
1.4. *Теоретическая ковариация* 43
1.5. *Выборочная дисперсия* 44
1.6. *Правила расчета дисперсии* 45
1.7. *Теоретическая дисперсия выборочного среднего* 47
1.8. *Коэффициент корреляции* 47
1.9. *Почему ковариация не является хорошей мерой связи?* 50
1.10. *Коэффициент частной корреляции* 52

2. Парный регрессионный анализ 53

2.1. *Модель парной линейной регрессии* 53
2.2. *Регрессия по методу наименьших квадратов* 55
2.3. *Регрессия по методу наименьших квадратов: два примера* 58
2.4. *Детальное рассмотрение остатков* 61
2.5. *Регрессия по методу наименьших квадратов с одной независимой переменной* 62
2.6. *Интерпретация уравнения регрессии* 64
2.7. *Качество оценки: коэффициент* R^2 69

3. Свойства коэффициентов регрессии и проверка гипотез 73

3.1. *Случайные составляющие коэффициентов регрессии* 73
3.2. *Эксперимент по методу Монте-Карло* 74
3.3. *Предположения о случайном члене* 79
3.4. *Несмещенность коэффициентов регрессии* 82
3.5. *Точность коэффициентов регрессии* 83
3.6. *Теорема Гаусса—Маркова* 87
3.7. *Проверка гипотез, относящихся к коэффициентам регрессии* 89
3.8. *Доверительные интервалы* 102
3.9. *Односторонние t-тесты* 104

3.10. *F-тест на качество оценивания* ... 109
3.11. *Взаимосвязи между критериями в парном регрессионном анализе* ... 111

4. Преобразования переменных ... 115

4.1. *Базисная процедура* ... 115
4.2. *Логарифмические преобразования* ... 119
4.3. *Случайный член* ... 125
4.4. *Нелинейная регрессия* ... 126
4.5. *Выбор функции: тесты Бокса—Кокса* ... 129
Приложение 4.1 ... 132

5. Множественный регрессионный анализ ... 134

5.1. *Иллюстрация: модель с двумя независимыми переменными* ... 134
5.2. *Вывод и интерпретация коэффициентов множественной регрессии* ... 137
5.3. *Множественная регрессия в нелинейных моделях* ... 141
5.4. *Свойства коэффициентов множественной регрессии* ... 146
5.5. *Мультиколлинеарность* ... 155
5.6. *Качество оценивания: коэффициент* R^2 ... 159

6. Спецификация переменных в уравнениях регрессии: предварительное рассмотрение ... 165

6.1. *Моделирование* ... 165
6.2. *Влияние отсутствия в уравнении переменной, которая должна быть включена* ... 166
6.3. *Влияние включения в модель переменной, которая не должна быть включена* ... 177
6.4. *Замещающие переменные* ... 182
6.5. *Проверка линейного ограничения* ... 188
6.6. *Как извлечь максимум информации из анализа остатков* ... 193
6.7. *Лаговые переменные* ... 196

7. Гетероскедастичность и автокоррелированность случайного члена ... 200

7.1. *Еще раз об условиях Гаусса—Маркова* ... 200
7.2. *Гетероскедастичность и ее последствия* ... 201
7.3. *Обнаружение гетероскедастичности* ... 204
7.4. *Что можно сделать в случае гетероскедастичности?* ... 210
7.5. *Автокорреляция и связанные с ней факторы* ... 217
7.6. *Обнаружение автокорреляции первого порядка: критерий Дарбина—Уотсона* ... 219
7.7. *Что можно сделать в отношении автокорреляции?* ... 222
7.8. *Автокорреляция с лаговой зависимой переменной* ... 227
7.9. *Автокорреляция как следствие неправильной спецификации модели* ... 229

Приложение 7.1 234
Приложение 7.2 237
Приложение 7.3 240
Приложение 7.4 241

8. Стохастические объясняющие переменные и ошибки измерения 243

8.1. *Стохастические объясняющие переменные* 243
8.2. *Последствия ошибок измерения* 247
8.3. *Критика М. Фридменом стандартной функции потребления* 253
8.4. *Инструментальные переменные* 259

9. Фиктивные переменные 262

9.1. *Иллюстрация использования фиктивной переменной* 262
9.2. *Общий случай* 270
9.3. *Множественные совокупности фиктивных переменных* 277
9.4. *Фиктивные переменные для коэффициента наклона* 280
9.5. *Тест Чоу* 282
Приложение 9.1 285

10. Моделирование динамических процессов 288

10.1. *Введение* 288
10.2. *Распределение Койка* 289
10.3. *Частичная корректировка* 291
10.4. *Адаптивные ожидания* 295
10.5. *Гипотеза Фридмена о постоянном доходе* 298
10.6. *Полиномиально распределенные лаги Алмон* 303
10.7. *Рациональные ожидания* 306
10.8. *Предсказание* 309
10.9. *Тесты на устойчивость* 315
Приложение 10.1 319

11. Оценивание систем одновременных уравнений 322

11.1. *Смещение при оценке одновременных уравнений* 322
11.2. *Структурная и приведенная формы уравнений* 325
11.3. *Косвенный метод наименьших квадратов (КМНК)* 327
11.4. *Инструментальные переменные (ИП)* 330
11.5. *Неидентифицируемость* 332
11.6. *Сверхидентифицированность* 336
11.7. *Двухшаговый метод наименьших квадратов (ДМНК)* 337
11.8. *Условие размерности для идентификации* 340
11.9. *Идентификация относительно стабильных зависимостей* 345
Приложение 11.1 348

12. Что дальше? 350

12.1. *Метод максимального правдоподобия (ММП)* 350
12.2. *Спецификация модели* 354
12.3. *Послесловие к функциям спроса* 363
Приложение А. Статистические таблицы 366
Приложение Б. Набор данных 374
Библиография 384
Именной указатель 387
Предметный указатель 389

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

ОБЗОР: СЛУЧАЙНЫЕ ПЕРЕМЕННЫЕ И ТЕОРИЯ ВЫБОРОК

В этой книге при рассмотрении методов оценивания большое внимание будет уделено следующим свойствам оценок: несмещенности, состоятельности и эффективности. Для читателя важно понимание этих свойств, и в книге предполагается, что он знаком с ними в пределах вводного курса статистики. В данной главе предлагается краткий обзор по этим вопросам.

Дискретная случайная переменная

Ваше интуитивное понимание вероятности почти наверняка соответствует задачам этой книги, и поэтому мы опустим традиционный раздел чистой теории вероятностей, хотя он мог бы быть весьма увлекательным. Многие люди непосредственно сталкивались с вероятностями, участвуя в лотереях и азартных играх, и их заинтересованность в том, чем они занимались, часто приводила к удивительно высокой практической компетентности, обычно при полном отсутствии формальной подготовки.

Мы начнем непосредственно с *дискретных случайных переменных. Случайная переменная* — это любая переменная, значение которой не может быть точно предсказано. *Дискретной* называется случайная величина, имеющая определенный набор возможных значений. Пример — сумма выпавших очков при бросании двух игральных костей. Пример случайной величины, не являющейся дискретной, — температура в комнате. Она может принять любое из непрерывного диапазона значений и является примером непрерывной случайной величины. К рассмотрению таких величин в этом обзоре мы перейдем позже.

Продолжая разговор о примере с двумя игральными костями, предположим, что одна из них зеленая, а другая — красная. Если их бросить, то возможны 36 элементарных исходов эксперимента, поскольку на зеленой кости может выпасть любое число от 1 до 6 и то же самое — на красной. Случайная переменная, определенная как их сумма, которую мы обозначим через x, может принимать только одно из 11 числовых значений — от 2 до 12. Взаимосвязь между исходами эксперимента и значениями случайной величины в данном случае показана на рис. О.1.

Красная	Зеленая					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Рис. О.1. Исходы в примере с двумя игральными костями

Предположив, что кости «правильные», мы можем воспользоваться рис. О.1 для определения вероятности каждого значения x. Поскольку на костях имеется 36 различных комбинаций, каждый исход имеет вероятность 1/36. Лишь одна из возможных комбинаций {зеленая = 1, красная = 1} дает сумму, равную 2, так что вероятность $x = 2$ равна 1/36. Чтобы получить сумму $x = 7$, нам потребуются сочетания {зеленая = 1, красная = 6}, либо {зеленая = 2, красная = 5}, либо {зеленая = 3, красная = 4}, либо {зеленая = 4, красная = 3}, либо {зеленая = 5, красная = 2}, либо {зеленая = 6, красная = 1}. В данном случае нас устроят 6 возможных исходов, и поэтому вероятность получения 7 равна 6/36. Все эти вероятности приведены в табл. О.1. Если все их сложить, то получится ровно 1. Это будет так, поскольку с вероятностью 100% рассматриваемая сумма примет одно из значений от 2 до 12.

Совокупность всех возможных значений случайной переменной описывается *генеральной совокупностью*, из которой извлекаются эти значения. В нашем случае генеральная совокупность — это набор чисел от 2 до 12.

Таблица О.1

Значения x	2	3	4	5	6	7	8	9	10	11	12
Вероятность	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Упражнение

О.1. Случайная переменная x определяется как разность между большим и меньшим числами, выпавшими при бросании двух костей. Если они равны между собой, то переменная x считается равной нулю. Найдите распределение вероятностей для x.

Математическое ожидание дискретной случайной величины

Математическое ожидание дискретной случайной величины — это взвешенное среднее всех ее возможных значений, причем в качестве весового коэффициента берется вероятность соответствующего исхода. Вы можете рассчитать его, перемножив все возможные значения случайной величины на их вероятности и просуммировав полученные произведения. Математически если случайная величина обозначена как x, то ее математическое ожидание обозначается как $E(x)$.

Предположим, что x может принимать n конкретных значений $(x_1, x_2, \ldots, x_n)$ и что вероятность получения x_i равна p_i. Тогда

$$E(x) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i. \qquad (О.1)$$

(Читатели, желающие освежить в памяти использование обозначений Σ, могут сделать это с помощью приложения О.1.)

В случае с двумя костями величинами от x_1 до x_n были числа от 2 до 12: $x_1 = 2$, $x_2 = 3$, ... , $x_{11} = 12$ и $p_1 = 1/36$, $p_2 = 2/36$, ... , $p_{11} = 1/36$. Математическое ожидание рассчитывается так:

$$\left(2 \times \frac{1}{36}\right) + \left(3 \times \frac{2}{36}\right) + \left(4 \times \frac{3}{36}\right) + \ldots + \left(11 \times \frac{2}{36}\right) + \left(12 \times \frac{1}{36}\right).$$

Если вычислить эту величину, то получится 7.

Прежде чем пойти дальше, рассмотрим еще более простой пример случайной переменной — число очков, выпадающее при бросании лишь одной игральной кости.

В данном случае возможны шесть исходов: $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, $x_5 = 5$, $x_6 = 6$. Каждый исход имеет вероятность 1/6, поэтому здесь

$$E(x) = \sum_{i=1}^{6} x_i p_i = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3{,}5. \qquad (О.2)$$

В данном случае математическим ожиданием случайной переменной является число, которое само по себе не может быть получено при бросании кости.

Математическое ожидание случайной величины часто называют ее *средним по генеральной совокупности*. Для случайной величины x это значение часто обозначается как μ.

Упражнение

О.2. Найдите математическое ожидание случайной величины x в упражнении О.1.

Математические ожидания функций дискретных случайных переменных

Пусть $g(x)$ — некоторая функция от x. Тогда $E\{g(x)\}$ — математическое ожидание $g(x)$ записывается как

$$E\{g(x)\} = \sum g(x_i)p_i, \qquad (O.3)$$

где суммирование производится по всем возможным значениям x. В табл. О.2 показана последовательность практического расчета *математического ожидания функции от x*.

Предположим, что x может принимать n различных значений от x_1 до x_n с соответствующими вероятностями от p_1 до p_n. В первой колонке записываются все возможные значения x. Во второй — записываются соответствующие вероятности. В третьей колонке рассчитываются значения функции для соответствующих величин x. В четвертой колонке перемножаются числа из колонок *2* и *3*. Ответ приводится в суммирующей строке колонки *4*.

Пример

Каково математическое ожидание величины x^2? Разумно ли предположить, что она равняется μ^2?

Рассмотрим пример с числами, выпадающими при бросании одной кости. Использовав схему, приведенную в табл. О.2, заполним табл. О.3.

Таблица О.2

x	Вероятность	Функция от x	Функция, взвешенная по вероятности
x_1	p_1	$g(x_1)$	$g(x_1)p_1$
x_2	p_2	$g(x_2)$	$g(x_2)p_2$
...	...	...	...
x_n	p_n	$g(x_n)$	$g(x_n)p_n$
Всего			$\sum g(x_i)p_i = E\{g(x)\}$

Таблица О.3

x_i	p_i	x_i^2	$x_i^2 p_i$
1	1/6	1	0,167
2	1/6	4	0,667
3	1/6	9	1,500
4	1/6	16	2,667
5	1/6	25	4,167
6	1/6	36	6,000
Всего			15,167

В четвертой ее колонке даны шесть значений x^2, взвешенных по соответствующим вероятностям, которые в данном примере все равняются $^1/_6$. (С вероятностью $^1/_6$ величина x^2 будет равна единице, поскольку это произойдет при $x = 1$, что имеет место в одном случае из шести, с вероятностью $^1/_6$ значение x^2 равняется 4, так как это произойдет при $x = 2$, и т. д.). По определению, величина $E(x^2)$ равна $\Sigma x_i^2 p_i$, она приведена как сумма в четвертой колонке и равна 15,167.

Математическое ожидание x, как уже было показано, равно 3,5, и 3,5 в квадрате равно 12,25. Таким образом, величина $E(x^2)$ не равна μ^2, и, следовательно, нужно аккуратно проводить различия между $E(x^2)$ и $\{E(x)\}^2$ (последнее равно произведению $E(x)$ на $E(x)$, то есть μ^2).

Упражнения

О.3. Пусть x — случайная переменная с математическим ожиданием μ, и λ — константа. Докажите, что математическое ожидание λx равно $\lambda\mu$.

О.4. Рассчитайте $E(x^2)$ для величины x, определенной, как в упражнении О.1.

Правила расчета математического ожидания

Существуют три правила, которые далее будут использоваться много раз. Эти правила практически самоочевидны, и они одинаково применимы для дискретных и непрерывных случайных переменных.

Правило 1. Математическое ожидание суммы нескольких переменных равно сумме их математических ожиданий. Например, если имеются три случайные переменные (x, y и z), то

$$E(x + y + z) = E(x) + E(y) + E(z). \tag{O.4}$$

Правило 2. Если случайная переменная умножается на константу, то ее математическое ожидание умножается на ту же константу. Если x — случайная переменная и a — константа, то

$$E(ax) = aE(x). \tag{O.5}$$

Правило 3. Математическое ожидание константы есть она сама. Например, если a — константа, то

$$E(a) = a. \tag{O.6}$$

Правило *2* уже доказано в упражнении O.3. Правило *3* тривиально, поскольку оно следует из определения константы. Хотя доказательство правила *1* довольно простое, мы его опустим.

Объединяя три правила вместе, можно упростить и более сложные выражения. Например, предположим, что вы хотите рассчитать $E(y)$, где

$$y = a + bx, \tag{O.7}$$

где a и b — константы. Следовательно,

$$\begin{aligned} E(y) &= E(a + bx) = \\ &= E(a) + E(bx), \text{ согласно правилу } 1, \\ &= a + bE(x), \text{ согласно правилам } 2 \text{ и } 3. \end{aligned} \tag{O.8}$$

Таким образом, вместо непосредственного вычисления $E(y)$ можно рассчитать $E(x)$ и получить $E(y)$ из уравнения O.8.

Упражнение

O.5. Пусть x — число очков, выпавшее при однократном бросании игральной кости. Рассчитайте возможные значения y, где y получается по формуле $y = x^2 + 3x - 2$, и, далее, рассчитайте $E(y)$. Покажите, что она равняется $E(x^2) + 3E(x) - 2$.

Независимость случайных переменных

Две случайные переменные x и y называются независимыми, если $E\{f(x)\,g(y)\}$ равняется $E\{f(x)\}E\{g(y)\}$ для любых функций $f(x)$ и $g(y)$. Из независимости следует как важный частный случай, что $E(xy)$ равно $E(x)E(y)$.

Теоретическая дисперсия дискретной случайной переменной

В этой книге нас будет интересовать одна из функций переменной x, ее *теоретическая дисперсия*, являющаяся полезной мерой разброса для вероятност-

ного распределения. Она определяется как *математическое ожидание* квадрата разности между величиной x и ее средним, т. е. величины $(x-\mu)^2$, где μ — математическое ожидание x. Дисперсия обычно обозначается как σ_x^2, и если ясно, о какой переменной идет речь, то нижний индекс может быть опущен. Мы иногда будем обозначать дисперсию как pop. var (x):

$$\sigma_x^2 = \text{pop.var}(x) = E\{(x-\mu)^2\} =$$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 p_i = (x_1-\mu)^2 p_1 + \ldots + (x_n-\mu)^2 p_n. \qquad \text{(O.9)}$$

Из σ_x^2 можно получить σ_x — *теоретическое стандартное отклонение* — столь же распространенную меру разброса для распределения вероятностей; стандартное отклонение случайной переменной есть квадратный корень из ее дисперсии.

Мы проиллюстрируем расчет дисперсии на примере с одной игральной костью. Поскольку $\mu = E(x) = 3{,}5$, то $(x-\mu)^2$ в этом случае равно $(x-3{,}5)^2$. Мы рассчитаем математическое ожидание величины $(x-3{,}5)^2$, используя схему, представленную в табл. О.2. Дополнительный столбец $(x-\mu)$ представляет определенный этап расчета $(x-\mu)^2$. Суммируя последний столбец в табл. О.4, получим значение дисперсии σ^2, равное 2,92. Следовательно, стандартное отклонение (σ), равно $\sqrt{2{,}92}$, то есть 1,71.

Таблица О.4

x_i	p_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(x_i-\mu)^2 p_i$
1	1/6	–2,5	6,25	1,042
2	1/6	–1,5	2,25	0,375
3	1/6	–0,5	0,25	0,042
4	1/6	0,5	0,25	0,042
5	1/6	1,5	2,25	0,375
6	1/6	2,5	6,25	1,042
Всего				2,92

Одним из важных приложений правил расчета математического ожидания является формула расчета теоретической дисперсии случайной переменной, которая может быть записана как

$$\sigma^2 = E(x^2) - \mu^2. \qquad \text{(O.10)}$$

Это выражение иногда оказывается более удобным, чем первоначальное определение. Доказательство дает хороший пример использования упомяну-

тых правил, но при первом чтении вы можете, если хотите, его пропустить. По этому определению,

$$\begin{aligned}\sigma^2 &= E\{(x-\mu)^2\} = E\{(x^2 - 2\mu x + \mu^2) = \\ &= E(x^2) + E(-2\mu x) + E(\mu^2), \text{ согласно правилу } 1, \\ &= E(x^2) - 2\mu E(x) + \mu^2, \quad \text{согласно правилам } 2 \text{ и } 3 \text{ и тому факту, что } -2\mu \text{ и } \mu^2 \text{ — константы,} \\ &= E(x^2) - 2\mu^2 + \mu^2, \quad \text{поскольку величины } E(x) \text{ и } \mu \text{ идентичны,} \\ &= E(x^2) - \mu^2. \end{aligned} \tag{O.11}$$

Таким образом, если вы хотите вычислить теоретическую дисперсию для x, то можете рассчитать математическое ожидание величины x^2 и вычесть из него μ^2.

Упражнение

О.6. Рассчитайте теоретическую дисперсию и стандартное отклонение величины x, определенной, как в упражнении О.1, использовав определение, заданное уравнением (О.9).

О.7. Использовав уравнение (О.10), найдите дисперсию случайной переменной x, определенной в упражнении О.1, и покажите, что результат получается тем же, что и в упражнении О.6. (Это займет совсем немного времени, потому что вы уже рассчитали μ в упражнении О.2 и $E(x^2)$ в упражнении О.4.)

Вероятность в непрерывном случае

С дискретными случайными переменными очень легко обращаться, поскольку они по определению принимают значения из некоторого конечного набора. Каждое из этих значений связано с определенной вероятностью, характеризующей его «вес». Если эти «веса» известны, то не составит труда рассчитать *теоретическое среднее* (математическое ожидание) и дисперсию.

Вы можете представить указанные «веса» как определенные количества «пластичной массы», равные вероятностям соответствующих значений. Сумма вероятностей и, следовательно, суммарный «вес» этой «массы» равен единице. Это показано на рис. О.2 для примера, где величина x есть сумма очков, выпавших при бросании двух игральных костей. Величина x принимает значения от 2 до 12, и для всех этих значений показано количество соответствующей «массы».

К сожалению, анализ в нашей книге проводится обычно для непрерывных случайных величин, которые могут принимать бесконечное число значений. Поскольку невозможно представить себе «пластичную массу», разделенную на бесконечное число частей, используем далее другой подход.

Проиллюстрируем наши рассуждения на примере температуры в комнате. Для определенности предположим, что она меняется в пределах от 55 до 75° по Фаренгейту, и вначале допустим, что все значения в этом диапазоне равновероятны.

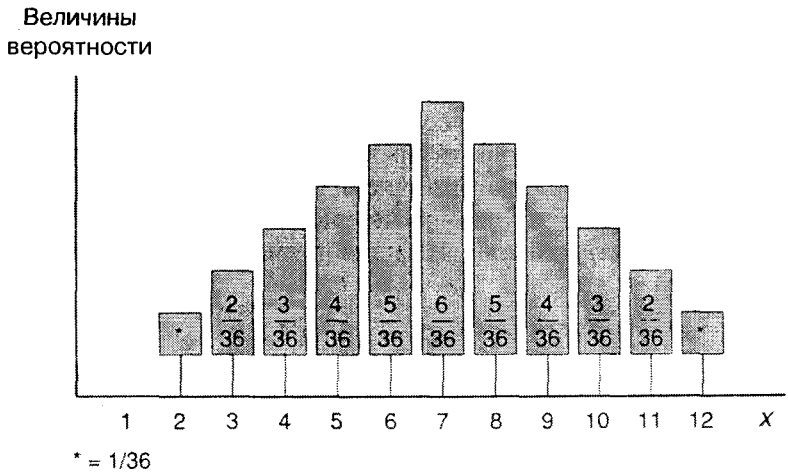


Рис. О.2. Дискретные вероятности (пример с двумя игральными костями)

Поскольку число различных значений, принимаемых показателем температуры, бесконечно, здесь бессмысленно пытаться разделить «пластичную массу» на малые части. Вместо этого можно «размазать» ее по всему диапазону. Поскольку все температуры от 55 до 75° F равновероятны, она должна быть «размазана» равномерно, как это показано на рис. О.3.

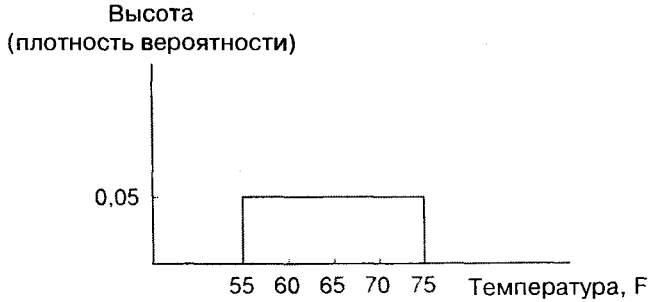


Рис. О.3

В этом примере, как и во всех остальных, мы будем полагать, что «пластичная масса размазана» на единичной площади. Это связано с тем, что совокупная вероятность всегда равняется единице. В данном случае наша «масса» покрыла прямоугольник, и, поскольку основание этого прямоугольника равно 20, его высота определяется из соотношения:

$$20 \times \text{Высота} = 1, \qquad (О.12)$$

так как произведение основания и высоты равно площади. Следовательно, высота равна 0,05, как это показано на рис. О.3.

Найдя высоту прямоугольника, мы можем ответить на вопросы типа: с какой вероятностью температура будет находиться в диапазоне от 65 до 70°F? Ответ определяется величиной «замазанной» площади (или, говоря более формально, *совокупной вероятностью*), лежащей в диапазоне от 65 до 70°F, пред-

ставленной заштрихованной фигурой на рис. О.4. Основание заштрихованного прямоугольника равно 5, его высота равна 0,05 и, соответственно, площадь — 0,25. Искомая вероятность равна 1/4, что в любом случае очевидно, поскольку промежуток от 65 до 70°F составляет 1/4 всего диапазона.

Высота заштрихованной площади представляет то, что формально называется *плотностью вероятности* в этой точке, и если эта высота может быть записана как функция значений случайной переменной, то эта функция называется *функцией плотности вероятности*. В нашем примере она записывается как $f(x)$, где x — температура, и

$$f(x) = 0{,}05; \qquad 55 \le x \le 75 \tag{О.13}$$

[$f(x) = 0$ для $x < 55$ или $x > 75$].

В качестве первого приближения функция плотности вероятности показывает вероятность нахождения случайной переменной внутри единичного интервала вокруг данной точки. В нашем примере эта функция всюду равна 0,05, откуда вытекает, что температура находится, например, между 60 и 61°F с вероятностью 0,05.

В нашем случае график функции плотности вероятности горизонтален, и ее указанная интерпретация точна, однако в общем случае эта функция непрерывно меняется, и ее интерпретация дает лишь приближение. Далее мы рассмотрим пример, когда эта функция непостоянна, поскольку не все температуры равновероятны. Предположим, что центральное отопление работает таким образом, что температура никогда не падает ниже 65°F, а в жаркие дни температура превосходит этот уровень, не превышая, как и ранее, 75°F. Мы будем считать, что плотность вероятности максимальна при температуре 65°F и далее она равномерно убывает до нуля при 75°F.

Общая «замазанная» площадь, как всегда, равна единице, поскольку совокупная вероятность равна единице. Площадь треугольника равна половине произведения основания на высоту, поэтому получаем:

$$\tfrac{1}{2} \times 10 \times \text{Высота} = 1, \tag{О.14}$$

и высота при 65°F равна 0,20.

Предположим вновь, что мы хотим знать вероятность нахождения температуры в промежутке между 65 и 70°F. Она представлена заштрихованной площадью на рис. О.6, и если вы немного помните геометрию, то сможете проверить, что она равна 0,75. Если вы предпочитаете процентное измерение, то

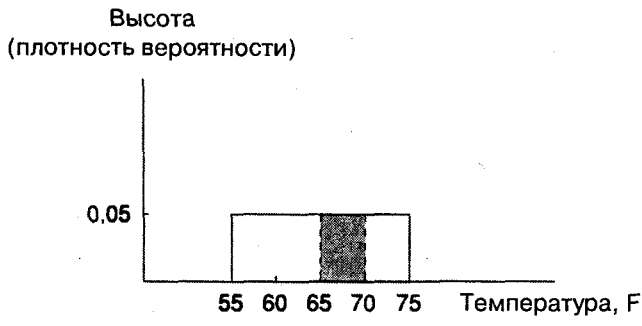


Рис. О.4

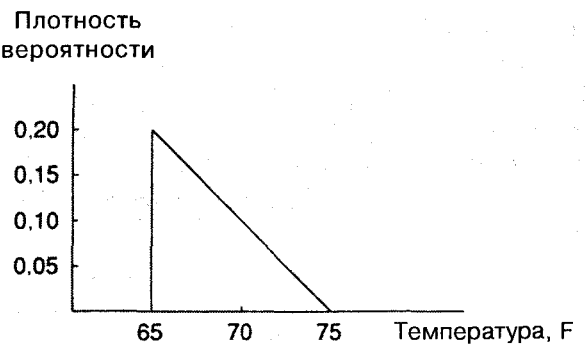


Рис. О.5

это означает, что с вероятностью 75% температура попадет в диапазон 65–70°F и только с вероятностью 25% — в диапазон 70–75°F.

В данном случае функция плотности вероятности записывается как $f(x)$, где

$$f(x) = 1{,}5 - 0{,}02x; \qquad 65 \le x \le 75. \tag{О.15}$$

(Вы можете проверить, что эта функция равна 0,20 при 65°F и нулю при 75°F.)

Прежде чем продолжить изложение, упомянем о хорошей и плохой новостях. «Плохая новость» — это то, что если вы хотите рассчитать вероятности для более сложных функций с криволинейными графиками, то элементарная геометрия становится неприменимой. Вообще говоря, вы должны воспользоваться интегральным исчислением или специальными таблицами (если последние существуют). Интегральное исчисление используется также и при определении математического ожидания и дисперсии непрерывной случайной величины.

«Хорошая новость» — в том, что специальные таблицы существуют для всех функций, которые будут интересовать нас на практике. Кроме того, математическое ожидание и дисперсия имеют практически тот же смысл для непрерывных случайных величин, что и для дискретных (формальные определения можно найти в приложении 0.2), и для них верны те же самые правила.

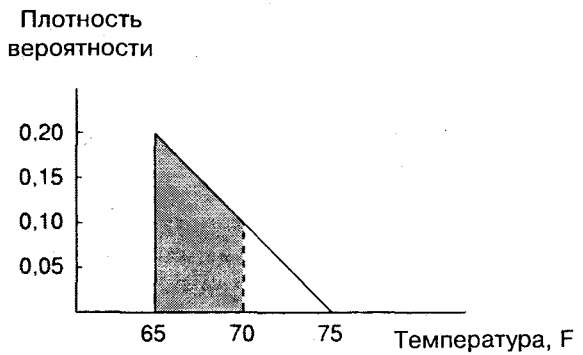


Рис. О.6

Постоянная и случайная составляющие случайной переменной

Часто вместо рассмотрения случайной величины как единого целого можно и удобно разбить ее на постоянную и чисто случайную составляющие, где постоянная составляющая всегда есть ее математическое ожидание. Если x — случайная переменная и μ — ее математическое ожидание, то декомпозиция случайной величины записывается следующим образом:

$$x = \mu + u, \tag{О.16}$$

где u — *чисто случайная составляющая* (в регрессионном анализе она обычно представлена *случайным членом*).

Конечно, можно было бы посмотреть на это по-другому и сказать, что случайная составляющая u определяется как разность между x и μ:

$$u = x - \mu. \tag{О.17}$$

Из определения следует, что математическое ожидание величины u равно нулю. Из уравнения (О.17) имеем:

$$E(u) = E(x - \mu) = E(x) - E(\mu) =$$

$$= E(x) - \mu = \mu - \mu = 0. \tag{О.18}$$

Поскольку весь разброс значений x обусловлен u, неудивительно, что теоретическая дисперсия x равна теоретической дисперсии u. Последнее нетрудно доказать. По определению,

$$\sigma_x^2 = E\{(x - \mu)^2\} = E\{u^2\} \tag{О.19}$$

и

$$\sigma_u^2 = E\{(u - \text{м.о.}(u))^2\} = E\{(u - 0)^2\} = E\{u^2\}. \tag{О.20}$$

Таким образом, σ^2 может быть эквивалентно определена как дисперсия x или u.

Обобщая, можно утверждать, что если x — случайная переменная, определенная по формуле (О.16), где μ — заданное число и u — случайный член с $E(u) = 0$ и pop. var $(u) = \sigma^2$, то математическое ожидание величины x равно μ, а дисперсия — σ^2.

Способы оценивания и оценки

До сих пор мы предполагали, что имеется точная информация о рассматриваемой случайной переменной, в частности — об ее распределении вероятностей (в случае дискретной переменной) или о функции плотности распределения (в случае непрерывной переменной). С помощью этой информации можно рассчитать теоретическое математическое ожидание, дисперсию и любые другие характеристики, в которых мы можем быть заинтересованы.

Однако на практике, за исключением искусственно простых случайных величин (таких, как число выпавших очков при бросании игральной кости), мы не знаем точного вероятностного распределения или плотности распреде-

ления вероятностей. Это означает, что неизвестны также и теоретическое математическое ожидание, и дисперсия. Мы, тем не менее, можем нуждаться в оценках этих или других теоретических характеристик генеральной совокупности.

Процедура оценивания всегда одинакова. Берется выборка из n наблюдений, и с помощью подходящей формулы рассчитывается оценка нужной характеристики. Нужно следить за терминами, делая важное различие между способом или формулой оценивания и рассчитанным по ней для данной выборки числом, являющимся значением оценки. *Способ оценивания* — это общее правило, или формула, в то время как *значение оценки* — это конкретное число, которое меняется от выборки к выборке[1].

В табл. О.5 приведены формулы оценивания для двух важнейших характеристик генеральной совокупности. *Выборочное среднее* $\bar{x}$ обычно дает оценку для математического ожидания, а формула s^2 в табл. О.5 — оценку дисперсии генеральной совокупности.

Таблица О.5

Характеристики генеральной совокупности		Формулы оценивания
Среднее,	μ	$\bar{x} = \frac{1}{n}\sum x_i$
Дисперсия,	σ^2	$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$

Отметим, что это *обычные* формулы оценки математического ожидания и дисперсии генеральной совокупности, однако не единственные. Возможно, вы настолько привыкли использовать $\bar{x}$ в качестве оценки для μ, что даже не задумывались об альтернативах. Конечно, не все формулы оценки, которые можно представить, одинаково хороши. Причина, по которой в действительности используется $\bar{x}$, в том, что эта оценка в наилучшей степени соответствует двум очень важным критериям — несмещенности и эффективности. Эти критерии будут рассмотрены ниже.

Оценки как случайные величины

Получаемая *оценка* представляет частный случай случайной переменной. Причина здесь в том, что сочетание значений x в выборке случайно, поскольку

[1] В русскоязычной литературе и способ оценивания, и значение оценки часто сокращенно называют просто оценкой. Иногда в дальнейшем мы тоже будем так поступать, если из контекста ясно, о чем идет речь. (*Прим. ред.*)

x — случайная переменная и, следовательно, случайной величиной является и функция набора ее значений. Возьмем, например, $\bar{x}$ — оценку математического ожидания:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n). \qquad (O.21)$$

Мы только что показали, что величина x в i-м наблюдении может быть разложена на две составляющие: постоянную часть μ и чисто случайную составляющую u_i:

$$x_i = \mu + u_i. \qquad (O.22)$$

Следовательно,

$$\bar{x} = \mu + \bar{u}, \qquad (O.23)$$

где $\bar{u}$ — выборочное среднее величин u_i.

Отсюда можно видеть, что $\bar{x}$, подобно x, имеет как фиксированную, так и чисто случайную составляющие. Ее фиксированная составляющая — μ, то есть математическое ожидание x, а ее случайная составляющая — $\bar{u}$, то есть среднее значение чисто случайной составляющей в выборке.

Функции плотности вероятности для x и $\bar{x}$ показаны на одинаковых графиках (рис. О.7). Как показано на рисунке, величина x считается нормально распределенной. Можно видеть, что распределения, как x, так и $\bar{x}$, симметричны относительно μ — теоретического среднего. Разница между ними в том, что распределение $\bar{x}$ уже и выше. Величина $\bar{x}$, вероятно, должна быть ближе к μ, чем значение единичного наблюдения x, поскольку ее случайная составляющая $\bar{u}$ есть среднее от чисто случайных составляющих $u_1, u_2, \ldots, u_n$ в выборке, которые, по-видимому, «гасят» друг друга при расчете среднего. Далее, теоретическая дисперсия величины $\bar{u}$ составляет лишь часть теоретической дисперсии u. В разделе 1.7 будет показано, что если pop. var $(u) = \sigma^2$, то pop. var $(\bar{u}) = \sigma^2/n$.

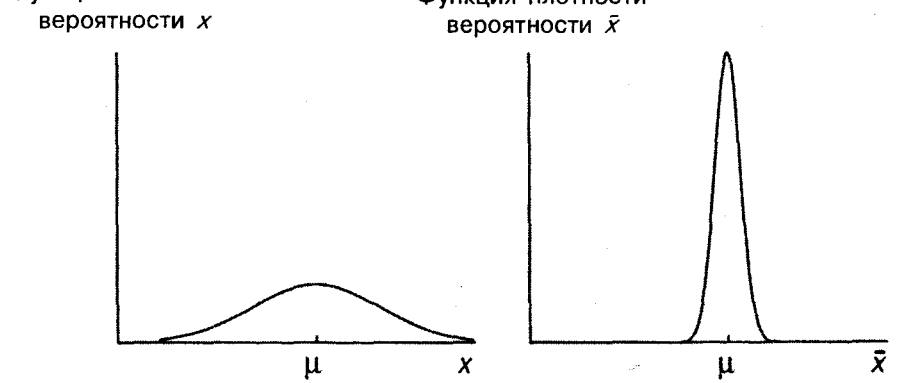


Рис. О.7. Сравнение функций плотности вероятности одиночного наблюдения и выборочного среднего

Величина s^2 — оценка теоретической дисперсии x — также является случайной переменной. Вычитая (О.23) из (О.22), имеем:

$$x_i - \bar{x} = u_i - \bar{u}. \tag{О.24}$$

Следовательно,

$$s^2 = \frac{1}{n-1}\sum\{(x_i - \bar{x})^2\} = \frac{1}{n-1}\sum\{(u_i - \bar{u})^2\}. \tag{О.25}$$

Таким образом, s^2 зависит от (и только от) чисто случайной составляющей наблюдений x в выборке. Поскольку эти составляющие меняются от выборки к выборке, также от выборки к выборке меняется и величина оценки s^2.

Несмещенность

Поскольку оценки являются случайными переменными, их значения лишь по случайному совпадению могут в точности равняться характеристикам генеральной совокупности. Обычно будет присутствовать определенная ошибка, которая может быть большой или малой, положительной или отрицательной, в зависимости от чисто случайных составляющих величин x в выборке.

Хотя это и неизбежно, на интуитивном уровне желательно, тем не менее, чтобы оценка в среднем за достаточно длительный период была аккуратной. Выражаясь формально, мы хотели бы, чтобы математическое ожидание оценки равнялось бы соответствующей характеристике генеральной совокупности. Если это так, то оценка называется *несмещенной*. Если это не так, то оценка называется *смещенной*, и разница между ее математическим ожиданием и соответствующей теоретической характеристикой генеральной совокупности называется *смещением*.

Начнем с выборочного среднего. Является ли оно несмещенной оценкой теоретического среднего? Равны ли $E(x)$ и μ? Да, это так, что непосредственно вытекает из (О.23).

Величина x включает две составляющие — μ и $\bar{u}$. Значение $\bar{u}$ равно средней чисто случайных составляющих величин x в выборке, и, поскольку математическое ожидание такой составляющей в каждом наблюдении равно нулю, математическое ожидание $\bar{u}$ равно нулю. Следовательно,

$$E(\bar{x}) = E(\mu + \bar{u}) = E(\mu) + E(\bar{u}) = \mu + 0 = \mu. \tag{О.26}$$

Тем не менее полученная оценка — не единственно возможная несмещенная оценка μ. Предположим для простоты, что у нас есть выборка всего из двух наблюдений — x_1 и x_2. Любое взвешенное среднее наблюдений x_1 и x_2 было бы несмещенной оценкой, если сумма весов равна единице. Чтобы показать это, предположим, что мы построили обобщенную формулу оценки:

$$Z = \lambda_1 x_1 + \lambda_2 x_2. \tag{О.27}$$

Математическое ожидание Z равно:

$$\begin{aligned} E(Z) &= E(\lambda_1 x_1 + \lambda_2 x_2) = E(\lambda_1 x_1) + E(\lambda_2 x_2) = \\ &= \lambda_1 E(x_1) + \lambda_2 E(x_2) = \lambda_1 \mu + \lambda_2 \mu = (\lambda_1 + \lambda_2)\mu. \end{aligned} \tag{О.28}$$

Если сумма λ_1 и λ_2 равна единице, то мы имеем $E(Z) = \mu$, и Z является несмещенной оценкой μ.

Таким образом, в принципе число несмещенных оценок бесконечно. Как выбрать одну из них? Почему в действительности мы всегда используем выборочное среднее с $\lambda_1 = \lambda_2 = 0{,}5$? Возможно, вы полагаете, что было бы несправедливым давать разным наблюдениям различные веса или что подобной асимметрии следует избегать в принципе. Мы, однако, не заботимся здесь о справедливости или о симметрии как таковой. В следующем разделе мы увидим, что имеется и более осязаемая причина.

До сих пор мы рассматривали только оценки теоретического среднего. Выше утверждалось, что величина s^2, определяемая в соответствии с табл. О.5, является оценкой теоретической дисперсии σ^2. Можно показать, что математическое ожидание s^2 равно σ^2, и эта величина является несмещенной оценкой теоретической дисперсии, если наблюдения в выборке независимы друг от друга. Доказательство этого математически несложно, но трудоемко, и поэтому оно вынесено в приложение О.3 в конце данного обзора.

Эффективность

Несмещенность — желательное свойство оценок, но это не единственное такое свойство. Еще одна важная их сторона — это надежность. Конечно, немаловажно, чтобы оценка была точной в среднем за длительный период, но, как однажды заметил Дж. М. Кейнс, «в долгосрочном периоде мы все умрем». Мы хотели бы, чтобы наша оценка с максимально возможной вероятностью давала бы близкое значение к теоретической характеристике, что означает желание получить функцию плотности вероятности, как можно более «сжатую» вокруг истинного значения. Один из способов выразить это требование — сказать, что мы хотели бы получить сколь возможно малую дисперсию.

Предположим, что мы имеем две оценки теоретического среднего, рассчитанные на основе одной и той же информации, что обе они являются несме-

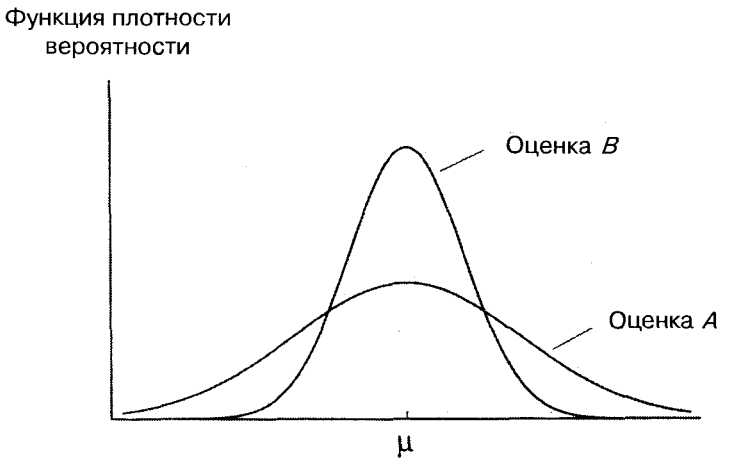


Рис. О.8. Эффективные и неэффективные оценки

щенными и что их функции плотности вероятности показаны на рис. О.8. Поскольку функция плотности вероятности для оценки B более «сжата», чем для оценки A, с ее помощью мы скорее получим более точное значение. Формально говоря, эта оценка более *эффективна*.

Важно заметить, что мы использовали здесь слово «скорее». Даже хотя оценка B более эффективна, это не означает, что она всегда дает более точное значение. При определенном стечении обстоятельств значение оценки A может быть ближе к истине. Однако вероятность того, что оценка A окажется более точной, чем B, составляет менее 50%.

Это напоминает вопрос о том, пользоваться ли ремнями безопасности при управлении автомобилем. Множество обзоров в разных странах показало, что значительно менее вероятно погибнуть или получить увечья в дорожном происшествии, если воспользоваться ремнями безопасности. В то же время не раз отмечались странные случаи, когда не сделавший этого индивид чудесным образом уцелел, но погиб бы, будучи пристегнут ремнями. Упомянутые обзоры не отрицают этого. В них лишь делается вывод, что преимущество на стороне тех, кто пользуется ремнями безопасности. Подобным же преимуществом обладает и *эффективная оценка*. (Неприятный комментарий: в тех странах, где пользование ремнями безопасности сделано обязательным, сократилось предложение для трансплантации почек людей, ставших жертвами аварий.)

Мы говорили о желании получить оценку как можно с меньшей дисперсией, и эффективная оценка — это та, у которой дисперсия минимальна. Сейчас мы рассмотрим дисперсию обобщенной оценки теоретического среднего и покажем, что она минимальна в том случае, когда оба наблюдения имеют равные веса.

Если наблюдения x_1 и x_2 независимы, теоретическая дисперсия обобщенной оценки равна:

$$\text{pop.var}(Z) = \text{pop.var}(\lambda_1 x_1 + \lambda_2 x_2) = (\lambda_1^2 + \lambda_2^2)\sigma^2. \qquad \text{(О.29)}$$

(Это можно показать, используя правила расчета дисперсии, рассматриваемые в главе 1.)

Мы уже выяснили, что для несмещенности оценки необходимо равенство единице суммы λ_1 и λ_2. Следовательно, для несмещенных оценок $\lambda_2 = (1 - \lambda_1)$ и

$$\lambda_1^2 + \lambda_2^2 = \lambda_1^2 + (1-\lambda_1)^2 = 2\lambda_1^2 - 2\lambda_1 + 1. \qquad \text{(О.30)}$$

Поскольку мы хотим выбрать λ_1 так, чтобы минимизировать дисперсию, нам нужно минимизировать при этом $(2\lambda_1{}^2 - 2\lambda_1 + 1)$. Эту задачу можно решить графически или с помощью дифференциального исчисления. В любом случае минимум достигается при $\lambda_1 = 0{,}5$. Следовательно, λ_2 также равно 0,5.

Итак, мы показали, что выборочное среднее имеет наименьшую дисперсию среди оценок рассматриваемого типа. Это означает, что оно имеет наиболее «сжатое» вероятностное распределение вокруг истинного среднего и, следовательно (в вероятностном смысле), наиболее точно. Строго говоря, выборочное среднее — это наиболее эффективная оценка среди всех несмещенных оценок. Конечно, мы показали это только для случая с двумя наблюдениями, но сделанные выводы верны для выборок любого размера, если наблюдения не зависят друг от друга.

Два заключительных замечания: во-первых, эффективность оценок можно сравнивать лишь тогда, когда они используют одну и ту же информацию, например один и тот же набор наблюдений нескольких случайных переменных. Если одна из оценок использует в 10 раз больше информации, чем другая, то она вполне может иметь меньшую дисперсию, но было бы неправильно считать ее более эффективной. Во-вторых, мы ограничиваем понятие эффективности сравнением распределений несмещенных оценок. Существуют определения эффективности, обобщающие это понятие на случай возможного сравнения смещенных оценок, но в этой книге мы будем придерживаться данного простого определения.

Упражнения

О.8. Рассчитайте дисперсию обобщенной оценки теоретического среднего для частного случая $\sigma^2 = 1$ и выборки из двух наблюдений, воспользовавшись уравнением (О.30) с величинами λ_1 от 0 до 1 при шаге 0,1. Нанесите полученные точки на график. Важно ли то, чтобы весовые коэффициенты λ_1 и λ_2 в точности равнялись друг другу?

О.9. Покажите, что при наличии n наблюдений условием того, чтобы обобщенная формула $(\lambda_1 x_1 + \ldots + \lambda_n x_n)$ давала несмещенную оценку μ, является $\lambda_1 + \ldots + \lambda_n = 1$.

О.10. Вообще говоря, при увеличении размера выборки дисперсия распределения оценки убывает. Правильно ли утверждать при этом, что оценка становится более эффективной?

Противоречия между несмещенностью и минимальной дисперсией

В данном обзоре мы уже выяснили, что для оценки желательна несмещен-

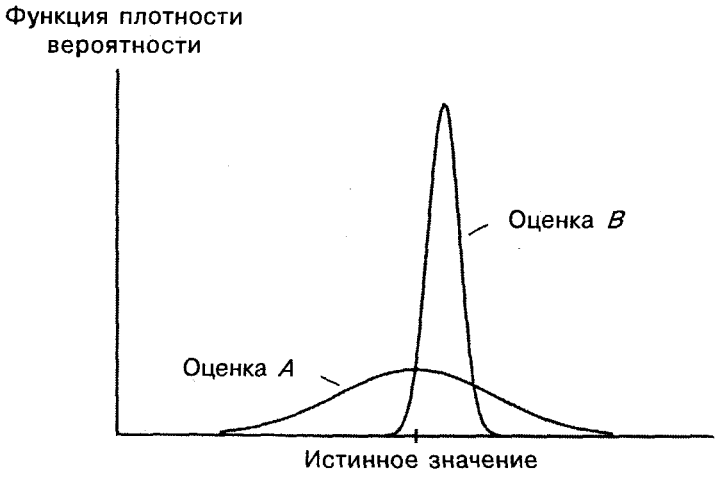


Рис. О.9. Какую оценку предпочесть: оценка *A* несмещенная, но у *B* меньше дисперсия

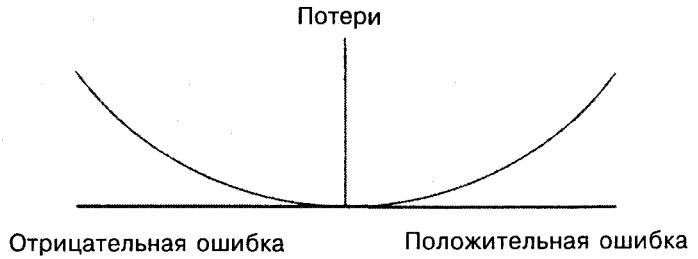


Рис. О.10. Функция потерь

ность и наименьшая возможная дисперсия. Эти критерии совершенно различны, и иногда они могут противоречить друг другу. Может случиться так, что имеются две оценки теоретической характеристики, одна из которых является несмещенной (*A* на рис. О.9), другая же смещена, но имеет меньшую дисперсию (*B*).

Оценка *A* хороша своей несмещенностью, но преимуществом оценки *B* является то, что ее значения практически всегда близки к истинному значению. Какую из них вы бы выбрали?

Данный выбор зависит от обстоятельств. Если возможные ошибки вас не очень тревожат при условии, что за длительный период они «погасят» друг друга, то, по-видимому, вы выберете *A*. С другой стороны, если для вас приемлемы малые ошибки, но неприемлемы большие, то вам следует выбрать *B*.

Формально говоря, выбор определяется *функцией потерь*, стоимостью сделанной ошибки как функцией ее размера. Обычно выбирают оценку, дающую наименьшее ожидание потерь, и делается это путем взвешивания функции потерь по функции плотности вероятности. (Если вы не любите риск, то можете также пожелать учесть дисперсию потерь.)

Типичным примером функции потерь, показанной квадратичной параболой на рис. О.10, может служить *квадрат ошибки*. Ее математическое ожидание, известное как *среднеквадратичная ошибка* (*MSE*), может быть разложено на составляющие:

$$MSE = \text{Дисперсия оценки} + \text{Квадрат смещения}. \qquad (О.31)$$

Чтобы показать это, предположим, что оценка Z используется для оценивания неизвестного значения параметра генеральной совокупности θ. Предположим, что математическое ожидание Z равно μ_Z. Оно будет равняться θ только в том случае, если Z — несмещенная оценка. В общем случае будет иметь место смещение, равное $(\mu_Z - \theta)$. Дисперсия Z равна $E\{(Z - \mu_Z)\}^2$. Величина *MSE* оценки Z может быть разложена на составляющие следующим образом:

$$\begin{aligned} MSE(Z) &= E\{(Z-\theta)^2\} = E\{([Z-\mu_z]+[\mu_z-\theta])^2\} = \\ &= E\{(Z-\mu_z)^2\} + 2E\{(Z-\mu_z)(\mu_z-\theta)\} + E\{(\mu_z-\theta)^2 = \\ &= \text{pop.var}(Z) + 2(\mu_z-\theta)E\{(Z-\mu_z)\} + \text{Квадрат смещения} = \\ &= \text{pop.var}(Z) + \text{Квадрат смещения}, \end{aligned} \qquad (О.32)$$

поскольку $E\{(Z-\mu_z)\} = E(Z) - \mu_z = 0$.

На рис. О.9 оценка *A* не имеет составляющей смещения, но имеет гораздо большую составляющую дисперсии, чем *B*, и поэтому она хуже по данному критерию.

Hiawatha Designs an Experiment [1]

M. G. Kendall

1. Hiawatha, mighty fighter,
   He could shoot ten arrows upwards
   Shoot them with such strength and swiftness
   That the last had left the bowstring
   Ere the first to earth descended
   This was commonly regarded
   As a feat of skill and cunning.

2. One or two sarcastic spirits
   Pointed out to him, however
   That it might be much more useful
   If he sometimes hit the target.
   Why not shoot a little straighter
   And employ a smaller sample?

3. Hiawatha, who at college
   Majored in applied statistics
   Consequently felt entitled
   To instruct his fellow men on
   Any subject whatsoever,
   Waxed exceedingly indignant
   Talked about the law of error
   Talked about truncated normals
   Talked of loss of information
   Talked about his lack of bias
   Pointed out that in the long run
   Independent observations
   Even though they missed the target
   Had an average point of impact
   Very near the spot he aimed at
   (With the possible exception
   Of a set of measure zero).

4. This, they said, was rather doubtful.
   Anyway, it did not matter
   What resulted in the long run;

[1] По согласованию с автором, мы решили сохранить в этом издании на языке оригинала приводимое им шуточное стихотворение М. Дж. Кендалла на мотивы «Песни о Гайавате». Читатель вполне может пропустить его без ущерба для понимания материала книги. (*Прим. ред.*)

Either he must hit the target
Much more often than at present
Or himself would have to pay for
All the arrows that he wasted.

5. Hiawatha, in a temper,
Quoted parts of R. A. Fisher
Quoted Yates and quoted Finney
Quoted yards of Oscar Kempthorne
Quoted reams of Cox and Cochran
Practically in extenso
Trying to impress upon them
That what actually mattered
Was to estimate the error.

6. One or two of them admitted
Such a thing might have it uses
Still, they said, he might do better
If he shot a little straighter.

7. Hiawatha, to convince them,
Organized a shooting contest
Laid out in the proper manner
Of designs experimental
Recommended in the textbooks
(mainly used for tasting tea, but
Sometimes used in other cases)
Randomized his shooting order
In factorial arrangements
Used in the theory of Galois
Fields of ideal polynomials
Got a nicely balanced layout
And successfully confounded
Second-order interactions.

8. All the other tribal marksmen
Ignorant, benighted creatures
Of experimental set-ups
Spent their time of preparation
Putting in a lot of practice
Merely shooting at a target.

9. Thus it happened in the contest
That their scores were most impressive
With one solitary exception
This (I hate to have to say it)
Was the score of Hiawatha
Who, as usual, shot his arrows
Shot them with great strength and swiftness
Managing to be unbiased

Not, however, with his salvo,
Managing to hit the target.

10. There, they said to Hiawatha
This is what we all expected.

11. Hiawatha, nothing daunted
Called for pen and called for paper
Did analyses of variance
Finally produced the figures
Showing beyond peradventure
Everybody else was biased
And the variance components
Did not differ from each other
Or from Hiawatha's
(this last point, one should acknowledge
Might have been much more convincing
If he hadn't been compelled to
Estimate his own component
From experimental plots in
Which the values all were missing
Still, they didn't understand it
So they couldn't raise objections
This is what so often happens
With analyses of variance).

12. All the same, his fellow tribesmen
Ignorant, benighted heathens
Took away his bow and arrows,
Said that though my Hiawatha
Was a brilliant statistician
He was useless as a bowman.
As for variance components
Several of the more outspoken
Made primeval observations
Hurtful of the finer feelings
Even of a statistician.

13. In a corner of the forest
Dwells alone my Hiawatha
Permanently cogitating
On the normal law of error
Wondering in idle moments
Whethering an increased precision
Might perhaps be rather better
Even at the risk of bias
If thereby one, now and then, could
Register upon the target.

From Kendall, 1959

О.11. Приведите примеры приложений, в которых вы могли бы: 1) предпочесть оценку типа *A* (рис. О.9); 2) предпочесть оценку типа *B* (рис. О.9).

О.12. Изобразите функцию потерь для прибытия в аэропорт позже (или раньше) времени окончания регистрации.

О.13. Имеются две оценки неизвестного параметра генеральной совокупности. Обязательно ли является более эффективной та из них, которая имеет меньшую дисперсию?

Влияние увеличения размера выборки на точность оценок

Будем по-прежнему предполагать, что мы исследуем случайную переменную x с неизвестным математическим ожиданием μ и теоретической дисперсией σ^2 и что для оценивания μ используется $\bar{x}$. Каким образом точность оценки $\bar{x}$ зависит от числа наблюдений n?

Ответ неудивителен: при увеличении n оценка $\bar{x}$, вообще говоря, становится более точной. В единичном эксперименте большая по размеру выборка необязательно даст более точную оценку, чем меньшая выборка, — всегда может присутствовать элемент везения, — но общая тенденция должна быть именно такой. Поскольку дисперсия $\bar{x}$ выражается формулой σ^2/n, она тем меньше, чем больше размер выборки и, значит, тем сильнее «сжата» функция плотности вероятности для $\bar{x}$.

Это показано на рис. О.11. Мы предполагаем, что x нормально распределена со средним 25 и стандартным отклонением 50. Если размер выборки равен 25, то стандартное отклонение величины $\bar{x}$, равное $\sigma / \sqrt{n}$, составит: $50 / \sqrt{25} = 10$. Если размер выборки равен 100, то это стандартное отклонение равно 5. На рис. О.11 показаны соответствующие функции плотности вероятности. Вторая ($n = 100$) выше первой в окрестности μ, что говорит о более высокой вероятности получения с ее помощью аккуратной оценки. За пределами этой окрестности вторая функция всюду ниже первой.

Чем больше размер выборки, тем у́же и выше будет график функции плотности вероятности для $\bar{x}$. Если n становится действительно большим, то график функции плотности вероятности будет неотличим от вертикальной прямой, соответствующей $\bar{x} = \mu$. Для такой выборки случайная составляющая x становится действительно очень малой, и поэтому $\bar{x}$ обязательно будет очень близкой к μ. Это вытекает из того факта, что стандартное отклонение $\bar{x}$, равное $\sigma / \sqrt{n}$, становится очень малым при больших n.

В пределе, при стремлении n к бесконечности, $\sigma / \sqrt{n}$ стремится к нулю и $\bar{x}$ стремится в точности к μ. Это можно записать математически:

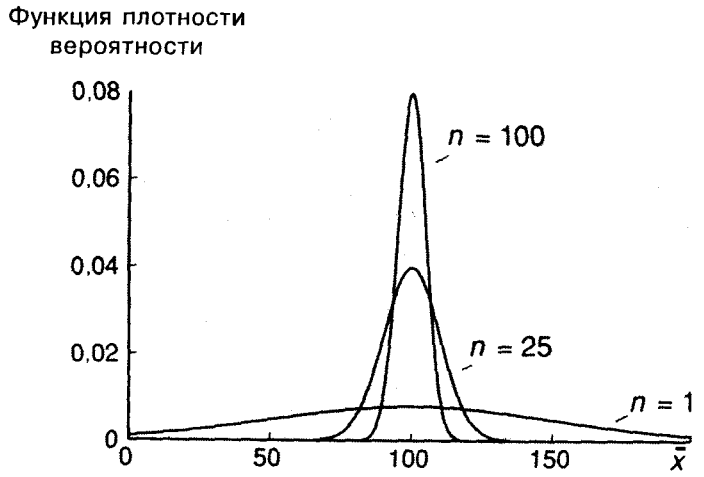


Рис. О.11. Влияние увеличения размера выборки на распределение $\bar{x}$.

$$\lim_{n \to \infty} \bar{x} = \mu. \tag{О.33}$$

Эквивалентный и более распространенный способ описания этого факта предлагает использование термина plim, где plim означает «предел по вероятности» и подчеркивает, что предел достигается в вероятностном смысле:

$$\text{plim}\ \bar{x} = \mu, \tag{О.34}$$

когда для любых сколь угодно малых ε и δ вероятность того, что $\bar{x}$ отличается от μ больше, чем на ε, будет меньшей δ при достаточно большом размере выборки.

Состоятельность

Вообще говоря, если предел оценки по вероятности равен истинному значению характеристики генеральной совокупности, то эта оценка называется *состоятельной*. Иначе говоря, состоятельной называется такая оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений.

В большинстве конкретных случаев в этой книге несмещенная оценка является и состоятельной. Для этого можно построить контрпримеры, но они, как правило, будут носить искусственный характер.

Иногда бывает, что оценка, смещенная на малых выборках, является состоятельной (иногда состоятельной может быть даже оценка, не имеющая на малых выборках конечного математического ожидания). На рис. О.12 показано, как при различных размерах выборки может выглядеть распределение вероятностей. Тот факт, что при увеличении размера выборки распределение становится симметричным вокруг истинного значения, указывает на асимп-

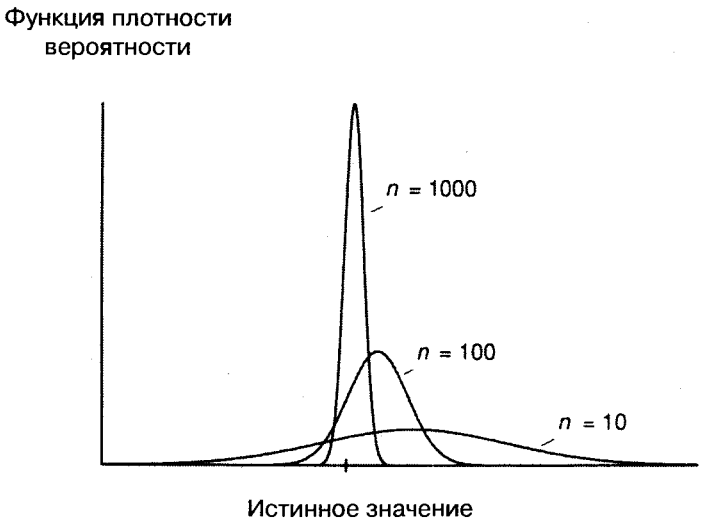


Рис. О.12. Состоятельная оценка, смещенная на малых выборках

тотическую несмещенность. То, что в конечном счете оно превращается в единственную точку истинного значения, говорит о состоятельности оценки.

Как мы увидим далее в этой книге, оценки типа показанных на рис. О.12 весьма важны в регрессионном анализе. Иногда невозможно найти оценку, несмещенную на малых выборках. Если при этом вы можете найти хотя бы состоятельную оценку, это может быть лучше, чем не иметь никакой оценки, особенно если вы можете предположить направление смещения на малых выборках.

Нужно, однако, иметь в виду, что состоятельная оценка в принципе может на малых выборках работать хуже, чем несостоятельная (например, иметь большую среднеквадратичную ошибку), и поэтому требуется осторожность. Подобно тому, как вы можете предпочесть смещенную оценку несмещенной, если ее дисперсия меньше, вы можете предпочесть состоятельную, но смещенную оценку несмещенной или несостоятельную оценку им обеим (также в случае меньшей дисперсии).

Полезное правило

Иногда оценка рассчитывается как отношение двух величин, имеющих случайные составляющие, например:

$$Z = X / Y, \tag{О.35}$$

где X и Y — величины, рассчитанные по данным выборки. Обычно трудно сказать что-либо определенное о математическом ожидании величины Z. Вообще говоря, она *не* равна частному от деления $E(X)$ на $E(Y)$. Если Y с некоторой вероятностью может равняться нулю, то математическое ожидание Z не может быть даже определено. Если, однако, X и Y стремятся к конечным величинам plim X и plim Y на больших выборках и plim Y не равен нулю, величина Z

будет стремиться к отношению $\text{plim}\, X/\text{plim}\, Y$. Следовательно, даже если нельзя сказать что-либо определенное о свойствах Z на малых выборках, мы иногда можем судить о ее состоятельности.

Предположим, например, что теоретические средние двух случайных переменных X и Y равны μ_X и μ_Y соответственно и что обе они подвержены случайным воздействиям, так что

$$X = \mu_X + u_X, \tag{O.36}$$

$$Y = \mu_Y + u_Y, \tag{O.37}$$

где u_X и u_Y — случайные составляющие с нулевым средним. Если мы по выборочным данным хотим оценить отношение μ_X/μ_Y, то оценка $Z = \overline{X} / \overline{Y}$ будет состоятельной, поскольку

$$\text{plim} Z = \text{plim}\overline{X} / \text{plim}\overline{Y} = \mu_X / \mu_Y, \tag{O.38}$$

и можно сказать, что Z является хорошей оценкой для больших выборок, хотя, возможно, для малых выборок о $E(Z)$ нельзя сказать ничего.

Упражнения

O.14. Является ли несмещенность необходимым или достаточным условием состоятельности?

O.15. Случайная величина X принимает значения *3* и *4* с равными вероятностями. Случайная величина Y принимает значения *1* и *2* также с равными вероятностями. Величины X и Y распределены независимо друг от друга. Переменная Z определяется как $Z = X/Y$ и имеет четыре возможных значения, каждое с вероятностью 0,25:

Таблица значений Z при данных значениях X и Y

X Y	3	4
1	3,0	4,0
2	1,5	2,0

Покажите, что $E(Z)$ *не* равно $E(X)/E(Y)$.

O.16. Величины $\overline{X}$ и $\overline{Y}$ являются выборочными средними X и Y, определенных как в предыдущем упражнении. Величина Z — оценка отношения теоретических средних и определяется как $\overline{X} / \overline{Y}$. К какому значению будет стремиться Z на большой выборке?

O.17. Выполните предыдущие два упражнения, предположив, что Y может принимать значения 0 и 1 с равными вероятностями.

Обозначения с использованием знака Σ: обзор

Обозначения с использованием знака Σ дают возможность быстрой и удобной записи ряда из подобных членов. Для чтения данной книги необходимо знакомство с такой записью, и здесь дается краткий обзор для тех, кому нужно освежить это в памяти.

Начнем с примера. Предположим, что объем выпуска лесопилки, измеренный в тоннах, за месяц i составляет q_i, причем q_1 — общий выпуск в январе, q_2 — общий выпуск в феврале и т. д. Обозначим годовой выпуск как Z. Тогда

$$Z = q_1 + q_2 + q_3 + q_4 + q_5 + q_6 + q_7 + q_8 + q_9 + q_{10} + q_{11} + q_{12}.$$

Можно «просуммировать» это выражение на словах, сказав, что Z есть сумма величин q_i (от q_1 до q_{12}). Очевидно, при определении Z нет необходимости записывать все 12 слагаемых. Иногда мы будем упрощать запись, представляя сумму в таком виде:

$$Z = q_1 + \ldots + q_{12},$$

имея в виду, что все пропущенные члены включаются в суммирование.

Запись с использованием знака Σ позволяет выразить эту сумму в удобной, аккуратной форме:

$$Z = \sum_{i=1}^{12} q_i.$$

Выражение справа от знака Σ говорит о том, какого вида члены суммируются, в данном случае это члены вида q_i. Под знаком Σ записывается индекс, который меняется при суммировании (в данном случае i), и его начальное значение (в данном случае 1). Таким образом, мы знаем, что первым слагаемым является q_1. Знак равенства указывает, что индекс i для первого слагаемого должен равняться единице.

Над знаком Σ записывается последнее значение i (в данном случае 12), и, таким образом, мы знаем, что последним слагаемым является q_{12}. Автоматически ясно, что все промежуточные члены между q_1 и q_{12} также должны быть включены в суммирование, и мы получаем удобно переписанное второе определение Z.

Предположим, что средняя цена за тонну продукции лесопилки за месяц i равна p_i. Общая стоимость выпуска за месяц i равна $p_i q_i$, а стоимость выпуска за год составляет V, где V рассчитывается по формуле:

$$V = p_1 q_1 + \ldots + p_{12} q_{12}.$$

Теперь мы суммируем члены вида $p_i q_i$, где нижний индекс i меняется от 1 до 12, и при использовании знака Σ это выражение может быть записано таким образом:

$$V = \sum_{i=1}^{12} p_i q_i.$$

Если c_i — общие издержки работы лесопилки за месяц i, то прибыль за месяц i будет равна $(p_i q_i - c_i)$ и, следовательно, общая прибыль за год (П) записывается как

$$\Pi = (p_1 q_1 - c_1) + \ldots + (p_{12} q_{12} - c_{12}),$$

что можно обобщить в виде:

$$\Pi = \sum_{i=1}^{12} (p_i q_i - c_i).$$

Заметим, что выражение для прибыли можно также переписать как разность общего дохода и общих издержек:

$$\Pi = (p_1 q_1 + \ldots + p_{12} q_{12}) - (c_1 + \ldots + c_{12}),$$

и с использованием знака Σ это выражение обобщается в виде:

$$\Pi = \sum_{i=1}^{12} p_i q_i - \sum_{i=1}^{12} c_i.$$

Если цена продукции в течение года постоянна и равна p, то выражение для стоимости годового выпуска можно упростить:

$$V = pq_1 + \ldots + pq_{12} = p(q_1 + \ldots + q_{12}) = p\sum_{i=1}^{12} q_i.$$

Следовательно,

$$\sum_{i=1}^{12} pq_i = p\sum_{i=1}^{12} q_i.$$

Если выпуск в каждом месяце постоянен и равен q, то выражение для годового выпуска также можно упростить:

$$Z = q_1 + \ldots + q_{12} = q + \ldots + q = 12q.$$

Следовательно, в этом случае

$$\sum_{i=1}^{12} q_i = 12q.$$

Мы проиллюстрировали три правила, которые могут быть записаны формально.

Правило суммирования 1 (проиллюстрировано разложением прибыли на общий доход минус общие издержки):

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

Правило суммирования 2 (проиллюстрировано выражением V в случае постоянной цены):

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i,$$

если a постоянно.

Правило суммирования 3 (проиллюстрировано выражением Z в случае постоянного объема выпуска):

$$\sum_{i=1}^{n} a = na,$$

если a постоянно.

Часто из контекста ясно, каковы начальный и конечный суммируемые члены. В этом случае выражение

$$\sum_{i=1}^{n} x_i$$

часто упрощается до Σx_i. Далее, часто столь же очевидно, какой индекс меняется при суммировании, и все выражение упрощается до Σx.

Иногда приходится выполнять суммирование внутри другого суммирования, что требует дальнейших пояснений. В принципе это нетрудно, но поскольку такие операции не выполняются в данной книге, мы не будем их рассматривать.

Приложение О.2

Математическое ожидание и дисперсия непрерывной случайной переменной

Определение математического ожидания непрерывной случайной переменной очень похоже на соответствующее определение дискретной случайной величины:

$$E(x) = \int xf(x)\,dx,$$

где интегрирование производится на всем интервале, где определена функция $f(x)$.

В обоих случаях разные возможные значения x взвешиваются по соответствующим им вероятностям. Для дискретной случайной величины суммирование осуществляется на основе последовательного перебора возможных значений x. В непрерывном случае это, конечно, производится на непрерывной основе, суммирование заменяется интегрированием, и значения вероятностей p_i заменяются значениями функции плотности вероятности $f(x)$. Принцип, однако, сохраняется тот же.

Дискретная	Непрерывная
$E(x) = \Sigma x_i p_i$	$E(x) = \int xf(x)\,dx$
Суммирование по всем возможным значениям	Интегрирование в области определения $f(x)$

В разделе, посвященном дискретным случайным величинам, показано, как рассчитать математическое ожидание функции $g(x)$ — функции случайной величины x. Берется список всех разных значений, которые может принимать $g(x)$, каждое взвешивается по соответствующей ему вероятности, и произведения суммируются.

Процедура для непрерывной случайной величины точно такая же, с той лишь разницей, что теперь она осуществляется на непрерывной основе, что означает суммирование путем интегрирования вместо Σ-суммирования. Для дискретной случайной величины $E\{g(x)\} = \Sigma g(x_i)p_i$, где суммирование производится по всем возможным значениям x. В непрерывном случае оно определяется как

$$E\{g(x)\} = \int g(x)f(x)\,dx.$$

где интегрирование производится по всей области определения $f(x)$.

Что касается дискретных случайных переменных, то здесь есть одна функция, которая особенно нас интересует, — теоретическая дисперсия. В обзоре она была определена как математическое ожидание $(x-\mu)^2$, где μ — теоретическое среднее (то же самое, что $E(x)$). Чтобы рассчитать дисперсию, нужно просуммировать значения $(x-\mu)^2$, взвешенные по соответствующим вероятностям, по всем возможным значениям x. Применительно к непрерывной случайной переменной это означает, что нужно вычислить σ^2 — теоретическую дисперсию x:

$$\sigma^2 = E\{(x-\mu)^2\} = \int (x-\mu)^2 f(x)\,dx.$$

В познавательных целях было бы полезным сравнить это равенство с уравнением (О.9), где дано аналогичное выражение для дискретной случайной переменной (переверните несколько страниц назад и проверьте).

Как и ранее, при расчете теоретической дисперсии вы можете вычислить теоретическое стандартное отклонение (σ) путем простого извлечения из нее квадратного корня.

Приложение О.3

Доказательство того, что s^2 — несмещенная оценка теоретической дисперсии

В табл. О.5 указано, что несмещенная оценка σ^2 рассчитывается по формуле s^2, где

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2.$$

Приступим к доказательству, переписав $(x_i - \bar{x})^2$ в более сложном, но полезном виде:

$$(x_i - \bar{x})^2 = \{(x_i - \mu) - (\bar{x} - \mu)\}^2 = (x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2$$

(при раскрытии скобок μ сократятся). Следовательно,

$$\sum(x_i - \bar{x})^2 = \sum(x_i - \mu)^2 - 2(\bar{x} - \mu)\sum(x_i - \mu) + n(\bar{x} - \mu)^2.$$

Первое слагаемое здесь есть сумма первых слагаемых предыдущего уравнения, записанная с использованием знака Σ. Аналогично второе слагаемое здесь есть сумма вторых слагаемых предыдущего уравнения, вновь с использованием знака Σ и того факта, что $(\bar{x} - \mu)^2$ — общий множитель. Перейдя к суммированию третьих членов предыдущего уравнения, отметим, что все они равны $(\bar{x} - \mu)^2$ и поэтому нет необходимости использовать знак Σ.

Вторая составляющая может быть переписана как $-2n(\bar{x} - \mu)^2$, поскольку

$$\sum(x_i - \mu) = \sum x_i - n\mu = n\bar{x} - n\mu = n(\bar{x} - \mu),$$

и мы получаем:

$$\sum(x_i - \bar{x})^2 = \sum(x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2 = \sum(x_i - \mu)^2 - n(\bar{x} - \mu)^2.$$

Беря в этом уравнении математические ожидания, имеем:

$$\begin{aligned} E\{\sum(x_i - \bar{x})^2\} &= E\{\sum(x_i - \mu)^2\} - nE\{(\bar{x} - \mu)^2\} = \\ &= E\{(x_i - \mu)^2\} + \ldots + E(x_n - \mu)^2\} - nE\{(\bar{x} - \mu)^2 = \\ &= n\,\text{pop.var}(x) - n\,\text{pop.var}(\bar{x}) = \\ &= n\sigma^2 - n(\sigma^2 / n) = (n-1)\sigma^2, \end{aligned}$$

используя тот факт, что теоретическая дисперсия $\bar{x}$ равна σ^2/n. Это доказывается в разделе 1.7. Следовательно,

$$E(s^2) = E\{\frac{1}{n-1}\sum(x_i - \bar{x})^2\} = \sigma^2$$

и поэтому s^2 — несмещенная оценка σ^2.

1

КОВАРИАЦИЯ, ДИСПЕРСИЯ И КОРРЕЛЯЦИЯ

В данной главе вводятся понятия ковариации и корреляции, которые подготовят почву для предстоящего изложения идей и понятий регрессионного анализа. Второй, не менее важной целью является демонстрация правил расчета выборочных дисперсии и ковариации и их выражений. Для закрепления практических навыков подробно разбирается несколько примеров. Они будут часто использоваться в последующих главах, и очень важно, чтобы вы хорошо их усвоили. Эти примеры существенно упрощают математические выкладки и значительно облегчают проведение анализа.

1.1. Выборочная ковариация

Выборочная ковариация является мерой взаимосвязи между двумя переменными. Данное понятие будет проиллюстрировано на простом примере. Со времен нефтяного кризиса 1973 г. реальная цена на бензин, т. е. цена бензина, отнесенная к уровню общей инфляции, значительно возросла, и это оказало заметное воздействие на потребительский спрос. Просматривая данные табл. Б.1, помещенной в приложении в конце книги, можно увидеть, что в период между 1963 и 1972 гг. потребительский спрос на бензин устойчиво повышался. Эта тенденция прекратилась в 1973 г., а затем последовали нерегулярные колебания спроса с незначительным его падением в целом. В табл. 1.1 приведены данные о потребительском спросе и реальных ценах после нефтяного кризиса. (Реальная цена вычислялась путем деления индекса номинальной цены на бензин, приведенного в табл. Б.2, на общий индекс потребительских цен из той же таблицы и умножения результата на 100. Индексы в табл. Б.2 основаны на данных 1972 г.; таким образом, индекс реальной цены в табл. 1.1 показывает повышение цены бензина относительно общей инфляции начиная с 1972 г.)

На рис. 1.1 эти данные показаны в виде диаграммы рассеяния. Можно видеть некоторую отрицательную связь между потребительским спросом на бензин и его реальной ценой.

Показатель выборочной ковариации позволяет выразить данную связь единым числом. Для его вычисления мы сначала находим средние (для рассматриваемого выборочного периода) значения цены и спроса на бензин. Обозначив цену через p и спрос — через y, мы, таким образом, определяем $\bar{p}$ и $\bar{y}$, кото-

Таблица 1.1

Потребительские расходы на бензин и его реальная цена в США

Год	Расходы (млдр. долл., цены 1972 г.)	Индекс реальных цен (1972=100)
1973	26,2	103,5
1974	24,8	127,0
1975	25,6	126,0
1976	26,8	124,8
1977	27,7	124,7
1978	28,3	121,6
1979	27,4	149,7
1980	25,1	188,8
1981	25,2	193,6
1982	25,6	173,9

рые для этой выборки оказываются равными соответственно 143,36 и 26,27. Затем для каждого года вычисляем отклонение величин p и y от средних и перемножаем их. Для первого года $(p - \bar{p})$ равно (103,5 − 143,36), или −39,86, и $(y - \bar{y})$ равно (26,2 − 26,27), или −0,07, а произведение $(p - \bar{p})(y - \bar{y})$ составит 2,79. Проделаем это для всех годов выборки и возьмем среднюю величину, она и будет выборочной ковариацией (как видно, не очень сложно вычисляемой).

Определение

При наличии n наблюдений двух переменных (x и y) выборочная ковариация между x и y задается формулой:

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\{(x_1 - \bar{x})(y_i - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})\}. \qquad (1.1)$$

Замечание 1. В разделе 1.4 мы определим также ковариацию генеральной совокупности. Для различения этих двух ковариаций мы используем обозначение Cov (x, y) с прописной буквы C применительно к выборочной ковариации и pop. cov (x, y) — для ковариации между x и y в генеральной совокупности. Иногда последнюю будет удобно обозначать как σ_{xy}. Аналогичные обозначения мы используем и для дисперсии: Var (x) — применительно к выборочной дисперсии и pop. var (x) — к дисперсии для генеральной совокупности (теоретической).

Замечание 2. В некоторых учебниках выборочная ковариация по аналогии с выборочной дисперсией определяется путем деления на $(n - 1)$ вместо n по

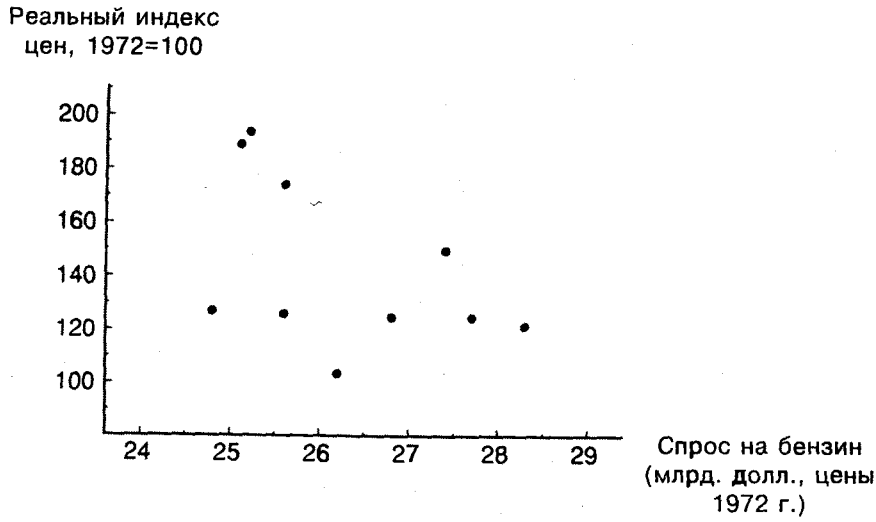


Рис. 1.1. Спрос на бензин в США, 1973–1982 гг.

причинам, которые будут объяснены в разделе 1.5. В примере с бензином детали проведенных вычислений для всей выборки приведены в табл. 1.2. Здесь в столбцах *2* и *3* представлены исходные данные для $\bar{p}$ и $\bar{y}$. В результирующих строках вычисляются p и y. В столбцах *4* и *5* рассчитываются $(p-\bar{p})$ и $(y-\bar{y})$ для каждого года, а в столбце *6* эти две величины перемножаются. В нижней клетке

Таблица 1.2

Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})(y-\bar{y})$
1973	103,5	26,2	-39,86	-0,07	2,79
1974	127,0	24,8	-16,36	-1,47	24,05
1975	126,0	25,6	-17,36	-0,67	11,63
1976	124,8	26,8	-18,56	0,53	-9,84
1977	124,7	27,7	-18,66	1,43	-26,68
1978	121,6	28,3	-21,76	2,03	-44,17
1979	149,7	27,4	6,34	1,13	7,16
1980	188,8	25,1	45,44	-1,17	-53,16
1981	193,6	25,2	50,24	-1.07	-53,76
1982	173,9	25,6	30,54	-0,67	-20,46
Сумма	1433,6	262,7			-162,44
Среднее	143,36	26,27			-16,24

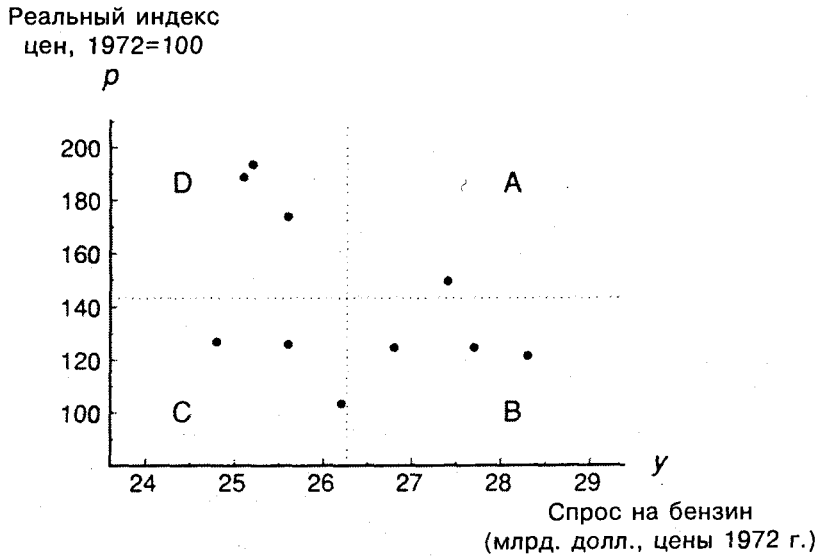


Рис. 1.2

последнего столбца определяется средняя величина (−16,24), она и является значением выборочной ковариации.

Вы должны заметить, что ковариация в данном случае отрицательна. Так это и должно быть. Отрицательная связь, как это имеет место в данном примере, выражается отрицательной ковариацией, а положительная связь — положительной ковариацией.

Имеет смысл рассмотреть причину этого. Рисунок 1.2 точно такой же, как и рис. 1.1, но здесь диаграмма рассеяния наблюдений делится на четыре части вертикальной и горизонтальной линиями, проведенными через $\bar{p}$ и $\bar{y}$ соответственно. Пересечение этих линий образует точку $(\bar{p}, \bar{y})$, которая показывает среднюю цену и средний спрос за период времени, соответствующий нашей выборке. Используя аналогию из физики, можно сказать, что эта точка является центром тяжести совокупности точек, представляющих наблюдения.

Для любого наблюдения, лежащего в квадранте *A*, значения реальной цены и спроса выше соответствующих средних значений. Для данных наблюдений как $(p-\bar{p})$, так и $(y-\bar{y})$ являются положительными, а поэтому должно быть положительным и $(p-\bar{p})(y-\bar{y})$. Наблюдение, таким образом, дает положительный вклад в ковариацию. Так, например, наблюдение за 1979 г. лежит в этом квадранте и $(p-\bar{p}) = 6{,}34$, $(y-\bar{y}) = 1{,}13$, а их произведение равно 7,16.

Далее рассмотрим квадрант *B*. Здесь наблюдения имеют реальную цену ниже средней и спрос выше среднего. Поэтому $(p-\bar{p})$ отрицательно, $(y-\bar{y})$ положительно, произведение $(p-\bar{p})(y-\bar{y})$ отрицательно, и наблюдение вносит отрицательный вклад в ковариацию. Например, наблюдение за 1978 г. имеет $(p-\bar{p}) = -21{,}76$, $(y-\bar{y}) = 2{,}03$ и $(p-\bar{p})(y-\bar{y})$, таким образом, равно −44,17.

В квадранте *C* как реальная цена, так и спрос ниже своих средних значе-

ний. Таким образом, $(p-\bar{p})$ и $(y-\bar{y})$ оба являются отрицательными, а $(p-\bar{p})(y-\bar{y})$ положительно. (В качестве примера см. наблюдение за 1974 г.)

Наконец, в квадранте D реальная цена выше средней, а спрос ниже среднего. Таким образом, $(p-\bar{p})$ положительно, $(y-\bar{y})$ отрицательно, поэтому $(p-\bar{p})(y-\bar{y})$ отрицательно, и в ковариацию, соответственно, вносится отрицательный вклад. (В качестве примера см. наблюдение за 1981 г.)

Поскольку выборочная ковариация является средней величиной произведения $(p-\bar{p})(y-\bar{y})$ для 20 наблюдений, она будет положительной, если положительные вклады будут доминировать над отрицательными, и отрицательной, если будут доминировать отрицательные вклады. Положительные вклады исходят из квадрантов A и C, и ковариация будет, скорее всего, положительной, если основной разброс пойдет по наклонной вверх. Точно так же отрицательные вклады исходят из квадрантов B и D. Поэтому если основное рассеяние идет по наклонной вниз, как в данном примере, то ковариация будет, скорее всего, отрицательной.

1.2. Несколько основных правил расчета ковариации

Есть несколько важных правил, которые вытекают непосредственно из определения ковариации. Поскольку они будут многократно использоваться в последующих главах, имеет смысл сформулировать их сейчас:

Правило 1

Если $y = v + w$, то $\mathrm{Cov}(x, y) = \mathrm{Cov}(x, y) + \mathrm{Cov}(x, w)$.

Правило 2

Если $y = az$, где a — константа, то $\mathrm{Cov}(x, y) = a\,\mathrm{Cov}(x, z)$.

Правило 3

Если $y = a$, где a — константа, то $\mathrm{Cov}(x, y) = 0$.

Сначала эти правила будут проиллюстрированы на примерах, и мы проверим их выполнение, после чего будут приведены доказательства. В большей части данной книги важнее понимать, что означают эти правила и как ими пользоваться, чем уметь доказывать их, но на самом деле доказательства нетрудны.

Демонстрация и доказательство правила 1

Допустим, что у нас есть данные по шести семьям (домохозяйствам), приведенные в табл. 1.3: общий годовой доход (x); расходы на питание и одежду (y);

Таблица 1.3

Семья	Доход семьи (x)	Расходы на питание и одежду (y)	Расходы на питание (v)	Расходы на одежду (w)	Вторая выборка: расходы семьи на питание и одежду (z)
1	3000	1100	850	250	2200
2	2500	850	700	150	1700
3	4000	1200	950	250	2400
4	6000	1600	1150	450	3200
5	3300	1000	800	200	2000
6	4500	1300	950	350	2600
Сумма	23300	7050	5400	1650	14100
Среднее	3883	1175	900	275	2350

расходы на питание (v) и расходы на одежду (w). Естественно, y равняется сумме v и w. Указанную в таблице величину z рассматривать пока не будем.

В табл. 1.4 величины $(x-\bar{x})$, $(y-\bar{y})$, $(v-\bar{v})$ и $(w-\bar{w})$ вычисляются для каждой семьи. Отсюда получаем $(x-\bar{x})(y-\bar{y})$, $(x-\bar{x})(v-\bar{v})$ и $(x-\bar{x})(w-\bar{w})$ для каждой семьи. $\mathrm{Cov}(x,y)$ получается как среднее величин $(x-\bar{x})(y-\bar{y})$ и равняется 266 250. Аналогично $\mathrm{Cov}(x,y)$ равна 157 500 и $\mathrm{Cov}(x,w) = 108\,750$. Мы проверили, что $\mathrm{Cov}(x,y)$ является суммой $\mathrm{Cov}(x,v)$ $\mathrm{Cov}(x,v)$ и $\mathrm{Cov}(x,w)$.

Таблица 1.4

Семья	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(v-\bar{v})$	$(x-\bar{x})(v-\bar{v})$	$(w-\bar{w})$	$(x-\bar{x})(w-\bar{w})$
1	-883	-75	66250	-50	44167	-25	22083
2	-1383	-325	449583	-200	276667	-125	172917
3	117	25	2917	50	5833	-25	-2917
4	2117	425	899583	250	529167	175	370416
5	-583	-175	102083	-100	58333	-75	43750
6	617	125	77083	50	30833	75	46250
Сумма			1597500		945000		652500
Среднее			266250		157500		108750

Легко показать, что именно так и должно быть. Рассмотрим i-ю семью. $(x_i - \bar{x})(y_i - \bar{y})$ — это ее вклад в величину Cov(x, y). Поскольку $y_i = v_i + w_i$ и $\bar{y} = \bar{v} + \bar{w}$, то

$$(x_i - \bar{x})(y_i - \bar{y}) = (x_i - \bar{x})(v_i + w_i - \bar{v} - \bar{w}) = (x_i - \bar{x})(v_i - \bar{v}) + (x_i - \bar{x})(w_i - \bar{w}), \qquad (1.2)$$

и, таким образом, мы показали, что вклад семьи i в Cov(x, y) является суммой ее вкладов в Cov(x, v) и Cov(x, w). То же самое справедливо для всех семей и, соответственно, для ковариации в целом.

Демонстрация и доказательство правила 2

В табл. 1.3 последняя колонка (z) дает расходы на питание и одежду для второго множества из 6 семей. Каждое наблюдение z фактически представляет собой удвоенное значение y. Предполагается, что значения величины x для второго набора семей являются такими же, как и ранее. Для вычисления Cov(x, z) нам, как и ранее, необходимы значения $(x - \bar{x})$, а также $(z - \bar{z})$ (табл. 1.5).

Таблица 1.5

Семья	$(x - \bar{x})$	$(z - \bar{z})$	$(x - \bar{x})(z - \bar{z})$
1	-883	-150	132500
2	-1383	-650	899167
3	117	50	5833
4	2117	850	1700167
5	-583	-350	204167
6	617	250	154167
Сумма			3195000
Среднее			532500

Из табл. 1.5 можно видеть, что Cov(x, z) равна 532 500, что в точности равно удвоенной Cov(x, y). Таким образом мы проверили, что Cov$(x, 2y)$ совпадает с 2Cov(x, y).

И снова легко видеть, почему так получается. Рассмотрим первую семью. Поскольку $z_1 = 2y_1$ и $\bar{z} = 2\bar{y}$, а $(x_1 - \bar{x})(z_1 - \bar{z})$ равно $(x_1 - \bar{x})(2y_1 - 2\bar{y})$ и, следовательно, равно $2(x_1 - \bar{x})(y_1 - \bar{y})$, то вклад первой семьи в величину Cov(x, z) в точности равен двойной величине ее вклада в Cov(x, y). То же самое справедливо для всех других семей. Средняя величина $(x - \bar{x})(z - \bar{z})$ поэтому равна уд-

военной средней величине $(x-\bar{x})(y-\bar{y})$ и, таким образом, $\mathrm{Cov}(x, z) = 2\mathrm{Cov}(x, y)$. Обобщая, получим, что если $z = ay$ (и отсюда $\bar{z} = a\bar{y}$), то

$$\mathrm{Cov}(x,z) = 2\mathrm{Cov}(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(z_i-\bar{z}) = \frac{1}{n}\sum(x_i-\bar{x})(ay_i-a\bar{y}) =$$

$$= \frac{a}{n}\sum(x_i-\bar{x})(y_i-\bar{y}) = a\mathrm{Cov}(x,y). \qquad (1.3)$$

Демонстрация и доказательство правила 3

Это совсем просто. Допустим, что каждая семья в выборке имеет по два взрослых человека, и предположим, что по недоразумению вы решили вычислить ковариацию между общим доходом (x) и числом взрослых в семье (a). Естественно, что $a_1 = a_2 = \ldots = a_6 = 2$. Таким образом, $\bar{a} = 2$. Отсюда для каждой семьи $(a-\bar{a}) = 0$ и, следовательно, $(x-\bar{x})(a-\bar{a}) = 0$. Поэтому $\mathrm{Cov}(x, a) = 0$.

Если вы настаиваете на построении обычно используемой в таких случаях таблицы, то она будет выглядеть как табл. 1.6.

Таблица 1.6

Семья	x	a	$(x-\bar{x})$	$(a-\bar{a})$	$(x-\bar{x})(a-\bar{a})$
1	3000	2	-883	0	0
2	2500	2	-1383	0	0
3	4000	2	117	0	0
4	6000	2	2117	0	0
5	3300	2	-583	0	0
6	4500	2	617	0	0
Сумма	23300	12			0
Среднее	3883	2			0

Дальнейшие выводы

Пользуясь этими основными правилами, вы можете упрощать значительно более сложные выражения с ковариациями. Например, если какая-то переменная равна сумме трех переменных — u, v и w, то, пользуясь правилом *1* и разбив y на две части (u и $v + w$), получим:

$$\mathrm{Cov}(x,y) = \mathrm{Cov}(x,u+v+w) = \mathrm{Cov}(x,u) + \mathrm{Cov}(x,v+w) \qquad (1.4)$$

и, снова воспользовавшись правилом *1*, имеем:

$$\mathrm{Cov}(x, y) = \mathrm{Cov}(x, u) + \mathrm{Cov}(x, v) + \mathrm{Cov}(x, w). \tag{1.5}$$

Другой пример: если $y = a + bz$, где a и b — константы, а z — переменная величина, то, пользуясь последовательно правилами *1*, *3* и *2*, получим:

$$\mathrm{Cov}(x, y) = \mathrm{Cov}(x, a) + \mathrm{Cov}(x, bz) = 0 + \mathrm{Cov}(x, bz) = b\mathrm{Cov}(x, z). \tag{1.6}$$

При наличии небольшой практики выполнить эти преобразования не составит труда.

1.3. Альтернативное выражение для выборочной ковариации

Выборочная ковариация между x и y определяется как

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{1.7}$$

Другим эквивалентным выражением является

$$\mathrm{Cov}(x, y) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \bar{x}\bar{y}. \tag{1.8}$$

Иными словами, ковариацию вы можете получить, вычисляя $x_1y_1 + x_2y_2 + \ldots + x_ny_n$, деля результат на n и вычитая затем из полученного значения величину $\bar{x}\bar{y}$. Это может оказаться более удобным, если вам придется рассчитывать ковариацию вручную. Но, конечно, на практике вы обычно будете делать это при помощи компьютерных программ.

Далее, для тех, кого это интересует, приводится доказательство эквивалентности указанных выражений. В случае возникновения затруднений его можно пропустить.

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} - \bar{x}\bar{y}) =$$

$$= \frac{1}{n}\{x_1y_1 - \bar{x}y_1 - x_1\bar{y} + \bar{x}\bar{y} + x_2y_2 - \bar{x}y_2 - x_2\bar{y} + \bar{x}\bar{y} + \ldots + x_ny_n - \bar{x}y_n - x_n\bar{y} + \bar{x}\bar{y}\}. \tag{1.9}$$

В результате сложения по столбцам, а также воспользовавшись тем, что $\sum x_i = n\bar{x}$ и $\sum y_i = n\bar{y}$, получим:

$$\mathrm{Cov}(x, y) = \frac{1}{n}\left\{\sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}\right\} = \frac{1}{n}\left\{\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right\}, \tag{1.10}$$

откуда и следует (1.8).

Упражнение

1.1. В некоторой бюрократической стране годовой доход каждого индивида y определяется по формуле:

$$y = 10000 + 500s + 200t,$$

где s — число лет обучения индивида; t — трудовой стаж (в годах); x — возраст индивида. Рассчитайте $\mathrm{Cov}(x, y)$, $\mathrm{Cov}(x, s)$ и $\mathrm{Cov}(x, t)$ для выборки из пяти индивидов, описанной ниже, и проверьте, что

$$\mathrm{Cov}(x, y) = 500\ \mathrm{Cov}(x, s) + 200\ \mathrm{Cov}(x, t).$$

Объясните аналитически, почему так происходит.

Индивид	Возраст (годы)	Годы обучения	Трудовой стаж	Доход
1	18	11	1	15700
2	29	14	6	18200
3	33	12	8	17600
4	35	16	10	20000
5	45	12	5	17000

1.4. Теоретическая ковариация

Если x и y — случайные величины, то *теоретическая ковариация* σ_{xy} определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$\mathrm{pop.cov}(x, y) = \sigma_{xy} = E\{(x - \mu_x)(y - \mu_y)\}, \tag{1.11}$$

где μ_x и μ_y — теоретические средние значения x и y соответственно.

Как вы и ожидаете, если теоретическая ковариация неизвестна, то для ее оценки может быть использована выборочная ковариация, вычисленная по ряду наблюдений. К сожалению, оценка будет иметь отрицательное смещение, так как

$$E\{\mathrm{Cov}(x, y)\}\frac{n-1}{n}\,\mathrm{pop.cov}(x, y). \tag{1.12}$$

Причина заключается в том, что выборочные отклонения измеряются по отношению к выборочным средним значениям величин x и y и имеют тенденцию к занижению отклонений от истинных средних значений. Очевидно, мы можем рассчитать несмещенную оценку путем умножения выборочной оценки на $n/(n-1)$. Доказательство соотношения (1.12) здесь не представлено, но

вы можете сами провести его, используя в качестве руководства приложение О.3 (предварительно ознакомьтесь с содержанием раздела 1.5). Правила для теоретической ковариации точно такие же, как и для выборочной ковариации, но их доказательства мы опускаем, поскольку для этого требуется интегральное исчисление.

Если x и y независимы, то их теоретическая ковариация равна нулю, поскольку

$$E\{(x-\mu_x)(y-\mu_y)\} = E(x-\mu_x)(y-\mu_y) = 0\times 0, \tag{1.13}$$

благодаря свойству независимости, отмеченному в обзоре, и факту, что $E(x)$ и $E(y)$ равняются соответственно μ_x и μ_y.

1.5. Выборочная дисперсия

До сих пор термин «дисперсия» использовался в смысле теоретической дисперсии (т. е. относящейся ко всей генеральной совокупности), как это и определялось в обзоре. Для целей, которые прояснятся при обсуждении регрессионного анализа, целесообразно ввести понятие *выборочной дисперсии* (при этом будет сделано три важных замечания). Для выборки из n наблюдений $x_1, \ldots, x_n$ выборочная дисперсия определяется как среднеквадратичное отклонение в выборке:

$$\mathrm{Var}(x) = \frac{1}{n}\sum(x_i - \bar{x})^2. \tag{1.14}$$

Сделаем следующие три замечания:

1. Определенная таким образом выборочная дисперсия представляет собой смещенную оценку теоретической дисперсии. В приложении О.3 показано, что s^2, определенная как

$$\frac{1}{n-1}\sum(x_i - \bar{x})^2,$$

является несмещенной оценкой σ^2. Отсюда следует, что ожидаемое значение величины Var (x) равно $[(n-1)/n]\sigma^2$ и что, следовательно, она имеет отрицательное смещение. Отметим, что если размер выборки n становится большим, то $(n-1)/n$ стремится к единице и, таким образом, математическое ожидание величины Var (x) стремится к σ^2. Можно легко показать, что ее предел по вероятности (plim) равен σ^2 и, следовательно, она является примером состоятельной оценки, которая смещена для небольших выборок.

2. Так как величина s^2 является несмещенной, то в некоторых работах ее часто определяют как выборочную дисперсию и либо избегают ссылок на Var (x), либо дают ей какое-то другое название. К сожалению, общепринятой договоренности по этому поводу до сих пор нет[1]. В каждой работе вам следует проверять определение.

[1] В русскоязычной литературе величина Var (x) обычно называется выборочной дисперсией, а s^2 — «исправленной», или несмещенной, выборочной дисперсией. (*Прим. ред.*)

3. Поскольку указанная договоренность отсутствует, отсутствует и договоренность относительно условного обозначения данного понятия, и для этого используются самые различные символы. В данной работе теоретическая (или генеральная) дисперсия переменной x обозначается как pop. var (x) или, если это удобно, σ_x^2. Если ясно, о какой переменной идет речь, то нижний индекс может быть опущен. Выборочная дисперсия всегда будет обозначаться как Var (x) с прописной буквы V.

Почему выборочная дисперсия в среднем занижает значение теоретической дисперсии? Причина заключается в том, что она вычисляется как среднеквадратичное отклонение от выборочного среднего, а не от истинного значения. Так как выборочное среднее автоматически находится в центре выборки, то отклонения от него в среднем меньше отклонений от теоретического среднего значения.

1.6. Правила расчета дисперсии

Существует несколько простых и очень полезных правил для расчета дисперсии, являющихся аналогами правил для ковариации, рассмотренных в разделе 1.2. Эти правила в равной степени можно использовать как для выборочной, так и для теоретической дисперсии.

Правило дисперсии 1

Если $y = v + w$, то $\text{Var}(y) = \text{Var}(v) + \text{Var}(w) + 2\text{Cov}(v,w)$.

Правило дисперсии 2

Если $y = az$, где a является постоянной, то $\text{Var}(y) = a^2\text{Var}(z)$.

Правило дисперсии 3

Если $y = a$, где a является постоянной, то $\text{Var}(y) = 0$.

Правило дисперсии 4

Если $y = v + a$, где a является постоянной, то $\text{Var}(y) = \text{Var}(v)$.

Во-первых, заметим, что дисперсия переменной x может рассматриваться как ковариация между двумя величинами x:

$$\text{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \text{Cov}(x,x). \qquad (1.15)$$

Учитывая это равенство, мы можем воспользоваться правилами расчета выборочной ковариации, чтобы вывести правила расчета дисперсии. Кроме того,

мы можем получить другую формулу для представления Var (x), используя соотношение (1.8) для выборочной ковариации:

$$\text{Var}(x) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] - \bar{x}^2. \tag{1.16}$$

Доказательство правила 1

Если $y = v + w$, то

$$\begin{aligned}
\text{Var}(y) &= \text{Cov}(y, y) = \text{Cov}(y,[v + w]) = \\
&= \text{Cov}([v + w],v) + \text{Cov}([v + w],w), \text{ по правилу ковариации } 1, \\
&= \text{Cov}(v, v) + \text{Cov}(w, v) + \text{Cov}(v, w) + \text{Cov}(w, w), \\
&\quad \text{по правилу ковариации } 1, \\
&= \text{Var}(v) + \text{Var}(w) + 2\,\text{Cov}(v, w).
\end{aligned} \tag{1.17}$$

Доказательство правила 2

Если $y = az$, где a является постоянной, то, дважды используя правило ковариации *2*, получим:

$$\begin{aligned}
\text{Var}(y) &= \text{Cov}(y, y) = \text{Cov}(y, az) = a\,\text{Cov}(y, z) = \\
&= a\,\text{Cov}(az, z) = a^2\,\text{Cov}(z, z) = a^2\,\text{Var}(z).
\end{aligned} \tag{1.18}$$

Доказательство правила 3

Если $y = a$, где a является постоянной, то по правилу ковариации *3* имеем:

$$\text{Var}(y) = \text{Cov}(a, a) = 0. \tag{1.19}$$

Действительно, если y — постоянная, то ее среднее значение является той же самой постоянной и $(y - \bar{y})$ равняется нулю для всех наблюдений. Следовательно, Var (y) равна нулю.

Доказательство правила 4

Если $y = v + a$, где a — постоянная, то по правилу ковариации *1*, используя затем правила *1* и *3* для дисперсии и правило *3* для ковариации, получаем:

$$\text{Var}(y) = \text{Var}(v + a) = \text{Var}(v) + \text{Var}(a) + 2\,\text{Cov}(v, a) = \text{Var}(v). \tag{1.20}$$

Теоретическая дисперсия подчиняется тем же самым правилам, но доказательства здесь вновь опускаются, поскольку они требуют применения интегрального исчисления.

Упражнение

1.2. Используя данные из упражнения 1.1, вычислите Var (y), Var (s) и Var (t) и проверьте, что

$$\text{Var}\,(y) = 250000\,\text{Var}\,(s) + 40000\,\text{Var}\,(t) + 200000\,\text{Cov}\,(s, t),$$

при этом результат объясните аналитически.

1.7. Теоретическая дисперсия выборочного среднего

Если две переменные независимы (и следовательно, их совокупная ковариация равняется нулю), то теоретическая дисперсия суммы этих переменных будет равна сумме их теоретических дисперсий:

$$\text{pop. var}\,(x + y) = \text{pop. var}\,(x) + \text{pop. var}\,(y) + 2\,\text{pop. cov}\,(x, y) =$$

$$= \text{pop. var}\,(x) + \text{pop. var}\,(y) = \sigma_x^2 + \sigma_y^2. \qquad (1.21)$$

Из данного результата можно получить более общее правило о том, что теоретическая дисперсия суммы любого числа переменных равняется сумме их дисперсий при условии, что наблюдения независимы друг от друга. При этом можно показать, что если случайная переменная x имеет дисперсию σ^2, то дисперсия выборочного среднего $\bar{x}$ будет равна σ^2/n, где n — число наблюдений в выборке:

$$\text{pop.var}\,(\bar{x}) = \text{pop.var}\left\{\frac{x_1 + \ldots + x_n}{n}\right\} = \frac{1}{n^2}\,\text{pop.var}\,(x_1 + \ldots + x_n) =$$

$$= \frac{1}{n^2}\{\text{pop. var}\,(x_1) + \ldots + \text{pop. var}\,(x_n)\} = \frac{1}{n^2}\{\sigma^2 + \ldots + \sigma^2\} = \frac{1}{n^2}\{n\sigma^2\} = \sigma^2 / n. \qquad (1.22)$$

Как было показано в обзоре, выборочное среднее является наиболее эффективной несмещенной оценкой теоретического среднего при условии, что наблюдения проводятся независимо друг от друга на основе одного и того же распределения.

1.8. Коэффициент корреляции

В этой главе большое внимание уделено ковариации. Это объясняется тем, что она весьма удобна с математической точки зрения, а вовсе не тем, что ковариация является особенно хорошим измерителем взаимосвязи между величинами. Мы рассмотрим ее недостатки в разделе 1.9. Более точной мерой зависимости является тесно связанный с ней *коэффициент корреляции.*

Подобно дисперсии и ковариации, коэффициент корреляции имеет две формы — теоретическую и выборочную. *Теоретический коэффициент корреляции* традиционно обозначается греческой буквой ρ, которая произносится как «ро» и соответствует латинской *«r»*. Для переменных x и y этот коэффициент определяется следующим образом:

$$\rho_{x,y} = \frac{\text{pop.cov}(x,y)}{\sqrt{\text{pop.var}(x)\text{pop.var}(y)}} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2\sigma_y^2}}. \tag{1.23}$$

Если x и y независимы, то ρ равно нулю, так как равна нулю теоретическая ковариация. Если между переменными существует положительная зависимость, то $\sigma_{x,y}$, а следовательно, и $\rho_{x,y}$ будут положительными. Если существует строгая положительная линейная зависимость, то $\rho_{x,y}$ примет максимальное значение, равное 1. Аналогичным образом при отрицательной зависимости $\rho_{x,y}$ будет отрицательным с минимальным значением -1.

Выборочный коэффициент корреляции r определяется путем замены теоретических дисперсий и ковариации в выражении (1.23) на их несмещенные оценки. Мы показали, что такие оценки могут быть получены умножением выборочных дисперсий и ковариации на $n/(n-1)$. Следовательно,

$$r_{x,y} = \frac{\frac{n}{n-1}\text{Cov}(x,y)}{\sqrt{\frac{n}{n-1}\text{Var}(x)\frac{n}{n-1}\text{Var}(y)}}. \tag{1.24}$$

Множители $n/(n-1)$ сокращаются, поэтому можно определить выборочную корреляцию как

$$r_{x,y} = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}. \tag{1.25}$$

Подобно величине ρ, r имеет максимальное значение, равное единице, которое получается при строгой линейной положительной зависимости между выборочными значениями x и y (когда на диаграмме рассеяния все точки находятся точно на восходящей прямой линии). Аналогичным образом r принимает минимальное значение -1, когда существует линейная отрицательная зависимость (точки лежат точно на нисходящей прямой линии). Величина $r = 0$ показывает, что зависимость между наблюдениями x и y в выборке отсутствует. Разумеется, тот факт, что $r = 0$, необязательно означает, что $\rho = 0$, и наоборот.

Иллюстрация

Для иллюстрации вычисления выборочного коэффициента корреляции мы используем пример о спросе на бензин из раздела 1.1. Данные представлены в табл. 1.1 и показаны на рис. 1.1. Мы уже вычислили Cov (p, y) (см. табл. 1.2), которая составляет $-16{,}24$, поэтому нам теперь необходимы только Var (p) и Var (y) (см. табл. 1.7).

В последних двух колонках табл. 1.7 можно найти, что Var (p) составляет 888,58 и Var (y) равна 1,33. Следовательно,

$$r = \frac{-16{,}24}{\sqrt{888{,}58 \times 1{,}33}} = \frac{-16{,}24}{34{,}38} = -0{,}47. \tag{1.26}$$

Таблица 1.7

Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})^2$	$(y-\bar{y})^2$
1	103,5	26,2	-39,86	-0,07	1588,82	0,01
2	127,0	24,8	-16,36	-1,47	267,65	2,16
3	126,0	25,6	-17,36	-0,67	301,37	0,45
4	124,8	26,8	-18,56	0,53	344,47	0,28
5	124,7	27,7	-18,66	1,43	348,20	2,05
6	121,6	28,3	-21,76	2,03	473,50	4,12
7	149,7	27,4	6,34	1,13	40,20	1,28
8	188,8	25,1	45,44	-1,17	2064,79	1,37
9	193,6	25,2	50,24	-1,07	2524,06	1,15
10	173,9	25,6	30,54	-0,67	932,69	0,45
Сумма	1433,6	262,7			8885,75	13,30
Среднее	143,36	26,27			888,58	1,33

Упражнения

1.3. На с. 50 представлены данные о темпах прироста численности занятых — e и темпах прироста производительности труда — p (выпуска продукции за один человеко-час) для промышленности 12 стран за период с 1953–1954 по 1963–1964 гг. (годовые экспоненциальные темпы прироста). Постройте диаграмму рассеяния и вычислите выборочный коэффициент корреляции между e и p. [Рекомендуется сделать его, используя уравнения (1.8) и (1.16) для выборочной ковариации и дисперсии, и сохранить вычисления, поскольку это сэкономит вам время при рассмотрении другого примера, представленного в главе 2.] Объясните полученные результаты и прокомментируйте возможные причины положительной корреляции между двумя переменными.

1.4. Пусть наблюдения двух случайных переменных x и y находятся на прямой линии:

$$y = a + bx.$$

Покажите, что $\mathrm{Cov}(x, y) = b\,\mathrm{Var}(x)$ и что $\mathrm{Var}(y) = b^2\,\mathrm{Var}(x)$, а следовательно, выборочный коэффициент корреляции равен 1, если наклон линии положителен, и -1, если этот наклон отрицателен.

1.5. Пусть переменная y определяется строгой линейной зависимостью:

$$y = a + bx,$$

и предположим, что для x, y и третьей переменной z получена выборка наблюдений. Покажите, что если коэффициент b положителен, то выборочный коэффициент корреляции для y и z должен быть таким же, как и для x и z.

Годовые темпы прироста продукции (%)		
	Занятость	*Производительность*
Австрия	2,0	4,2
Бельгия	1,5	3,9
Канада	2,3	1,3
Дания	2,5	3,2
Франция	1,9	3,8
Италия	4,4	4,2
Япония	5,8	7,8
Нидерланды	1,9	4,1
Норвегия	0,5	4,4
ФРГ	2,7	4,5
Великобритания	0,6	2,8
США	0,8	2,6

Источник: Kaldor, 1966.

1.9. Почему ковариация не является хорошей мерой связи?

Коэффициент корреляции является более подходящим измерителем зависимости, чем ковариация. Основная причина этого заключается в том, что ковариация зависит от единиц, в которых измеряются переменные x и y, в то время как коэффициент корреляции есть величина безразмерная. Это будет показано для случая выборочного коэффициента корреляции, доказательство для теоретического коэффициента корреляции будет оставлено для самостоятельного упражнения.

Возвращаясь к примеру со спросом на бензин, мы исследуем, что может случиться, когда при вычислении индекса реальных цен в качестве базового года используется 1980 г. вместо 1972 г. В этом случае ковариация изменится, а коэффициент корреляции — нет.

При использовании 1972 г. в качестве базового года индекс реальных цен для 1980 г. составил 188,8. Если теперь принять этот индекс за 100 для 1980 г., то нужно пересчитать ряды путем умножения на коэффициент $100/188{,}8 = 0{,}53$. Новые ряды представлены во второй колонке табл. 1.8 и будут обозначены через P. Величина P численно меньше, чем p.

Так как каждое отдельное наблюдение ряда цен было пересчитано с коэффициентом 0,53, то отсюда следует, что и среднее значение за выборочный период $(\bar{P})$ пересчитывается с этим коэффициентом. Следовательно, в году t

$$P_t - \bar{P} = 0{,}53p_t - 0{,}53\bar{p} = 0{,}53(p_t - \bar{p}). \tag{1.27}$$

Это означает, что в году t

$$(P - \bar{P})(y - \bar{y}) = 0{,}53(p - \bar{p})(y - \bar{y}), \tag{1.28}$$

и, следовательно, $\mathrm{Cov}(P, y) = 0{,}53\,\mathrm{Cov}(p, y)$. Однако на коэффициент корреляции это изменение не повлияет. Коэффициент корреляции для P и y будет равен:

$$r_{p,y} = \frac{\mathrm{Cov}(P, y)}{\sqrt{\mathrm{Var}(P)\mathrm{Var}(y)}}. \tag{1.29}$$

Таблица 1.8

Семья	P	y	$P - \bar{P}$	$y - \bar{y}$	$(P - \bar{P})^2$	$(y - \bar{y})^2$	$(P - \bar{P})(y - \bar{y})$
1973	54,82	26,2	−21,11	−0,07	445,73	0,01	1,48
1974	67,27	24,8	−8,67	−1,47	75,09	2,16	12,74
1975	66,74	25,6	−9,20	−0,67	84,55	0,45	6,16
1976	66,10	26,8	−9,38	0,53	96,64	0,28	−5,21
1977	66,05	27,7	−9,88	1,43	97,68	2,05	−14,13
1978	64,41	28,3	−11,53	2,03	132,84	4,12	−23,40
1979	79,29	27,4	3,36	1,13	11,28	1,28	3,80
1980	100,00	25,1	24,07	−1,17	579,26	1,37	−28,16
1981	102,54	25,2	26,61	−1,07	708,10	1,15	−28,47
1982	92,11	25,6	16,18	−0,67	261,66	0,45	−10,84
Сумма	759,32	262,7			2492,28	13,30	−86,04
Среднее	75,93	26,27			249,23	1,33	−8,60

Числитель (верхняя часть дроби) был умножен на 0,53, но на ту же величину был умножен и знаменатель (нижняя часть), так как $\mathrm{Var}(P) = (0{,}53)^2\,\mathrm{Var}(p)$. (Необходимо иметь в виду, что, когда вы умножаете переменную величину на постоянную, ее дисперсия умножается на эту постоянную в квадрате.) Знаменатель умножается на 0,53, а не на $(0{,}53)^2$, так как из $\mathrm{Var}(P)$ извлекается квадратный корень.

Упражнения

1.6. Вычислите коэффициент корреляции для P и y, используя данные табл. 1.8, и проверьте, что он будет таким же, как и коэффициент корреляции для p и y.

1.7. Покажите, что теоретический коэффициент корреляции останется неизменным при изменении единицы измерения одной из переменных.

1.10. Коэффициент частной корреляции

Анализ критериев значимости для коэффициента корреляции будет дан в главе 3, где эти показатели рассматриваются вместе с критериями значимости коэффициентов регрессии. Будет выяснено, что коэффициент корреляции в примере со спросом на бензин незначимо отличается от нуля, что кажется неправдоподобным с точки зрения здравого смысла.

Одна из причин получения такого результата заключается в очень небольшом размере выборки. Возможно, что при большем размере выборки мы смогли бы показать, что коэффициент корреляции значимо отличается от нуля. Здесь, однако, есть и еще одна причина для получения отрицательного результата: мы не учитывали влияние увеличения дохода на потребительский спрос в целом и на спрос на бензин в частности. Положительный эффект увеличения дохода в основном компенсировал отрицательный эффект роста цен, и, таким образом, спрос на бензин оставался стабильным. Следующий этап исследования состоит в выделении влияния этих двух факторов. Мы можем сделать это, используя так называемый коэффициент частной корреляции, который определяется следующим образом:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}, \qquad (1.30)$$

где $r_{xy.z}$ — коэффициент частной корреляции между x и y в случае постоянства воздействия величины z, а r_{xy}, r_{xz} и r_{yz} — обычные коэффициенты корреляции между x и y, x и z, y и z соответственно.

В примере со спросом на бензин мы можем вычислить корреляцию между ценой и располагаемым личным доходом и между спросом и доходом, используя данные для нужных лет табл. Б.1 . Результаты приблизительно составят 0,84 и 0,02. Подставляя эти значения в уравнение (1.30), мы оценим частный коэффициент корреляции для реальной цены и спроса как −0,91, что является намного более приемлемым результатом.

2

ПАРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

В данной главе показано, как, используя соответствующие данные, можно получить количественное выражение гипотетического линейного соотношения между двумя переменными. В главе объясняется важный принцип регрессионного анализа — метод наименьших квадратов, а также выводятся формулы, выражающие коэффициенты регрессии.

Большинство студентов, изучающих вводный курс эконометрики, уже знакомы с основами дифференциального исчисления, поэтому вывод коэффициентов регрессии не будет для них сложным. Тем, кто не прошел этот курс, следует пропустить раздел 2.3 и доказательство в разделе 2.5. Им придется принять формулы расчета коэффициентов на веру, однако в общих чертах они смогут понять, как были получены эти выражения.

2.1. Модель парной линейной регрессии

Коэффициент корреляции показывает, что две переменные связаны друг с другом, однако он не дает представления о том, каким образом они связаны. Рассмотрим более подробно те случаи, для которых мы предполагаем, что одна переменная зависит от другой.

Сразу же отметим, что не следует ожидать получения точного соотношения между какими-либо двумя экономическими показателями, за исключением тех случаев, когда оно существует по определению. В учебниках по экономической теории эта проблема обычно решается путем приведения соотношения, как если бы оно было точным, и предупреждения читателя о том, что это аппроксимация. В статистическом анализе, однако, факт неточности соотношения признается путем явного включения в него случайного фактора, описываемого случайным остаточным членом.

Начнем с рассмотрения простейшей модели:

$$y = \alpha + \beta x + u. \tag{2.1}$$

Величина y, рассматриваемая как *зависимая переменная*, состоит из двух составляющих: 1) неслучайной составляющей $\alpha + \beta x$, где x выступает как *объясняющая (или независимая) переменная*, а постоянные величины α и β — как параметры уравнения; 2) случайного члена u.

На рис. 2.1 показано, как комбинация этих двух составляющих определяет величину y. Показатели x_1, x_2, x_3 и x_4 — это четыре гипотетических значения объясняющей переменной. Если бы соотношение между y и x было точным, то соответствующие значения y были бы представлены точками Q_1, Q_2, Q_3, Q_4 на прямой. Наличие случайного члена приводит к тому, что в действительности значение y получается другим. Предполагалось, что случайный член возмущения положителен в первом и четвертом наблюдениях и отрицателен в двух других. Поэтому если отметить на графике реальные значения y при соответствующих значениях x, то мы получим точки P_1, P_2, P_3, P_4.

Следует подчеркнуть, что точки P — это единственные точки, отражающие реальные значения переменных на рис. 2.1. Фактические значения α и β и, следовательно, положения точек Q неизвестны, так же как и фактические значения случайного члена. Задача регрессионного анализа состоит в получении оценок α и β и, следовательно, в определении положения прямой по точкам P.

Очевидно, что чем меньше значения u, тем легче эта задача. Действительно, если бы случайный член отсутствовал вовсе, то точки P совпали бы с точками Q и точно показали бы положение прямой. В этом случае достаточно было бы просто построить эту прямую и определить значения α и β.

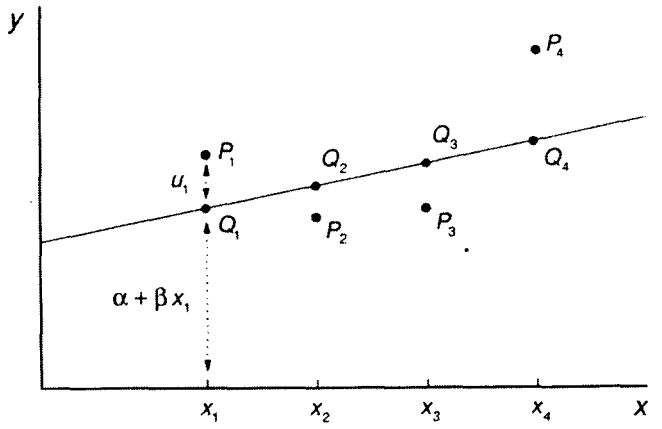


Рис. 2.1. Истинная зависимость между y и x

Почему же существует случайный член? Имеется несколько причин.

1. *Невключение объясняющих переменных.* Соотношение между y и x почти наверняка является очень большим упрощением. В действительности существуют другие факторы, влияющие на y, которые не учтены в формуле (2.1). Влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой. Часто происходит так, что имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем этого сделать потому, что не знаем, как их измерить, например психологические факторы. Возможно, что существуют также другие факторы, которые мы можем измерить, но которые оказывают такое слабое влияние, что их не стоит учитывать. Кроме того, могут быть факторы, которые являются существенными, но которые мы

из-за отсутствия опыта таковыми не считаем. Объединив все эти составляющие, мы получаем то, что обозначено как u. Если бы мы точно знали, какие переменные присутствуют здесь, и имели возможность точно их измерить, то могли бы включить их в уравнение и исключить соответствующий элемент из случайного члена. Проблема состоит в том, что мы никогда не можем быть уверены, что входит в данную совокупность, а что — нет.

2. *Агрегирование переменных.* Во многих случаях рассматриваемая зависимость — это попытка объединить вместе некоторое число микроэкономических соотношений. Например, функция суммарного потребления — это попытка общего выражения совокупности решений отдельных индивидов о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между совокупными расходами и доходом является лишь аппроксимацией. Наблюдаемое расхождение при этом приписывается наличию случайного члена.

3. *Неправильное описание структуры модели.* Структура модели может быть описана неправильно или не вполне правильно. Здесь можно привести один из многих возможных примеров. Если зависимость относится к данным о временном ряде, то значение y может зависеть не от фактического значения x, а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между y и x существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайного члена.

4. *Неправильная функциональная спецификация.* Функциональное соотношение между y и x математически может быть определено неправильно. Например, истинная зависимость может не являться линейной, а быть более сложной. Нелинейные зависимости будут рассмотрены в главе 4. Безусловно, надо постараться избежать возникновения этой проблемы, используя подходящую математическую формулу, но любая самая изощренная формула является лишь приближением, и существующее расхождение вносит вклад в остаточный член.

5. *Ошибки измерения.* Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не будут соответствовать точному соотношению, и существующее расхождение будет вносить вклад в остаточный член.

Остаточный член является суммарным проявлением всех этих факторов. Очевидно, что если бы вас интересовало только измерение влияния x на y, то было бы значительно удобнее, если бы остаточного члена не было. Если бы он отсутствовал, мы бы знали, что любое изменение y от наблюдения к наблюдению вызвано изменением x, и смогли бы точно вычислить β. Однако в действительности каждое изменение y отчасти вызвано изменением u, и это значительно усложняет жизнь. По этой причине u иногда описывается как шум.

2.2. Регрессия по методу наименьших квадратов

Допустим, что вы имеете четыре наблюдения для x и y, представленные на рис. 2.1, и перед вами поставлена задача — определить значения α и β в уравнении (2.1). В качестве грубой аппроксимации вы можете сделать это, отложив четыре точки P и построив прямую, в наибольшей степени соответствующую

этим точкам. Это сделано на рис. 2.2. Отрезок, отсекаемый прямой на оси y, представляет собой оценку α и обозначен a, а угловой коэффициент прямой представляет собой оценку β и обозначен b.

С самого начала необходимо признать, что вы никогда не сможете рассчитать истинные значения α и β при попытке построить прямую и определить положение линии регрессии. Вы можете получить только оценки, и они могут быть хорошими или плохими. Иногда оценки могут быть абсолютно точными, но это возможно лишь в результате случайного совпадения, и даже в этом случае у вас не будет способа узнать, что оценки абсолютно точны.

Это справедливо и при использовании более совершенных методов. Построение линии регрессии на глаз является достаточно субъективным. Более того, как мы увидим в дальнейшем, это просто невозможно, если переменная y зависит не от одной, а от двух или более независимых переменных. Возникает вопрос: существует ли способ достаточно точной оценки α и β алгебраическим путем?

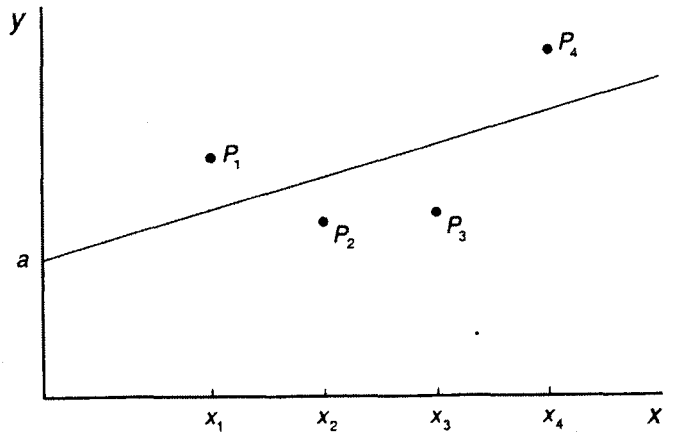


Рис. 2.2. Прямая, построенная по точкам

Первым шагом является определение *остатка* для каждого наблюдения. За исключением случаев чистого совпадения, построенная вами линия регрессии не пройдет точно ни через одну точку наблюдения. Например, на рис. 2.3 при $x = x_1$ соответствующей ему точкой на линии регрессии будет R_1 со значением y, которое мы обозначим $\hat{y}_1$ вместо фактически наблюдаемого значения y_1. Величина $\hat{y}_1$ описывается как расчетное значение y, соответствующее x_1. Разность между фактическим и расчетным значениями $(y_1 - \hat{y}_1)$, определяемая отрезком P_1R_1, описывается как остаток в первом наблюдении. Обозначим его e_1. Соответственно, для других наблюдений остатки будут обозначены как e_2, e_3 и e_4.

Очевидно, что мы хотим построить линию регрессии таким образом, чтобы эти остатки были минимальными. Очевидно также, что линия, строго соответствующая одним наблюдениям, не будет соответствовать другим, и наоборот. Необходимо выбрать какой-то критерий подбора, который будет одновременно учитывать величину всех остатков.

Существует целый ряд возможных критериев, одни из которых «работают» лучше других. Например, бесполезно минимизировать сумму остатков. Сумма будет автоматически равна нулю, если вы сделаете α равным $\bar{y}$, а β равным нулю, получив горизонтальную линию $y = \bar{y}$. В этом случае положительные остатки точно уравновесят отрицательные, но строгой зависимости при этом не будет.

Один из способов решения поставленной проблемы состоит в минимизации суммы квадратов остатков S. Для рис. 2.3 верно такое соотношение:

$$S = e_1^2 + e_2^2 + e_3^2 + e_4^2. \tag{2.2}$$

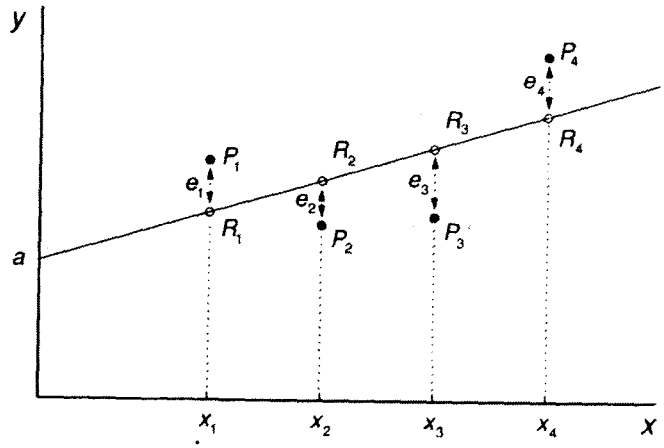


Рис. 2.3. Построенная по точкам линия регрессии, показывающая остатки

Величина S будет зависеть от выбора a и b, так как они определяют положение линии регрессии. В соответствии с этим критерием, чем меньше S, тем строже соответствие. Если $S = 0$, то получено абсолютно точное соответствие, так как это означает, что все остатки равны нулю. В этом случае линия регрессии будет проходить через все точки, однако, вообще говоря, это невозможно из-за наличия случайного члена.

Существуют и другие достаточно разумные решения, однако при выполнении определенных условий *метод наименьших квадратов* дает несмещенные и эффективные оценки α и β. По этой причине метод наименьших квадратов является наиболее популярным в вводном курсе регрессионного анализа. В данной работе рассматривается *обычный метод наименьших квадратов* (МНК, или OLS — ordinary least squares). В последующих разделах будут рассмотрены другие его варианты, которые могут быть использованы для решения некоторых специальных проблем.

2.3. Регрессия по методу наименьших квадратов: два примера[1]

Пример 1

Приведем действительно простой пример всего с двумя наблюдениями для того, чтобы продемонстрировать механизм процесса: как показано на рис. 2.4, наблюдаемое значение $y = 3$, когда $x = 1$, и $y = 5$ при $x = 2$.

Оценим коэффициенты a и b уравнения

$$\hat{y} = a + bx. \tag{2.3}$$

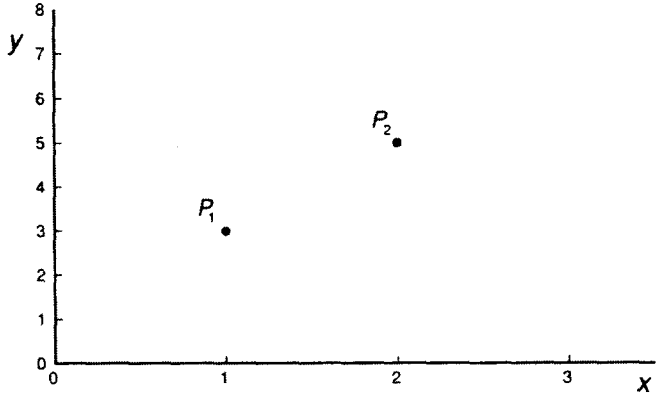


Рис. 2.4. Пример с двумя наблюдениями

Таблица 2.1

x	y	$\hat{y}$	e
1	3	$a + b$	$3 - a - b$
2	5	$a + 2b$	$5 - a - 2b$

Очевидно, что при наличии всего двух наблюдений мы можем получить точное соответствие, проведя линию регрессии через две точки, однако сделаем вид, что мы этого не понимаем. Вместо этого придем к тому же выводу, используя метод регрессии.

Если $x = 1$, то $\hat{y} = (a + b)$ в соответствии с уравнением регрессии. Если $x = 2$, то $\hat{y} = a + 2b$. Следовательно, мы можем составить табл. 2.1.

Значение $\hat{y}_1$ (величина y в точке R_1 на рис. 2.3) равно $(a + b)$, а значение $\hat{y}_2 = a + 2b$. Следовательно, остаток e_1 для первого наблюдения, который определяется как $(y_1 - \hat{y}_1)$, равен $(3 - a - b)$, а остаток e_2, который определяется как $(y_2 - \hat{y}_2)$, равен $(5 - a - 2b)$. Следовательно,

[1] Тем, кто не знаком с дифференциальным исчислением, этот раздел можно пропустить.

$$S = e_1^2 + e_2^2 = (3 - a - b)^2 + (5 - a - 2b)^2 =$$
$$= (9 + a^2 + b^2 - 6a - 6b + 2ab) + (25 + a^2 + 4b^2 - 10a - 20b + 4ab) =$$
$$= 2a^2 + 5b^2 + 6ab - 16a - 26b + 34. \quad (2.4)$$

Теперь мы хотим выбрать такие значения a и b, чтобы значение S было минимальным. Для этого мы используем дифференциальное исчисление и находим значения a и b, удовлетворяющие следующим соотношениям:

$$\frac{\partial S}{\partial a} = 0; \quad \frac{\partial S}{\partial b} = 0; \quad (2.5)$$

и

$$\frac{\partial S}{\partial a} = 4a + 6b - 16; \quad \frac{\partial S}{\partial b} = 10b + 6a - 2b. \quad (2.6)$$

Таким образом, мы имеем:

$$2a + 3b - 8 = 0 \quad (2.7)$$

и

$$3a + 5b - 13 = 0. \quad (2.8)$$

Решив эти два уравнения, получим $a = 1$ и $b = 2$. Следовательно, уравнение регрессии будет иметь следующий вид:

$$\hat{y} = 1 + 2x. \quad (2.9)$$

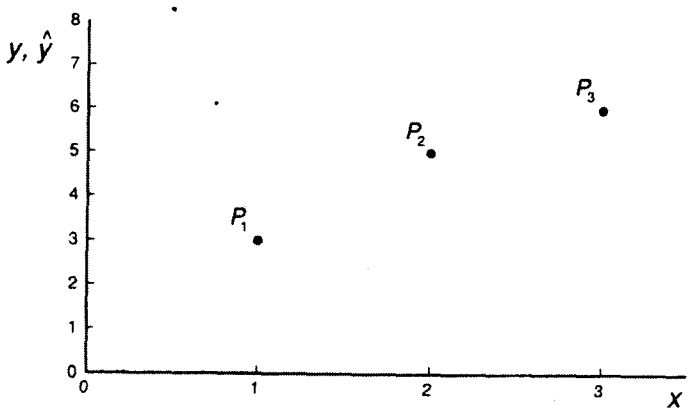


Рис. 2.5. Пример с тремя наблюдениями

Для того чтобы проверить, что мы пришли к правильному выводу, вычислим остатки:

$$e_1 = 3 - a - b = 3 - 1 - 2 = 0; \quad (2.10)$$

$$e_2 = 5 - a - 2b = 5 - 1 - 4 = 0. \quad (2.11)$$

Таким образом, оба остатка равны нулю, что означает, что линия регрессии проходит точно через обе точки, что мы, разумеется, знали с самого начала. Если у вас всего два наблюдения, то проводите прямую через эти две точки. В данном случае в проведении регрессионного анализа нет необходимости.

Пример 2

Используем пример, рассмотренный в предыдущем разделе, и добавим третье наблюдение: $y = 6$ при $x = 3$. Три наблюдения, показанные на рис. 2.5, не лежат на одной прямой, поэтому точное соответствие получить невозможно. В этом случае для вычисления положения прямой мы должны использовать регрессию по методу наименьших квадратов.

Начнем с задания стандартного уравнения

$$\hat{y} = a + bx. \tag{2.12}$$

Для значений x, равных 1, 2 и 3, расчетные значения y равны соответственно $(a + b)$, $(a + 2b)$ и $(a + 3b)$; они приведены в табл. 2.2.

Таблица 2.2

x	y	$\hat{y}$	e
1	3	$a + b$	$3 - a - b$
2	5	$a + 2b$	$5 - a - 2b$
3	6	$a + 3b$	$6 - a - 3b$

Следовательно,

$$S = e_1^2 + e_2^2 + e_3^2 = (3 - a - b)^2 + (5 - a - 2b)^2 + (6 - a - 3b)^2 =$$

$$= (9 + a^2 + b^2 - 6a - 6b + 2ab) + (25 + a^2 + 4b^2 - 10a - 20b + 4ab) +$$

$$+(36 + a^2 + 9b^2 - 12a - 36b + 6ab) = 3a^2 + 14b^2 + 12ab - 28a - 62b + 70. \tag{2.13}$$

Условия $\partial S/\partial a = 0$ и $\partial S/\partial b = 0$ дают:

$$6a + 12b - 28 = 0 \tag{2.14}$$

и

$$28b + 12a - 62 = 0. \tag{2.15}$$

Решая эти уравнения, получим $a = 1{,}67$ и $b = 1{,}50$. Следовательно, уравнение регрессии имеет следующий вид:

$$\hat{y} = 1{,}67 + 1{,}50x. \tag{2.16}$$

2.4. Детальное рассмотрение остатков

После построения линии регрессии стоит более детально рассмотреть общее выражение для остатка в каждом наблюдении. Логика этого рассмотрения является достаточно простой. Однако на первый взгляд она может показаться абстрактной, поэтому полезно графическое представление.

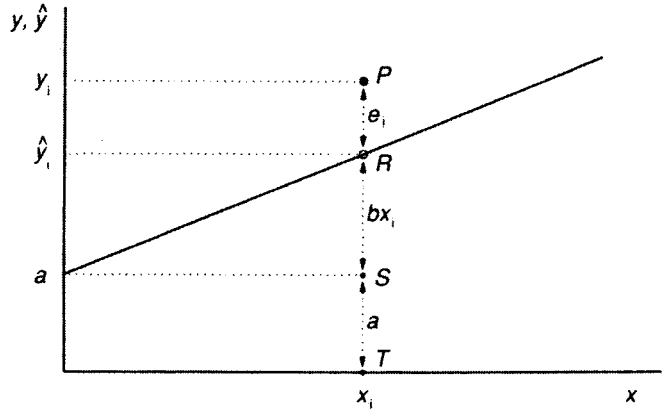


Рис. 2.6

На рис. 2.6 линия регрессии

$$\hat{y} = a + bx \tag{2.17}$$

построена по выборке наблюдений. Для того чтобы не загромождать график, показано только одно такое наблюдение: наблюдение i, представленное точкой P с координатами (x_i, y_i).

Когда $x = x_i$ линия регрессии предсказывает значение $y = \hat{y}_i$, что соответствует точке R на графике, где

$$\hat{y}_i = a + bx_i. \tag{2.18}$$

Используя условные обозначения, принятые на рис. 2.6, это уравнение можно переписать следующим образом:

$$RT = ST + RS, \tag{2.19}$$

так как отрезок ST равен a, а отрезок RS равен bx_i.

Остаток PR — это разность между PT и RT:

$$PR = PT - RT = PT - ST - RS. \tag{2.20}$$

Используя обычную математическую запись, представим формулу (2.20) в следующем виде:

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i. \tag{2.21}$$

Если бы в примере, показанном на графике, мы выбрали несколько большее значение a или несколько большее значение b, то прямая прошла бы ближе

к P, и остаток e_i был бы меньше. Однако это повлияло бы на остатки всех других наблюдений, и это необходимо учитывать. Минимизируя сумму квадратов остатков, мы попытаемся найти некоторое равновесие между ними.

2.5. Регрессия по методу наименьших квадратов с одной независимой переменной

Рассмотрим случай, когда имеется n наблюдений двух переменных x и y. Предположив, что y зависит от x, мы хотим подобрать уравнение

$$\hat{y} = a + bx. \tag{2.22}$$

Расчетное значение зависимой переменной $\hat{y}_i$ и остаток e_i для наблюдения i заданы уравнениями (2.18) и (2.21). Мы хотим выбрать a и b, чтобы минимизировать величину S:

$$S = \sum e_i^2 = e_1^2 + \ldots + e_n^2. \tag{2.23}$$

Можно обнаружить, что величина S минимальна, когда

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \tag{2.24}$$

и

$$a = \bar{y} - b\bar{x}. \tag{2.25}$$

Варианты выражения для b

Так как

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y}) = \frac{1}{n}\sum xy - \bar{x}\bar{y}; \tag{2.26}$$

и

$$\mathrm{Var}(x) = \frac{1}{n}\sum (x - \bar{x})^2 = \frac{1}{n}\sum x^2 - \bar{x}^2, \tag{2.27}$$

мы можем получить следующие выражения для b:

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{\frac{1}{n}\sum (x - \bar{x})^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}; \tag{2.28}$$

$$b = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\frac{1}{n}\sum x^2 - \bar{x}^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}. \tag{2.29}$$

В дальнейшем в тексте будет использоваться первоначальное определение $b = \text{Cov}(x, y)/\text{Var}(x)$ и это выражение, вероятно, легче всего запомнить. На практике для вычисления коэффициентов регрессии используется компьютер, поэтому нет смысла запоминать альтернативные выражения. Зная определения выборочной дисперсии и ковариации, вы всегда сможете вывести эти выражения.

Вывод выражений для a и b[1]

Вывод выражений для a и b будет осуществляться в соответствии с той же процедурой, которая использовалась в двух примерах в разделе 2.3, и предлагается сравнивать общий вариант с примерами на каждом этапе. Начнем с того, что выразим квадрат i-го остатка через a и b и наблюдения значений x и y:

$$e_i^2 = (y_i - \hat{y}_i)^2 = (y_i - a - bx_i)^2 = y_i^2 + a^2 + b^2x_i^2 - 2ay_i + 2abx_i - 2bx_iy_i. \tag{2.30}$$

Суммируя по всем n наблюдениям, запишем S в виде:

$$S = \sum y_i^2 + na^2 + b^2\sum x_i^2 - 2a\sum y_i + 2ab\sum x_i - 2b\sum x_iy_i. \tag{2.31}$$

Заметим, что данное выражение для S является квадратичной формой по a и b, и ее коэффициенты определяются выборочными значениями x и y. Мы можем влиять на величину S, только задавая значения a и b. Значения x и y, которые определяют положение точек на диаграмме рассеяния, уже не могут быть изменены после того, как мы взяли определенную выборку. Полученное уравнение представляет собой обобщенный вариант уравнений (2.4) и (2.13).

Условия первого порядка для минимума, то есть $\partial S/\partial a = 0$ и $\partial S/\partial b = 0$, принимают вид:

$$\frac{\partial S}{\partial a} = 2an - 2\sum y_i + 2b\sum x_i = 0; \tag{2.32}$$

$$\frac{\partial S}{\partial b} = 2b\sum x_i^2 + 2a\sum x_i - 2\sum x_iy_i = 0. \tag{2.33}$$

Эти уравнения известны как нормальные уравнения для коэффициентов регрессии и представляют собой обобщенные варианты уравнений (2.7), (2.8), (2.14) и (2.15) в двух примерах. Уравнение (2.32) позволяет выразить a через $\bar{y}$, $\bar{x}$ и пока неизвестное b. Подставив $n\bar{y}$ вместо Σy_i и $n\bar{x}$ вместо Σx_i, получим:

$$2an - 2n\bar{y} + 2bn\bar{x} = 0. \tag{2.34}$$

Следовательно,

$$a = \bar{y} - b\bar{x}. \tag{2.35}$$

Подставив выражение для a в уравнение (2.33) и помня, что Σx_i равно $n\bar{x}$, имеем:

[1] Те, кто не знаком с дифференциальным исчислением, могут пропустить следующую часть данного раздела.

$$2b\sum x_i^2 + 2n\bar{x}\bar{y} - 2bn\bar{x}^2 - 2\sum x_i y_i = 0. \tag{2.36}$$

После деления на $2n$ и перегруппировки получим:

$$b\left\{\frac{1}{n}\sum x_i^2 - \bar{x}^2\right\} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}. \tag{2.37}$$

С учетом формул (2.26) и (2.27) это выражение можно переписать в следующем виде:

$$b\mathrm{Var}(x) = \mathrm{Cov}(x, y), \tag{2.38}$$

и, таким образом, мы получим уравнение (2.24). Найдя из этого выражения b, выразим затем a из уравнения (2.25). Тот, кто знаком с условиями второго порядка, без труда сможет убедиться, что они удовлетворены.

Во втором числовом примере, приводимом в разделе 2.3, $\mathrm{Cov}\,(x, y) = 1{,}0$; $\mathrm{Var}\,(x) = 0{,}67$; $\bar{y} = 4{,}67$; $\bar{x} = 2{,}00$. Следовательно,

$$b = 1{,}00 / 0{,}67 = 1{,}5; \tag{2.39}$$

$$a = \bar{y} - b\bar{x} = 4{,}67 - 1{,}5(2{,}00) = 1{,}67, \tag{2.40}$$

что подтверждает исходные вычисления.

2.6. Интерпретация уравнения регрессии

Существуют два этапа интерпретации уравнения регрессии. Первый этап состоит в словесном истолковании уравнения так, чтобы это было понятно человеку, не являющемуся специалистом в области статистики. На втором этапе необходимо решить, следует ли ограничиться этим или провести более детальное исследование зависимости.

Оба этапа чрезвычайно важны. Второй этап мы рассмотрим несколько позже, а пока обратим основное внимание на первый этап. Это будет проиллюстрировано моделью регрессии для функции спроса, т. е. регрессией между расходами потребителя на питание (y) и располагаемым личным доходом (x) по данным, приведенным в табл. Б.1 для США за период с 1959 по 1983 г. Данные представлены в виде графика (рис. 2.7).

Предположим, что истинная модель описывается следующим выражением:

$$y = \alpha + \beta x + u, \tag{2.41}$$

и оценена регрессия

$$\hat{y} = 55{,}3 + 0{,}093x. \tag{2.42}$$

Полученный результат можно истолковать следующим образом. Коэффициент при x (коэффициент наклона) показывает, что если x увеличивается на одну единицу, то y возрастает на 0,093 единицы. Как x, так и y измеряются в миллиардах долларов в постоянных ценах; таким образом, коэффициент наклона показывает, что если доход увеличивается на 1 млрд. долл., то расходы на

питание возрастают на 93 млн. долл. Другими словами, из каждого дополнительного доллара дохода 9,3 цента будут израсходованы на питание.

Что можно сказать о постоянной в уравнении? Формально говоря, она показывает прогнозируемый уровень y, когда $x = 0$. Иногда это имеет ясный смысл, иногда нет. Если $x = 0$ находится достаточно далеко от выборочных значений x, то буквальная интерпретация может привести к неверным результатам; даже если линия регрессии довольно точно описывает значения наблюдаемой выборки, нет гарантии, что так же будет при экстраполяции влево или вправо.

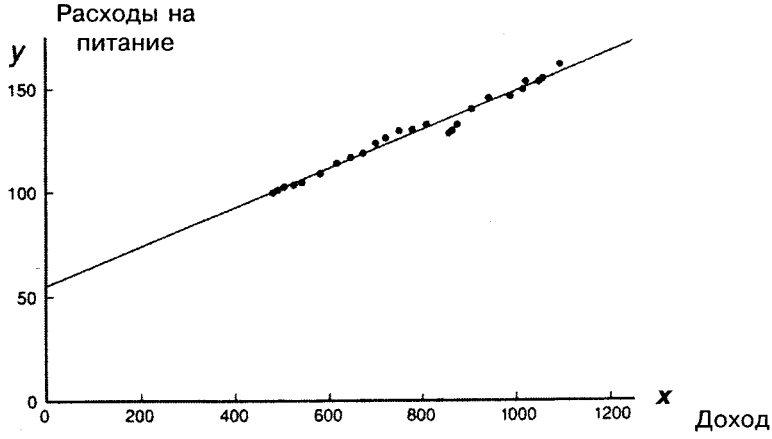


Рис. 2.7. Регрессионная зависимость расходов на питание от доходов (США, 1959–1983 гг.)

В рассматриваемом случае экстраполяция к вертикальной оси приводит к выводу о том, что если доход был бы равен нулю, то расходы на питание составили бы 55,3 млрд. долл. Такое толкование может быть правдоподобным в отношении отдельного человека, так как он может израсходовать на питание накопленные или одолженные средства. Однако оно не имеет никакого смысла применительно к совокупности. В данном случае константа выполняет единственную функцию: она позволяет определить положение линии регрессии на графике. Пример постоянной, которая имеет ясный смысл, приведен в упражнении 2.1.

При интерпретации уравнения регрессии чрезвычайно важно помнить о трех вещах. Во-первых, a является лишь оценкой α, а b — оценкой β. Поэтому вся интерпретация в действительности представляет собой лишь оценку. Во-вторых, уравнение регрессии отражает только общую тенденцию для выборки. При этом каждое отдельное наблюдение подвержено воздействию случайностей. В-третьих, верность интерпретации зависит от правильности спецификации уравнения.

В сущности, мы построили довольно наивную зависимость для функции спроса. Мы будем неоднократно возвращаться к этому в следующих разделах, уточняя как определение, так и статистические методы, используемые для оценки коэффициентов уравнения. В то же время читателю рекомендуется, начиная с упражнения 2.4, проводить параллельные эксперименты для определения функций спроса для других товаров, представленных в табл. Б.1.

После оценивания регрессии возникает следующий вопрос: существуют ли какие-либо средства определения точности оценок? Этот очень важный вопрос будет рассмотрен в следующем разделе. Мы же сначала рассмотрим более подробно роль остаточного члена и его влияние на оценки α и β.

Интерпретация линейного уравнения регрессии

Представим простой способ интерпретации коэффициентов линейного уравнения регрессии

$$\hat{y} = a + bx,$$

когда y и x — переменные с простыми, естественными единицами измерения.

Во-первых, можно сказать, что увеличение x на одну единицу (в единицах измерения переменной x) приведет к увеличению значения y на b единиц (в единицах измерения переменной y). Вторым шагом является проверка, каковы действительно единицы измерения x и y, и замена слова «единица» фактическим количеством. Третьим шагом является проверка возможности более простого выражения результата, который может оказаться не вполне удобным. В примере, приведенном в данном разделе, в качестве единицы измерения для x и y использовались миллиарды долларов, что позволило произвести очевидные упрощения.

Постоянная a дает прогнозируемое значение y (в единицах y), если $x = 0$. Это может иметь или не иметь ясного смысла в зависимости от конкретной ситуации.

Упражнения[1]

2.1. Регрессионная зависимость расходов на питание y (основанная на тех же данных, на которых уже строилась описанная в тексте функция спроса) от времени, определенного как $t = 1$ для 1959 г., $t = 2$ для 1960 г. и т. д., задана уравнением:

$$\hat{y} = 95{,}3 + 2{,}53t.$$

Интерпретируйте результаты оценивания регрессии и сравните их с аналогичными результатами в случае с моделью регрессии для функции спроса, рассматриваемой в тексте. Обратите внимание, что в данном случае постоянная имеет простое толкование.

2.2. Регрессионная зависимость расходов на оплату жилья от располагаемо-

[1] Упражнение 2.4 особенно важно в том смысле, что оно начинает серию регрессий для функций спроса, которые будут оцениваться читателем на протяжении всей книги. Если это упражнение выполняется группой студентов, то преподаватель должен дать студентам задания с разными товарами. Более подробную информацию об имеющихся данных можно получить в приложении Б.

го личного дохода в соответствии с табл. Б.1, где обе величины измерены в миллиардах долларов за период с 1959 по 1983 г., может быть формализована в виде:

$$\hat{y} = -27{,}6 + 0{,}178x.$$

Регрессионная зависимость расходов на оплату жилья от времени, определенная так же, как в упражнении 2.1, может быть представлена таким образом:

$$\hat{y} = 48{,}9 + 4{,}84t.$$

Дайте экономическое толкование этих регрессий. Они предполагают различные объяснения для одних и тех же данных по переменной y. В какой степени они могут быть согласованы?

2.3. Постройте уравнение регрессии между p и e по данным из упражнения 1.3, сначала используя все 12 наблюдений, а затем исключив наблюдение для Японии, и дайте экономическую интерпретацию[1].

2.4. В табл. Б.1 приведены ежегодные данные о потребительских расходах и располагаемых личных доходах для США на период с 1959 по 1983 г. Выберите один товар — не продукты питания и не жилье, — обозначьте его как y и оцените регрессию между y и x, где x — располагаемый личный доход, используя данные за 25 лет. Дайте интерпретацию коэффициентов регрессии[2].

2.5. Оцените регрессии между характеристиками товара и временем, как это сделано в упражнении 2.1. Дайте соответствующую интерпретацию и сравните ее с интерпретацией регрессии, полученной в упражнении 2.4.

2.6. Два человека строят временной тренд для одного и того же набора из 25 наблюдений переменной y, используя модель:

$$y = \alpha + \beta t + u,$$

где t — время (последовательно принимающее значения от 1 до 25), а u — случайный член. Первый получает уравнение:

$$\hat{y} = 6{,}70 + 1{,}79t.$$

Второй по ошибке оценивает регрессию между t и y и приходит к такому уравнению:

$$\hat{t} = -0{,}25 + 0{,}44y.$$

Из этого уравнения он получает:

$$\hat{y} = 0{,}57 + 2{,}27t.$$

Объясните наличие расхождения между данным уравнением и уравнением, полученным первым исследователем.

2.7. Как изменился бы результат оценивания регрессии в упражнении 2.1, если бы в качестве t использовались фактические даты (1959–1983 гг.), а не числа от 1 до 25?

2.8. Исследователь изучает зависимость между совокупным спросом на ус-

[1] Не следует начинать вычисление коэффициентов регрессии сначала, так как вы уже выполнили большую часть арифметических расчетов в упражнении 1.3.

[2] Преподавателю необходимо иметь в виду, что если это групповое занятие, то учащимся следует дать задания оценить регрессию для разных видов товаров, помимо продуктов питания и жилья.

луги (y) и совокупным располагаемым личным доходом (x) по данным для американской экономики (обе величины измерены в миллиардах долларов в постоянных ценах), используя ежегодные данные временных рядов и модель:

$$y = \alpha + \beta x + u.$$

1. Исследователь получает уравнение, проводя регрессионный анализ с помощью обычного метода наименьших квадратов. Предполагая, что обе величины y и x могут быть существенно занижены в системе национальных счетов из-за стремления людей уклониться от уплаты налогов, исследователь принимает два альтернативных метода уточнения заниженных оценок.
2. Исследователь добавляет в каждом году 90 млрд. долл. к показателю y и 200 млрд. долл. к показателю x.
3. Исследователь увеличивает значения как для x, так и для y на 10% за каждый год.

Оцените влияние корректировок (2) и (3) на результаты оценивания регрессии.

2.9. Исследователь имеет ежегодные данные о временных рядах для совокупной заработной платы (W), совокупной прибыли (Π), и совокупного дохода (Y) для страны за период в n лет. По определению

$$Y = W + \Pi.$$

Используя обычный метод наименьших квадратов, получаем уравнение регрессии:

$$\hat{W} = a_0 + a_1 Y;$$

$$\hat{\Pi} = b_0 + b_1 Y.$$

Покажите, что коэффициенты регрессии будут автоматически удовлетворять следующим уравнениям:

$$a_1 + b_1 = 1;$$

$$a_0 + b_0 = 0.$$

Объясните на интуитивном уровне, почему это должно быть именно так.

2.10. Исследователь считает, что в нестохастической части истинной модели y пропорционален x:

$$y = \beta x + u.$$

Выведите на основании исходных принципов формулу для b, оценки МНК для b. Покажите, что в этом случае (2.31) можно записать в следующем виде:

$$S = \sum y_i^2 + b^2 \sum x_i^2 - 2b \sum x_i y_i,$$

и что, следовательно,

$$b = \sum x_i y_i \,/\, \sum x_i^2.$$

2.11. Выведите из исходных предпосылок оценку МНК для α в модели:

$$y = \alpha + u.$$

Другими словами, y представляет собой просто сумму постоянной величины и случайного члена. Снова вначале определите S, а затем продифференцируйте.

2.7. Качество оценки: коэффициент R^2

Цель регрессионного анализа состоит в объяснении поведения зависимой переменной y. В любой данной выборке y оказывается сравнительно низким в одних наблюдениях и сравнительно высоким — в других. Мы хотим знать, почему это так. Разброс значений y в любой выборке можно суммарно описать с помощью выборочной дисперсии Var (y). Мы должны уметь рассчитывать величину этой дисперсии.

В парном регрессионном анализе мы пытаемся объяснить поведение y путем определения регрессионной зависимости y от соответственно выбранной независимой переменной x. После построения уравнения регрессии мы можем разбить значение y_i в каждом наблюдении на две составляющих — $\hat{y}_i$ и e_i:

$$y_i = \hat{y}_i + e_i. \tag{2.43}$$

Величина $\hat{y}_i$ — расчетное значение y в наблюдении i — это то значение, которое имел бы y при условии, что уравнение регрессии было правильным, и отсутствии случайного фактора. Это, иными словами, величина y, спрогнозированная по значению x в данном наблюдении. Тогда остаток e_i есть расхождение между фактическим и спрогнозированным значениями величины y. Это та часть y, которую мы не можем объяснить с помощью уравнения регрессии.

Используя (2.43), разложим дисперсию y:

$$\mathrm{Var}(y) = \mathrm{Var}(\hat{y} + e) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(e) + 2\mathrm{Cov}(\hat{y}, e). \tag{2.44}$$

Далее, оказывается, что Cov ($\hat{y}_i$, e) должна быть равна нулю (см. упражнение 2.12). Следовательно, мы получаем:

$$\mathrm{Var}(y) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(e). \tag{2.45}$$

Это означает, что мы можем разложить Var (y) на две части: Var ($\hat{y}$) — часть, которая «объясняется» уравнением регрессии в вышеописанном смысле, и Var (e) — «необъясненную» часть[1].

Согласно (2.45), Var ($\hat{y}$)/Var (y) — это часть дисперсии y, объясненная уравнением регрессии. Это отношение известно как коэффициент детерминации, и его обычно обозначают R^2:

$$R^2 = \frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}, \tag{2.46}$$

что равносильно

[1] Слова «объясненный» и «необъясненный» взяты в кавычки, так как объяснение, в сущности, может быть мнимым. В действительности y может зависеть от какой-то другой переменной z, и x может действовать как величина, замещающая z (более подробно об этом см. в главе 6). Поэтому вместо слова «объясненный» здесь лучше употреблять выражение «представляющийся объясненным».

$$R^2 = 1 - \frac{\text{Var}(e)}{\text{Var}(y)}. \qquad (2.47)$$

Максимальное значение коэффициента R^2 равно единице. Это происходит в том случае, когда линия регрессии точно соответствует всем наблюдениям, так что $\hat{y}_i = y_i$ для всех i и все остатки равны нулю. Тогда Var $(\hat{y})$ = Var (y), Var $(e) = 0$ и $R^2 = 1$.

Если в выборке отсутствует видимая связь между y и x, то коэффициент R^2 будет близок к нулю.

При прочих равных условиях желательно, чтобы коэффициент R^2 был как можно больше. В частности, мы заинтересованы в таком выборе коэффициентов a и b, чтобы максимизировать R^2. Не противоречит ли это нашему критерию, в соответствии с которым a и b должны быть выбраны таким образом, чтобы минимизировать сумму квадратов остатков? Нет, легко показать, что эти критерии эквивалентны, если (2.47) используется как определение коэффициента R^2. Отметим сначала, что

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i, \qquad (2.48)$$

откуда, беря среднее значение e_i по выборке и используя уравнение (2.25), получим:

$$\bar{e} = \bar{y} - a - b\bar{x} = \bar{y} - [\bar{y} - b\bar{x}] - b\bar{x} = 0. \qquad (2.49)$$

Следовательно,

$$\text{Var}(e) = \frac{1}{n}\sum (e_i - \bar{e})^2 = \frac{1}{n}\sum e_i^2. \qquad (2.50)$$

Отсюда следует, что принцип минимизации суммы квадратов остатков эквивалентен минимизации дисперсии остатков при условии выполнения (2.25). Однако если мы минимизируем Var (e), то при этом в соответствии с (2.47) автоматически максимизируется коэффициент R^2.

Альтернативное представление коэффициента R^2

На интуитивном уровне представляется очевидным, что чем больше соответствие, обеспечиваемое уравнением регрессии, тем больше должен быть коэффициент корреляции для фактических и прогнозных значений y, и наоборот. Покажем, что R^2 фактически равен квадрату такого коэффициента корреляции между y и $\hat{y}_i$, который мы обозначим $r_{y,\hat{y}}$ (заметим, что Cov $(e, \hat{y}) = 0$; см. упражнение 2.12):

$$r_{y,\hat{y}} = \frac{\text{Cov}(y, \hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}} = \frac{\text{Cov}(\{\hat{y} + e\}, \hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}} = \frac{\text{Cov}(\hat{y}, \hat{y}) + \text{Cov}(e, \hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}} =$$

$$= \frac{\text{Var}(\hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}} = \frac{\sqrt{\text{Var}(\hat{y})}}{\sqrt{\text{Var}(y)}} = \sqrt{R^2}. \qquad (2.51)$$

Пример вычисления коэффициента R^2

Вычисление коэффициента R^2 выполняется на компьютере в рамках программы оценивания регрессии, поэтому данный пример приведен лишь в целях иллюстрации. Будем использовать простейший пример с тремя наблюдениями, описанный в разделе 2.3, где уравнение регрессии

$$\hat{y} = 1{,}6667 + 1{,}5000x \tag{2.52}$$

построено по наблюдениям x и y, приведенным в табл. 2.3. В таблице также даны $\hat{y}_i$ и e_i для каждого наблюдения, вычисленные с помощью уравнения (2.52), и все остальные данные, необходимые для вычисления Var (y), Var ($\hat{y}$) и Var (e). (Заметим, что $\bar{e}$ должно быть равно нулю, так что величина Var (e) = $(1/n)\sum e_i^2$.)

Таблица 2.3

Наблюдения	x	y	$\hat{y}$	e	$y-\bar{\hat{y}}$	$\hat{y}-\bar{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{\hat{y}})^2$	e^2
1	1	3	3,1667	−0,1667	−1,6667	−1,5	2,7778	2,25	0,0278
2	2	5	4,6667	0,3333	0,3333	0,0	0,1111	0,00	0,1111
3	3	6	6,1667	−0,1667	1,3333	1,5	1,7778	2,25	0,0278
Сумма	6	14	14	0			4,6667	4,50	0,1667
Среднее	2	4,6667	4,6667	0			1,5556	1,50	0,0556

Из табл. 2.3 можно видеть, что Var (y) = 1,5556, Var ($\hat{y}$) = 1,5000 и Var (e) = 0,0556. Заметим, что Var (y) = Var ($\hat{y}$) + Var (e), как это и должно быть. На основании этих значений мы можем вычислить коэффициент R^2, используя уравнение (2.46) или (2.47):

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{1{,}5000}{1{,}5556} = 0{,}96; \tag{2.53}$$

$$R^2 = 1 - \frac{\text{Var}(e)}{\text{Var}(y)} = 1 - \frac{0{,}0556}{0{,}5556} = 0{,}96; \tag{2.54}$$

Упражнения

2.12. Докажите, что Cov ($\hat{y}$, e) должна быть равна нулю, используя равенства $\hat{y} = a + bx$, $e = y - a - bx$ и ковариационные правила.

2.13. Используя данные, приведенные в табл. 2.3, вычислите коэффициент корреляции между y и $\hat{y}$ и убедитесь, что значение коэффициента R^2, полученное путем возведения его в квадрат, является таким же, как в нашем примере.

2.14. Значения коэффициента R^2 для регрессионных зависимостей (1) расходов на продукты питания и (2) расходов на жилье от располагаемого личного дохода [см. уравнение (2.42) и упражнение 2.2] составили, соответственно, 0,98 и 0,99. Какой вывод можно сделать на основании этих значений (если какой-либо вывод здесь возможен)?

2.15. Каково значение коэффициента R^2 в регрессии между характеристиками выбранного вами товара и располагаемым личным доходом? Прокомментируйте это.

3

СВОЙСТВА КОЭФФИЦИЕНТОВ РЕГРЕССИИ И ПРОВЕРКА ГИПОТЕЗ

С помощью регрессионного анализа мы можем получить оценки параметров зависимости. Однако они являются лишь *оценками*. Поэтому возникает вопрос о том, насколько они надежны. Дадим сначала общий ответ, изучив условия несмещенности и факторы, определяющие дисперсию оценок. Основываясь на этом, мы будем совершенствовать способы проверки совместимости регрессионной оценки с конкретной априорной гипотезой об истинном значении оцениваемого параметра. И следовательно, мы будем строить доверительный интервал для истинного значения, который представляет собой множество всех возможных гипотетических значений, не противоречащих результатам экспериментов. Будет также показано, каким образом можно проверить, является ли качество подбора кривой более высоким, чем при чисто случайном подборе.

3.1. Случайные составляющие коэффициентов регрессии

Коэффициент регрессии, вычисленный методом наименьших квадратов, — это особая форма случайной величины, свойства которой зависят от свойств остаточного члена в уравнении. Мы продемонстрируем это сначала теоретически, а затем посредством контролируемого эксперимента. В частности, мы увидим, какое значение для оценки коэффициентов регрессии имеют некоторые конкретные предположения, касающиеся остаточного члена.

В ходе рассмотрения мы постоянно будем иметь дело с моделью парной регрессии, в которой y связан с x следующей зависимостью:

$$y = \alpha + \beta x + u, \tag{3.1}$$

и на основе n выборочных наблюдений будем оценивать уравнение регрессии.

$$\hat{y} = a + bx. \tag{3.2}$$

Мы также будем предполагать, что x — это неслучайная экзогенная переменная. Иными словами, ее значения во всех наблюдениях можно считать заранее заданными и никак не связанными с исследуемой зависимостью.

Во-первых, заметим, что величина y состоит из двух составляющих. Она вклю-

чает неслучайную составляющую ($\alpha + \beta x$), которая не имеет ничего общего с законами вероятности (α и β могут быть неизвестными, но тем не менее это постоянные величины), и случайную составляющую u.

Отсюда следует, что, когда мы вычисляем b по обычной формуле:

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \tag{3.3}$$

b также содержит случайную составляющую. Cov (x, y) зависит от значений y, а y зависит от значений u.

Если случайная составляющая принимает разные значения в n наблюдениях, то мы получаем различные значения y и, следовательно, разные величины Cov (x, y) и b.

Теоретически мы можем разложить b на случайную и неслучайную составляющие. Воспользовавшись соотношением (3.1), а также правилом *1* расчета ковариации из раздела 1.2, получим:

$$\text{Cov}(x, y) = \text{Cov}(x, [\alpha + \beta x + u]) = \text{Cov}(x, \alpha) + \text{Cov}(x, \beta x) + \text{Cov}(x, u). \tag{3.4}$$

По ковариационному правилу *3,* ковариация Cov (x, α) равна нулю. По ковариационному правилу *2,* ковариация Cov (x, βx) равна βCov (x, x). Причем Cov (x, x) это тоже, что и Var (x). Следовательно, мы можем записать:

$$\text{Cov}(x, y) = \beta \text{Var}(x) + \text{Cov}(x, u), \tag{3.5}$$

и, таким образом,

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \beta + \frac{\text{Cov}(x, u)}{\text{Var}(x)}. \tag{3.6}$$

Итак, мы показали, что коэффициент регрессии b, полученный по любой выборке, представляется в виде суммы двух слагаемых: 1) постоянной величины, равной истинному значению коэффициента β; 2) случайной составляющей, зависящей от Cov (x, u), которой обусловлены отклонения коэффициента b от константы β. Аналогичным образом можно показать, что a имеет постоянную составляющую, равную истинному значению α, плюс случайную составляющую, которая зависит от случайного фактора u.

Следует заметить, что на практике мы не можем разложить коэффициенты регрессии на составляющие, так как не знаем истинных значений α и β или фактических значений u в выборке. Они интересуют нас потому, что при определенных предположениях позволяют получить некоторую информацию о теоретических свойствах a и b.

3.2. Эксперимент по методу Монте-Карло

По-видимому, никто точно не знает, почему *эксперимент по методу Монте-Карло* называется именно так. Возможно, это название имеет какое-то отношение к известному казино как символу действия законов случайности.

Основное понятие будет объяснено посредством аналогии. Предположим, что свинья обучена находить трюфели. Это дикорастущие земляные грибы, встре-

чающиеся во Франции и Италии и считающиеся деликатесом. Они дороги, так как их трудно найти, и хорошая свинья, обученная поиску трюфелей, стоит дорого. Проблема состоит в том, чтобы узнать, насколько хорошо свинья ищет трюфели. Она может находить их время от времени, но возможно также, что большое количество трюфелей она пропускает. В случае действительной заинтересованности вы могли бы выбрать участок земли, закопать трюфели в нескольких местах, отпустить свинью и посмотреть, сколько грибов она обнаружит. Посредством такого контролируемого эксперимента можно было бы непосредственно оценить степень успешности поиска.

Какое отношение это имеет к регрессионному анализу? Проблема в том, что мы никогда не знаем истинных значений α и β (иначе зачем бы мы использовали регрессионный анализ для их оценки?). Поэтому мы не можем сказать, хорошие или плохие оценки дает наш метод. Эксперимент по методу Монте-Карло — это искусственный контролируемый эксперимент, дающий возможность такой проверки. Простейший возможный эксперимент по методу Монте-Карло состоит из трех частей. Во-первых:

1) выбираются истинные значения α и β;

2) в каждом наблюдении выбирается значение x;

3) используется некоторый процесс генерации случайных чисел (или берется последовательность из таблицы случайных чисел) для получения значений случайного фактора u в каждом из наблюдений.

Во-вторых, в каждом наблюдении генерируется значение y с использованием соотношения (3.1) и значений α, β, x и u.

В-третьих, применяется регрессионный анализ для оценивания параметров a и b с использованием только полученных указанным образом значений y и данных для x. При этом вы можете видеть, являются ли a и b хорошими оценками α и β, и это позволит почувствовать пригодность метода построения регрессии.

На первых двух шагах проводится подготовка к применению регрессионного метода. Мы полностью контролируем модель, которую создаем, и знаем истинные значения параметров, потому что сами их определили. На третьем этапе мы определяем, может ли поставленная нами задача решаться с помощью метода регрессии, т. е. могут ли быть получены хорошие оценки для α и β при использовании только данных об y и x. Заметим, что проблема возникает вследствие включения случайного фактора в процесс получения y. Если бы этот фактор отсутствовал, то точки, соответствующие значениям каждого наблюдения, лежали бы точно на прямой (3.1) и точные значения α и β можно было бы очень просто определить по значениям y и x.

Произвольно положим $\alpha = 2$ и $\beta = 0{,}5$, так что истинная зависимость имеет вид:

$$y = 2 + 0{,}5x + u. \tag{3.7}$$

Предположим для простоты, что имеется 20 наблюдений и что x принимает значения от 1 до 20. Для случайной остаточной составляющей u будем использовать случайные числа, взятые из нормально распределенной совокупности с нулевым средним и единичной дисперсией. Нам потребуется набор из 20 значений, обозначим их rn_1, ..., rn_{20}. Случайный член u_1 в первом наблюдении просто равен rn_1 и т. д.

Зная значения x и u в каждом наблюдении, можно вычислить значения y, используя уравнение (3.7); это сделано в табл. 3.1. Теперь при оценивании регрессионной зависимости y от x получим:

$$\hat{y} = 1{,}63 + 0{,}54x. \tag{3.8}$$

В данном случае оценка a приняла меньшее значение (1,63) по сравнению с α (2,00), а b немного выше β (0,54 по сравнению с 0,50). Расхождения вызваны совместным влиянием случайных членов в 20 наблюдениях.

Очевидно, что одного эксперимента такого типа едва ли достаточно для оценки качества метода регрессии. Он дал довольно хорошие результаты, но, возможно, это лишь счастливый случай. Для дальнейшей проверки повторим эксперимент с *тем же* истинным уравнением (3.7) и с *теми же* значениями x, но с *новым* набором случайных чисел для остаточного члена, взятых из того же распределения (нулевое среднее и единичная дисперсия). Используя эти значения и значения x, получим новый набор значений y.

В целях экономии места таблица с новыми значениями u и y не приводится. Вот результат оценивания регрессии между новыми значениями y и x:

$$\hat{y} = 2{,}52 + 0{,}48x. \tag{3.9}$$

Таблица 3.1

x	u	y	x	u	y
1	−0,59	1,91	11	1,59	9,09
2	−0,24	2,76	12	−0,92	7,08
3	−0,83	2,67	13	−0,71	7,79
4	0,03	4,03	14	−0,25	8,75
5	−0,38	4,12	15	1,69	11,19
6	−2,19	2,81	16	0,15	10,15
7	1,03	6,53	17	0,02	10,52
8	0,24	6,24	18	−0,11	10,89
9	2,53	9,03	19	−0,91	10,59
10	−0,13	6,87	20	1,42	13,42

Второй эксперимент также был успешным. Теперь a оказалось больше α, а b — несколько меньше β. В табл. 3.2 приведены оценки a и b при 10-кратном повторении эксперимента с использованием разных наборов случайных чисел в каждом варианте.

Можно заметить, что, несмотря на то что в одних случаях оценки принимают заниженные значения, а в других — завышенные, в целом значения a и

Таблица 3.2

Эксперимент	*a*	*b*
1	1,63	0,54
2	2,52	0,48
3	2,13	0,45
4	2,14	0,50
5	1,71	0,56
6	1,81	0,51
7	1,72	0,56
8	3,18	0,41
9	1,26	0,58
10	1,94	0,52

b группируются вокруг истинных значений α и β, равных соответственно 2,00 и 0,50. При этом хороших оценок получено больше, чем плохих. Например, фиксируя значения *b* при очень большом числе повторений эксперимента, можно построить таблицу частот и получить аппроксимацию функции плотности вероятности, показанную на рис. 3.1. Это нормальное распределение со средним 0,50 и стандартным отклонением 0,0388.

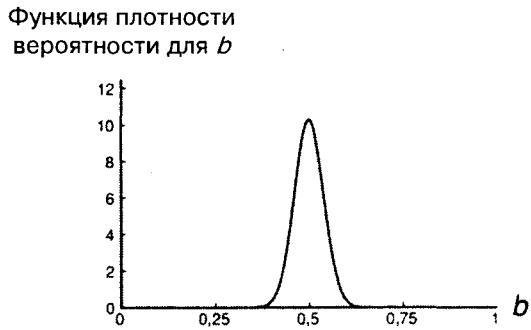


Рис. 3.1. Распределение *b* в эксперименте по методу Монте-Карло

Выше говорилось, что расхождения между коэффициентами регрессии и истинными значениями параметров вызваны случайным членом *u*. Отсюда следует, что чем больше элемент случайности, тем, вообще говоря, менее точными являются оценки.

Этот вывод будет проиллюстрирован с помощью второй серии экспериментов по методу Монте-Карло, связанной с первой. Мы будем использовать те же значения α и β, что и раньше, те же значения *x* и тот же источник случайных чисел для генерирования случайного члена, но теперь будем брать

Таблица 3.3

x	u'	y	x	u'	y
1	−1,18	1,32	11	3,18	10,68
2	−0,48	2,52	12	−1,84	6,16
3	−1,66	1,84	13	−1,42	7,08
4	0,06	3,94	14	−0,50	8,50
5	−0,76	3,74	15	3,38	12,88
6	−4,38	0,62	16	0,30	10,30
7	2,06	7,56	17	0,04	10,54
8	0,48	6,48	18	−0,22	10,78
9	5,06	11,56	19	−1,82	9,68
10	−0,26	6,74	20	2,84	14,84

значения случайного члена в каждом наблюдении. Последний выразим через u'_i $(i = 1, 2, ..., n)$, значения которого равны удвоенному случайному числу: $u'_1 = 2rn_1, ..., u'_{20} = 2rn_{20}$. Фактически мы используем в точности ту же выборку случайных чисел, что и раньше, но на этот раз удвоим их значения. Теперь на основе данных табл. 3.1 рассчитаем табл. 3.3. Далее, оценивая регрессию между y и x, получим уравнение:

$$\hat{y} = 1{,}26 + 0{,}58x. \tag{3.10}$$

Это уравнение гораздо менее точно, чем уравнение (3.8).

Таблица 3.4

Эксперимент	*a*	*b*
1	1,26	0,58
2	3,05	0,45
3	2,26	0,39
4	2,28	0,50
5	1,42	0,61
6	1,61	0,52
7	1,44	0,63
8	4,37	0,33
9	0,52	0,65
10	1,88	0,55

В табл. 3.4 приведены результаты всех 10 экспериментов при $u' = 2rn$. Мы будем называть это серией экспериментов II, а первоначальную серию экспериментов, результаты которых приведены в табл. 3.2, — серией I. При сравнении табл. 3.2 и 3.4 можно видеть, что значения a и b во второй таблице являются значительно более неустойчивыми, хотя в них по-прежнему нет систематической тенденции к занижению или завышению значений оценок.

Детальное исследование позволяет обнаружить важную особенность. В серии I значение b в эксперименте *1* было равно 0,54, и завышение оценки составило 0,04. В серии II значение b в эксперименте *1* равнялось 0,58 и завышение составило 0,08, т. е. оно было ровно вдвое больше, чем раньше. То же самое повторяется для каждого из 9 других экспериментов, а также для коэффициента регрессии a в каждом эксперименте. Удвоение случайного члена в каждом наблюдении приводит к удвоению ошибок в значениях коэффициентов регрессии.

Этот результат следует непосредственно из разложения b в соответствии с уравнением (3.6). В серии I случайная ошибка в b задается в виде $\mathrm{Cov}(x, u)/\mathrm{Var}(x)$. В серии II она представлена как $\mathrm{Cov}(x, u')/\mathrm{Var}(x)$, и

$$\frac{\mathrm{Cov}(x,u')}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x,2u)}{\mathrm{Var}(x)} = 2\frac{\mathrm{Cov}(x,u)}{\mathrm{Var}(x)}. \qquad (3.11)$$

Увеличение неточности отражено в функции плотности вероятности для b в серии II, показанной на рис. 3.2. Эта функция вновь симметрична относительно истинного значения 0,50, однако если вы сравните ее с функцией, изображенной на рис. 3.1, то увидите, что данная кривая более полога и широка. Удвоение значений u привело к удвоению стандартного отклонения распределения.

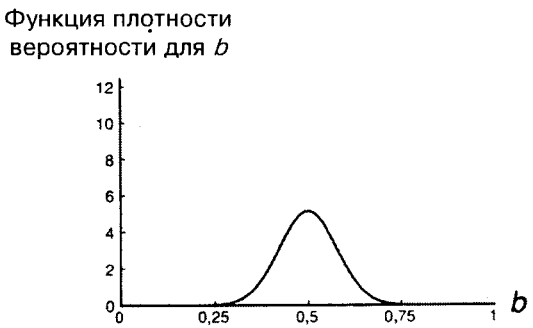


Рис. 3.2. Распределение b при удвоении стандартного отклонения u

3.3. Предположения о случайном члене

Итак, очевидно, что свойства коэффициентов регрессии существенным образом зависят от свойств случайной составляющей. В самом деле, для того чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможных результаты, случайный член должен удовлетворять четырем условиям, известным как условия Гаусса—Мар-

кова. Не будет преувеличением сказать, что именно понимание важности этих условий отличает компетентного исследователя, использующего регрессионный анализ, от некомпетентного. Если эти условия не выполнены, исследователь должен это сознавать. Если корректирующие действия возможны, то аналитик должен быть в состоянии их выполнить. Если ситуацию исправить невозможно, исследователь должен быть способен оценить, насколько серьезно это может повлиять на результаты.

Рассмотрим теперь эти условия одно за другим, объясняя кратко, почему они имеют важное значение. Три последних условия будут также подробно рассмотрены в следующих главах.

1-е условие Гаусса—Маркова: $E(U_i) = 0$ для всех наблюдений

Первое условие состоит в том, что математическое ожидание случайного члена в любом наблюдении должно быть равно нулю. Иногда случайный член будет положительным, иногда отрицательным, но он не должен иметь систематического смещения ни в одном из двух возможных направлений.

Фактически если уравнение регрессии включает постоянный член, то обычно бывает разумно предположить, что это условие выполняется автоматически, так как роль константы состоит в определении любой систематической тенденции в y, которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

2-е условие Гаусса—Маркова: pop. var (u_i) постоянна для всех наблюдений

Второе условие состоит в том, что дисперсия случайного члена должна быть постоянна для всех наблюдений. Иногда случайный член будет больше, иногда меньше, однако не должно быть априорной причины для того, чтобы он порождал большую ошибку в одних наблюдениях, чем в других.

Эта постоянная дисперсия обычно обозначается σ_u^2, или часто в более краткой форме σ^2, а условие записывается следующим образом:

$$\text{pop. var}\,(u_i) = \sigma_u^2 \text{ для всех } i. \tag{3.12}$$

Так как $E(u_i) = 0$ и pop. var $(u_i) = E(u_i^2)$, условие можно переписать в виде:

$$E(u_i^2) = \sigma_u^2 \text{ для всех } i. \tag{3.13}$$

Величина σ_u, конечно, неизвестна. Одна из задач регрессионного анализа состоит в оценке стандартного отклонения случайного члена.

Если рассматриваемое условие не выполняется, то коэффициенты регрессии, найденные по обычному методу наименьших квадратов, будут неэффективны, и можно получить более надежные результаты путем применения модифицированного метода регрессии. Это будет рассмотрено в главе 7.

3-е условие Гаусса—Маркова: pop. cov (u_i, u_j) = 0 ($i \neq j$)

Это условие предполагает отсутствие систематической связи между значениями случайного члена в любых двух наблюдениях. Например, если случайный член велик и положителен в одном наблюдении, это не должно обусловливать систематическую тенденцию к тому, что он будет большим и положительным в следующем наблюдении (или большим и отрицательным, или малым и положительным, или малым и отрицательным). Случайные члены должны быть абсолютно независимы друг от друга.

В силу того, что $E(u_i) = E(u_j) = 0$, данное условие можно записать следующим образом:

$$E(u_i u_j) = 0 \ (i \neq j). \tag{3.14}$$

Если это условие не будет выполнено, то регрессия, оцененная по обычному методу наименьших квадратов, вновь даст неэффективные результаты. В главе 7 рассматриваются возникающие здесь проблемы и пути их преодоления.

4-е условие Гаусса—Маркова: случайный член должен быть распределен независимо от объясняющих переменных

В большинстве глав книги мы будем в сущности использовать более сильное предположение о том, что объясняющие переменные не являются стохастическими, т. е. не имеют случайной составляющей. Значение любой независимой переменной в каждом наблюдении должно считаться экзогенным, полностью определяемым внешними причинами, не учитываемыми в уравнении регрессии.

Если это условие выполнено, то теоретическая ковариация между независимой переменной и случайным членом равна нулю. Так как $E(u_i) = 0$, то

$$\text{pop.cov}(x_i, u_i) = E\{(x_i - \bar{x})(u_i)\} = E(x_i u_i) - \bar{x}E(u_i) = E(x_i u_i). \tag{3.15}$$

Следовательно, данное условие можно записать также в виде:

$$E(x_i u_i) = 0. \tag{3.16}$$

В главах 8 и 11 рассматриваются два важных случая, в которых данное условие не выполнено, и последствия этого.

Предположение о нормальности

Наряду с условиями Гаусса—Маркова обычно также предполагается нормальность распределения случайного члена. Читатели должны знать о нормальном распределении из вводного курса статистики. Дело в том, что если случайный член *и* нормально распределен, то так же будут распределены и коэффициенты регрессии. Это условие пригодится нам позже в данной главе, когда потребу-

ется проводить проверку гипотез и определять доверительные интервалы для α и β, используя результаты построения регрессии.

Предположение о нормальности основывается на *центральной предельной теореме*. В сущности, теорема утверждает, что если случайная величина является общим результатом взаимодействия большого числа других случайных величин, ни одна из которых не является доминирующей, то она будет иметь приблизительно нормальное распределение, даже если отдельные составляющие не имеют нормального распределения.

Случайный член u определяется несколькими факторами, которые не входят в явной форме в уравнение регрессии. Поэтому даже если мы ничего не знаем о распределении этих факторов (или даже об их сущности), мы имеем право предположить, что они нормально распределены. В любом случае вряд ли вы столкнетесь здесь с проблемами.

3.4. Несмещенность коэффициентов регрессии

На основании уравнения (3.6) можно показать, что b будет несмещенной оценкой β, если выполняется 4-е условие Гаусса—Маркова:

$$E\{b\} = E\left\{\beta + \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}\right\} = \beta + E\left\{\frac{\mathrm{Cov}(x, u)}{\mathrm{Var}(x)}\right\}, \tag{3.17}$$

так как β — константа. Если мы примем сильную форму 4-го условия Гаусса—Маркова и предположим, что x — неслучайная величина, мы можем также считать $\mathrm{Var}(x)$ известной константой и, таким образом,

$$E\{b\} = \beta + \frac{1}{\mathrm{Var}(x)} E\{\mathrm{Cov}(x, u)\}. \tag{3.18}$$

Далее, если x — неслучайная величина, то $E\{\mathrm{Cov}(x, u)\} = 0$ и, следовательно,

$$E\{b\} = \beta. \tag{3.19}$$

Таким образом, b — несмещенная оценка β. Можно получить тот же результат со слабой формой 4-го условия Гаусса—Маркова (которая допускает, что переменная x имеет случайную ошибку, но предполагает, что она распределена независимо от u); это показано в главе 8.

За исключением того случая, когда случайные факторы в n наблюдениях в точности «гасят» друг друга, что может произойти лишь при случайном совпадении, b будет отличаться от β в каждом конкретном эксперименте. Однако с учетом соотношения (3.19) не будет систематической ошибки, завышающей или занижающей оценку. То же самое справедливо и для коэффициента a. Используем уравнение (2.35):

$$a = \bar{y} - b\bar{x}. \tag{3.20}$$

Следовательно,

$$E\{a\} = E\{\bar{y}\} - \bar{x}E\{b\}. \tag{3.21}$$

Поскольку y определяется уравнением (3.1),

$$E\{y_i\} = \alpha + \beta x_i + E\{u_i\} = \alpha + \beta x_i, \tag{3.22}$$

так как $E\{u_i\} = 0$, если выполнено 1-е условие Гаусса—Маркова. Следовательно,

$$E\{\bar{y}\} = \alpha + \beta \bar{x}. \tag{3.23}$$

Подставив это выражение в (3.21) и воспользовавшись тем, что $E\{b\} = \beta$, получим:

$$E\{a\} = (\alpha + \beta \bar{x}) - \beta \bar{x} = \alpha. \tag{3.24}$$

Таким образом, a — это несмещенная оценка α при условии выполнения 1-го и 4-го условий Гаусса—Маркова. Безусловно, для любой конкретной выборки фактор случайности приведет к расхождению оценки и истинного значения.

3.5. Точность коэффициентов регрессии

Рассмотрим теперь теоретические дисперсии оценок a и b. Они задаются следующими выражениями (доказательства для эквивалентных выражений можно найти в работе Дж. Томаса [Thomas, 1983, section 8.3.3]):

$$\text{pop.var}(a) = \frac{\sigma_u^2}{n}\left\{1 + \frac{\bar{x}^2}{\text{Var}(x)}\right\} \quad \text{и} \quad \text{pop.var}(b) = \frac{\sigma_u^2}{n\text{Var}(x)}. \tag{3.25}$$

Из уравнения (3.25) можно сделать три очевидных заключения. Во-первых, дисперсии a и b прямо пропорциональны дисперсии остаточного члена σ^2. Чем больше фактор случайности, тем хуже будут оценки при прочих равных условиях. Это уже было проиллюстрировано в экспериментах по методу Монте-Карло в разделе 3.2. Оценки в серии II были гораздо более неточными, чем в серии I, и это произошло потому, что в каждой выборке мы удвоили случайный член. Удвоив u, мы удвоили его стандартное отклонение и, следовательно, удвоили стандартные отклонения a и b. Во-вторых, чем больше число наблюдений, тем меньше дисперсии оценок. Это также имеет определенный смысл. Чем большей информацией вы располагаете, тем более точными, вероятно, будут ваши оценки. В-третьих, чем больше дисперсия x, тем меньше будет дисперсия коэффициентов регрессии. В чем причина этого? Напомним, что (1) коэффициенты регрессии вычисляются на основании предположения, что наблюдаемые изменения y происходят вследствие изменений x, но (2) в действительности они лишь *отчасти* вызваны изменениями x, а *отчасти* вариациями u. Чем меньше дисперсия x, тем больше, вероятно, будет относительное влияние фактора случайности при определении отклонений y и тем более вероятно, что регрессионный анализ может оказаться неверным. В действительности, как видно из уравнения (3.25), важное значение имеет не *абсолютная*, а *относительная* величина σ_u^2 и $\text{Var}(x)$.

На практике мы не можем вычислить теоретические дисперсии a или b, так как σ_u^2 неизвестно, однако мы можем получить оценку σ_u^2 на основе остатков. Очевидно, что разброс остатков относительно линии регрессии будет отражать неизвестный разброс u относительно линии $y = \alpha + \beta x$, хотя в общем остаток и случайный член в любом данном наблюдении не равны друг другу. Следовательно, выборочная дисперсия остатков Var (e), которую мы можем измерить, сможет быть использована для оценки σ_u^2, которую мы получить не можем.

Прежде чем пойти дальше, задайте себе следующий вопрос: какая прямая будет ближе к точкам, представляющим собой выборку наблюдений по x и y: истинная прямая $y = \alpha + \beta x$ или линия регрессии $\hat{y} = a + bx$? Ответ будет таков: линия регрессии, потому что по определению она строится таким образом, чтобы свести к минимуму сумму квадратов расстояний между ней и значениями наблюдений. Следовательно, разброс остатков у нее меньше, чем разброс значений u, и Var (e) имеет тенденцию занижать оценку σ_u^2. Действительно, можно показать, что математическое ожидание Var (e), если имеется всего одна независимая переменная, равно $[(n-2)/n]\,\sigma_u^2$. Однако отсюда следует, что если определить s_u^2 как

$$s_u^2 = \frac{n}{n-2}\,\mathrm{Var}(e), \tag{3.26}$$

то σ_u^2 будет представлять собой несмещенную оценку σ_u^2 (см. доказательство в работе Дж. Томаса).

Используя уравнения (3.25) и (3.26), можно получить оценки теоретических дисперсий для a и b и после извлечения квадратного корня — оценки их стандартных отклонений. Вместо слишком громоздкого термина «оценка стандартного отклонения функции плотности вероятности» коэффициента регрессии будем использовать термин *«стандартная ошибка»* коэффициента регрессии, которую в дальнейшем мы будем обозначать в виде сокращения «с. о.» Таким образом, для парного регрессионного анализа мы имеем:

$$\text{с.о.}(a) = \sqrt{\frac{s_u^2}{n}\left\{1 + \frac{\bar{x}^2}{\mathrm{Var}(x)}\right\}} \quad \text{и} \quad \text{с.о.}(b) = \sqrt{\left\{\frac{s_u^2}{n\mathrm{Var}(x)}\right\}}. \tag{3.27}$$

Если воспользоваться компьютерной программой оценивания регрессии, то стандартные ошибки будут подсчитаны автоматически одновременно с оценками a и b.

Полученные соотношения будут проиллюстрированы экспериментами по методу Монте-Карло, описанными в разделе 3.2. В серии I u определялось на основе случайных чисел, взятых из генеральной совокупности с нулевым средним и единичной дисперсией ($\sigma_u^2 = 1$), а x представлял собой набор чисел от 1 до 20. Можно легко вычислить Var (x), которая равна 33,25. Следовательно,

$$\text{pop.var}(a) = \frac{1}{20}\left\{1 + \frac{10{,}5^2}{33{,}25}\right\} = 0{,}2158 \tag{3.28}$$

и

$$\text{pop.var}(b) = \frac{1}{20 \times 33{,}25} = 0{,}001504. \qquad (3.29)$$

Таким образом, истинное стандартное отклонение для b равно $\sqrt{0{,}001504} = 0{,}039$. Какие же результаты получены вместо этого компьютером в 10 экспериментах серии I? Он должен был вычислить стандартную ошибку, используя уравнение (3.27); результаты этих расчетов для 10 экспериментов представлены в табл. 3.5. Как видите, большинство оценок достаточно хороши.

Таблица 3.5

Эксперимент	*с.о.(b)*	*Эксперимент*	*с.о.(b)*
1	0,043	6	0,044
2	0,041	7	0,039
3	0,038	8	0,040
4	0,035	9	0,033
5	0,027	10	0,033

Следует подчеркнуть один основной момент. Стандартная ошибка дает только общую оценку степени точности коэффициента регрессии. Она позволяет вам получить некоторое представление о кривой функции плотности вероятности, как показано на рис. 3.1. Однако она *не* несет информации о том, находится ли полученная оценка в середине распределения и, следовательно, является точной или в «хвосте» распределения и, таким образом, относительно неточна.

Чем больше дисперсия случайного члена, тем, очевидно, больше будет выборочная дисперсия остатков и, следовательно, существеннее стандартные ошибки коэффициентов в уравнении регрессии, что позволяет с высокой вероятностью заключить, что полученные коэффициенты неточны. Однако это всего лишь *вероятность*. Возможно, что в какой-то конкретной выборке воздействия случайного фактора в различных наблюдениях будут взаимно погашены и в конечном итоге коэффициенты регрессии будут точны. Проблема состоит в том, что, вообще говоря, нельзя утверждать, произойдет это или нет.

Упражнения

В тех случаях, когда результат какой-то игры, требующей определенного умения, измеряется числом, повышение уровня игры, достигаемое постоянной практикой, можно представить графически с помощью так называемой кривой обучения. Это особенно наглядно для видеоигр, когда играющий в реальном времени управляет объектом, который атакует и защищается от других объектов, управляемых программой. Тот, кто первый раз участвует в та-

кой игре, обычно проигрывает уже через несколько секунд. Чем больше вы будете играть, тем скорее привыкнете к игре и тем бо́льшее количество очков вы будете набирать, хотя очевидно, что могут иметь место некоторые отклонения, вызванные фактором случайности. Предположим, что количество очков определяется кривой обучения

$$y = 500 + 100x + u,$$

где y — результат очередной игры, x — число игр, проведенных игроком до текущей игры (порядковый номер текущей игры минус единица), и u — случайный член.

В следующей таблице приведены результаты первых 20 игр нового игрока: x автоматически изменяется от 0 до 19; в качестве значений u были взяты числа, полученные с помощью генератора нормально распределенных случайных чисел с нулевым средним и единичной дисперсией, которые были затем умножены на 400; величина y определялась через значения x и u в соответствии с линейной кривой обучения.

Наблюдение	*x*	*u*	*y*	*Наблюдение*	*x*	*u*	*y*
1	0	-236	264	11	10	636	2136
2	1	-96	504	12	11	-368	1232
3	2	-332	368	13	12	-284	1416
4	3	12	812	14	13	-100	1700
5	4	-152	748	15	14	676	2576
6	5	-876	124	16	15	60	2060
7	6	412	1512	17	16	8	2108
8	7	96	1296	18	17	-44	2156
9	8	1012	2312	19	18	-364	1936
10	9	-52	1348	20	19	-568	2968

Оценивая регрессию между y и x, получим уравнение (в скобках указаны стандартные ошибки):

$$\hat{y} = 369 + 116{,}8x.$$
$$(190) \quad (17{,}1)$$

3.1. Почему постоянный член в этом уравнении не равен 500, а коэффициент перед x не равен 100?

3.2. Каковы значения стандартных ошибок?

3.3. Эксперимент повторяется с 9 другими новыми игроками (в каждом случае случайный член получают путем умножения на 400 разных наборов из 20 случайных чисел), а результаты оценивания регрессии для всех 10 игроков при-

ведены в следующей таблице. Почему постоянный член, коэффициент при x и их стандартные ошибки меняются от выборки к выборке?

Игрок	*Постоянная*	*с.о. постоянной*	*Коэффициент при x*	*с.о. коэффициента при x*
1	369	190	116,8	17,1
2	699	184	90,1	16,5
3	531	169	78,5	15,2
4	555	158	99,5	14,2
5	407	120	122,6	10,8
6	427	194	104,3	17,5
7	412	175	123,8	15,8
8	613	192	95,8	17,3
9	234	146	130,1	13,1
10	485	146	109,6	13,1

3.4. Дисперсия x равна 33,25, а дисперсия u равна 160 000. Используя уравнение (3.25), покажите, что стандартное отклонение функции плотности вероятности коэффициента при x равно 15,5. Являются ли приведенные в таблице стандартные ошибки хорошими оценками стандартного отклонения?

3.6. Теорема Гаусса—Маркова

В обзоре мы рассматривали оценки неизвестного математического ожидания μ случайной величины x по данным выборочных наблюдений. Хотя мы интуитивно использовали в качестве оценки для μ выборочное среднее x, было показано, что оно является лишь одной из бесконечного числа возможных несмещенных оценок этого параметра. Причина предпочтения выборочного среднего всем другим оценкам состоит в том, что при определенных предположениях оно является наиболее эффективным.

Аналогичные рассуждения применимы и к коэффициентам регрессии. Мы увидим, что оценки по обычному методу наименьших квадратов являются не только несмещенными оценками коэффициентов регрессии, но и наиболее эффективными в том случае, если выполнены условия Гаусса—Маркова. С другой стороны, если условия Гаусса—Маркова не выполнены, то, вообще говоря, можно найти оценки, которые будут более эффективными по сравнению с оценками, полученными обычным методом наименьших квадратов.

В данной работе не приводится общее рассмотрение этих вопросов. Мы дадим лишь иллюстрацию. Предположим, что мы имеем зависимость, заданную

уравнением (3.1), и сосредоточим внимание на оценках для β. Человек, не знакомый с регрессионным анализом, увидев диаграмму разброса для выборки наблюдений, может попытаться получить оценку тангенса угла наклона путем простого объединения первого и последнего наблюдений и деления прироста высоты на горизонтальный отрезок между ними, как показано на рис. 3.3. Оценка b в этом случае будет определяться следующим образом:

$$b = \frac{y_n - y_1}{x_n - x_1}. \tag{3.30}$$

Каковы свойства этой оценки? Сначала мы исследуем, является ли она несмещенной. Используя уравнение (3.1) применительно к первому и последнему наблюдениям, получим:

$$y_1 = \alpha + \beta x_1 + u_1; \tag{3.31}$$

$$y_n = \alpha + \beta x_n + u_n. \tag{3.32}$$

Следовательно,

$$b = \frac{\beta x_n + u_n - \beta x_1 - u_1}{x_n - x_1} = \beta + \frac{u_n - u_1}{x_n - x_1}. \tag{3.33}$$

Таким образом, мы разложили «наивную» оценку на две составляющие: истинное значение и остаточный член. Это разложение выполнено подобно тому, как это сделано в разделе 3.1 для оценки МНК. Однако остаточный член является другим. Предполагая $E(u) = 0$, мы имеем, что математические ожидания, как u_1, так и u_n, равны нулю, но тогда математическое ожидание остаточного члена в уравнении (33.3) также равно нулю. Таким образом, несмотря на то что эта оценка столь «наивна», она является несмещенной.

Это, разумеется, не единственная оценка, которая наряду с оценкой, полученной методом МНК, обладает свойством несмещенности. Вы можете получить еще одну оценку такого типа путем объединения двух произвольно выбранных наблюдений, а если вы хотите рассмотреть менее «наивные» процедуры, то здесь открываются поистине безграничные возможности.

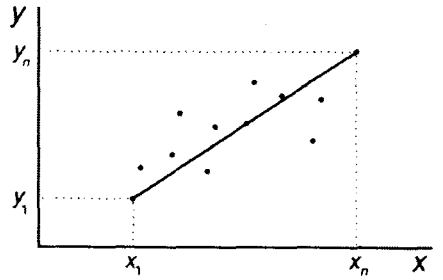


Рис. 3.3. «Наивная» оценка b

Интуитивно легко понять, что мы не предпочтем «наивную» оценку типа (3.30) оценке МНК. В отличие от оценки МНК, в которой учитывается каждое наблюдение, в «наивной» оценке берется только первое и последнее наблюдения и не используется большая часть имеющейся в выборке информации. «Наивная» оценка зависит от значений остаточного члена u в данных двух

наблюдениях, тогда как оценка по методу наименьших квадратов объединяет все значения остаточного члена и более эффективно использует возможность того, что эти значения в некоторой степени взаимно «погашаются».

При сравнении с менее «наивными» оценками превосходство оценки МНК в эффективности может быть не столь очевидным. Тем не менее в том случае, если условия Гаусса—Маркова для остаточного члена выполнены, коэффициенты регрессии, построенной обычным методом наименьших квадратов, будут *наилучшими линейными несмещенными оценками* (best linear unbiased estimators, или BLUE): несмещенными, как уже было показано; линейными, так как они являются линейными функциями значений y; наилучшими, так как они являются наиболее эффективными в классе всех несмещенных линейных оценок. Теорема Гаусса—Маркова доказывает это (краткое изложение, не использующее матричной алгебры, дано в работе Дж. Томаса [Thomas, 1983, section 8.3]).

Упражнения

3.5. Исследователь обоснованно считает, что зависимость между двумя переменными x и y задается уравнением (3.1). Используя выборку из n наблюдений, он оценивает β, вычисляя среднее значение y, деленное на среднее значение x. Проанализируйте свойства этой оценки. Что изменится, если предположить, что $\alpha = 0$?

3.6. Исследователь обоснованно считает, что зависимость между двумя переменными x и y задается уравнением (3.1). Используя выборку из n наблюдений временного ряда, он оценивает β как $\mathrm{Cov}(y, t)/\mathrm{Cov}(x, t)$, где t — переменная времени, которая по определению равна единице в первом наблюдении, двум — во втором и т. д. Проанализируйте свойства этой оценки. (Можно показать, что ее теоретическая дисперсия равна теоретической дисперсии соответствующей оценки МНК, деленной на $r^2_{x,t}$, где $r_{x,t}$ — коэффициент корреляции между x и t.)

3.7. Проверка гипотез, относящихся к коэффициентам регрессии

С чего начинается статистическое исследование — с теоретического построения гипотез или с эмпирического анализа? В действительности, теория и практика взаимно обогащают друг друга, и подобные вопросы не задаются. Поэтому вопрос о проверке гипотез мы будем рассматривать с двух точек зрения. С одной стороны, мы можем предположить, что сначала формулируется гипотеза, и цель эксперимента заключается в выяснении ее применимости. Это приведет к проверке гипотезы о значимости. С другой стороны, мы можем сначала провести эксперимент и затем определить, какие из теоретических гипотез соответствуют результатам эксперимента. Это приводит к построению доверительных интервалов.

Вам уже известна логика, лежащая в основе построения критериев значи-

мости и доверительных интервалов и описанная в вводном курсе статистики. Поэтому вы уже знакомы с большинством понятий, используемых в регрессионном анализе. Однако один вопрос может оказаться для вас новым — это использование односторонних критериев. Такие критерии применяются в регрессионном анализе очень часто. В самом деле, они являются, или должны быть, более обычными здесь, чем двусторонние критерии, традиционно используемые в учебниках. Поэтому важно, чтобы вы поняли целесообразность их применения, и путь к этому лежит через последовательный ряд небольших аналитических шагов. Ни один из них не должен представлять трудности, но следует иметь в виду, что если вы попытаетесь сократить путь или, еще хуже, сделаете попытку свести всю процедуру к механическому использованию нескольких формул, вы столкнетесь с большими трудностями.

Формулирование нулевой гипотезы

Начнем с допущения о том, что формулирование гипотезы предшествует эксперименту и что вы уже имеете в виду некоторую гипотетическую связь или зависимость. Например, можно считать, что темпы общей инфляции в экономике ($\dot{p}$, в процентах) зависят от темпов инфляции, вызванной ростом заработной платы ($\dot{w}$, в процентах), и что эта зависимость описывается линейным уравнением:

$$\dot{p} = \alpha + \beta\dot{w} + u, \tag{3.34}$$

где α и β — параметры, а u — случайный член. Далее можно построить гипотезу о том, что без учета эффектов, вносимых случайным членом, общая инфляция равна инфляции, вызванной ростом заработной платы. В этих условиях можно сказать, что гипотеза, которую вы собираетесь проверить, считается *нулевой*, обозначается H_0 и состоит в том, что $\beta = 1$. Мы также определяем *альтернативную гипотезу*, которая обозначается H_1 и представляет собой заключение, даваемое в том случае, если экспериментальная проверка указала на ложность H_0. В данном случае эта гипотеза состоит в том, что $\beta \neq 1$. Две гипотезы сформулированы с использованием следующих обозначений:

$$H_0: \beta = 1;$$

$$H_1: \beta \neq 1.$$

В этом конкретном случае, если действительно считать, что общая инфляция равна инфляции, вызванной ростом заработной платы, мы делаем попытку защитить нулевую гипотезу H_0, подвергнув ее максимально строгой проверке и надеясь, что она не будет опровергнута. Однако на практике более обычным является построение нулевой гипотезы, которая затем будет проверяться с помощью альтернативной гипотезы, которая предполагается верной. Например, рассмотрим простую функцию спроса:

$$y = \alpha + \beta x + u, \tag{3.35}$$

где y — величина спроса, скажем, на продукты питания, а x — доход. Исходя из вполне разумных теоретических оснований, вы предполагаете, что спрос

на продукты питания зависит от дохода, но ваша гипотеза недостаточно «сильна», чтобы можно было определить конкретное значение для β. Тем не менее вы можете установить наличие зависимости величины y от x, используя для этого обратную процедуру, когда в качестве нулевой гипотезы принимается утверждение о том, что величина y *не* зависит от x, т. е. что $\beta = 0$. Альтернативная гипотеза заключается в том, что $\beta \neq 0$, иными словами, что значение x *влияет* на величину y. Если можно отвергнуть нулевую гипотезу, вы таким образом устанавливаете наличие зависимости, по крайней мере в общих чертах. С использованием введенной системы обозначений нулевая и альтернативная гипотезы соответственно примут вид:

$$H_0: \beta = 0 \quad \text{и} \quad H_1: \beta \neq 0.$$

Последующее рассмотрение касается модели парной регрессии (3.1). Оно будет относиться только к коэффициенту наклона β, но точно такие же процедуры применимы и к постоянному члену α. Возьмем общий случай, в котором в нулевой гипотезе утверждается, что β равно некоторому конкретному значению, скажем, β_0, и альтернативная гипотеза состоит в том, что β не равно этому значению ($H_0: \beta = \beta_0$; $H_1: \beta \neq \beta_0$). Вы можете предпринять попытку отклонить или подтвердить нулевую гипотезу в зависимости от того, что вам необходимо в данном случае. Будем предполагать, что четыре условия Гаусса—Маркова выполняются.

Вывод следствий гипотезы

Если гипотеза H_0 верна, то оценки β, полученные в ходе регрессионного анализа, будут иметь распределение с математическим ожиданием β_0 и дисперсией $\sigma_u^2/[n \operatorname{Var}(x)]$ [см. уравнение (3.25)]. Теперь мы вводим допущение, что остаточный член u имеет нормальное распределение. Если это так, то величина b будет также нормально распределена, как показано на рис. 3.4. Сокращение

Рис. 3.4. Структура нормального распределения оценки b, выраженной через стандартные отклонения от математического ожидания

«s. d.» на рисунке соответствует величине стандартного отклонения оценки b, т. е. $\sqrt{\frac{\sigma_u^2}{n\mathrm{Var}(x)}}$. Учитывая структуру нормального распределения, большинство оценок параметра β будет находиться в пределах двух стандартных отклонений от β_0 (если верна гипотеза $H_0: \beta = \beta_0$).

Сначала мы допустим, что знаем значение стандартного отклонения величины b. Это наиболее неправдоподобное допущение, и мы позднее отбросим его. На практике же значение этого отклонения (так же как и неизвестные значения параметров α и β) подлежит оценке. Можно, тем не менее, упростить рассмотрение, предположив, что точное значение отклонения известно, и, следовательно, имея возможность построить график (рис. 3.4).

Проиллюстрируем это на примере модели общей инфляции (3.34). Предположим, что некоторым образом мы знаем, что стандартное отклонение величины b составляет 0,1. Тогда если нулевая гипотеза $H_0: \beta = 1$ верна, то оценки коэффициентов регрессии будут распределены так, как это показано на рис. 3.5. Из этого рисунка можно видеть, что при справедливости нулевой гипотезы оценки будут находиться приблизительно между 0,8 и 1,2.

Сопоставимость, случайность и уровень значимости

Теперь приступим к главному. Предположим, что мы взяли фактическую выборку из наблюдений общей инфляции и инфляции, вызванной ростом заработной платы, и построили оценку β, используя для этого регрессионный анализ. Если оценка близка 1,0, мы должны быть полностью удовлетворены нулевой гипотезой, так как она и результат оценивания для выборки совместимы друг с другом. Но с другой стороны, предположим, что оценка значительно отличается от 1,0. Допустим, например, что она равна 0,7. Это составит три стандартных отклонения вниз от 1,0. Вероятность того, что отличие от среднего до-

Рис. 3.5. Пример распределения величины b (модель связи общей инфляции и инфляции, вызванной ростом заработной платы)

стигнет трех стандартных отклонений в положительную или отрицательную сторону, составляет лишь 0,0027, т. е. очень низка. Исходя из этого вызывающего беспокойство результата, вы можете прийти к одному из двух выводов.

1. Вы можете продолжать считать, что нулевая гипотеза $\beta = 1{,}0$ верна и что эксперимент дал случайный результат. Вы допускаете, что вероятность получения такого низкого значения для β является очень небольшой, но, тем не менее, она имеет место в 0,27% случаев, и вы допускаете, что это именно тот случай.

2. Вы можете сделать вывод о том, что гипотеза противоречит результату оценивания регрессии. Вы не удовлетворяетесь объяснением, данным в пункте *1*, так как вероятность очень мала, и понимаете, что наиболее правдоподобным объяснением является то, что величина β вовсе не равняется 1,0. Другими словами, вы принимаете альтернативную гипотезу $H_1: \beta \neq \beta_0$.

Каким образом вы определите, когда необходимо выбрать первый вывод, а когда — второй? Очевидно, что чем меньше вероятность построения регрессии, подобной той, которую вы получили при условии правильности гипотезы, тем больше вероятность отказа от гипотезы и выбор второго вывода. Насколько малой должна быть указанная вероятность для выбора второго вывода?

На этот вопрос нет и не может быть определенного ответа. В большинстве работ по экономике за критический уровень берется 5 или 1%. Если выбирается уровень 5%, то переключение на второй вывод происходит в том случае, когда при истинности нулевой гипотезы вероятность получения столь экстремального значения b составляет менее 5%. В этом случае говорят, что нулевая гипотеза должна быть отвергнута при 5-процентном *уровне значимости*.

Это происходит в том случае, когда величина b отстоит от β_0 более чем на 1,96 стандартного отклонения. Если вы посмотрите на таблицу нормального распределения (табл. A.1 в конце книги), то увидите, что вероятность того, что величина b будет превосходить среднее значение на более чем 1,96 стандартного отклонения, составляет 2,5% и, аналогичным образом, вероятность того, что эта величина будет более чем на 1,96 стандартного отклонения ниже среднего значения, также будет 2,5%. Общая вероятность того, что данная величина отстоит от математического ожидания более чем на 1,96 стандартного отклонения, составляет, таким образом, 5%.

Можно обобщить это решающее правило в математической форме, сказав, что нулевая гипотеза отвергается, если

$$Z > 1{,}96 \text{ или } Z < -1{,}96, \tag{3.36}$$

где Z — число стандартных отклонений между регрессионной оценкой и гипотетическим значением β:

$$Z = \frac{\textit{Разница между оценкой регрессии и гипотетическим значением}}{\textit{Стандартное отклонение величины } b} = \frac{b - \beta_0}{\text{s.d.}(b)}. \tag{3.37}$$

Нулевая гипотеза не будет отвергнута, если

$$-1{,}96 < Z < 1{,}96. \tag{3.38}$$

Это условие можно записать с помощью величин b и β_0, подставив выражение для Z из уравнения (3.37):

$$-1{,}96 < \frac{b - \beta_0}{\text{s.d.}(b)} < 1{,}96. \tag{3.39}$$

Умножив все части неравенства на стандартное отклонение величины b, можно получить:

$$-1{,}96 \text{ s.d. } (b) < b - \beta_0 < 1{,}96 \text{ s.d. } (b), \tag{3.40}$$

а из этого уравнения можно получить следующее:

$$\beta_0 - 1{,}96 \text{ s.d. } (b) < b < \beta_0 + 1{,}96 \text{ s.d. } (b). \tag{3.41}$$

Уравнение (3.41) дает множество значений для величины b, которые не приводят к отказу от конкретной нулевой гипотезы о том, что $\beta = \beta_0$. Это множество значений получило название *области принятия гипотезы* для b при 5-процентном уровне значимости.

В нашем примере, где s.d. $(b) = 0{,}1$, можно отвергнуть гипотезу при уровне значимости в 5%, если величина b находится выше или ниже гипотетического среднего значения на величину более 0,196, т. е. выше 1,196 или ниже 0,804. Таким образом, область принятия гипотезы включает значения величины b от 0,804 до 1,196. Это показано незаштрихованной областью на рис. 3.6.

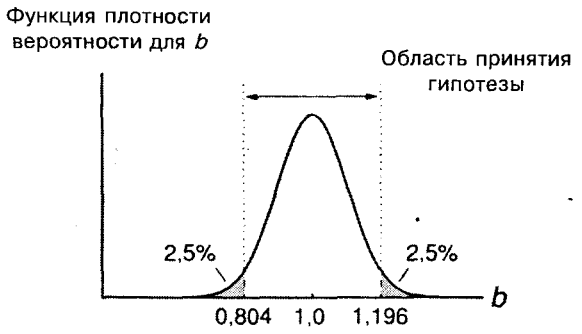


Рис. 3.6. Область принятия гипотезы для величины b при 5-процентном уровне значимости

Аналогичным образом считается, что нулевая гипотеза должна быть отвергнута при уровне значимости в 1%, если гипотеза подразумевает, что вероятность получения столь экстремального значения для величины b составляет менее 1%. Это происходит, когда величина b отстоит на более чем 2,58 стандартного отклонения вверх или вниз от гипотетического значения β, т. е. когда

$$Z > 2{,}58 \text{ или } Z < -2{,}58. \tag{3.42}$$

Возвращаясь к таблице нормального распределения, можно видеть, что вероятность того, что величина b более чем на 2,58 стандартного отклонения превысит свое математическое ожидание, составляет 0,5% и та же самая вероятность будет для варианта, что b окажется ниже своего математического ожидания на более чем 2,58 стандартного отклонения. Таким образом, общая вероятность получения столь экстремальных значений составляет 1%. В нашем при-

мере вы отвергнете нулевую гипотезу о том, что $\beta = 1{,}0$, если оценка коэффициента регрессии будет находиться выше 1,258 или ниже 0,742.

Можно задаться вопросом: почему исследователи обычно представляют свои результаты при уровнях значимости в 5 и 1%? Почему недостаточно ограничиться только одним уровнем? Причина заключается в том, что обычно делается попытка найти баланс между риском допущения *ошибок I и II рода. Ошибка I рода имеет место в том случае, когда вы отвергаете истинную нулевую гипотезу. Ошибка II рода возникает, когда вы не отвергаете ложную гипотезу.*

Ошибки I и II рода в повседневной жизни

Проблема, как избежать ошибок I и II рода, известна всем. Типичным примером этого является расследование уголовного преступления. Если за нулевую гипотезу принять вариант, что подсудимый невиновен, то ошибка I рода происходит, когда суд присяжных признает его виновным. Ошибка II рода имеет место в том случае, когда суд присяжных ошибочно оправдывает виновного подсудимого.

Вполне очевидно, что чем ниже критическая вероятность, тем меньше риск получения ошибок I рода. Если вы используете уровень значимости, равный 5%, то вы отвергнете истинную гипотезу в 5% случаев. Если уровень значимости составляет 1%, вы совершите ошибку I рода в 1% случаев. Таким образом, в этом отношении однопроцентный уровень значимости более надежен. Если вы отвергли гипотезу на данном уровне, вы почти наверняка были вправе сделать это. Именно по этой причине однопроцентный уровень значимости описывается как *«более высокий»* в сравнении с 5-процентным уровнем.

В то же время если нулевая гипотеза ложна, то чем выше уровень значимости, тем шире область принятия гипотезы, тем выше вероятность того, что вы не отвергнете ее, и тем выше риск допущения ошибки II рода. Таким образом, вы оказываетесь перед дилеммой. Если вы будете настаивать на очень высоком уровне значимости, то столкнетесь с относительно высоким риском допущения ошибки II рода, когда гипотеза окажется ложной. Если вы выбираете низкий уровень значимости, то оказываетесь перед относительно высоким риском допущения ошибки I рода, если гипотеза истинна.

Большинство людей выбирают достаточно простую форму обеспечения гарантий и осуществляют проверку на обоих уровнях значимости, представляя результаты каждой такой проверки. На самом деле часто нет необходимости непосредственно ссылаться на оба результата. Так как величина *b* должна быть более «экстремальной» для гипотезы, отвергаемой при однопроцентном уровне значимости, но не при 5-процентном, и если вы отклоняете ее при однопроцентном уровне, то из этого автоматически следует, что вы отклоните ее и при уровне значимости в 5%, и нет необходимости упоминать об этом. Если же вы не отвергаете гипотезу при уровне значимости в 5%, то из этого автоматически следует, что вы не отвергнете ее и при однопроцентном уровне зна-

чимости, и вновь нет смысла об этом говорить. Только в одном случае вы должны представить оба результата: если гипотеза отвергается на 5-процентном, но не на однопроцентном уровне значимости.

Что происходит, когда стандартное отклонение величины b неизвестно?

До сих пор мы считали, что стандартное отклонение величины b известно. Однако на практике это допущение нереально. Это можно показать на примере стандартной ошибки для величины b, взятой из уравнения (3.27). Это приводит к двум изменениям процедуры проверки гипотез. Во-первых, величина Z определяется на основе использования стандартной ошибки с.о. (b) вместо стандартного отклонения s.d. (b) и носит название t-статистики:

$$t = \frac{b - \beta_0}{\text{с.о.}(b)}. \tag{3.43}$$

Во-вторых, критические уровни t определяются величиной, имеющей так называемое t-распределение вместо нормального распределения. Мы не будем вдаваться в причины этого или даже описывать t-распределение математически. Достаточно будет сказать, что оно родственно нормальному распределению, а его точная форма зависит от числа *степеней свободы* в регрессии, и оно все лучше аппроксимируется нормальным распределением по мере увеличения числа степеней свободы. Вы, конечно, уже встречали понятие t-распределения во вводном курсе статистики. В табл. А.2 в конце книги представлены критические значения для t, сгруппированных по уровням значимости и числу степеней свободы.

Оценивание каждого параметра в уравнении регрессии поглощает одну степень свободы в выборке. Отсюда число степеней свободы равняется количеству наблюдений в выборке минус количество оцениваемых параметров. Параметрами являются постоянный член (при условии, что он введен в модель регрессии) и коэффициенты при независимых переменных. В рассматриваемом случае парной регрессии оцениваются только два параметра α и β, поэтому число степеней свободы составляет $n - 2$. Следует подчеркнуть, что, когда мы перейдем к множественному регрессионному анализу, потребуется более общее выражение.

Критическое значение t, которое мы обозначим как $t_{крит}$, заменит число 1,96 в уравнении (3.39). Таким образом, условие того, что оценка регрессии не должна приводить к отказу от нулевой гипотезы $\beta = \beta_0$, будет следующим:

$$-t_{крит} < \frac{b - \beta_0}{\text{с.о.}(b)} < t_{крит}. \tag{3.44}$$

Примеры

В разделе 2.6 функция расходов на питание оценивалась как зависимость от личного располагаемого дохода на основании совокупных ежегодных данных для

США за 25-летний срок (1959–1983 гг.) и уравнение регрессии было представлено формулой (2.42):

$$y = \underset{(2,4)}{55,3} + \underset{(0,003)}{0,093x} \qquad (3.45)$$

Цифры, указанные в скобках, являются стандартными ошибками.

Предположим, что одна из задач оценивания регрессии состояла в подтверждении догадки о том, что уровень расходов на питание зависит от размера дохода. Соответственно, мы формулируем нулевую гипотезу о том, что величина β равняется нулю, и затем пытаемся опровергнуть ее. Соответствующая *t-статистика*, вычисленная по формуле (3.43), есть оценка коэффициента, деленная на ее стандартную ошибку:

$$t = \frac{b - \beta_0}{\text{с.о.}(b)} = \frac{b - 0}{\text{с.о.}(b)} = \frac{0,093}{0,003} = 31,0. \qquad (3.46)$$

Так как в выборку включено 25 наблюдений и мы оценили два параметра, то число степеней свободы составляет 23. Критическое значение для t при 5-процентном уровне значимости с 23 степенями свободы равняется 2,069. Причем t-статистика не лежит между значениями 2,069 и −2,069. Следовательно, неравенство (3.44) не выполняется и мы отвергаем нулевую гипотезу, сделав вывод о том, что величина β в действительности отличается от нуля и, следовательно, размер дохода *влияет* на уровень расходов на питание.

Если этот критерий описать словами, то верхний и нижний 2,5-процентные «хвосты» t-распределения начинаются со стандартного отклонения 2,069 вверх и вниз от его математического ожидания, равного нулю. Коэффициент регрессии, который по оценкам находится в пределах 2,069 стандартного отклонения от гипотетического значения, не приводит к отказу от последнего. В рассматриваемом случае расхождение будет эквивалентно 31,0 стандартного отклонения, и мы приходим к выводу о том, что результат оценивания регрессии противоречит нулевой гипотезе.

Конечно, в том, что мы используем уровень значимости в 5% в качестве основы для проверки гипотезы, существует 5-процентный риск допущения ошибки I рода. В этом случае мы могли бы снизить риск до 1% за счет применения уровня значимости в 1%. Критическое значение для t при однопроцентном уровне значимости с 23 степенями свободы составляет 2,807. Используя это число в соотношении (3.44), мы видим, что можно легко отказаться от нулевой гипотезы также и при этом уровне значимости.

Процедура установления взаимосвязи между зависимой и объясняющей переменными путем формулирования, а затем отклонения нулевой гипотезы о том, что $\beta = 0$, используется очень часто. Соответственно, большая часть, если не все программы регрессии, автоматически выводят t-статистику для этого специального случая; иными словами, коэффициент делится на его стандартную ошибку. Данное отношение часто обозначается как «t-статистика».

Если, однако, нулевая гипотеза определяет некоторое ненулевое значение величины β, то необходимо использовать более общее выражение (3.43), а t-статистика вычисляется вручную. Например, вновь рассмотрим модель регрессии между общей инфляцией и инфляцией, вызванной ростом заработной

платы (3.34), и предположим, что выбранное уравнение регрессий оказалось следующим (в скобках указаны стандартные ошибки):

$$\dot{p} = -1{,}21 + 0{,}82\dot{w}. \qquad (3.47)$$
$$(0{,}05)\quad(0{,}10)$$

Если теперь исследовать гипотезу о том, что общая инфляция в долгосрочном периоде будет равна инфляции, вызванной ростом заработной платы, то нулевая гипотеза будет состоять в том, что коэффициент при $\dot{w}$ равен 1,0. Соответствующая t-статистика примет вид:

$$t = \frac{b - \beta_0}{\text{с.о.}(b)} = \frac{0{,}82 - 1{,}00}{0{,}10} = -1{,}8. \qquad (3.48)$$

Если в выборке содержится, скажем, 20 наблюдений, то число степеней свободы составит 18, а критическое значение для t при 5-процентном уровне значимости будет 2,101. В этом случае t-статистика лежит между 2,101 и −2,101, поэтому мы не отвергаем нулевую гипотезу. Оценка, равная 0,82, лежит ниже нашего гипотетического значения 1,00, но не настолько ниже, чтобы исключить возможность правильности нулевой гипотезы.

Терминология принятия (отклонения) гипотезы

В этом разделе было показано, что следует отклонить нулевую гипотезу, если t-статистика больше, чем $t_{крит}$, или меньше, чем $-t_{крит}$, и не следует отклонять эту гипотезу, если t-статистика находится между $-t_{крит}$ и $t_{крит}$. Почему «не отклонять», к чему это усложнение? Не было бы проще сказать, что вы принимаете гипотезу, если t-статистика находится между $-t_{крит}$ и $t_{крит}$?

Аргументом против использования термина «принять» является то, что вы способны «принять» несколько взаимоисключающих гипотез в одно и то же время. Так, в примере с зависимостью между общей инфляцией и инфляцией, вызванной ростом заработной платы, вы не могли бы отклонить нулевую гипотезу H_0: $\beta = 0{,}9$ или нулевую гипотезу H_0: $\beta = 0{,}8$. Логично сказать, что вы не отклоняете эти нулевые гипотезы, а также нулевую гипотезу H_0: $\beta = 1{,}0$, рассмотренную выше, но практически бессмысленно заявлять, что вы одновременно принимаете все три гипотезы. В следующем разделе вы увидите, что можно определить целый ряд гипотез, которые не могут быть отклонены в результате данного эксперимента. Поэтому было бы неосторожно выбрать одну из них как «принятую».

Описание результатов проверок по t-критерию

Предположим, что вы имеете теоретическую зависимость

$$y = \alpha + \beta x + u,$$

и нулевая и альтернативная гипотезы заданы в виде $H_0: \beta = \beta_0$, $H_1: \beta \neq \beta_0$. Если для β по выборочным данным получена оценка b, то области принятия и отклонения гипотез для 5-процентного и однопроцентного уровней значимости могут быть в общем представлены левой частью рис. 3.7.

Правая часть рис. 3.7 показывает те же самые области для конкретного примера модели регрессии между общей инфляцией и инфляцией, вызванной ростом заработной платы; при этом нулевая гипотеза будет иметь вид $\beta = 0$. Нулевая гипотеза не будет отклонена при уровне значимости в 5%, если величина b находится в пределах 2,101 стандартной ошибки от единицы, т. е. в диапазоне от 0,79 до 1,21, и она не будет отклонена при уровне значимости в 1%, если величина b находится в пределах 2,878 стандартного отклонения от единицы, т. е. в диапазоне от 0,71 до 1,29.

Из рис. 3.7 можно видеть, что существует три типа зон принятия решений.

1. Зона, где величина b настолько далека от гипотетического значения β, что нулевая гипотеза отклоняется как при 5-процентном, так и при однопроцентном уровнях значимости.

2. Зона, где величина b достаточно далека от гипотетического значения β, чтобы нулевая гипотеза была отклонена при 5-процентном, но не при однопроцентном уровне значимости.

3. Зона, где величина b достаточно близка к гипотетическому значению β, чтобы нулевая гипотеза не была отклонена ни при одном из двух рассматриваемых уровнях значимости.

На основании схемы можно проверить, что если нулевая гипотеза отклоняется при однопроцентном уровне значимости, то она автоматически отклоняется и при 5-процентном уровне. Следовательно, в случае *1* необходимо заявить лишь об отклонении гипотезы при однопроцентном уровне значимости. Заявлять об ее отклонении при 5-процентном уровне нет необходимости. Это равнозначно тому, чтобы сделать заявление о возможности взятия прыгуном высоты в 2 м, а затем в качестве дополнения заявить о его возможности взять высоты в 1 и 1,5 м.

Аналогичным образом для случая *3* вам необходимо сделать только заявление о том, что в этом конкретном случае гипотеза не будет отклонена при 5-процентном уровне значимости. Отсюда автоматически следует, что она не будет отклонена и при однопроцентном уровне. А дополнение к этому заявлению имело бы тот же эффект, как если бы к заявлению о том, что прыгун в высоту не может взять высоты в 1 и 1,5 м, было добавлено утверждение о его неспособности взять высоту в 2 м.

Лишь в случае *2* необходимо (и желательно) представить результаты обеих проверок.

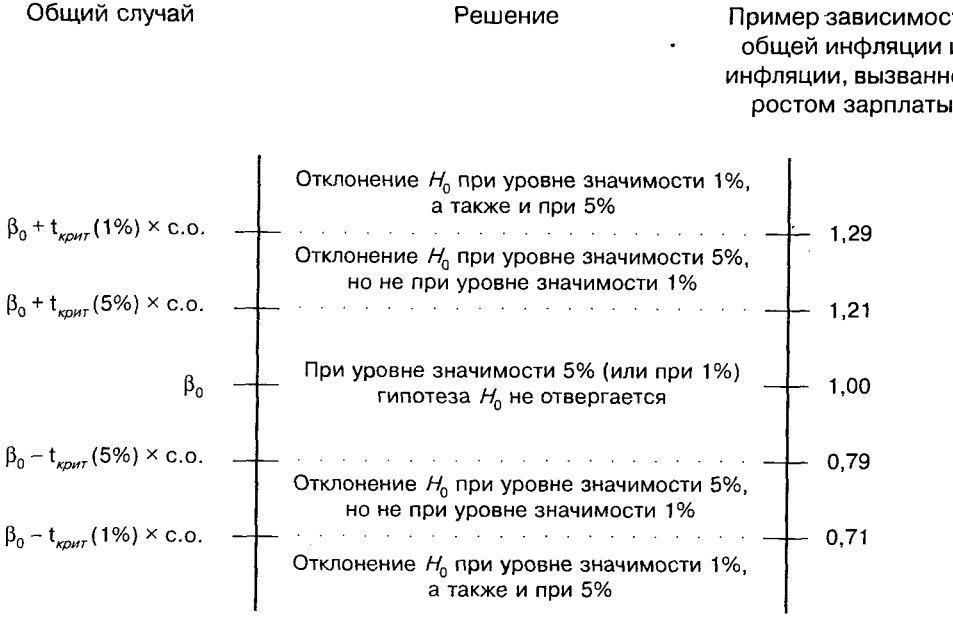


Рис. 3.7. Представление результатов проверки гипотез по *t*-критерию (выводы, заключенные в скобках, представлять не требуется)

Заключительное замечание относительно представления результатов оценивания регрессии состоит в том, что в некоторых работах в скобках под коэффициентом вместо стандартной ошибки приводится *t*-статистика. Вы должны внимательно следить за этим, и, когда представляете результаты, это должно быть сделано предельно понятно.

Упражнения

3.7. Приведите еще несколько примеров для повседневно встречающихся случаев, когда при принятии решений могут возникнуть ошибки I и II рода.

3.8. Перед началом курса обучения 36 студентов были подвергнуты тесту на проверку способностей. Результаты теста и курса обучения (по типу «прошел»/«не прошел») представлены в таблице на с. 101.

Как вы думаете, полезен ли тест на проверку способностей для принятия на курс, и если это так, то как бы вы определили проходной балл? (Рассмотрите вариант компромисса между ошибками I и II рода, связанными с выбором проходного балла.)

3.9. Стандартная ошибка коэффициента при *t* в упражнении 2.1 составила 0,08. Проверьте нулевую гипотезу о том, что истинное значение коэффициента равно нулю:

1) при 5-процентном уровне значимости;

2) при однопроцентном уровне значимости.

3.10. В упражнении 3.9 следует представить только результат проверки при однопроцентном уровне значимости. Почему?

Студент	Балл на тестировании	Результат обучения	Студент	Балл на тестировании	Результат обучения
1	30	не прошел	19	51	не прошел
2	29	прошел	20	45	не прошел
3	33	не прошел	21	22	не прошел
4	62	прошел	22	30	прошел
5	59	не прошел	23	40	не прошел
6	63	прошел	24	26	не прошел
7	80	прошел	25	9	не прошел
8	32	не прошел	26	36	прошел
9	60	прошел	27	61	прошел
10	76	прошел	28	79	не прошел
11	13	не прошел	29	57	не прошел
12	41	прошел	30	46	прошел
13	26	не прошел	31	70	не прошел
14	43	прошел	32	31	прошел
15	43	не прошел	33	68	прошел
16	68	прошел	34	62	прошел
17	63	прошел	35	56	прошел
18	42	не прошел	36	36	прошел

3.11. Предположим, что вы проверили нулевую гипотезу при обоих уровнях значимости в 5% и в 1%. При каких условиях вы представите:

1) только результат проверки при однопроцентном уровне;

2) только результат проверки при 5-процентном уровне;

3) результаты при обоих уровнях значимости?

3.12. Уравнения регрессии между расходами на коммунальные услуги и (1) располагаемым личным доходом и (2) временем в упражнении 2.2 имеют вид (стандартные ошибки указаны в скобках):

$$\hat{y} = -27{,}6 + 0{,}178x; \qquad \hat{y} = 48{,}9 + 4{,}84x.$$
$$(3{,}4)\quad(0{,}004) \qquad\qquad (1{,}5)\quad(0{,}10)$$

Выполните t-тест для проверки значимости коэффициентов там, где это необходимо. Четко сформулируйте проверяемые нулевые гипотезы и их альтернативы. Кроме того, объясните, почему вы формулируете сначала нулевые гипотезы.

3.13. Выполните аналогичные t-тесты для проверки коэффициентов регрес-

сий, оцененных вами в упражнениях 2.4 и 2.5, четко формулируя нулевые и альтернативные гипотезы.

3.14. Предположим, что по принятой гипотезе 10% предельного дохода расходуется на питание. Проверьте эту гипотезу, используя результат оценивания регрессии, представленной в уравнении (3.45).

3.8. Доверительные интервалы

До сих пор мы предполагали, что гипотеза предшествует эмпирическим исследованиям. Однако это необязательно. Очень часто гипотеза и эксперимент взаимодействуют, в этом отношении типичным примером является регрессия расходов на питание. Вначале мы оцениваем регрессию, потому что в соответствии с экономической теорией ожидаем, что размер дохода влияет на уровень расходов на питание. Результат оценивания регрессии подтвердил это интуитивное ожидание в том смысле, что мы отвергли нулевую гипотезу $\beta = 0$. Но после этого возникло ощущение некоторой пустоты, поскольку на основе этой гипотезы нельзя выдвинуть предположения о том, что значение β равняется некоторому конкретному числу. Теперь, однако, мы можем двинуться в противоположном направлении и задаться вопросом о том, какие гипотезы совместимы с результатом оценивания регрессии.

Вполне очевидно, что гипотеза о том, что $\beta = 0{,}093$, будет совместимой, так как гипотеза и результаты эксперимента совпадают. Кроме того, совместимыми будут и гипотезы о том, что $\beta = 0{,}09229$ и $\beta = 0{,}09301$, так как разница между гипотезой и результатом эксперимента будет небольшой. Вопрос состоит в том, насколько сильно гипотетическое значение может отличаться от результата эксперимента, прежде чем они станут несовместимыми и мы должны будем отклонить нулевую гипотезу.

Можно ответить на этот вопрос, используя предшествующие рассуждения. Из уравнения (3.44) видно, что коэффициент регрессии b и гипотетическое значение β будут несовместимыми, если выполняются условия:

$$\frac{b-\beta}{\text{с.о.}(b)} > t_{крит} \quad \text{или} \quad \frac{b-\beta}{\text{с.о.}(b)} > -t_{крит}, \tag{3.49}$$

т. е. если

$$b-\beta > \text{с.о.}(b) \times t_{крит} \quad \text{или} \quad b-\beta < -\text{с.о.}(b) \times t_{крит}, \tag{3.50}$$

что соответствует

$$b - \text{с.о.}(b) \times t_{крит} > \beta \quad \text{или} \quad b + \text{с.о.}(b) \times t_{крит} < \beta. \tag{3.51}$$

Отсюда следует, что гипотетическое значение β является совместимым с результатом оценивания регрессии, если одновременно выполнены условия:

$$b - \text{с.о.}(b) \times t_{крит} < \beta \quad \text{и} \quad \beta < b + \text{с.о.}(b) \times t_{крит}, \tag{3.52}$$

т. е. если величина β удовлетворяет двойному неравенству:

$$b - \text{с.о.}(b) \times t_{крит} < \beta < b + \text{с.о.}(b) \times t_{крит}. \tag{3.53}$$

Любое гипотетическое значение β, которое удовлетворяет соотношению (3.53), будет автоматически совместимо с оценкой b, иными словами, не будет опровергаться ею. Множество всех этих значений, определенных как интервал между нижней и верхней границами неравенства, известно как *доверительный интервал* для величины β.

Отметим, что посредине доверительного интервала лежит сама величина b. Границы интервала одинаково отстоят от b. Отметим также, что, так как значение $t_{крит}$ зависит от выбора уровня значимости, границы будут также зависеть от этого выбора. Если принимается 5-процентный уровень значимости, то соответствующим доверительным интервалом считается 95-процентный интервал. Если выбирается однопроцентный уровень, то получают доверительный интервал в 99% и т. д.

Так как $t_{крит}$ будет больше для однопроцентного уровня, чем для 5-процентного (при любом данном числе степеней свободы), то, следовательно, интервал в 99% будет шире интервала в 95%. Так как посредине обоих интервалов лежит величина b, то интервал в 99% включает все гипотетические значения β в 95-процентном доверительном интервале, а также дополнительные промежутки с той и другой стороны.

Пример

При оценивании регрессии между расходами на питание и доходом величина b составила 0,093, ее стандартная ошибка 0,003, а $t_{крит}$ при 5-процентном уровне значимости 2,069. Отсюда соответствующий 95-процентный доверительный интервал составляет:

$$0{,}093 - 0{,}003 \times 2{,}069 < \beta < 0{,}093 + 0{,}003 \times 2{,}069, \tag{3.54}$$

иными словами,

$$0{,}087 < \beta < 0{,}099. \tag{3.55}$$

Поэтому мы отвергаем гипотетические значения только свыше 0,099 и ниже 0,087. Любые гипотезы, не выходящие за рамки этих пределов, не будут опровергаться полученным результатом оценивания регрессии.

Упражнения

3.15. Вычислите 99-процентный доверительный интервал для β в предыдущем упражнении и объясните, почему некоторые значения не были включены в 95-процентный доверительный интервал.

3.16. Вычислите 95-процентный доверительный интервал для β в примере с зависимостью между общей инфляцией и инфляцией, вызванной ростом заработной платы. Какой вывод вы можете сделать из этого вычисления?

3.17. Вычислите 95- и 99-процентные доверительные интервалы для коэффициента наклона в уравнении регрессии между расходами на коммунальные услуги и располагаемым личным доходом, представленной в упражнении 3.12.

3.18. Вычислите 95- и 99-процентные доверительные интервалы для коэффициента наклона в регрессии, оцененной в упражнении 2.4.

Вторая интерпретация доверительного интервала

Когда вы строите доверительный интервал, числа, которые вы определяете в качестве его верхнего и нижнего пределов, содержат случайные составляющие, которые зависят от значений случайного члена в наблюдениях выборки. Например, неравенство (3.53) включает верхний предел:

$$b + \text{с. о. } (b) \times t_{крит}.$$

Как b, так и с. о. (b) частично определяются значениями случайного члена, то же происходит и с нижним пределом. Мы надеемся, что доверительный интервал будет включать истинное значение параметра, но иногда он будет настолько искажен случайными факторами, что это будет не так.

Какова же вероятность того, что доверительный интервал будет включать истинное значение параметра? Легко показать, используя элементарную теорию вероятностей, что в случае 95-процентного доверительного интервала данная вероятность составит 95%, если модель правильно определена и условия Гаусса—Маркова выполняются. Аналогичным образом в случае 99-процентного доверительного интервала данная вероятность будет 99%.

Оцененный коэффициент [например, b в неравенстве (3.53)] обеспечивает точечную оценку рассматриваемого параметра, но при этом вероятность того, что истинное значение будет в точности равно этой оценке, ничтожно мала. Доверительный интервал дает так называемую «интервальную оценку» параметра, т. е. диапазон значений, который будет включать истинное значение с высокой, заранее определенной вероятностью — 95% в случае 95-процентного и 99% в случае 99-процентного доверительного интервала. Именно эта интерпретация и дает название доверительному интервалу (более подробно этот вопрос рассматривается в работе Т. Уоннакотта и Р. Уоннакотта [Wonnacott, Wonnacott, 1985, глава 8]).

3.9. Односторонние *t*-тесты

Рассмотрение t-тестов мы начали с нулевой гипотезы $H_0: \beta = \beta_0$ и провели проверку возможности ее отклонения при коэффициенте регрессии, равном b. Если бы мы отклонили эту гипотезу, то косвенно приняли бы альтернативную гипотезу $H_1: \beta \neq \beta_0$.

До сих пор альтернативная гипотеза была лишь простым отрицанием нулевой гипотезы. Если, однако, можно сформулировать альтернативную гипотезу более конкретно, то следует и усовершенствовать процедуру проверки. Проведем исследование трех случаев: первый случай — весьма частный, когда

существует единственное альтернативное истинное значение β, которое мы обозначим β_1; второй случай — если β не равно β_0, то оно должно быть больше β_0; и третий случай — если величина β не равна β_0, то она должна быть меньше β_0.

$$H_0: \beta = \beta_0, \quad H_1: \beta = \beta_1$$

В этом случае по каким-то причинам существуют только два возможных истинных значения коэффициента при x — β_0 и β_1. Для определенности допустим, что β_1 больше, чем β_0. Предположим, что мы хотим проверить гипотезу H_0 при 5-процентном уровне значимости и используем для этого обычную процедуру, которая уже рассматривалась в этой главе. Мы находим границы для верхнего и нижнего 2,5-процентных «хвостов» t-распределения, считая, что H_0 верна, и обозначим их как A и B на рис. 3.8. Гипотеза H_0 отклоняется, если коэффициент регрессии b оказывается правее точки B или левее точки A.

Далее, если значение b находится справа от B, то оно намного лучше совместимо с гипотезой H_1, чем с гипотезой H_0; вероятность его нахождения справа от B, если истинна гипотеза H_1, намного больше, чем при истинности гипотезы H_0. Здесь у нас не должно быть сомнений в том, чтобы отклонить гипотезу H_0 и принять гипотезу H_1.

Если, однако, b находится слева от A, то используемая процедура проверки приведет нас к неверному заключению. Последняя требует отклонить гипотезу H_0 и, следовательно, принять гипотезу H_1, несмотря на то что при истинности гипотезы H_1 вероятность нахождения b слева от A ничтожно мала. Мы даже не построили кривой функции плотности вероятности, соответствующей гипотезе H_1. Если такое значение b получается только один раз на миллион случаев при истинности гипотезы H_1, но в 2,5% случаев при истинности гипотезы H_0, то здесь намного логичнее считать, что истинной является гипотеза H_0. Конечно, в одном случае из миллиона вы сделаете ошибку, но в остальных случаях вы будете правы.

Следовательно, мы отклоним гипотезу H_0, только если b оказывается в верхнем 2,5-процентном «хвосте» распределения, т. е. справа от B. Это означает, что теперь мы выполняем проверку гипотезы с односторонним критерием, сократив в результате вероятность допущения ошибки I рода до 2,5%. Поскольку уровень значимости определен как вероятность допущения ошибки I рода, то он теперь также составляет 2,5%.

Как уже отмечалось, экономисты обычно предпочитают проверку гипотез с пяти- и однопроцентным уровнями значимости проверкам с 2,5-процентным уровнем. Если вы хотите провести проверку с 5-процентным уровнем значимости, то вам следует переместить точку B влево так, чтобы получить 5% вероятности в «хвосте» распределения и увеличить вероятность допущения ошибки I рода до 5%. (*Вопрос.* Почему намеренно выбирается увеличение вероятности допущения ошибки I рода? *Ответ.* Потому что одновременно сокращается вероятность допущения ошибки II рода, т. е. вероятность того, что нулевая гипотеза не будет отклонена, когда она является ложной.)

Если стандартное отклонение величины b известно (что практически маловероятно), а распределение нормально, то точка B будет находиться в Z стандартных отклонениях вправо от β_0, где Z определяется из соотношения $A(Z) =$

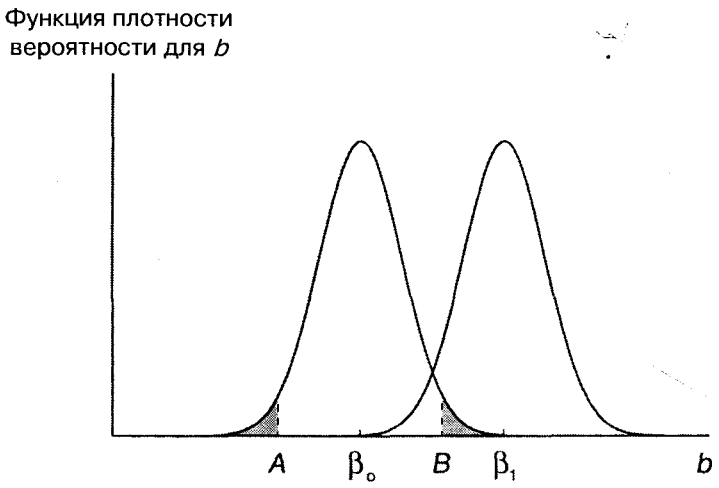


Рис. 3.8. Распределение величины b в соответствии с гипотезами H_0 и H_1

= 0,9500 по табл. А.1. Соответствующее значение для Z равно 1,64. Если стандартное отклонение неизвестно и оценивается как стандартная ошибка величины b, то мы должны использовать t-распределение. Можно найти критическое значение t по табл. А.2 для соответствующего числа степеней свободы в колонке, относящейся к 5%.

Аналогично если вы хотите выполнить проверку с однопроцентным уровнем значимости, то вы перемещаете B вправо до той точки, где «хвост» распределения содержит 1% вероятности. Если вам пришлось вычислить стандартную ошибку величины b на основе выборочных данных, то нужно найти критическое значение t в колонке, соответствующей 1%.

В проведенном анализе мы допустили, что β_1 больше, чем β_0. Очевидно, что если оно будет меньше β_0, то можно использовать ту же самую логику для проведения односторонней проверки, выбрав левый «хвост» распределения в качестве критической области гипотезы H_0.

Мощность критерия

В данном конкретном случае мы можем вычислить вероятность допущения ошибки II рода, т. е. принятия ложной гипотезы. Предположим, что мы приняли ложную гипотезу $H_0: \beta = \beta_0$ и что на самом деле истинна альтернативная гипотеза $H_1: \beta = \beta_1$. Если вернуться к рис. 3.8, то мы примем гипотезу H_0, если коэффициент регрессии выборки b оказывается слева от точки B. Так как гипотеза H_1 истинна, то вероятность того, что b будет отстоять слева от B, описывается областью, которая находится слева от B под кривой функции плотности вероятности, соответствующей гипотезе H_1. Что необходимо сделать — это вычислить t-статистику для точки B, считая, что $\beta = \beta_1$, и использовать таблицу t-распределения для нахождения вероятности того, что b больше, чем на t стандартных ошибок, будет отстоять слева от β_1.

Если эту вероятность обозначить γ, то мощность критерия, определенная как

вероятность недопущения ошибки II рода, составляет $(1-\gamma)$. Очевидно, необходим компромисс между мощностью критерия и уровнем значимости. Чем выше уровень значимости, тем дальше вправо будет сдвинута точка B, тем больше будет γ и тем меньше — мощность критерия.

Используя односторонний критерий вместо двустороннего, можно получить большую мощность при любом уровне значимости. Как мы уже видели, при переходе к одностороннему критерию с 5-процентным уровнем значимости точка B на рис. 3.8 сдвигается влево, и тем самым сокращается вероятность принятия ложной гипотезы H_0. Нужно, однако, помнить, что выигрыш в мощности будет получен только в условиях, когда использование одностороннего критерия оправдано.

$$H_0: \beta = \beta_0, \quad H_1: \beta > \beta_0$$

Мы рассмотрели случай, когда альтернативная гипотеза H_1 включала конкретное гипотетическое значение β_1 (при $\beta_1 > \beta_0$). Вполне понятно, что логика, которая привела нас к использованию одностороннего критерия, применима и в более общем случае, когда гипотеза H_1 выражается в виде $\beta_1 > \beta_0$ без указания какого-либо конкретного значения.

Мы по-прежнему хотели бы исключить левый «хвост» распределения из критической области гипотезы, так как низкое значение для b более вероятно получить при гипотезе $H_0: \beta = \beta_0$, чем при гипотезе $H_1: \beta > \beta_0$, а следовательно, это будет говорить в пользу гипотезы H_0, а не против нее. Поэтому мы вновь предпочтем односторонний критерий проверки гипотезы двустороннему, рассматривая правый «хвост» распределения как область непринятия гипотезы. Отметим, что так как β_1 не определено, у нас теперь нет возможности вычислить мощность такого критерия.

$$H_0: \beta = \beta_0, \quad H_1: \beta < \beta_0$$

Аналогичным образом, если альтернативная гипотеза представлена в виде $H_1: \beta < \beta_0$, мы предпочтем проверку, основанную на одностороннем критерии, использующем левый «хвост» распределения в качестве области отклонения гипотезы.

Проверки с использованием одностороннего критерия очень важны на практике при решении экономических задач. Как мы уже видели, обычно для установления того, что независимая переменная действительно оказывает влияние на зависимую переменную, формулируется нулевая гипотеза $H_0: \beta = 0$, которую затем пытаются опровергнуть. Очень часто гипотеза оказывается достаточно обща, чтобы утверждать, что если x оказывает влияние на y, то это влияние имеет определенную направленность.

Если у нас есть достаточно веские причины считать, что это влияние, скажем, положительно, то можно использовать альтернативную гипотезу $H_1: \beta > 0$ вместо более общей гипотезы $H_1: \beta \neq 0$. Это является преимуществом, поскольку критическое значение t для отклонения гипотезы H_0 при проверке по одностороннему критерию будет меньшим, что облегчает отклонение нулевой гипотезы и установление наличия зависимости.

Примеры

При оценивании регрессии расходов на продукты питания мы имели 23 степени свободы. При использовании однопроцентного уровня значимости и двустороннего критерия проверки гипотезы критическое значение $t = 2{,}807$. Целесообразно предположить, что размер дохода имеет положительное влияние на уровень расходов на питание. Но тогда для проверки гипотезы можно воспользоваться преимуществами одностороннего критерия, для которого критическое значение имеет меньшую величину, равную 2,500. Причем t-статистика, вычисленная по выборке, равнялась 31,0; поэтому в данном случае улучшение не имеет значения. Оценка коэффициента настолько велика по отношению к стандартной ошибке, что мы отклоняем нулевую гипотезу независимо от того, на каком критерии основывается проверка гипотезы — двустороннем или одностороннем.

В примере зависимости между общей инфляцией и инфляцией, вызванной ростом заработной платы, возможность использования для проверки одностороннего критерия имеет смысл. Нулевая гипотеза состояла в том, что инфляция, вызванная ростом заработной платы, полностью отражена в общей инфляции, и мы имеем $H_0: \beta = 1{,}0$. Было бы целесообразно принять альтернативную гипотезу о том, что $\beta < 1$ из-за повышения производительности труда, которое может привести к более низкому уровню общей инфляции по сравнению с инфляцией, вызванной ростом заработной платы, т. е. $H_1: \beta < 1{,}0$. В результате расчетов получим коэффициент регрессии, равный 0,82, и стандартную ошибку 0,10; тогда t-статистика для нулевой гипотезы составит $-1{,}80$. Это значение не столь велико, чтобы отвергнуть гипотезу H_0 при 5-процентном уровне значимости и использовании двустороннего критерия (критическое значение составляет 2,10). Однако если мы используем односторонний критерий проверки, то критическое значение уменьшится до 1,73, и теперь мы *можем* отклонить нулевую гипотезу. Другими словами, можно сделать вывод о том, что общая инфляция будет значимо ниже инфляции, вызванной ростом заработной платы.

Упражнения

3.19. С помощью одностороннего критерия выполните проверку значимости коэффициента наклона модели регрессии между расходами на коммунальные услуги и располагаемым личным доходом, представленной в упражнении 3.12. Обоснуйте целесообразность использования одностороннего критерия.

3.20. Используя результаты упражнений 2.4 и 2.5, проверьте значимость коэффициентов регрессии. Примените односторонний t-критерий в тех случаях, когда это необходимо. Обоснуйте целесообразность использования одностороннего критерия.

3.10. *F*-тест на качество оценивания

Даже если между y и x отсутствует зависимость, по любой данной выборке наблюдений может показаться, что такая зависимость существует, возможно и слабая. Только по случайному стечению обстоятельств выборочная ковариация будет *в точности* равна нулю. Следовательно, только чисто случайно коэффициент корреляции и коэффициент R^2 будут *в точности* равны нулю.

Это представляет для нас проблему. Как узнать, действительно ли полученное при оценке регрессии значение коэффициента R^2 отражает истинную зависимость или оно появилось случайно?

В принципе можно было бы принять следующую процедуру. Сформулируем в качестве нулевой гипотезы утверждение, что связь между y и x отсутствует, и найдем значение коэффициента, которое может быть превышено в 5% случаев. Затем используем эту цифру в качестве критического значения для проверки гипотезы при 5-процентном уровне значимости. Если этот уровень превышается, то мы отклоняем нулевую гипотезу. Если он не превышен, то эта гипотеза принимается.

Такая проверка, подобно t-тесту для коэффициента регрессии, не служит доказательством. Действительно, при 5-процентном уровне значимости имеется риск допущения ошибки I рода (отклонения нулевой гипотезы, когда она истинна) в 5% случаев, но можно, конечно, снизить этот риск за счет использования более высокого уровня значимости, например в 1%. Тогда критическое значение может быть случайно превышено только в 1% случаев, поэтому оно выше критического значения для проверки гипотезы при 5-процентном уровне значимости.

Каким образом можно определить критическое значение коэффициента R^2 при любом уровне значимости? Здесь возникает небольшая проблема. У нас нет таблицы критических значений коэффициента R^2. Традиционная процедура состоит в использовании косвенного подхода и выполнения так называемого F-теста, основанного на анализе дисперсии (теория, лежащая в основе этого подхода, описывается в работе А. Муда и Ф. Грейбилла [Mood, Graybill, 1963]).

Предположим, что, как и прежде, можно разложить дисперсию зависимой переменной на «объясненную» и «необъясненную» составляющие, воспользовавшись уравнением (2.45):

$$\text{Var}\,(y) = \text{Var}\,(\hat{y}) + \text{Var}\,(e). \tag{3.56}$$

Используя определение выборочной дисперсии и умножив на n обе части уравнения (3.56), можно представить его следующим образом:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum e^2. \tag{3.57}$$

(Напомним, что $\bar{e} = 0$ и выборочное среднее значение $\hat{y}$ равняется выборочному среднему y.)

Левая часть уравнения представляет собой общую сумму квадратов отклонений (*TSS*) зависимой переменной от ее выборочного среднего значения. Первый член в правой части уравнения является объясненной суммой квадратов

(*ESS*), а второй член — необъясненной суммой квадратов отклонений (*RSS*), который может быть просто назван *S*:

$$TSS = ESS + RSS. \tag{3.58}$$

F-статистика для проверки качества оценивания регрессии записывается как отношение объясненной суммы квадратов (в расчете на одну независимую переменную к остаточной сумме квадратов) в расчете на одну степень свободы:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}}, \tag{3.59}$$

где k — число независимых переменных.

После деления на *TSS* числителя и знаменателя соотношения (3.59) *F*-статистика может быть эквивалентно выражена на основе коэффициента R^2:

$$F = \frac{(ESS / TSS) / k}{(RSS / TSS) / (n-k-1)} = \frac{R^2 / k}{(1-R^2) / (n-k-1)}. \tag{3.60}$$

В данном контексте $k = 1$ и, таким образом, уравнение (3.60) принимает вид:

$$F = \frac{R^2}{(1-R^2) / (n-2)}. \tag{3.61}$$

После вычисления критерия F по значению коэффициента R^2 вы отыскиваете величину $F_{крит}$ — критическое значение F в соответствующей таблице. Если $F > F_{крит}$, то вы отклоняете нулевую гипотезу и делаете вывод о том, что имеющееся «объяснение» поведения величины y лучше, чем можно было бы получить чисто случайно.

В табл. А.3 представлены критические значения F при уровнях значимости в 5 и 1%. В каждом случае критическое значение зависит от числа независимых переменных k, которое находится в верхней строке таблицы, и от числа степеней свободы $(n - k - 1)$, которое включено в ее крайний левый столбец. В данном контексте рассматривается случай парной регрессии, когда $k = 1$, и мы должны использовать первую колонку таблицы.

В примере с расходами на питание коэффициент R^2 составил 0,9775. Поскольку было 25 наблюдений, *F*-статистика равняется:

$$R^2 / \{(1 - R^2) / 23\} = 0{,}9775 / (0{,}0225 / 23) = 999{,}2.$$

При однопроцентном уровне значимости критическое значение критерия F (первая колонка, ряд 23) составляет 7,88. Поэтому в данном конкретном примере у нас не остается никаких сомнений относительно того, что нулевую гипотезу следует отклонить. Другими словами, полученное значение коэффициента R^2 столь высоко, что мы отклоняем предположение о том, что оно могло появиться случайно. На практике *F*-статистика всегда вычисляется вместе с коэффициентом R^2, поэтому нет необходимости использовать уравнение (3.60).

Какие же проблемы возникают при использовании этого косвенного подхода? Почему бы не иметь таблицу критических значений коэффициента R^2? Ответ заключается в том, что таблица значений критерия F является полезной для многих способов проверки дисперсии, одним из которых выступает расчет коэффициента R^2. Вместо специализированной таблицы для каждого конкретного случая намного удобнее (или, по меньшей мере, экономнее) иметь одну обобщенную таблицу, делая при необходимости преобразования типа (3.60).

Конечно, при необходимости можно вывести и критические значения R^2. Критическое значение R^2 связано с критическим значением F следующим уравнением:

$$F_{крит} = \frac{R^2_{крит} / k}{(1 - R^2_{крит}) / (n - k - 1)}, \tag{3.62}$$

из которого следует, что

$$R^2_{крит} = \frac{kF_{крит}}{kF_{крит} + (n - k - 1)}. \tag{3.63}$$

В примере с расходами на питание критическое значение F при уровне значимости в 1% составило 7,88. Следовательно, в этом случае при $k = 1$

$$R^2_{крит} = \frac{7{,}88}{30{,}88} = 0{,}26. \tag{3.64}$$

В нашем примере величина R^2 намного выше 0,26, поэтому непосредственное сравнение величины R^2 с его критическим значением подтверждает вывод о том, что в результате F-теста мы должны отклонить нулевую гипотезу.

Упражнения

3.21. В упражнении 3.12 значение коэффициента R^2 в модели регрессии между расходами на коммунальные услуги и располагаемым личным доходом составило (с точностью до четырех десятичных разрядов) 0,9875. Вычислите соответствующую F-статистику и проверьте, что она равна 1814,7, т. е. результату, выданному компьютером. Выполните F-тест при уровнях значимости в 5 и 1%. Есть ли необходимость представлять результаты проверки на обоих уровнях?

3.22. Аналогичным образом, используя результат упражнения 2.4, вычислите F-статистику на основе значения коэффициента R^2 и проверьте, что она не противоречит расчетам, выполненным на компьютере. Проведите соответствующий F-тест.

3.11. Взаимосвязи между критериями в парном регрессионном анализе

Теперь мы выведем некоторые зависимости между критерием F для коэффициента R^2, критерием t для коэффициента при x и критерием t для коэф-

фициента корреляции между x и y в парной регрессионной модели. Начнем с последнего из них.

t-тест для коэффициента корреляции

В разделе 1.8 мы определили выборочный коэффициент корреляции для двух переменных x и y:

$$r_{x,y} = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}. \tag{3.65}$$

Даже если переменные x и y вообще не коррелированы и теоретический коэффициент корреляции ρ равен нулю, вы будете связаны известным ограничением и неизбежно получите в расчетах *некоторую* величину выборочного коэффициента корреляции. Для конкретной выборки $r_{x,y}$ может точно равняться нулю только чисто случайно, и можно ли утверждать, что значение $r_{x,y}$ действительно указывает на наличие зависимости, или же оно появилось случайно?

Как обычно, ответом будет формулирование нулевой гипотезы о том, что зависимости нет, а затем — попытка ее опровергнуть. Для проверки гипотетической линейной зависимости между x и y, т. е. единственного типа зависимости, который будет рассматриваться в данной книге, справедлива следующая процедура.

Первый шаг состоит в вычислении t-статистики для r:

$$t = r\sqrt{\frac{n-2}{1-r^2}}. \tag{3.66}$$

Выбрав уровень значимости, скажем, в 5%, вы находите критическое значение t с $(n-2)$ степенями свободы. Если величина t превышает его критическое значение (в положительную или отрицательную сторону), вы отклоняете нулевую гипотезу о том, что $\rho = 0$, и заключаете, что нашли линейную зависимость (положительную или отрицательную).

Доказательство этого вам уже знакомо. Если нулевая гипотеза верна, то величина t будет превышать его критическое значение (в положительную или отрицательную сторону) только в 5% случаев. Это означает, что при выполнении проверки вероятность допущения ошибки I рода, отклоняющей нулевую гипотезу, когда она фактически верна, составляет 5%.

Возможно, что риск допущения такой ошибки в 5% случаев слишком велик для вас. Тогда вы можете сократить степень риска, осуществляя расчеты при уровне значимости в 1%. Критическое значение t теперь будет выше, чем до сих пор, поэтому вам потребуется более высокая (положительная или отрицательная) t-статистика для отклонения нулевой гипотезы, а это означает, что вам потребуется более высокое значение r.

Обоснованность этого t-теста зависит от того, удовлетворяют ли y и x некоторым условиям. Достаточно будет, чтобы величина y была связана с величиной x с помощью модели парной регрессии рассматриваемого типа. Данный тест

справедлив только для нулевой гипотезы об отсутствии зависимости. Если вы хотите проверить гипотезу о том, что теоретический коэффициент корреляции вместо нуля равен некоторому другому значению, то должны будете использовать более сложную процедуру.

Зависимость между критериями

Мы увидим, что в случае парного регрессионного анализа (и только парного регрессионного анализа) t-критерий для гипотезы $\rho_{x,y} = 0$, F-критерий для коэффициента R^2 и t-критерий для гипотезы $\beta = 0$ эквивалентны друг другу. Мы начнем с определения зависимости между первыми двумя тестами.

В разделе 2.7 было показано, что коэффициент R^2 может интерпретироваться как квадрат коэффициента корреляции между $\hat{y}$ и y, то есть $r^2_{\hat{y},y}$. Теперь, в случае парной регрессии, $\hat{y}$ является линейной функцией x, поэтому коэффициент корреляции между $\hat{y}$ и y должен совпадать с коэффициентом корреляции между x и y, то есть с $r_{x,y}$ (см. упражнение 1.5). Следовательно, в парном регрессионном анализе (и только в парном регрессионном анализе) R^2 должен быть равен квадрату коэффициента корреляции между x и y. Мы докажем это, непосредственно начиная с уравнения (2.46), т. е. с определения коэффициента R^2:

$$R^2 = \frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} = \frac{b^2\mathrm{Var}(x)}{\mathrm{Var}(y)}, \tag{3.67}$$

поскольку $\mathrm{Var}(\hat{y}) = \mathrm{Var}(a + bx) = b^2\mathrm{Var}(x)$. Делая замену для b, получим:

$$R^2 = \left\{\frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}\right\}^2 \frac{\mathrm{Var}(x)}{\mathrm{Var}(y)} = \frac{\mathrm{Cov}(x,y)^2}{\mathrm{Var}(x)\mathrm{Var}(y)}. \tag{3.68}$$

Из уравнений (3.61) и (3.66) можно видеть, что F-статистика для коэффициента R^2 является в точности квадратом t-статистики для $r_{x,y}$. Как и следовало ожидать, критическое значение F будет равно квадрату критического значения t-статистики при любом уровне значимости, и эти два теста всегда дают один и тот же результат. Другими словами, коэффициент корреляции между x и y будет указывать на значимую зависимость, если и только если уровень R^2 в регрессии между y и x будет говорить о такой зависимости.

Более того, можно показать, что величина b будет значимо отличаться от нуля при использовании t-теста, если и только если F-тест значим при данном уровне значимости. Используя тот факт, что $\mathrm{Var}(\hat{y})$ равняется $b^2\mathrm{Var}(x)$, и оба определения коэффициента R^2 в уравнениях (2.46) и (2.47), мы можем переписать выражение для стандартной ошибки величины b:

$$\text{с.о.}(b) = \sqrt{\frac{\mathrm{Var}(e)}{(n-2)\mathrm{Var}(x)}} = b\sqrt{\frac{\mathrm{Var}(e)}{(n-2)\mathrm{Var}(\hat{y})}} = b\sqrt{\frac{1-R^2}{(n-2)R^2}} = b\sqrt{\frac{1-r^2_{x,y}}{(n-2)r^2_{x,y}}}. \tag{3.69}$$

Следовательно,

$$\frac{b}{\text{с.о.}(b)} = r_{x,y}\sqrt{\frac{n-2}{1-r_{x,y}^2}}, \qquad (3.70)$$

и мы показали, что t-статистика для проверки гипотезы $\beta = 0$ такая же, как и t-статистика для проверки гипотезы $\rho_{x,y} = 0$.

Таким образом, при наличии только одной независимой переменной t-критерий для гипотезы $\beta = 0$, t-критерий для гипотезы $\rho_{x,y} = 0$ и F-критерий для коэффициента R^2 эквивалентны. Как мы увидим далее, при использовании более одной независимой переменной это утверждение перестает быть справедливым.

Упражнения

3.23. Проверьте, что F-статистика в регрессии, оцененной вами в упражнении 2.4, равняется квадрату t-статистики для коэффициента b и что критическое значение F при уровне значимости в 1% равняется квадрату критического значения t.

3.24. В упражнении 2.6 оба исследователя получили для своих регрессий значения коэффициента R^2, равные 0,79. Было ли это совпадением?

4

ПРЕОБРАЗОВАНИЯ ПЕРЕМЕННЫХ

Многие экономические процессы наилучшим образом описываются нелинейными соотношениями, например нелинейными функциями спроса и производственными функциями. В данной главе мы сначала определим, что понимается под линейным регрессионным анализом, а затем покажем, как он может быть применен для некоторых явно нелинейных соотношений. Далее мы рассмотрим, что необходимо делать в тех случаях, когда невозможно использовать линейные методы. В заключение рассматривается метод, позволяющий установить статистическое различие между линейными и нелинейными соотношениями.

4.1. Базисная процедура

Одним из недостатков линейного регрессионного анализа, как это следует из самого названия, является то, что он может быть применен только к линейным уравнениям. В случае простого регрессионного анализа речь идет об уравнениях вида

$$y = \alpha + \beta_1 x_1, \tag{4.1}$$

состоящих из постоянной величины (которая может и отсутствовать), независимой переменной, умноженной на некоторый коэффициент, и из случайного остаточного члена возмущения, которым мы можем временно пренебречь. В общем случае линейное уравнение выглядит так, что каждый объясняющий элемент, за исключением постоянной величины, записан в виде произведения переменной и коэффициента:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \tag{4.2}$$

Уравнения вида

$$y = \alpha + \frac{\beta}{x} \tag{4.3}$$

и

$$y = \alpha x^{\beta} \tag{4.4}$$

являются нелинейными. Выбрав значения для α и β и построив графики, мы обнаружим, что оба они представлены кривыми.

Зависимости (4.3) и (4.4) считаются приемлемыми для описания кривых Энгеля, характеризующих соотношение между спросом на определенный то-

вар (y) и общей суммой дохода (x). Как можно определить параметры α и β в каждом уравнении, зная значения y и x?

В конечном счете в обоих случаях можно применить линейный регрессионный анализ, для этого потребуется лишь небольшая подготовка. Во-первых, заметим, что уравнения (4.1) и (4.2) являются линейными в двух смыслах. Правая часть линейна по переменным, если определить их в представленном виде, а не как функции. Следовательно, она состоит из взвешенной суммы переменных, а параметры являются весами. Например, в уравнении (4.1) имеется просто x_1, а не $\log(x_1)$. Правая часть также линейна по параметрам, так как она состоит из взвешенной суммы параметров, а переменные x в данном случае являются весами.

Для целей линейного регрессионного анализа важное значение имеет только второй тип линейности. Нелинейность по переменным всегда можно обойти путем использования соответствующих определений. Например, предположим, что соотношение имеет вид:

$$y = \alpha + \beta_1 x_1^2 + \beta_2 \sqrt{x_2} + \ldots \tag{4.5}$$

Если определить $z_1 = x_1^2$, $z_2 = \sqrt{x_2}$ и т. д., то соотношение примет следующий вид:

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \ldots, \tag{4.6}$$

и теперь оно является линейным как по переменным, так и по параметрам. Такой тип преобразований является лишь косметическим, и обычно уравнения регрессии записываются с нелинейными выражениями относительно переменных. Это позволяет избежать лишних обозначений.

С другой стороны, уравнение типа (4.4) является нелинейным как по параметрам, так и по переменным, и его нельзя преобразовать только путем замены определений. (Не следует думать, что его можно преобразовать в линейное, если определить $z = x^\beta$ и подставить x^β вместо z; поскольку β неизвестно, мы не сможем рассчитать выборочное значение z.) Проблема преобразования нелинейных по параметрам соотношений будет рассмотрена в следующем разделе.

В случае (4.3), однако, единственное, что нам нужно сделать, — это определить $z = (1/x)$. Тогда уравнение (4.3) примет вид:

$$y = \alpha + \beta z, \tag{4.7}$$

и оно будет линейным, в этом случае мы без всяких проблем оценим регрессию между y и z. Постоянный член в уравнении регрессии будет представлять собой оценку α, а коэффициент при z — оценку β.

Пример

Допустим, вы исследуете соотношение между ежегодным потреблением бананов и годовым доходом, и наблюдения приведены в табл. 4.1, где собраны наблюдения для 10 семей (забудем пока о z).

На рис. 4.1 представлено облако точек, соответствующих наблюдениям, а также график уравнения регрессии между y и x:

$$\hat{y} = 5{,}09 + 0{,}73x; \quad R^2 = 0{,}64. \tag{4.8}$$
$$\text{(с.о.)}\ (1{,}23)\quad (0{,}20)$$

Из рис. 4.1 видно, что график уравнения регрессии не вполне соответствует точкам наблюдений, несмотря на то что коэффициент при x существенно отличается от нуля при однопроцентном уровне значимости. Очевидно, что точки наблюдений лежат на кривой, тогда как уравнение регрессии характеризуется прямой. В данном случае нетрудно заметить, что функциональная зависимость между y и x определена неправильно. В том случае, если вы не можете представить зависимость в графическом виде (например, если вы используете множественный регрессионный анализ), понять, что где-то допущена ошибка, можно с помощью анализа остатков. В данном случае значения остатков приведены в табл. 4.2.

Таблица 4.1

Семья	*Бананы (в фунтах) (y)*	*Доход (в 10000 долл.) (x)*	*(z)*
1	1,93	1	1,000
2	7,13	2	0,500
3	8,78	3	0,333
4	9,69	4	0,250
5	10,09	5	0,200
6	10,42	6	0,167
7	10,62	7	0,143
8	10,71	8	0,125
9	10,79	9	0,111
10	11,13	10	0,100

Положительные или отрицательные, большие или малые остатки должны чередоваться случайным образом. Здесь же, как видно из таблицы, сначала остатки отрицательны, затем они становятся положительными, достигают максимума, а потом снова уменьшаются и становятся отрицательными: это представляется достаточно сомнительным.

В данном примере значения y и x были получены с помощью метода Монте-Карло, истинное соотношение имеет вид:

$$y = 12 - \frac{10}{x} + \text{Случайный член}, \tag{4.9}$$

x принимает целые значения от 1 до 10, а значения случайного члена получают

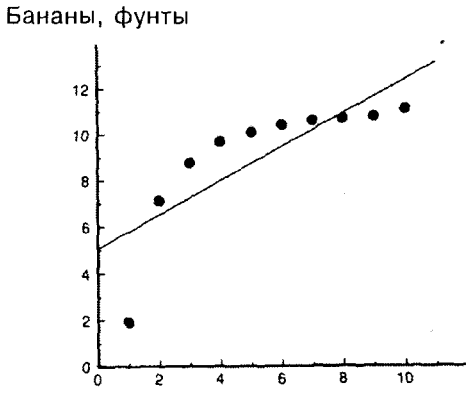


Рис. 4.1. Регрессионная зависимость расходов на бананы от годового дохода

с помощью нормально распределенных случайных чисел со средним значением 0 и среднеквадратичным отклонением 0,1.

Если мы знаем это и определим $z = 1/x$, то уравнение примет линейный вид (4.7). Значение z для каждой семьи уже подсчитано в табл. 4.1. Оценив регрессию между y и z, получим:

$$\hat{y} = 12{,}08 - 10{,}08z; \quad R^2 = 0{,}9989. \tag{4.10}$$
$$\text{(с.о.)} \quad (0{,}04) \quad (0{,}12)$$

Подставив $z = 1/x$, имеем:

$$\hat{y} = 12{,}08 - \frac{10{,}08}{x}. \tag{4.11}$$

Таблица 4.2

Семья	*y*	$\hat{y}$	*e*
1	1,93	5,82	–3,90
2	7,13	6,56	0,57
3	8,78	7,29	1,49
4	9,69	8,03	1,67
5	10,09	8,76	1,33
6	10,42	9,50	0,93
7	10,62	10,23	0,39
8	10,71	10,97	–0,26
9	10,79	11,70	–0,91
10	11,13	12,43	–1,31

С учетом высокого качества оцененного уравнения (4.10) неудивительно, что соотношение (4.11) близко к истинному уравнению (4.9). На рис. 4.2А и 4.2Б показаны регрессионная зависимость и точки наблюдений для y, x и z. Улучшение качества уравнения, измеряемого с помощью коэффициента R^2, отражено в более полном соответствии графиков. Сравните рис. 4.1 и 4.2Б.

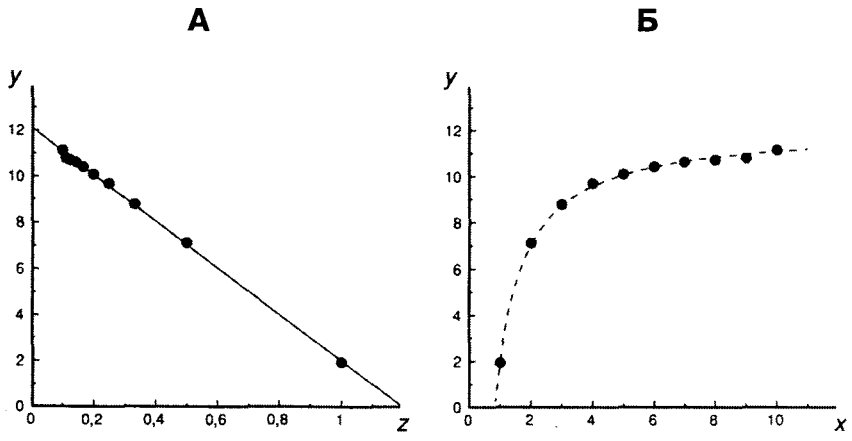


Рис. 4.2. *А* — регрессионная зависимость y от $1/x$; *Б* — расчетная линия для величин y и x

4.2. Логарифмические преобразования

Рассмотрим далее функции вида (4.4), которые являются нелинейными как по параметрам, так и по переменным:

$$y = \alpha x^{\beta}. \tag{4.4}$$

Мы обнаружим, что соотношение (4.4) может быть преобразовано в линейное уравнение путем использования *логарифмов*, безусловно, знакомых вам из курса математики. Возможно, при изучении этого курса вам казалось, что логарифмы имеют чисто академический интерес и неприменимы на практике. В эконометрике, однако, они просто необходимы, поэтому если вы не уверены в своих знаниях, то вам следует их освежить в памяти. Далее приведена таблица основных свойств логарифмов, которая вам поможет. Если вы не имеете достаточного опыта работы с логарифмами, не волнуйтесь, вы легко приобретете нужные навыки.

Применение логарифмов

Основные правила гласят:

1. Если $y = xz$, то $\log y = \log x + \log z$.

2. Если $y = x/z$, то $\log y = \log x - \log z$.

3. Если $y = x^n$, то $\log y = n \log x$.

Эти правила могут применяться вместе для преобразования более сложных выражений. Например, возьмем уравнение (4.4). Если $y = \alpha x^{\beta}$, то по правилу *1*:

$\log y = \log \alpha + \log x^{\beta}$ и по правилу 3

$= \log \alpha + \beta \log x$.

До сих пор мы не определили, по какому основанию берем логарифм — e или 10. В данной работе мы будем использовать в качестве основания число e, т. е. натуральные логарифмы. Теперь это считается стандартом в эконометрике. Некоторые обозначают натуральный логарифм с помощью символа ln, а не log, однако в этом уже нет необходимости. Никто больше не использует логарифмы по основанию 10. Таблицы десятичных логарифмов широко использовались для умножения или деления больших чисел до начала 1970-х гг. Однако с изобретением карманных калькуляторов они стали не нужны.

Для натуральных логарифмов справедливо еще одно правило:

4. Если $y = e^x$, то $\log y = x$.

Выражение e^x, которое часто записывается как $\exp(x)$, известно также как антилогарифм x.[1] Можно сказать, что $\log(e^x)$ является логарифмом антилогарифма x, и так как логарифм и антилогарифм взаимно уничтожаются, неудивительно, что $\log(e^x)$ превращается просто в x.

Используя приведенные выше правила, уравнение (4.4) можно преобразовать в линейное путем *логарифмирования* его обеих частей. Если соотношение (4.4) верно, то

$$\log y = \log \alpha x^{\beta} = \log \alpha + \beta \log x. \tag{4.12}$$

Если обозначить $y' = \log y$, $z = \log x$ и $\alpha' = \log \alpha$, то уравнение (4.12) можно переписать в следующем виде:

$$y' = \alpha' + \beta z. \tag{4.13}$$

Процедура оценивания регрессии теперь будет следующей. Сначала вычис-

[1] Операция перехода от x к $\exp(x)$ часто называется *потенцированием*. (*Прим. ред.*)

лим y' и z для каждого наблюдения путем взятия логарифмов от исходных значений. Вы можете сделать это на компьютере с помощью имеющейся статистической программы. Затем оценим регрессионную зависимость y' от z. Коэффициент при z будет представлять собой непосредственно оценку β. Постоянный член является оценкой α', т. е. $\log \alpha$. Для получения оценки α необходимо взять *антилогарифм*, т. е. вычислить $\exp(\alpha')$.

Моделирование эластичности

Функции вида (4.4) часто встречаются в экономике. Когда вы видите такую функцию, то можете сразу сказать, что эластичность y по x равна β. Например, в разделе 4.1 отмечалось, что это общая форма кривых Энгеля, y представляет собой спрос на товар, x — доход, а β — эластичность спроса по доходу.

Докажем указанное свойство *эластичности*. Независимо от математической связи между y и x или определения величин y и x, эластичность y по x рассчитывается как относительное изменение y на единицу относительного изменения x:

$$\text{Эластичность} = \frac{dy / y}{dx / x}.$$

Таким образом, если, например, y — это спрос, а x — доход, то данное выражение определяет эластичность спроса на данный товар по доходу.

Выражение для эластичности можно переписать в следующем виде: $(dy/dx)/(y/x)$. Для примера с функцией спроса его можно представить как отношение предельной склонности к потреблению товара к средней склонности к потреблению данного товара.

Если соотношение между y и x имеет вид (4.4), то

$$\text{Эластичность} = \frac{dy / dy}{y / x} = \frac{\beta y / x}{y / x} = \beta.$$

Таким образом, например, если имеется кривая Энгеля вида

$$y = 0{,}01x^{0{,}30},$$

то это означает, что эластичность спроса по доходу равна 0,3. Если вы хотите объяснить это кому-нибудь, кто не знаком с экономической терминологией, то наиболее просто будет сказать, что изменение x (дохода) на 1% вызывает изменение y (спроса) на 0,3%.

Функция вида (4.4) может также применяться к кривым спроса, где y — это спрос на товар, x — цена товара, а β — это эластичность спроса по цене. (На практике обычно такая функция спроса объединяется с кривой Энгеля, в результате чего получается зависимость спроса одновременно от дохода и цены. Мы вернемся к этому вопро-

су, когда будем рассматривать модели множественной регрессии в главе 5.)

Что произойдет, если математическая связь между y и x не соответствует уравнению (4.4)? Что можно в этом случае сказать об эластичности? Это можно понять на основе базовых принципов. Предположим, имеется обычное линейное уравнение:

$$y = \alpha + \beta x.$$

В данном случае dy/dx равно β; следовательно, эластичность определяется следующим образом:

$$\text{Эластичность} = \frac{dy/dx}{y/x} = \frac{\beta}{y/x} = \frac{\beta x}{y}.$$

В этом случае значение эластичности в любой точке будет зависеть не только от значения β, но также и от значений y и x в данной точке.

Таким образом, два основных достоинства математической формы (4.4) состоят в следующем:

1. Если эластичность y по x постоянна, то это единственная математическая форма, которая обладает данным свойством. Это, безусловно, означает, что если вы считаете, что эластичность *не* постоянна, то данное соотношение *не* следует моделировать с помощью уравнения (4.4).

2. Вы можете получить прямую регрессионную оценку эластичности путем оценивания зависимости $\log y$ от $\log x$. Эта оценка, конечно, будет достоверна только в том случае, если зависимость определяется уравнением (4.4). Если зависимость линейна, то правильная процедура будет состоять в оценивании линейной регрессии между y и x и последующем вычислении $\beta x/y$.

Показательные функции

Показательные (или экспоненциальные) функции — это функции вида:

$$y = \alpha e^{\beta x}. \tag{4.14}$$

Наиболее общим их приложением является случай, когда предполагается, что переменная y имеет постоянный темп прироста во времени, в этом случае вместо x обычно используется время (t), а вместо β — постоянный темп прироста (r):

$$y = \alpha e^{rt}. \tag{4.15}$$

Моделирование экспоненциальных временных трендов

Если зависимость y от t задана уравнением вида (4.15), то *абсолютный* прирост y за единицу времени (dy/dt), определяется как

$$\frac{dy}{dt} = r\alpha e^{rt} = ry.$$

Следовательно, *относительный* прирост y за единицу времени можно записать как

$$\frac{dy\ /\ dt}{y} = \frac{ry}{y} = r.$$

Следует помнить, что оценка r, которую вы получаете при оценивании регрессии (4.17), представляет собой оценку темпа прироста *в абсолютном выражении*. Обычно говорят о *процентных* темпах прироста, что означает умножение полученной оценки на 100. Следовательно, если оценка составляет 0,053, это означает, что темп прироста в процентах будет 5,3% за период.

Если имеются значения y для нескольких временных периодов (1, ..., T), то параметры α и r можно оценить, если прологарифмировать (по основанию e) обе части уравнения (4.15):

$$\log y = \log \alpha + \log (e^{rt}) = \log \alpha + rt. \qquad (4.16)$$

Заметим, что $\log (e^{rt})$ — это просто rt; следовательно, мы просто берем логарифм антилогарифма rt. Если определить $y' = \log y$ и $\alpha' = \log \alpha$, то из соотношения (4.16) получим:

$$y' = \alpha' + rt. \qquad (4.17)$$

Таким образом, оценивая регрессию между логарифмом y и t, мы непосредственно получаем оценку темпа прироста r. Обычно оценка α имеет второстепенное значение, но если она представляет для вас интерес, то можно получить ее, потенцируя оценку α' (т. е. беря ее антилогарифм).

Пример

Кривая Энгеля была построена для расходов на питание в США за период с 1959 по 1983 г. с использованием тех же данных, что и в уравнении (2.42), однако вместо линейной функции в данном случае использовалась нелинейная (4.4), приведенная к линейному виду, как в соотношении (4.12), путем взятия логарифмов. Преобразованное выражение имело вид:

$$\widehat{\log y} = 1{,}20 + 0{,}55 \log x. \qquad (4.18)$$

Выполнив обратные преобразования, получим:

$$\hat{y} = e^{1,20}x^{0,55} = 3,32x^{0,55}. \qquad (4.19)$$

Если уравнение (4.4) представляет собой правильную формулу зависимости (в действительности, это, безусловно, сильно упрощено), то полученный результат предполагает, что эластичность спроса на продукты питания по доходу составляет 0,55, что означает, что увеличение личного располагаемого дохода на 1% приведет к увеличению расходов на питание на 0,55%. Коэффициент 3,32 не имеет простого толкования. Он помогает прогнозировать значения y при заданных значениях x, приводя их к единому масштабу.

Те же данные о расходах на питание были использованы для оценивания экспоненциального временного тренда типа (4.15), также приведенного к линейному виду путем взятия логарифмов [см. уравнение (4.16)]. Оцененная зависимость имеет вид:

$$\log \hat{y} = 4,58 + 0,020\, t. \qquad (4.20)$$

Выполнив обратные преобразования, получим:

$$\hat{y} = e^{4,58}e^{0,020t} = 97,5\, e^{0,020t}. \qquad (4.21)$$

Уравнение показывает, что расходы на продукты питания в течение выборочного периода росли с темпом 2% в год. В этом случае постоянный множитель имеет толкование, так как он «прогнозирует», что в момент $t = 0$, т. е. в 1958 г. общие расходы на питание составили 97,5 млрд. долл. Такой прогноз, безусловно, не имеет важного значения, так как мы легко можем найти в справочниках действительные расходы на питание в 1958 г.

Упражнения

4.1. Данные о расходах на оплату жилья в упражнении 2.2 были связаны (1) с располагаемым личным доходом и (2) с экспоненциальным временным трендом в соответствии с моделями (4.4) и (4.15), что дало следующие результаты:

$$\log \hat{y} = -3,48 + 1,230 \log x;$$

$$\log \hat{y} = 4,08 + 0,045\, t.$$

Дайте интерпретацию этих двух уравнений.

4.2. Оцените аналогичные регрессии для товара, выбранного вами в упражнении 2.4, и дайте интерпретацию полученных коэффициентов.

4.3. Выведите выражение для эластичности спроса по доходу для кривой Энгеля, используя (4.3) в качестве модели, и покажите, что при отрицательном β эта эластичность уменьшается с увеличением x. Считаете ли вы, что такая ситуация может быть реальной? Если да, то для какого вида товара возможна такая функциональная форма?

4.3. Случайный член

До сих пор ничего не было сказано о том, как осуществленные преобразования влияют на *случайный член*. В приведенных выше рассуждениях все это вышло за рамки рассмотрения.

Основное требование здесь состоит в том, чтобы случайный член в преобразованном уравнении присутствовал в виде слагаемого ($+u$) и удовлетворял условиям Гаусса—Маркова. В противном случае коэффициенты регрессии, полученные по методу наименьших квадратов, не будут обладать обычными свойствами и проводимые для них тесты окажутся недостоверными.

Например, желательно, если мы учитываем случайное воздействие, чтобы уравнение (4.7) имело следующий вид:

$$y = \alpha + \beta z + u. \tag{4.22}$$

Если это так, то исходное (т. е. непреобразованное) уравнение (4.3) будет иметь вид:

$$y = \alpha + \frac{\beta}{x} + u. \tag{4.23}$$

В данном конкретном случае, если в исходном уравнении случайный член является аддитивным и условия Гаусса—Маркова выполнены, то это также будет верно для преобразованного уравнения. В этом случае проблем нет.

Что произойдет, если мы используем модель вида (4.4)? Регрессионная модель после приведения к линейному виду путем логарифмирования будет представлять собой уравнение (4.13), и оно должно будет также включать случайный член возмущения, который является аддитивным и удовлетворяет условиям Гаусса—Маркова:

$$y' = \alpha' + \beta z + u. \tag{4.24}$$

Если вернуться к исходному уравнению, это означает, что формулу (4.4) следует переписать в следующем виде:

$$y = \alpha x^{\beta} v, \tag{4.25}$$

где v и u связаны соотношением $\log v = u$. Следует помнить, что мы приводим уравнение (4.25) к линейному виду путем логарифмирования его обеих частей. В этом случае мы получаем соотношение:

$$\log y = \log \alpha + \beta \log x + \log v, \tag{4.26}$$

которое представляет собой уравнение (4.24) с соответствующими изменениями определений.

Следовательно, для получения аддитивного случайного члена в уравнении регрессии мы должны начать с мультипликативного случайного члена в исходном уравнении.

Случайный член v изменяет выражение αx^{β} путем увеличения или уменьшения его в случайной *пропорции*, а не на случайную величину. Заметим, что $u = 0$, если $\log v = 0$, что происходит при $v = 1$. Случайная составляющая в оцениваемом уравнении (4.24) будет равна нулю, если $v = 1$. Это имеет смысл, так как если $v = 1$, то оно никак не изменяет значение αx^{β}.

Для того чтобы были применимы t- и F-критерии, величина u должна иметь нормальное распределение. Это означает, что $\log v$ должен иметь нормальное распределение, что возможно только при логарифмически нормальном распределении v. Что произошло бы, если предположить, что случайный член в исходном уравнении является аддитивным, а не мультипликативным?

$$y = \alpha x^{\beta} + u. \tag{4.27}$$

Ответ таков, что, когда вы берете логарифм, невозможно математическим путем упростить выражение $\log(\alpha x^{\beta} + u)$. Наше преобразование не ведет к линеаризации. В этом случае следует использовать метод оценивания нелинейной регрессии, например метод, рассмотренный в разделе 4.5.

Упражнения

4.4. Логарифмические регрессии между (1) расходами на продукты питания или (2) на оплату жилья и личным располагаемым доходом имели следующий вид (в скобках приведены среднеквадратичные ошибки):

$$\widehat{\log y} = \underset{(0,11)}{1,20} + \underset{(0,02)}{0,55} \log x; \quad R^2 = 0,98; \tag{1}$$

$$\widehat{\log y} = \underset{(0,16)}{-3,48} + \underset{(0,02)}{1,23} \log x; \quad R^2 = 0,99. \tag{2}$$

Выполните соответствующие статистические тесты и определите 95-процентный доверительный интервал для эластичности по доходу в каждом случае.

4.5. Выполните соответствующие статистические тесты для логарифмической кривой Энгеля, построенной вами в упражнении 4.2. Определите 95-процентный доверительный интервал для эластичности по доходу.

4.4. Нелинейная регрессия

Предположим, вы считаете, что переменная y связана с переменной x следующим соотношением:

$$y = \alpha + \beta x^{\gamma} + u, \tag{4.28}$$

и хотите получить оценки α, β и γ, имея значения y и x. Уравнение (4.28) не может быть преобразовано в уравнение линейного вида, поэтому в этом случае невозможно применение обычной процедуры оценивания регрессии.

Тем не менее для получения оценок параметров мы по-прежнему можем применить принцип минимизации суммы квадратов отклонений. Далее рассмотрим кратко этот метод не потому, что вам придется применять его самостоятельно (исследователи, как правило, консультируются со специалистом по эконометрике), а потому, что это позволит вам лучше понять идеи, которые лежат в основе регрессионного анализа.

Процедуру лучше всего описать как последовательность шагов.

1. Принимаются некоторые правдоподобные исходные значения параметров.
2. Вычисляются предсказанные значения y по фактическим значениям x с использованием этих значений параметров.
3. Вычисляются остатки для всех наблюдений в выборке и, следовательно, S — сумма квадратов остатков.
4. Вносятся небольшие изменения в одну или более оценку параметров.
5. Вычисляются новые предсказанные значения y, остатки и S.
6. Если S меньше, чем прежде, то новые оценки параметров лучше прежних и их следует использовать в качестве новой отправной точки.
7. Шаги *4*, *5* и *6* повторяются вновь до тех пор, пока не окажется невозможным внести такие изменения в оценки параметров, которые привели бы к уменьшению S.
8. Делается вывод о том, что величина S минимизирована и конечные оценки параметров являются оценками по методу наименьших квадратов.

Пример

Вернемся к примеру с бананами, рассмотренному в разделе 4.1, где y и x связаны следующей зависимостью:

$$y = \alpha + \frac{\beta}{x} + u. \tag{4.29}$$

Для большей простоты предположим, что мы знаем, что $\alpha = 12$; следовательно, нам нужно определить только один неизвестный параметр. Предположим, мы поняли, что зависимость имеет вид (4.29), однако не можем догадаться, что следует применить преобразования, рассмотренные в разделе 4.1. Вместо этого мы применяем нелинейную регрессию.

На рис. 4.3 показаны значения S, которые будут получены при любом возможном выборе b при значениях y и x, приведенных в табл. 4.1. Предположим, что мы начнем, приняв b равным $-6{,}0$. Уравнение при этом примет вид:

$$y = 12 - \frac{6}{x}. \tag{4.30}$$

Вычислим предсказанные значения y и остатки и на основании последних вычислим значение $S = 24{,}02$. Затем подставим $b = -7$. Теперь величина S равна 13,40, т. е. она уменьшилась. Значит, мы движемся в правильном направлении. Подставим $b = -8$; тогда $S = 5{,}87$. Продолжим дальше. При $b = -9$ значение S равно 1,44; при $b = -10$ значение S составит 0,12; при $b = -11$ оно будет 1,89.

Очевидно, что, выбрав $b = -11$, мы перестарались, так как значение S вновь начало расти. Будем двигаться назад, но более мелкими шагами, например по 0,1, беря значения $-10{,}9$; $-10{,}8$ и т. д. Продолжим движение назад до тех пор, пока опять не будет «перебора», затем вновь начнем двигаться вперед еще более мелкими шагами (например, равными 0,01). Каждый раз, когда будет наблюдаться «перебор», будем изменять направление на противоположное, сокра-

щая размер шага. Будем продолжать делать это до тех пор, пока не достигнем требуемой точности вычисления оценки β. Последовательность шагов для данного примера показана в табл. 4.3.

Процесс, показанный в табл. 4.3, был прекращен после 32 итераций, к этому времени стало очевидно, что оценка находится между −9,92 и −9,94. Ясно, что в результате дальнейшего продолжения итерационного процесса могла бы быть получена более высокая точность.

Заметим, что, хотя полученная оценка очень близка к истинному значению

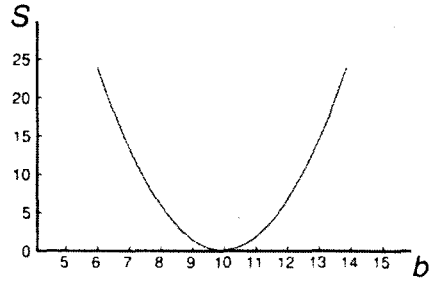


Рис. 4.3. Сумма квадратов отклонений как функция от *b*

10, она не совпадает с оценкой, полученной для уравнения (4.10). В принципе оба набора результатов должны быть одинаковыми, так как и тот и другой минимизируют сумму квадратов отклонений. Расхождение вызвано тем, что мы были не совсем честны в нелинейном случае. Мы предположили, что *a* равно истинному значению 12, а не оценили его. Если бы мы действительно не смогли найти преобразование, которое позволяет использовать линейный регрессионный анализ, то нам бы пришлось использовать нелинейный метод и искать наилучшие значения *a* и *b* одновременно, и тогда мы получили бы оценку *a*, равную 12,08, и оценку *b*, равную −10,08, как и в уравнении (4.10).

Таблица 4.3

b	*S*	*b*	*S*	*b*	*S*	*b*	*S*
−6,0	24,0233	−10,7	1,0309	−9,9	0,1093	−9,87	0,1133
−7,0	13,3971	−10,6	0,8072	−9,8	0,1335	−9,88	0,1116
−8,0	5,8706	−10,5	0,6145	−9,81	0,1297	−9,89	0,1103
−9,0	1,4435	−10,4	0,4528	−9,82	0,1262	−9,90	0,1093
−10,0	0,1160	−10,3	0,3221	−9,83	0,1230	−9,91	0,1085
−11,0	1,8880	−10,2	0,2224	−9,84	0,1201	−9,92	0,1081
−10,9	1,5713	−10,1	0,1537	−9,85	0,1175	−9,93	0,1080
−10,8	1,2856	−10,0	0,1160	−9,86	0,1152	−9,94	0,1082

Основной недостаток нелинейной регрессии состоит в том, что она оценивается значительно медленнее, чем линейная регрессия, особенно в том случае, когда приходится оценивать несколько параметров. Разработаны итерационные процедуры, которые с математической точки зрения являются значительно более сложными по сравнению с теми, которые рассмотрены в данном примере, однако до недавнего времени даже при использовании этих процедур высокая стоимость компьютерных расчетов служила тормозом к применению нелинейной регрессии. В течение последних нескольких лет ситуация стала меняться благодаря небывалому росту мощности и быстродействия компьютеров. Вследствие этого к нелинейным методам стал проявляться больший интерес, а для рядовых пользователей разработано «дружественное» программное обеспечение.

4.5. Выбор функции: тесты Бокса—Кокса

Возможность построения нелинейных моделей, как с помощью их приведения к линейному виду, так и путем использования нелинейной регрессии, значительно повышает универсальность регрессионного анализа, но и усложняет задачу исследователя. Нужно спросить себя, будете ли вы начинать с линейной зависимости *или с нелинейной и если с последней*, то какого типа.

Если вы ограничиваетесь парным регрессионным анализом, то можете построить график наблюдений y и x как диаграмму разброса, и это поможет вам принять решение. В примере в разделе 4.2 было очевидно, что зависимость является нелинейной, и не потребовалось бы большого труда, чтобы убедиться, что уравнение вида (4.3) дает почти точное соответствие. Однако обычно все оказывается не так просто. Часто несколько разных нелинейных функций приблизительно соответствуют наблюдениям, если они лежат на некоторой кривой. Однако в случае множественного регрессионного анализа невозможно даже построить график.

При рассмотрении альтернативных моделей с одним и тем же определением зависимой переменной процедура выбора достаточно проста. Наиболее разумным является оценивание регрессии на основе всех вероятных функций, которые можно вообразить, и выбор функции, в наибольшей степени объясняющей изменения зависимой переменной. Если две или более функции подходят примерно одинаково, то вы должны представить результаты для каждой из них.

Из примера в разделе 4.1 видно, что линейная функция объясняет 64% дисперсии y, а гиперболическая функция (4.3) — 99,9%. В этом примере мы без колебаний выбираем последнюю. Однако если разные модели используют разные функциональные формы, то проблема выбора модели становится более сложной, так как нельзя непосредственно сравнить коэффициенты R^2 или суммы квадратов отклонений. В частности — и это наиболее общий пример для данной проблемы, — нельзя сравнить эти статистики для линейного и логарифмического вариантов модели.

Например, линейная регрессия между расходами на жилье и личным располагаемым доходом для США (см. упражнение 2.2) имела коэффициент $R^2 = 0{,}985$, а сумма квадратов отклонений (СКО) была равна 385,2. Для двой-

ной логарифмической версии модели, когда логарифмы берутся по обеим осям (см. упражнение 4.1), соответствующие значения были равны 0,9915 и 0,02. Во втором случае, СКО значительно меньше, но это ничего не решает. Значения $\log y$ значительно меньше соответствующих значений y, поэтому неудивительно, что остатки также значительно меньше. Величина R^2 безразмерна, однако в двух уравнениях она относится к разным понятиям. В одном уравнении она измеряет объясненную регрессией долю дисперсии y, а в другом — объясненную регрессией долю дисперсии $\log y$. Если для одной модели коэффициент R^2 значительно больше, чем для другой, то вы сможете сделать оправданный выбор без особых раздумий, однако, если значения R^2 для двух моделей приблизительно равны, то проблема выбора существенно усложняется.

В этом случае следует использовать стандартную процедуру, известную под названием *теста Бокса—Кокса* (Box, Cox, 1964). Если вы хотите только сравнить модели с использованием y и $\log y$ в качестве зависимой переменной, то можно использовать вариант теста, разработанный Полом Зарембкой (Zarembka, 1968). Данный тест предполагает такое преобразование масштаба наблюдений y, при котором обеспечивалась бы возможность непосредственного сравнения СКО в линейной и логарифмической моделях. Процедура включает следующие шаги:

1. Вычисляется среднее геометрическое значений y в выборке. (Оно совпадает с экспонентой среднего арифметического $\log y$, поэтому если вы уже оценили логарифмическую регрессию и регрессионная программа выдает вам распечатку среднего значения зависимой переменной, то необходимо вычислить лишь экспоненту от этого значения.)

2. Пересчитываются наблюдения y, они делятся на это значение, то есть

$$y_i^* = y_i/(\text{Среднее геометрическое } y),$$

где y_i^* — пересчитанное значение для i-го наблюдения.

3. Оценивается регрессия для линейной модели с использованием y^* вместо y в качестве зависимой переменной и для логарифмической модели с использованием $\log(y^*)$ вместо $\log y$; во всех других отношениях модели должны оставаться неизменными. Теперь значения СКО для двух регрессий сравнимы, и, следовательно, модель с меньшей суммой квадратов отклонений обеспечивает лучшее соответствие.

4. Для того чтобы проверить, не обеспечивает ли одна из моделей значимо лучшее соответствие, можно вычислить величину $(T/2)\log Z$, где T — число наблюдений, отношение значений СКО в пересчитанных регрессиях, и взять ее абсолютное значение (т. е. игнорировать знак «минус», если он имеется). Эта статистика имеет распределение χ^2 с одной степенью свободы. Если она превышает критическое значение χ^2 при выбранном уровне значимости, то делается вывод о наличии значимой разницы в качестве оценивания.

Пример

Тест будет выполнен как для данных о расходах на продукты питания, так и для данных о расходах на жилье в США. Логарифмические регрессии для этих двух видов благ [уравнение (4.18), упражнение 4.1] показали, что средние значения log y составляют 4,8422 для расходов на питание и 4,6662 для расходов на жилье. Масштабирующие множители равны $e^{4,8422}$ и $e^{4,6662}$ соответственно. В табл. 4.4 приведены значения СКО для линейной и двойной логарифмической регрессии, при этом использованы пересчитанные данные для двух видов благ.

Таблица 4.4

	Расходы на питание	*Расходы на жилье*
Линейная регрессия	0,0119	0,0341
Логарифмическая регрессия	0,0119	0,0221

Из табл. 4.4 видно, что для регрессии расходов на питание соответствие одинаково хорошо в обоих случаях. В случае расходов на жилье логарифмическая регрессия дает более точное соответствие. Логарифм отношения значений СКО для двух регрессий равен здесь 0,4337, и, следовательно, после умножения на 12,5 тестовая статистика составляет 5,42. Критический уровень χ^2 с одной степенью свободы составляет 3,84 при 5-процентном уровне значимости и 6,64 — при однопроцентном уровне (см. табл. А.4), так что в данном случае соответствие будет значимо различным для двух регрессий только при 5-процентном уровне. Эти результаты могут показаться несколько неожиданными, так как можно предположить, что с точки зрения теории модель с логарифмами является более совершенной. Однако период выборки настолько мал, что кривизна функции Энгеля, вероятно, не успеет проявиться, поэтому линейная функция может обеспечить почти столь же хорошее соответствие, как и нелинейная функция [1].

Упражнение

4.6. Оцените еще раз линейную и логарифмическую регрессии для вашего товара, выполнив сначала пересчет по методу Заремб ки, а затем проверьте, имеется ли значимое различие в их качестве.

[1] Регрессии, пересчитанные по методу Заремб ки, могут быть использованы только для того, чтобы решить, какую предпочесть модель. Не надо обращать внимание на коэффициенты, важны только значения СКО. Коэффициенты следует определять непосредственно из непересчитанного варианта выбранной модели.

Приложение 4.1

Более общий тест Бокса—Кокса[1]

Исходная процедура Бокса—Кокса является более общей, чем вариант, описанный в разделе 4.5. Дж. Бокс и Д. Кокс заметили, что y и $\log y$ — это специальные случаи функции $(y^{\lambda} - 1)/\lambda$, из которой получается функция y, когда $\lambda = 1$, и функция $\log y$ (предельный случай), когда λ стремится к нулю. Нет оснований предполагать, что одно из этих значений λ является оптимальным, а есть смысл попробовать целый ряд значений с тем, чтобы определить, какое из них дает минимальное значение СКО (после выполнения пересчета по методу Зарембки). Эта процедура известна под названием *решетчатого поиска*. Для нее нет специальных возможностей в типовых эконометрических компьютерных программах, но тем не менее выполнить ее нетрудно. Если вы предполагаете использовать 10 значений λ, то необходимо задать в регрессионном пакете 10 новых зависимых переменных, используя функциональную форму и различные значения λ, после предварительного пересчета по методу Зарембки. Затем вы находите регрессию между каждой из них и независимыми переменными. В табл. 4.5 приведены результаты оценивания регрессий для расходов на питание и жилье для различных значений λ. Для оценивания регрессий личный располагаемый доход был преобразован так же, как y, за исключением пересчета по методу Зарембки. Такое преобразование не обязательно, при желании вы можете оставить переменную (или переменные) в правой части в линейной форме или же произвести для них одновременный отдельный решетчатый поиск другого значения λ.

Таблица 4.5

λ	*Сумма квадратов отклонений*	
	продукты питания	*жилье*
1,00	0,0119	0,0341
0,75	0,0117	0,0304
0,50	0,0117	0,0272
0,25	0,0118	0,0244
0,00	0,0119	0,0221
-0,25	0,0122	0,0202
-0,50	0,0127	0,0187
-0,75	0,0132	0,0178
-1,00	0,0139	0,0174

[1] Данное приложение содержит материал повышенной сложности, и его в принципе можно пропустить.

Результаты показывают, что оптимальное значение λ для продуктов питания составляет приблизительно 0,5, что говорит о примерно одинаковой приемлемости линейной и логарифмической регрессий. В случае расходов на жилье регрессия обратных величин переменных на первый взгляд дает более точное соответствие по сравнению с линейной и логарифмической регрессией. Однако, как будет видно из следующих разделов, рассматриваемая простая спецификация модели имеет столько недостатков, что детальное исследование оптимальной математической формы на этом этапе не гарантировано.

Наряду с получением точечной оценки для λ можно также получить доверительный интервал, однако данная процедура выходит за рамки этой книги. (Если вас интересует этот вопрос, обратитесь к работе Дж. Спицера [Spitzer, 1982, pp. 307–313].)

5

МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

В этой главе регрессионный анализ по методу наименьших квадратов обобщается для случая, когда в модели регрессии вместо одной независимой переменной используется несколько независимых переменных. Рассматриваются два новых вопроса. Один из них касается проблемы разграничения эффектов различных независимых переменных. Эта проблема в случае ее обострения известна под названием мультиколлинеарности. Другой вопрос состоит в оценке объединенной объясняющей способности независимых переменных в противоположность их отдельным предельным эффектам.

5.1. Иллюстрация: модель с двумя независимыми переменными

Множественный регрессионный анализ является развитием парного регрессионного анализа применительно к случаям, когда зависимая переменная гипотетически связана с более чем одной независимой переменной. Большая часть анализа будет непосредственным расширением парной регрессионной модели, но здесь мы сталкиваемся с двумя новыми проблемами. Во-первых, при оценке влияния данной независимой переменной на зависимую переменную нам придется решать проблему разграничения ее воздействия и воздействий других независимых переменных. Во-вторых, мы должны будем решить проблему спецификации модели. Часто предполагается, что несколько переменных могут оказывать влияние на зависимую переменную, с другой стороны, некоторые переменные могут не подходить для модели. Мы должны решить, какие из них следует включить в уравнение регрессии, а какие — исключить из него. Вторая проблема будет рассмотрена в главе 6. В данной главе мы полагаем, что спецификация модели правильна. В большинстве ситуаций мы ограничимся основным случаем, где используются только две независимые переменные.

Начнем с рассмотрения примера, в котором определяются факторы совокупного спроса на продукты питания. Расширим первоначальную модель, включив учет влияния ценовых изменений на спрос, и допустим, что истинную зависимость можно выразить следующим образом:

$$y = \alpha + \beta_1 x + \beta_2 p + u, \tag{5.1}$$

где y — общая величина расходов на питание, x — располагаемый личный доход, а p — цена продуктов питания. Это, разумеется, является значительным упрощением как с точки зрения состава независимых переменных, включенных в зависимость, так и с точки зрения математической формулы связи. Кроме того, мы неявно предполагаем наличие лишь прямой связи за счет допущения о том, что расходы на питание не влияют на доход и цену. Это могло бы быть в том случае, если бы цены определялись на мировом рынке, но в большинстве ситуаций более реально допустить, что расходы на продукты и их цены определяются совместно в результате взаимодействия предложения и спроса. Проблемы, которые возникают в таких моделях, будут рассмотрены в главе 11.

Для геометрической иллюстрации этой зависимости необходима трехмерная диаграмма с отдельными осями для y, x и p (рис. 5.1). Основание диаграммы содержит оси для x и p, и если пренебречь текущим влиянием случайного члена, то наклонная плоскость над ним показывает величину y, соответствующую любому сочетанию x и p, измеренную расстоянием по вертикали от данной точки до этой плоскости. Так как расходы на питание могут увеличиваться с ростом доходов и уменьшаться с увеличением цены, изображение на диаграмме было построено на основе допущения о том, что величина β_1 является положительной, а β_2 — отрицательной. Конечно, нереально было бы предполагать, что одна из величин x и p могла бы быть равной нулю, и структуру диаграммы можно описать следующим образом. Если бы обе величины x и p оказались равными нулю, то величина y равнялась бы α. При сохранении $p = 0$ уравнение (5.1) означает, что для любого положительного дохода величина y будет равна $(\alpha + \beta_1 x)$, и на рисунке приращение $\beta_1 x$ обозначено как «чистый эффект дохода». При сохранении $x = 0$ уравнение означает, что для любой положительной цены величина y будет равной $(\alpha + \beta_2 p)$, приращение $\beta_2 p$ на рисунке обозначено как «чистый эффект цены». Поскольку β_2 на практике является отрицательной величиной, отрицательным будет и этот эффект. Показан также комбинированный эффект дохода и цены $(\beta_1 x + \beta_2 p)$.

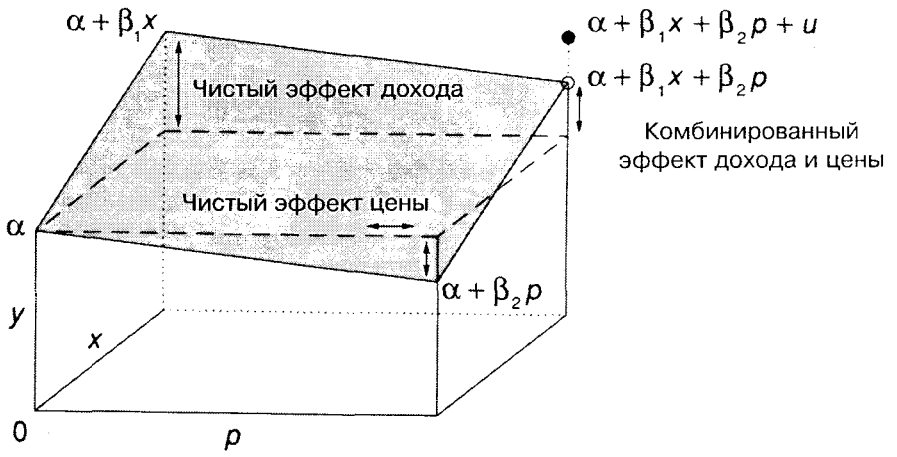


Рис. 5.1. Истинная модель с двумя независимыми переменными: расход как функция дохода и цены

Итак, до сих пор мы пренебрегали случайным членом. Если он отсутствует на данный момент в уравнении (5.1), то значения y в выборке наблюдений для y, x и p будут находиться точно на наклонной плоскости и будет довольно просто вывести точные значения β_1 и β_2 (это не так просто сделать геометрически, если вы не имеете достаточно большого опыта построения трехмерных моделей, однако это довольно просто сделать алгебраическим путем).

Учет случайного члена приводит к тому, что фактические значения y будут лежать несколько выше или несколько ниже значений, соответствующих наклонной плоскости. Следовательно, теперь мы имеем трехмерный аналог для двухмерной задачи, показанной на рис. 2.2. Вместо нахождения линии, соответствующей двухмерному рассеянию точек, мы теперь должны расположить плоскость так, чтобы она соответствовала трехмерному рассеянию. Уравнение для выбранной плоскости будет иметь вид:

$$\hat{y} = a + b_1 x + b_2 p, \tag{5.2}$$

и ее расположение будет зависеть от выбора величин a, b_1 и b_2, являющихся, соответственно, оценками α, β_1 и β_2.

Используя данные для США за 1959–1983 гг. из табл. Б.1 и Б.2 по затратам на питание, располагаемому личному доходу и ценам, мы получим уравнение регрессии:

$$\begin{array}{lllll} \hat{y} = & 116{,}7 + 0{,}112x - 0{,}739p; & & R^2 = 0{,}99, & \\ \text{(с.о.)} & (9{,}6) \quad (0{,}003) & (0{,}114) & & \end{array} \tag{5.3}$$

где y и x измерены в долларах США в постоянных ценах 1972 г., а p является индексом относительной цены, вычисленным путем деления неявного дефлятора цен продуктов питания на неявный дефлятор общих расходов (равный 100 в 1972 г.) и умноженным на 100.

Полученное уравнение следует интерпретировать следующим образом. При каждом увеличении располагаемого личного дохода на 1 млрд. долл. (при сохранении постоянных цен) расходы на питание увеличатся на 112 млн. долл. На каждую единицу увеличения индекса цен (при сохранении постоянных доходов) эти расходы уменьшатся на 739 млн. долл. Чистый эффект в любой момент времени будет зависеть не только от этих коэффициентов, но также от размеров изменений x и p.

Например, в период 1975–1980 гг. располагаемый личный доход увеличился на 145,8 млрд. долл., и, согласно уравнению (5.3), это привело к увеличению расходов на питание на 16,3 млрд. долл. В течение указанного периода индекс цен упал со 111,9 до 109,7, т. е. на 2,2 пункта, и это привело к дальнейшему увеличению y на 1,6 млрд. долл. Совместный эффект, прогнозируемый уравнением (5.3), таким образом, составил увеличение затрат на питание в размере 17,9 млрд. долл. Как видно из табл. Б.1, фактическое увеличение оказалось несколько больше, а именно 20,3 млрд. долл.

Даже если бы спецификация модели оказалась правильной (разумеется, это является большим упрощением), то между прогнозируемым изменением и полученным результатом будет наблюдаться расхождение. Прежде всего оценки β_1 и β_2 подвержены влиянию ошибки выборки. Кроме того, фактические уровни затрат на питание в 1975 и 1980 гг. определялись не только экономической зависимостью, но и случайным членом и в тот и другой годы, а следовательно,

измеренное приращение в течение этого периода имеет, наряду с экономической составляющей, также и случайную составляющую.

Упражнение

5.1. Вам необходимо рассчитать индекс относительных цен для выбранного вами товара для использования в упражнении 5.3, которое является продолжением упражнения 2.4. Рассчитайте его путем деления дефлятора цен для вашего товара из табл. Б.2 на дефлятор общих расходов и умножения на 100. Постройте график рассчитанного индекса. Можете ли вы дать экономическое объяснение изменений относительного индекса цен в течение указанного периода?

5.2. Вывод и интерпретация коэффициентов множественной регрессии

Как и в случае парной регрессии, мы так выбираем значения коэффициентов регрессии, чтобы обеспечить наилучшее соответствие наблюдениям в надежде получить оптимальные оценки для неизвестных истинных значений параметров. Как и прежде, оценка оптимальности соответствия определяется минимизацией S, т. е. суммы квадратов отклонений:

$$S = e_1^2 + \ldots + e_n^2, \tag{5.4}$$

где e_i является остатком в наблюдении i, разницей между фактическим значением y в этом наблюдении и значением $\hat{y}$, прогнозируемым по уравнению регрессии:

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}, \tag{5.5}$$

$$e_i = y_i - \hat{y}_i = y_i - a - b_1 x_{1i} - b_2 x_{2i}. \tag{5.6}$$

Используя уравнение (5.6), мы можем записать:

$$S = \Sigma e^2 = \Sigma (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2. \tag{5.7}$$

Необходимые условия первого порядка для минимума, то есть $\partial S/\partial a = 0$, $\partial S/\partial b_1 = 0$ и $\partial S/\partial b_2 = 0$, дают следующие уравнения:

$$\partial S/\partial a = -2 \Sigma (y_i - a - b_1 x_{1i} - b_2 x_{2i}) = 0; \tag{5.8}$$

$$\partial S/\partial b_1 = -2 \Sigma x_{1i} (y_i - a - b_1 x_{1i} - b_2 x_{2i}) = 0; \tag{5.9}$$

$$\partial S/\partial b_2 = -2 \Sigma x_{2i} (y_i - a - b_1 x_{1i} - b_2 x_{2i}) = 0. \tag{5.10}$$

Следовательно, мы имеем три уравнения с тремя неизвестными: a, b_1 и b_2. Первое уравнение можно легко перегруппировать для выражения величины a через b_1, b_2 и данные наблюдений для x и y:

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2. \tag{5.11}$$

Используя это выражение и два других уравнения, путем некоторых преобразований можно получить следующее выражение для b_1:

$$b_1 = \frac{\mathrm{Cov}(x_1, y)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \{\mathrm{Cov}(x_1, x_2)\}^2}. \tag{5.12}$$

Аналогичное выражение для b_2 можно получить путем перестановки x_1 и x_2 в уравнении (5.12).

Цель данного обсуждения состоит в выделении двух основных моментов. Во-первых, принципы, лежащие в основе вычисления коэффициентов регрессии, в случаях множественной и парной регрессии не различаются. Во-вторых, используемые при этом формулы будут разными, поэтому не следует пытаться использовать выражения, выведенные для парной регрессии, в случае множественной регрессии. Отметим также, что вычисление формул регрессии при двух независимых переменных является более трудоемкой задачей, чем при одной переменной, и вам придется использовать компьютер.

Общая модель

В предыдущем примере мы имели только две независимые переменные. В тех случаях, когда этих переменных более двух, уже невозможно дать геометрическое представление того, что происходит, но развитие алгебраических выкладок в принципе вполне очевидно. Допустим, что переменная y связана с k независимыми переменными x_1, ..., x_k неизвестной истинной зависимостью:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + u. \tag{5.13}$$

Оценим уравнение для данного множества n наблюдений для y, x_1, ... , x_k по методу наименьших квадратов:

$$\hat{y} = a + b_1 x_1 + \ldots + b_k x_k. \tag{5.14}$$

Это вновь означает минимизацию суммы квадратов разностей, а отклонение в наблюдении i выражается как

$$e_i = y_i - \hat{y}_i = y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki}. \tag{5.15}$$

Уравнение (5.15) является обобщением уравнения (5.6). Теперь мы выбираем a, b_1, ... , b_k так, чтобы свести к минимуму S — сумму квадратов отклонений Σe_i^2. Мы получаем $(k + 1)$ условий первого порядка $\partial S/\partial a = 0$, $\partial S/\partial b_1 = 0$, ..., $\partial S/\partial b_k = 0$, что дает $(k + 1)$ уравнение для нахождения $(k + 1)$ неизвестных. Можно легко показать, что первое из этих уравнений позволяет получить аналог для уравнения (5.11), относящегося к случаю с двумя независимыми переменными:

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_k \bar{x}_k. \tag{5.16}$$

Выражения для b_1, b_2, ..., b_k становятся очень сложными, и математика не будет здесь представлена в явном виде. Вычисления целесообразнее сделать с помощью матричной алгебры, но для этого в книге не приводится теоретических

или практических приложений. Для практических примеров вычисления вручную неприемлемы, и для нахождения решений следует использовать компьютер.

Интерпретация коэффициентов множественной регрессии

Множественный регрессионный анализ позволяет разграничить влияние независимых переменных, допуская при этом возможность их коррелированности. Коэффициент регрессии при каждой переменной x дает оценку ее влияния на величину y в случае неизменности влияния на нее всех остальных переменных x.

Это может быть продемонстрировано двумя способами. Один из них состоит в выяснении того, что если модель правильно специфицирована и выполнены условия Гаусса—Маркова, то оценки получаются несмещенными. Это будет сделано в следующей главе для случая, когда имеются только две независимые переменные. Второй способ состоит в оценивании регрессионной зависимости y от одной из независимых переменных, устранив перед этим возможность замещения этой переменной любой другой независимой переменной и показав далее, что оценка ее коэффициента в данном случае совпадет с оценкой коэффициента множественной регрессии. Этот способ будет описан для случая регрессии с двумя независимыми переменными. Предположим, что

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u, \tag{5.17}$$

и допустим, что величины β_1 и β_2 положительны и x_1 и x_2 положительно коррелированы.

Что произойдет, если оценить парную регрессию между y и x_1? По мере увеличения x_1 (1) y имеет тенденцию к росту, поскольку коэффициент β_1 положителен; (2) x_2 имеет тенденцию к росту, так как x_1 и x_2 положительно коррелированы; (3) y получит ускорение из-за увеличения x_2 и благодаря тому, что коэффициент β_2 положителен. Другими словами, изменения y будут преувеличивать влияние текущих значений x_1, так как отчасти они будут связаны с изменениями x_2.

Это показано на рис. 5.2. Стрелки, проведенные сплошной линией, показывают непосредственные воздействия x_1 и x_2 на y. Если x_2 не включается в рассмотрение, то часть изменений y за счет изменений x_2 будет приписана x_1, если переменная x_1 может замещать x_2, действуя подобно ей. В результате оценка значения β_1 будет смещена. Расчет величины смещения будет представлен в разделе 6.2.

Предположим, однако, что вы устранили возможность замещения величиной x_1 величины x_2. Допустим, что можно разложить переменную x_1 на две составляющие:

$$x_1 = \tilde{x}_1 + \hat{x}_1, \tag{5.18}$$

где $\hat{x}_1$ — составляющая, способная замещать x_2, и $\tilde{x}_1$ — оставшаяся часть. Парная регрессионная зависимость y от $\tilde{x}_1$ дает оценку влияния x_1, не искаженного тем, что данная переменная частично выступает в качестве замещающей

для x_2. Мы покажем, что оценка β_1, полученная таким образом, является идентичной оценке коэффициента множественной регрессии (5.12).

Для регрессионной зависимости y от $\tilde{x}_1$ коэффициент наклона составит $\tilde{b}$, где

$$\tilde{b}_1 = \frac{\mathrm{Cov}(\tilde{x}_1, y)}{\mathrm{Var}(\tilde{x}_1)} = \frac{\mathrm{Cov}(x_1, y) - \mathrm{Cov}(\hat{x}_1, y)}{\mathrm{Var}(x_1) + \mathrm{Var}(\hat{x}_1) - 2\mathrm{Cov}(x_1, \hat{x}_1)}, \tag{5.19}$$

так как $\tilde{x}_1$ равно $(x_1 - \hat{x}_1)$.

Чтобы определить величину $\hat{x}_1$, можно оценить регрессию между x_1 и x_2 на основе данных об этих показателях, получив зависимость:

$$\hat{x}_1 = c + dx_2, \tag{5.20}$$

Следовательно, величина $\hat{x}_1$ является составляющей x_1, которая может быть спрогнозирована с помощью x_2. Величина d определяется следующим выражением:

$$d = \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_2)}. \tag{5.21}$$

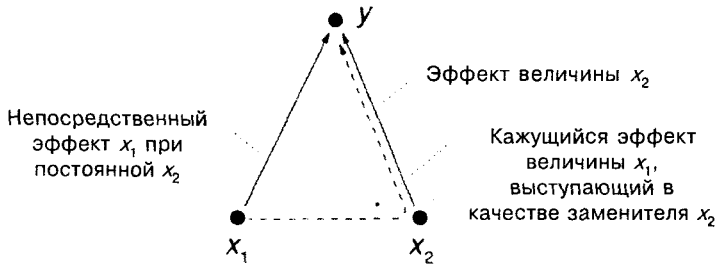


Рис. 5.2

Подставляя формулу (5.20) в уравнение (5.19), получим:

$$\tilde{b}_1 = \frac{\mathrm{Cov}(x_1, y) - \mathrm{Cov}(\{c + dx_2\}, y)}{\mathrm{Var}(x_1) + \mathrm{Var}(c + dx_2) - 2\mathrm{Cov}(x_1, \{c + dx_2\})} =$$

$$= \frac{\mathrm{Cov}(x_1, y) - d\mathrm{Cov}(x_2, y)}{\mathrm{Var}(x_1) + d^2\mathrm{Var}(x_2) - 2d\mathrm{Cov}(x_1, x_2)}. \tag{5.22}$$

(Отметим, что c исключается из выражений ковариации и дисперсии, так как эта величина является постоянной.) Подставив выражение для d из соотношения (5.21) и перегруппировав члены, получим выражение для b_1, представленное уравнением (5.12).

Итак, мы показали для случая с двумя переменными, что оценки множественной регрессии совершенно идентичны оценкам, которые мы могли бы получить при использовании двухэтапной процедуры с исключением перекре-

стных эффектов. Полученный результат может быть обобщен для случая k переменных.

Упражнения

5.2. Индекс относительной цены коммунальных услуг был получен путем деления неявного ценового дефлятора из табл. Б.2 на дефлятор общих расходов и умножения на 100. Оценка множественной регрессии между расходами на коммунальные услуги, располагаемым личным доходом и индексом относительных цен дает следующий результат:

$$\hat{y} = -43{,}4 + 0{,}181x + 0{,}137p.$$

Дайте экономическую интерпретацию этого результата. Почему он не может вас удовлетворить?

5.3. Дополнение к упражнению 2.4. Оцените множественную регрессию между расходами на ваш товар, располагаемым личным доходом и индексом цен, построенным в упражнении 5.1, и дайте интерпретацию результатов.

5.4. Используя формулу (5.21), замените d в уравнении (5.22) и покажите, что можно получить выражение для b_1, представленное уравнением (5.12).

5.3. Множественная регрессия в нелинейных моделях

В главе 4 было показано, что линейные модели регрессии могут быть описаны как линейные в двух отношениях: как линейные по переменным и как линейные по параметрам. Для линейного регрессионного анализа требуется линейность только по параметрам, поскольку нелинейность по переменным может быть устранена с помощью изменения определений. В качестве иллюстрации покажем, что зависимость

$$y = \alpha + \beta_1 x_1^2 + \beta_2 \sqrt{x_2} + \ldots \tag{4.5}$$

может быть переписана в форме, которая будет линейной по переменным:

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \ldots, \tag{4.6}$$

путем простого определения $z_1 = x_1^2$, $z_2 = \sqrt{x_2}$ и т. д. Если случайный член (не показанный явно в уравнении) удовлетворял условиям Гаусса—Маркова в начальном уравнении, то он будет им удовлетворять и в переписанном уравнении. Следовательно, в качестве примера мы могли бы оценить квадратичную зависимость:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + u, \tag{5.23}$$

записав $z = x^2$ и оценив регрессию между y, x и z. Включая более высокие степени для x, мы в принципе могли бы оценить коэффициенты многочлена любого нужного нам вида.

Нелинейность по параметрам является более серьезной проблемой. Если, однако, правая часть модели состоит из членов вида x^{β} или $e^{\beta x}$, умноженных друг на друга, а случайный член является мультипликативным, то модель может быть линеаризована посредством логарифмирования ее обеих частей. Следовательно, например, функция спроса

$$y = \alpha x^{\beta} p^{\gamma} v, \tag{5.24}$$

где y — расходы на товар, x — доход, p — относительная цена, а v — случайный член, может быть преобразована в форму, которая является линейной по параметрам:

$$\log y = \log \alpha + \beta \log x + \gamma \log p + \log v. \tag{5.25}$$

Если вы оцениваете регрессию между данными для $\log y$, $\log x$ и $\log p$, то коэффициент при $\log x$ будет непосредственной оценкой β — эластичности спроса по доходу, а коэффициент при $\log p$ будет оценкой γ — эластичности спроса по цене.

Пример 1. Функция спроса

Логарифмическая регрессия между расходами на питание и располагаемым личным доходом для США была оценена на основе тех же данных, которые использовались для уравнения (5.3), и был получен следующий результат (в скобках указаны стандартные ошибки):

$$\begin{array}{lll} \widehat{\log y} = 2{,}82 + 0{,}64 \log x - 0{,}48 \log p; & R^2 = 0{,}99; & \\ \quad (0{,}42) \quad (0{,}03) \qquad (0{,}12) & F = 820{,}1. & \end{array} \tag{5.26}$$

Уравнение регрессии показывает, что эластичность спроса по доходу составляет 0,64, а эластичность спроса по цене — 0,48, и оба коэффициента значимо отличаются от нуля при однопроцентном уровне значимости.

Пример 2. Производственная функция Кобба—Дугласа

В 1927 г. Пол Дуглас, экономист по образованию, обнаружил, что если нанести на одну и ту же диаграмму графики логарифмов показателей реального объема выпуска *(Y)*, капитальных затрат *(K)* и затрат труда *(L)*, то расстояния от точек графика показателей выпуска до точек графиков показателей затрат труда и капитала будут составлять постоянную пропорцию. Затем он обратился к математику Чарльзу Коббу с просьбой найти математическую зависимость, обладающую такой особенностью, и Кобб предложил следующую функцию:

$$Y = AK^{\alpha} L^{1-\alpha}. \tag{5.27}$$

Эта функция была предложена примерно 30 годами раньше Филипом Уикстидом (Wicksteed), как было указано Ч. Коббом и П. Дугласом в их классической работе (Cobb, Douglas, 1929), но они были первыми, кто использовал для ее построения эмпирические данные, представленные в табл. 5.1. Авторы не описывают, каким образом они на самом деле подобрали функцию, но

предположительно они использовали начальную форму регрессионного анализа, так как они ссылались на «теорию наименьших квадратов». По их оценке, $\alpha = {}^{1}/_{4}$.

Если повторить их вычисления, используя регрессионный анализ, то нельзя сразу провести линеаризацию путем логарифмирования обеих частей уравнения, поскольку тогда мы получим две различные оценки α. Коэффициент при $\log K$ дает нам одну оценку, а коэффициент при $\log L$, который является оценкой $(1 - \alpha)$, позволит нам вычислить другую оценку. Вместо этого мы разделим обе стороны уравнения на величину L и перепишем функцию следующим образом:

$$Y/L = A(K/L)^{\alpha}v \tag{5.28}$$

(включая случайный член v). В этой форме функция может быть интерпретирована как соотношение выпуска на одного работника к капитальным затратам на одного работника, и теперь мы проведем ее линеаризацию, взяв логарифмы:

$$\log (Y/L) = \log A + \alpha \log (K/L) + \log v. \tag{5.29}$$

Использовав для оценивания этого уравнения данные из табл. 5.1, получим (стандартные ошибки указаны в скобках):

Таблица 5.1

Индексы реального объема производства, реальных капитальных затрат и реальных затрат труда (промышленность США, 1899–1922 гг.) (1899 = 100)

Год	*Y*	*K*	*L*	*Год*	*Y*	*K*	*L*
1899	100	100	100	1911	153	216	145
1900	101	107	105	1912	177	226	152
1901	112	114	110	1913	184	236	154
1902	122	122	118	1914	169	244	149
1903	124	131	123	1915	189	266	154
1904	122	138	116	1916	225	298	182
1905	143	149	125	1917	227	335	196
1906	152	163	133	1918	223	366	200
1907	151	176	138	1919	218	387	193
1908	126	185	121	1920	231	407	193
1909	155	198	140	1921	179	417	147
1910	159	208	144	1922	240	431	161

Источник: Cobb, Douglas (1928).

$$\widehat{\log} Y/L = 0{,}02 + 0{,}25 \log K/L; \qquad R^2 = 0{,}63; \qquad (5.30)$$
$$(0{,}02)\ \ (0{,}04) \qquad\qquad F = 38{,}0.$$

что подтверждает вычисления Ч. Кобба. Формула Кобба—Дугласа, конечно, является частным случаем более общей формулы:

$$Y = AK^{\alpha}L^{\beta}v, \qquad (5.31)$$

где показатели эластичности выпуска по затратам капитала и труда не связаны между собой. Оценив его с использованием тех же самых данных, мы получим (стандартные ошибки указаны в скобках):

$$\widehat{\log} Y = -0{,}18 + 0{,}23 \log K + 0{,}81 \log L; \qquad R^2 = 0{,}96; \qquad (5.32)$$
$$(0{,}43)\ \ (0{,}06) \qquad\ (0{,}15) \qquad\qquad F = 236{,}1.$$

Это указывает на то, что эластичность выпуска продукции по затратам капитала составляет 0,23, что очень близко к предыдущей оценке, а эластичность по затратам труда составляет 0,81, что несколько выше предыдущей оценки, равной 0,75.

Упражнения

5.5. Оценка логарифмической регрессии между расходами на жилищные услуги, располагаемым личным доходом и относительной ценой этих услуг с использованием данных, приведенных в упражнении 5.2, дает следующий результат:

$$\widehat{\log} y = -1{,}60 + 1{,}18 \log x - 0{,}34 \log p.$$

Дайте интерпретацию этого уравнения. Сравните ее с интерпретацией, данной для упражнения 5.2. В каком смысле она лучше?

5.6. Оцените аналогичную логарифмическую регрессию между расходами на товар, выбранный вами в упражнении 2.4, располагаемым личным доходом и относительной ценой товара. Дайте интерпретацию результата.

Свойства производственной функции Кобба—Дугласа

В рассмотрении эластичности выпуска продукции, эффекта от масштаба производства и прогнозируемых долей производственных факторов мы будем использовать более общую форму функции и пренебрегать случайным членом.

Эластичность выпуска продукции

Эластичность выпуска продукции по капиталу и труду равна соответственно α и β, так как

$$\frac{\partial Y / \partial K}{Y / K} = \frac{A(\alpha[K^{\alpha-1}])L^{\beta}}{AK^{\alpha-1}L^{\beta}} = \alpha,$$

и аналогичным образом легко показать, что $(\partial Y/\partial L)/(Y/L)$ равно β. Следовательно, увеличение затрат капитала на 1% приведет к росту выпуска продукции на α процентов, а увеличение затрат труда на 1% приведет к росту выпуска на β процентов. Можно предположить, что обе величины α и β находятся между нулем и единицей. Они должны быть положительными, так как увеличение затрат производственных факторов должно вызывать рост выпуска. В то же время, вероятно, они будут меньше единицы, так как разумно предположить, что уменьшение эффекта от масштаба производства приводит к более медленному росту выпуска продукции, чем затрат производственных факторов, если другие факторы остаются постоянными.

Эффект от масштаба производства

Если α и β в сумме превышают единицу, то говорят, что функция имеет возрастающий эффект от масштаба производства (это означает, что если K и L увеличиваются в некоторой пропорции, то Y растет в большей пропорции). Если их сумма равна единице, то это говорит о постоянном эффекте от масштаба производства (Y увеличивается в той же пропорции, что и K и L). Если их сумма меньше, чем единица, то имеет место убывающий эффект от масштаба производства (Y увеличивается в меньшей пропорции, чем K и L).

Например, предположим, что K и L удваиваются. Тогда новый уровень выпуска (Y') записывается следующим образом:

$$Y' = A(2K)^{\alpha}(2L)^{\beta} = A2^{\alpha}K^{\alpha}2^{\beta}L^{\beta} = 2^{\alpha+\beta}AK^{\alpha}L^{\beta} = 2^{\alpha+\beta}Y.$$

Если $\alpha + \beta = 1{,}2$, а $2^{\alpha+\beta} = 2{,}30$, то Y увеличивается больше чем в 2 раза.
Если $\alpha + \beta = 1{,}0$, а $2^{\alpha+\beta} = 2$, то удвоение K и L приводит к удвоению Y.
Если $\alpha + \beta = 0{,}8$, а $2^{\alpha+\beta} = 1{,}74$, то Y увеличивается меньше чем в 2 раза.

В своей первой статье Ч. Кобб и П. Дуглас описывали функцию в виде соотношение (5.27), т. е. они изначально предполагали постоянную отдачу от масштаба. Впоследствии они ослабили это допущение, предпочитая оценивать степень отдачи от масштаба производства.

Прогнозируемые доли производственных факторов

В соответствии с допущением о конкурентности рынков факторов производства a и b имеют дальнейшую интерпретацию как прогнозируемые доли дохода, полученного соответственно за счет капитала и труда. Если рынок труда имеет конкурентный характер, то ставка заработной платы (w) будет равна предельному продукту труда ($\partial Y/\partial L$):

$$w = \frac{\partial Y}{\partial L} = AK^{\alpha}\beta L^{\beta-1} = \frac{\beta Y}{L}.$$

Следовательно, общая сумма заработной платы (wL) будет равна βY, а доля труда в общем выпуске продукции (wL/Y) составит постоянную величину β. Аналогичным образом норма прибыли ρ выражается через $\partial Y/\partial K$:

$$\rho = \frac{\partial Y}{\partial K} = A\alpha K^{\alpha-1}L^{\beta} = \frac{\alpha Y}{K},$$

и, следовательно, общая прибыль (ρK) будет равна αY, а доля прибыли будет постоянной величиной α. В своей первой работе Ч. Кобб и П. Дуглас подтвердили, что доля труда[1] действительно составила примерно $^3/_4$, как и прогнозировалось оцененной ими функцией.

Существует ряд проблем по применению такой функции, особенно в тех случаях, когда она используется для экономики в целом. В частности, даже в тех случаях, когда между выпуском продукции, производственным оборудованием и трудом в производственном процессе существует технологическая зависимость, то совершенно необязательно, что подобная зависимость существует тогда, когда указанные факторы комбинируются в масштабах экономики в целом. Во-вторых, даже если такая зависимость для экономики в целом существует, то нет никаких оснований считать, что она будет иметь простую форму.

5.4. Свойства коэффициентов множественной регрессии

Как и в случае парного регрессионного анализа, коэффициенты регрессии должны рассматриваться как случайные переменные специального вида, случайные компоненты которых обусловлены наличием в модели случайного члена. Каждый коэффициент регрессии вычисляется как функция значений y и независимых переменных в выборке, а y в свою очередь определяется независимыми переменными и случайным членом. Отсюда следует, что коэффициенты регрессии действительно определяются значениями независимых переменных и случайным членом, а их свойства существенно зависят от свойств последнего.

Мы продолжаем считать, что выполняются условия Гаусса—Маркова, а именно: 1) математическое ожидание u в любом наблюдении равно нулю; 2) теоретическая дисперсия его распределения одинакова для всех наблюдений; 3) теоретическая ковариация его значений в любых двух наблюдениях равняется нулю; 4) распределение u независимо от распределения любой объясняющей переменной. Первые три условия идентичны условиям для парного регрессионного анализа, а четвертое условие является обобщением своего аналога. На данный момент мы примем усиленный вариант четвертого условия, допустив, что независимые переменные являются нестохастическими.

[1] Речь идет о доле труда в США. (*Прим. ред.*)

Существуют еще два практических требования. Во-первых, нужно иметь достаточное количество данных для проведения линии регрессии, что означает наличие стольких (независимых) наблюдений, сколько параметров необходимо оценить. Во-вторых, как мы увидим далее в этом разделе, между независимыми переменными не должно существовать строгой линейной зависимости.

Несмещенность

Мы покажем, что b_1 является несмещенной оценкой β_1 для случая с двумя объясняющими переменными. Доказательство можно легко обобщить, используя матричную алгебру для любого числа объясняющих переменных. Как видно из уравнения (5.12), величина b_1 является функцией от x_1, x_2 и y в свою очередь y определяется по x_1, x_2 и u. Следовательно, величина b_1 фактически зависит от значений x_1, x_2 и u в выборке (поняв суть преобразований, можно опустить детали математических выкладок):

$$b_1 = \frac{\mathrm{Cov}(x_1, y)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \{\mathrm{Cov}(x_1, x_2)\}^2} =$$

$$= \frac{1}{\Delta}\{\mathrm{Cov}(x_1, \{\alpha + \beta_1 x_1 + \beta_2 x_2 + u\})\mathrm{Var}(x_2) -$$

$$-\mathrm{Cov}(x_2, \{\alpha + \beta_1 x_1 + \beta_2 x_2 + u\})\mathrm{Cov}(x_1, x_2)\} =$$

$$= \frac{1}{\Delta}\{[\beta_1 \mathrm{Var}(x_1) + \beta_2 \mathrm{Cov}(x_1, x_2) + \mathrm{Cov}(x_1, u)]\mathrm{Var}(x_2) -$$

$$-[\beta_1 \mathrm{Cov}(x_1, x_2) + \beta_2 \mathrm{Var}(x_2) + \mathrm{Cov}(x_2, u)]\mathrm{Cov}(x_1, x_2)\} =$$

$$= \frac{1}{\Delta}\{\beta_1 \Delta + \mathrm{Cov}(x_1, u)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, u)\mathrm{Cov}(x_1, x_2)\} =$$

$$= \beta_1 + \frac{1}{\Delta}\{\mathrm{Cov}(x_1, u)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, u)\mathrm{Cov}(x_1, x_2)\}, \qquad (5.33)$$

где Δ равно $\mathrm{Var}(x_1)\,\mathrm{Var}(x_2) - \{\mathrm{Cov}(x_1, x_2)\}^2$. Отсюда величина b_1 имеет две составляющие: истинное значение β_1 и составляющую ошибки. Перейдя к математическому ожиданию, получим:

$$E(b_1) = \beta_1 + \frac{1}{\Delta}\{\mathrm{Var}(x_2)E[\mathrm{Cov}(x_1, u)] - \mathrm{Cov}(x_1, x_2)E[\mathrm{Cov}(x_2, u)]\} = \beta_1, \qquad (5.34)$$

при допущении, что выполняется четвертое условие Гаусса—Маркова.

Точность коэффициентов множественной регрессии

В теореме Гаусса—Маркова для множественного регрессионного анализа доказывается, что, как и для парной регрессии, обычный метод наименьших квадратов (МНК) дает наиболее эффективные линейные оценки в том смысле, что на основе той же самой выборочной информации невозможно найти другие не-

смещенные оценки с меньшими дисперсиями при выполнении условий Гаусса—Маркова. Мы не будем доказывать эту теорему, но исследуем факторы, регулирующие возможную точность коэффициентов регрессии. В общем случае можно сказать, что коэффициенты регрессии, скорее всего, являются более точными:

1) чем больше число наблюдений в выборке;
2) чем больше дисперсия выборки объясняющих переменных;
3) чем меньше теоретическая дисперсия случайного члена;
4) чем меньше связаны между собой объясняющие переменные.

Первые три из желательных условий повторяют то, на чем мы уже останавливались в случае парного регрессионного анализа. Лишь четвертое условие является новым. Сначала мы рассмотрим случай с двумя независимыми переменными и затем перейдем к более общему случаю.

Две независимых переменных

Если истинная зависимость имеет вид:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u, \tag{5.35}$$

и вы получили уравнение регрессии

$$\hat{y} = a + b_1 x_1 + b_2 x_2, \tag{5.36}$$

использовав необходимые данные, то теоретическая дисперсия вероятностного распределения для b_1 будет описываться выражением:

$$\text{pop. var}\,(b_1) = \frac{\sigma_u^2}{n\text{Var}\,(x_1)} \dot{\times} \frac{1}{1 - r_{x_1,x_2}^2}, \tag{5.37}$$

где σ_u^2 — теоретическая дисперсия величины u. Аналогичное выражение можно получить для теоретической дисперсии величины b_2, заменив Var (x_1) на Var (x_2).

Из уравнения (5.37) можно видеть, что, как и в случае парного регрессионного анализа, желательно, чтобы величины n и Var (x_1) были большими, а величина σ_u^2 — малой. Однако теперь мы получили еще и член $(1 - r_{x_1,x_2}^2)$, и вполне очевидно, что желательно иметь слабую корреляцию между x_1 и x_2.

Этому легко дать интуитивное объяснение. Предположим, что истинная зависимость имеет вид:

$$y = 2 + 3x_1 + x_2 + u. \tag{5.38}$$

Предположим, что между x_1 и x_2 существует нестрогая линейная зависимость:

$$x_2 = 2x_1 - 1, \tag{5.39}$$

и допустим, что величина x_1 увеличивается на одну единицу в каждом наблюдении. Тогда x_2 увеличится на две единицы, а y — на пять единиц, как показано, например, в табл. 5.2.

Таблица 5.2

Приблизительное значение			*Приблизительное значение*		
x_1	x_2	y	*приращения* x_1	*приращения* x_2	*приращения* y
10	19	51	1	2	5
11	21	56	1	2	5
12	23	61	1	2	5
13	25	66	1	2	5
14	27	71	1	2	5
15	29	76	1	2	5

При рассмотрении этих данных можно прийти к любому из следующих выводов:

1) величина y определяется уравнением (5.38) (правильное утверждение);
2) величина x_2 не имеет отношения к данному случаю, и величина y определяется зависимостью:

$$y = 1 + 5x_1 + u;$$

3) величина x_1 не имеет отношения к данному случаю, и величина y определяется зависимостью:

$$y = 3{,}5 + 2{,}5x_2 + u.$$

В действительности этими возможностями дело не ограничивается. Любая зависимость, которая является средним взвешенным условий (2) и (3), также будет соответствовать описанным данным. Условие (1) можно рассматривать как среднее взвешенное условий (2) с коэффициентом 0,6 и (3) с коэффициентом 0,4.

При использовании регрессионного анализа или любого другого метода применительно к данному случаю трудно провести различие между этими возможностями, и полученные оценки будут очень чувствительными по отношению к случайному члену и могут содержать значительные ошибки. Дисперсии коэффициентов регрессии будут большими, что, очевидно, является другим способом выражения того же самого.

Если истинная зависимость (5.39) была строгой, то при оценивании представляется совершенно невозможным провести различие между всеми вероятными зависимостями, поскольку каждая из них будет одинаково хорошо соответствовать данным. Вы даже не сможете вычислить коэффициенты регрессии, так как и числитель и знаменатель уравнения (5.12) будут равны нулю.

Если между x_1 и x_2 существует нестрогая линейная зависимость, то коэффициент корреляции r_{x_1,x_2} будет близким к единице, если зависимость положи-

тельна, и к минус единице, если зависимость отрицательна, и в обоих случаях $r^2_{x_1,x_2}$ будет близким к единице. В результате знаменатель второго члена в уравнении (5.37) будет близок к нулю, а теоретические дисперсии b_1 и b_2 будут большими числами. В предельном случае наличия строгой линейной зависимости дисперсии будут стремиться к бесконечности.

Отметим, что отсюда не следует *автоматически*, что величины b_1 и b_2 будут иметь большие теоретические дисперсии, если между x_1 и x_2 существует нестрогая линейная зависимость. Дисперсии зависят также от n и σ^2_u, как и в случае парного регрессионного анализа. Если n велико, а σ^2_u — мало, то теоретические дисперсии b_1 и b_2 могут быть небольшими, несмотря на нестрогую линейную зависимость. Если имеется большой объем информации (n велико), а случайный фактор является относительно незначимым (σ^2_u мало), то все еще можно разграничить влияние x_1 и x_2 на величину y.

Общий случай

Мы не будем выводить выражения для дисперсий коэффициентов регрессии в общем случае. Подобно выражениям для самих коэффициентов, их лучше всего рассчитывать с помощью матричной алгебры.

Вместо этого будет показан один важный момент на основе эксперимента по методу Монте-Карло. Согласно условию (4), желательно, чтобы независимые переменные не были тесно связаны. Чтобы исследовать это, оценим множественную регрессию три раза. Во-первых, если независимые переменные не слишком тесно связаны, то результаты оценивания регрессии будут надежными. Во-вторых, при более тесной зависимости между переменными результаты регрессии будут содержать ошибки. И в-третьих, при той же самой тесной корреляции между независимыми переменными, но при меньшей дисперсии случайного члена результаты оценивания регрессии значительно улучшаются.

Это показывает, что тесная корреляция между независимыми переменными может привести к неудовлетворительным результатам, но это не происходит автоматически. Это зависит также от дисперсии случайного члена.

Предположим, что заработная плата y в некоторой стране определяется числом лет обучения (S), стажем работы (X), возрастом (A), а также случаем. Базовая заработная плата составляет 10000, к которым добавляется 1500 за каждый год обучения сверх минимальных 10 лет, 500 — за каждый год работы и 25 — за каждый прожитый год. Кроме того, существует случайный фактор u:

$$y = 10\,000 + 1500\,(S - 10) + 500X + 25A + u. \qquad (5.40)$$

В результате упрощения это уравнение проводится к виду:

$$y = -5000 + 1500S + 500X + 25A + u. \qquad (5.41)$$

Первые четыре колонки табл. 5.3 представляют данные воображаемой выборки из 20 индивидов. Цифры для срока обучения, стажа работы и возраста были взяты произвольно. Значения u определялись на основе выборки из 20 нормально распределенных случайных чисел с нулевым математическим ожида-

нием и единичной дисперсией, которые умножались на 2000. Полученные в результате из уравнения (5.41) значения y показаны в пятой колонке табл. 5.3.

Допустив, что обучение начинается с 6 лет, можно получить неравенство:

$$X \leq A - S - 5. \tag{5.42}$$

Таблица 5.3

Индивид	*S*	*X*	*A*	*u*	*y*	*A–S–5*	*X'*	*y'*	*y''*
1	10	20	45	–1740	19385	30	28	23385	24951
2	10	5	23	1880	14955	8	6	15455	13763
3	10	19	36	760	21160	21	17	20160	19476
4	11	15	50	1300	21550	34	28	28050	26880
5	11	16	42	1880	22430	26	21	24930	23238
6	11	8	30	640	16890	14	10	17890	17314
7	11	4	21	3520	17545	5	4	17545	14377
8	12	10	34	–3540	15310	17	15	17810	20996
9	12	8	27	1720	19395	10	8	19395	17847
10	12	18	38	2680	25630	21	19	26130	23710
11	13	6	25	–5220	12905	7	6	12905	17603
12	13	10	46	2840	23490	28	25	30990	28434
13	14	10	38	–1100	20850	19	16	23850	24840
14	14	2	22	–340	17210	3	2	17210	17516
15	15	8	32	1000	23300	12	9	23800	22900
16	16	5	49	20	22745	28	23	31745	31727
17	16	4	28	–780	20920	7	6	21920	22622
18	17	7	33	3140	27965	11	8	28465	25639
19	18	3	27	–380	23795	4	3	23795	24137
20	19	3	32	40	25840	8	6	27340	27304

В табл. 5.3 показана величина $(A - S - 5)$, и можно видеть, что данные для X соответствуют ей, но зависимость между A, S и X является довольно слабой. Многие из индивидов, вполне очевидно, посвящают часть своего трудоспособного возраста другим занятиям.

Оценив регрессию между y, S, X и A, получаем следующий результат:

$$\hat{y} = -4063 + 1409S + 481X + 50A. \tag{5.43}$$
$$\text{(с.о.)} \quad (4140) \quad (280) \quad (175) \quad (88)$$

Эксперимент был повторен с теми же данными для S и A и такими же значениями u, но с другим набором данных для X, который значительно лучше согласован с показателем $(A - S - 5)$. Эти данные обозначены в табл. 5.3 как X', а результирующие значения y обозначены как y'. Так как наше неравенство сейчас в каждом случае почти превращается в равенство, то можно наблюдать нестрогую линейную зависимость между независимыми переменными. Оценивая регрессию между y', S, X' и A, теперь получаем:

$$\hat{y} = -7524 + 781S - 207X' + 664A. \qquad (5.44)$$
(с.о.) (4204) (529) (538) (476)

Результаты оценки регрессии теперь действительно весьма плохи.

Наконец, эксперимент был повторен еще раз при сохранении тех же самых значений S, A и X', но с получением значений u путем умножения случайных чисел на 200 вместо 2000. Результирующие значения y показаны в табл. 5.3 как y''. Оценивая регрессию между y'', S, X' и A, получаем:

$$\hat{y} = -5252 + 1428S + 429X + 89A. \qquad (5.45)$$
(с.о.) (420) (53) (54) (48)

За исключением коэффициента при A, эти результаты являются вполне удовлетворительными, несмотря на существование нестрогой линейной зависимости между независимыми переменными.

Конечно, нельзя придавать слишком большое значение результатам единственного набора экспериментов. Каждый из трех вариантов расчетов был выполнен еще 9 раз с использованием тех же данных для S, A, X и X', но при различных наборах случайных чисел для получения величины u. Результаты экспериментов обобщаются в табл. 5.4.

Таблица 5.4

	Первый вариант (слабая связь)				*Второй вариант (тесная связь)*				*Третий вариант (тесная связь, низкий σ_u^2)*			
	Постоянная	*S*	*X*	*A*	*Постоянная*	*S*	*X*	*A*	*Постоянная*	*S*	*X*	*A*
1	-4063	1409	481	50	-7524	781	-207	664	-5252	1428	429	89
2	-4905	1560	508	3	-8093	892	-218	636	-5309	1439	428	86
3	-9718	1812	597	33	-3147	2790	1684	-971	-4815	1629	618	-75
4	2584	935	347	53	3947	1744	1193	-609	-4105	1524	569	-38
5	-3754	1485	334	43	-4106	1998	854	-327	-4911	1550	535	-10
6	-7628	1591	637	15	-2595	2051	1168	-522	-4759	1555	567	-30
7	-8812	1712	754	-8	-4986	1590	679	-74	-4999	1509	518	15
8	-7760	1791	636	-26	-3701	2128	1034	-446	-4870	1563	553	-22
9	-1326	1281	533	3	-722	1288	547	-27	-4572	1479	509	20
10	-8910	1847	835	-107	-7361	985	-28	476	-5236	1449	447	70

При рассмотрении табл. 5.4 мы сосредоточим внимание на коэффициентах при S и X. Коэффициент при A и постоянная считаются ненадежными в любом случае: коэффициент при A потому, что его истинное значение близко к нулю, а постоянная потому, что точка, определяемая условиями $S=0$, $X=0$, $A=0$, весьма удалена от диапазона выборки.

В первом варианте коэффициенты при S и X находятся в целом в нужном диапазоне. Во втором варианте они безнадежно неточны, а в третьем — они весьма хороши. Результаты экспериментов обобщаются в табл. 5.5.

Отметим, что здесь не наблюдается смещения, характеризуемого тенденцией коэффициентов систематически оказываться выше или ниже их истинных значений, даже во втором варианте, где результаты весьма неточны. Во втором варианте средние значения коэффициентов при S и X соответственно составили 1624 и 671, что не так далеко от истинных значений.

Стандартные ошибки коэффициентов регрессии

Стандартная ошибка коэффициента множественной регрессии имеет такой же смысл, как и в парном регрессионном анализе, в том плане, что она является оценкой стандартного отклонения распределения коэффициента регрессии вокруг его истинного значения (см. раздел 3.5). Как и в парном регрессионном анализе, формула для стандартной ошибки может быть выведена на основе выражения дисперсии распределения, замены σ_u^2 на несмещенную оценку и извлечения квадратного корня. Как и прежде, значимость выражения, полученного таким образом, зависит от правильной спецификации модели и выполнения условий Гаусса—Маркова для случайного члена.

Таблица 5.5

Дисперсия случайного члена	*Линейная зависимость между независимыми переменными*	
	Слабая зависимость	*Тесная зависимость*
Низкая	Надежная	Приемлемая
Высокая	Приемлемая	Ненадежная

Например, если имеются только две независимые переменные, то теоретическая дисперсия коэффициента регрессии b_1 выражается уравнением (5.37). Можно показать, что в этом случае несмещенная оценка величины σ_u^2 может быть получена путем умножения величины Var (e), представляющей собой выборочную дисперсию остатков, на $n/(n-3)$. Следовательно,

$$\text{с.о.}(b_1) = \sqrt{\frac{s_u^2}{n\text{Var}(x_1)} \times \frac{1}{1-r_{x_1,x_2}^2}} = \sqrt{\frac{(n/n-3)\text{Var}(e)}{n\text{Var}(x_1)} \times \frac{1}{1-r_{x_1,x_2}^2}} =$$

$$= \sqrt{\frac{\text{Var}(e)}{n-3\text{Var}(x_1)} \times \frac{1}{1-r_{x_1,x_2}^2}}. \qquad (5.46)$$

Соответствующее выражение для стандартной ошибки b_2 можно получить путем перестановки индексов.

Когда имеется более двух независимых переменных, намного удобнее выразить стандартные ошибки, так же как и сами коэффициенты регрессии, с помощью матричной алгебры.

В начале этого раздела были сформулированы четыре условия, выполнение которых позволяет получать достаточно надежные оценки коэффициентов регрессии, при этом третье и четвертое условия исследовались непосредственно на основе экспериментов по методу Монте-Карло. Каждое условие отражено в выражениях для дисперсий коэффициентов регрессии, представленных в уравнении (5.37), и каждое в свою очередь отражено в соотношении (5.46).

В частности, тесная линейная связь между двумя объясняющими переменными приведет к получению значения $r^2_{x_1,x_2}$, близкого к единице, а следовательно, стандартные ошибки (при прочих равных условиях) будут относительно большими, что отражает вероятную неточность коэффициентов регрессии, что мы уже наблюдали ранее. Например, можно заметить, что стандартные ошибки в уравнении (5.44), где наблюдалась тесная линейная связь между *S, X'* и *A*, намного больше, чем стандартные ошибки в уравнении (5.43), где эта связь была слабой.

Кроме того, целесообразно сравнить стандартные ошибки в уравнениях (5.44) и (5.45). В первом из них величина *u* получалась путем умножения случайных чисел на 2000. Во втором — эти числа умножались на 200. В результате оценки регрессии в уравнении (5.45) были намного точнее, о чем свидетельствуют их гораздо меньшие ошибки. Коэффициенты регрессии оказались в 10 раз точнее (если рассмотреть различие между оценкой и истинным значением), а стандартные ошибки составили лишь $^1/_{10}$ прежнего размера.

t-тесты и доверительные интервалы

t-тесты для коэффициентов множественной регрессии выполняются так же, как это делается в парном регрессионном анализе. Отметим, что критический уровень *t* при любом уровне значимости зависит от числа степеней свободы, которое равно $(n - k - 1)$: число наблюдений минус число оцененных параметров (один коэффициент для каждой независимой переменной и постоянный член). Доверительные интервалы определяются точно так же, как и в парном регрессионном анализе, в соответствии с указанным примечанием относительно числа степеней свободы.

Упражнения

5.7. Линейная и логарифмическая регрессии (упражнения 5.2 и 5.5) между расходами на жилищные услуги, располагаемым личным доходом и относительной ценой этих услуг имели вид (в скобках указаны стандартные ошибки):

$$\hat{y} = \underset{(48,4)}{-43,4} + \underset{(0,009)}{0,181}x - \underset{(0,421)}{0,137}p; \qquad R^2 = 0,99;$$

$$\widehat{\log y} = \underset{(1,75)}{-1,60} + \underset{(0,05)}{1,18}\log x - \underset{(0,31)}{0,34}\log p; \qquad R^2 = 0,99.$$

Выполните соответствующие t-тесты и сформулируйте ваши выводы.

5.8. Выполните аналогичные t-тесты для коэффициентов линейной и логарифмической регрессий, оцененных в упражнениях 5.3 и 5.6.

5.9. Отметим, что первая часть уравнения (5.46) может быть переписана в виде:

$$\text{с.о.}(b_1) = \frac{s_u}{\sqrt{n}\sqrt{\text{Var}(x_1)}} \times \frac{1}{\sqrt{1 - r^2_{x_1,x_2}}}.$$

Используя это выражение, объясните вариации в стандартных ошибках оценок эластичности расходов по цене в логарифмических регрессиях для расходов на питание, жилье, лекарства и отдых (в каждом случае независимыми переменными служили доход и соответствующий индекс цен).

	s_u	$\sqrt{Var(\log p)}$	$r_{\log x,\ \log p}$	*С.о. ценовой эластичности*
Питание	0,018	0,056	0,85	0,121
Жилье	0,031	0,043	–0,89	0,314
Лекарства	0,037	0,155	–0,96	0,160
Отдых	0,037	0,060	–0,27	0,128

5.5. Мультиколлинеарность

Мультиколлинеарность — это понятие, которое используется для описания проблемы, когда нестрогая линейная зависимость между объясняющими переменными приводит к получению ненадежных оценок регрессии. Разумеется, такая зависимость совсем необязательно дает неудовлетворительные оценки. Если все другие условия благоприятствуют, т. е. если число наблюдений и выборочные дисперсии объясняющих переменных велики, а дисперсия случайного члена — мала, то в итоге можно получить вполне хорошие оценки.

Итак, мультиколлинеарность должна вызываться сочетанием нестрогой зависимости и одного (или более) неблагоприятного условия, и это — вопрос степени выраженности явления, а не его вида. Оценка любой регрессии будет страдать от нее в определенной степени, если только все независимые переменные не окажутся абсолютно некоррелированными. Рассмотрение данной проблемы начинается только тогда, когда это серьезно влияет на результаты оценки регрессии.

Эта проблема является обычной для регрессий временных рядов, т. е. когда

данные состоят из ряда наблюдений в течение какого-то периода времени. Если две или более независимые переменные имеют ярко выраженный временной тренд, то они будут тесно коррелированы, и это может привести к мультиколлинеарности.

Что можно предпринять в этом случае?

Различные методы, которые могут быть использованы для смягчения мультиколлинеарности, делятся на две категории: к первой категории относятся попытки повысить степень выполнения четырех условий, обеспечивающих надежность оценок регрессии; ко второй категории относится использование внешней информации. Если сначала использовать возможные непосредственно получаемые данные, то, очевидно, было бы полезным увеличить число наблюдений. Если вы применяете данные временных рядов, то это можно сделать путем сокращения продолжительности каждого периода времени. Например, при оценивании уравнений функции спроса в упражнениях 5.3 и 5.6 можно перейти с использования ежегодных данных на поквартальные данные. После этого вместо 25 наблюдений их станет 100. Это настолько очевидно и так просто сделать, что большинство исследователей, использующих временные ряды, почти автоматически применяют поквартальные данные, если они имеются, вместо ежегодных данных, даже если проблема мультиколлинеарности не стоит, просто для сведения к минимуму теоретических дисперсий коэффициентов регрессии. В таком подходе существуют, однако, и потенциальные проблемы. Можно привнести или усилить автокорреляцию (см. главу 7), но она может быть нейтрализована. Кроме того, можно привнести (или усилить) смещение, вызванное ошибками измерения (см. главу 8), если поквартальные данные измерены с меньшей точностью, чем соответствующие ежегодные данные. Эту проблему не так просто решить, но она может оказаться несущественной.

Если вы используете данные перекрестной выборки и находитесь на стадии планирования исследования, то можно увеличить точность оценок регрессии и ослабить проблему мультиколлинеарности просто за счет большего расхода средств на увеличение размера выборки. Однако такой подход имеет уменьшающуюся предельную отдачу, поскольку стандартные отклонения коэффициентов регрессии обратно пропорциональны величине $\sqrt{n}$, в то время как расходы прямо пропорциональны n.

Столь же важно, если вы используете данные перекрестной выборки и находитесь на стадии планирования исследования, максимизировать дисперсию наблюдений независимых переменных в выборке, например путем расслоения выборки. (Анализ теории и методов организации выборок, см., например, в работах Л. Киша [Kish, 1965] или К. Мозера и Г. Калтона [Moser, Kalton, 1979].)

Далее, можно сократить величину σ_u^2. Случайный член включает в себя объединенный эффект всех переменных, оказывающих влияние на величину y, которые не включены явно в уравнение регрессии. Если вы допускаете мысль

о том, что важная переменная могла быть опущена и, следовательно, оказывает влияние на u, то можно сократить величину σ_u^2, если добавить эту переменную в уравнение регрессии.

Если, однако, новая переменная линейно связана с одной или несколькими переменными, уже включенными в уравнение, то ее введение может еще больше усугубить проблему мультиколлинеарности. Мы вернемся к обсуждению этого вопроса, который представляет большую практическую важность, в конце следующей главы после рассмотрения ошибок спецификации.

Наконец, об использовании самого простого метода. Если вы действительно имеете возможность собрать дополнительные данные, то нужно постараться получить выборку, в которой независимые переменные слабо связаны между собой (конечно, это легче сказать, чем сделать).

Существуют два типа внешней информации, которая может оказаться полезной: теоретические ограничения и внешние эмпирические оценки. Теоретическое ограничение представляет собой допущение, касающееся величины коэффициента или некоторой связи между коэффициентами. Поясним это на примере.

При построении производственной функции с использованием данных временных рядов (как это было сделано в разделе 5.3) следует иметь в виду, что на выпуск продукции, наряду с изменениями в капитальных и трудовых затратах, вероятно, будет оказывать влияние технический прогресс. Если вы имеете дело с агрегированными данными, то невозможно количественно оценить технический прогресс, и проще всего включить экспоненциальный временной тренд в уравнение, записав функцию Кобба—Дугласа, например, в виде:

$$Y = AK^{\alpha}L^{\beta}e^{rt}v, \tag{5.47}$$

где Y, K и L имеют те же определения, что и в разделе 5.3; t — время; r — темп прироста выпуска благодаря техническому прогрессу. Оценив это соотношение по данным табл. 5.1, получим (стандартные ошибки указаны в скобках):

$$\begin{array}{lllll} \widehat{\log Y} = 2{,}81 - 0{,}53 \log K + 0{,}91 \log L + 0{,}047t; & & R^2 = 0{,}97; \\ \quad (1{,}38) \quad (0{,}34) \qquad (0{,}14) \qquad (0{,}021) & & F = 189{,}8. \end{array} \tag{5.48}$$

Со всей очевидностью этот результат показывает, что эластичность выпуска продукции по затратам капитала отрицательна, что означает снижение выпуска при увеличении затрат капитала. Уравнение также показывает темп прироста выпуска продукции за счет технического прогресса порядка 4,7% в год, что является неправдоподобно высокой оценкой для рассматриваемого периода. Здесь можно предположить, что по крайней мере отчасти проблема связана с мультиколлинеарностью, так как коэффициент корреляции между $\log K$ и t составляет 0,997, а стандартная ошибка коэффициента при $\log K$ в 5 раз больше, чем в уравнении без величины t (5.32).

Отсюда появляется желание ввести ограничения на эффект от масштаба, рассматривая его как постоянную величину, что позволит переписать уравнение только с двумя независимыми переменными, имеющими временной тренд, вместо трех и с капиталовооруженностью труда в качестве объясняющей переменной вместо затрат капитала. Этот показатель по-прежнему тесно коррелирован с временем (коэффициент корреляции составляет 0,96), но степень коррелированности уже не так предельно высока. Оценив уравнение (5.28) с экспо-

ненциальным временным трендом, мы получим (стандартные ошибки указаны в скобках):

$$\widehat{\log} Y/L = -0{,}11 + 0{,}11 \log K/L + 0{,}006t; \quad R^2 = 0{,}65; \quad F = 19{,}5. \tag{5.49}$$
$$(0{,}03) \quad (0{,}15) \qquad (0{,}006)$$

Оценки величин α и r, хотя и незначимо отличаются от нуля, теперь более реалистичны, чем раньше, а стандартные ошибки — намного меньше, чем в уравнении (5.48). Тот факт, что величина r незначимо отличается от нуля, подтверждает вывод Ч. Кобба и П. Дугласа о том, что темп увеличения общей производительности факторов в рассматриваемый период был очень низким. Очевидно, что обоснованность этой процедуры зависит от правильности введенного ограничения, поэтому сначала нужно статистически проверить ограничение, что рассматривается в следующей главе.

Наконец, можно использовать внешние оценки. Предположим, что вы решили воспользоваться уравнением (5.24) в качестве формулы для функции спроса, но имеется проблема мультиколлинеарности, так как располагаемый личный доход и цена имеют ярко выраженные временные тренды, а следовательно, тесно коррелированы. Предположим, однако, что вы также имеете перекрестные статистические данные для y и x, полученные из другой выборки. Если допустить, что все домохозяйства в проводимом анализе платили за данный товар одинаковую цену, то модель примет вид:

$$\log y' = \log \alpha' + \beta' \log x' + u'. \tag{5.50}$$

Получив оценку b'_1 для β'_1 при оценивании регрессионной зависимости y' от x', вы можете подставить ее в уравнение (5.24). Теперь определяется новая переменная $\widetilde{\log} y$, равная $(\log y - b'_1 \log x)$, описывающая спрос, скорректированный на изменения дохода. После этого уравнение (5.25) принимает вид:

$$\widetilde{\log} y = \log \alpha + \beta_2 \log p + u. \tag{5.51}$$

Рассчитав $\widetilde{\log} y$ для каждого наблюдения, вы оцениваете его регрессионную зависимость от $\log p$, и, так как здесь имеется только одна независимая переменная, мультиколлинеарность автоматически исключается.

При использовании этого метода могут возникнуть две проблемы, которые необходимо учитывать. Во-первых, оценка величины β_2 зависит от точности оценки величины β'_1, которая, безусловно, подвержена влиянию ошибки выборки. Во-вторых, вы допускаете, что коэффициент при доходе имеет одинаковый смысл для случаев временных рядов и перекрестных выборок, что, конечно, может быть и не так. Для большинства товаров краткосрочная и долгосрочная эластичность спроса по доходу может значительно различаться. Одна из причин этого состоит в том, что характер расходов подвержен влиянию инерции, которое в краткосрочном периоде может превзойти эффекты дохода. Другая причина заключается в том, что изменение уровня дохода может оказать на расходы как непосредственное (в виде изменения бюджетного ограничения), так и косвенное влияние (за счет изменения образа жизни), причем косвенное влияние происходит намного медленнее, чем прямое. В качестве первого приближения обычно считается, что регрессии для временных рядов, особенно с небольшими периодами выборки, дают показатели

краткосрочной эластичности, в то время как регрессии с использованием данных перекрестных выборок дают показатели долгосрочной эластичности. (Более подробно этот и другие связанные с ним вопросы рассматриваются в работе Э. Ку и Дж. Мейера [Kuh, Meyer, 1957, pp. 380–393].)

Упражнение

5.10. Оцените логарифмическую регрессию расходов на выбранный вами продукт, включив в уравнение временной тренд (наряду с доходом и относительной ценой). Есть ли признаки мультиколлинеарности? Улучшились ли результаты?

5.6. Качество оценивания: коэффициент R^2

Как и в парном регрессионном анализе, коэффициент детерминации R^2 определяет долю дисперсии y, объясненную регрессией, и эквивалентно определяется как величина Var $(\hat{y})$/ Var (y), как $\{1 - \text{Var}\,(e)/\text{Var}\,(y)\}$ или как квадрат коэффициента корреляции между y и $\hat{y}$. Этот коэффициент никогда не уменьшается (а обычно он увеличивается) при добавлении еще одной переменной в уравнение регрессии, если все ранее включенные объясняющие переменные сохраняются. Для иллюстрации этого предположим, что вы оцениваете регрессионную зависимость y от x_1 и x_2 и получаете уравнение вида:

$$\hat{y} = a + b_1 x_1 + b_2 x_2. \tag{5.52}$$

Далее, предположим, что вы оцениваете регрессионную зависимость y только от x_1, в результате получив следующее:

$$\hat{y} = a^* + b_1^* x_1. \tag{5.53}$$

Это уравнение можно переписать в виде:

$$\hat{y} = a^* + b_1^* x_1 + 0x_2. \tag{5.54}$$

Если сравнить уравнения (5.52) и (5.54), то коэффициенты в первом из них свободно определялись с помощью метода наименьших квадратов на основе данных для y, x_1 и x_2 при обеспечении наилучшего качества оценки. Однако в уравнении (5.54) коэффициент при x_2 был произвольно установлен равным нулю, и оценивание не будет оптимальным, если только по случайному совпадению величина b_2 не окажется равной нулю, когда оценки будут такими же. (В этом случае величина a^* будет равна a, а величина b_1^* будет равна b_1.) Следовательно, обычно коэффициент R^2 будет выше в уравнении (5.52), чем в уравнении (5.54), и он никогда не станет ниже. Конечно, если новая переменная на самом деле не относится к этому уравнению, то увеличение коэффициента R^2 будет, вероятно, незначительным.

Вы можете решить, что поскольку коэффициент R^2 измеряет долю дисперсии, совместно объясненной независимыми переменными, то можно определить отдельный вклад каждой независимой переменной и таким образом получить меру ее относительной важности. Было бы очень удобно, если бы это стало

возможным. К сожалению, такое разложение невозможно, если независимые переменные коррелированы, поскольку их объясняющая способность будет перекрываться. Эта проблема рассматривается в разделе 6.2.

F-тесты

В разделе 3.10 *F*-тест использовался для анализа дисперсии. Теперь, когда мы используем регрессионный анализ для деления дисперсии зависимой переменной на «объясненную» и «необъясненную» составляющие, можно построить *F*-статистику:

$$F = \frac{ESS / k}{RSS / (n - k - 1)}, \tag{5.55}$$

где *ESS* — объясненная сумма квадратов отклонений; *RSS* — остаточная (необъясненная) сумма квадратов; k — число степеней свободы, использованное на объяснение. С помощью этой статистики можно выполнить *F*-тест для определения того, действительно ли объясненная сумма квадратов больше той, которая может иметь место случайно. Для этого нужно найти критический уровень *F* в колонке, соответствующей k степеням свободы, и в ряду, соответствующем $(n - k - 1)$ степеням свободы, в той или иной части табл. А.3.

Чаще всего *F*-тест используется для оценки того, значимо ли объяснение, даваемое уравнением в целом. Кроме того, с помощью *F*-статистик можно выполнить ряд дополнительных тестов, что также будет рассмотрено ниже.

Уравнение в целом

При осуществлении *F*-теста для уравнения в целом проверяется, превышает ли коэффициент R^2 то значение, которое может быть получено случайно. Проверим, является ли значимой совместная объясняющая способность k независимых переменных; тест для этого может быть описан как проверка нулевой гипотезы:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0. \tag{5.56}$$

В определенном смысле этот тест дополняет *t*-тесты, которые используются для проверки значимости вклада отдельных случайных переменных, когда проверяется каждая из гипотез $\beta_1 = 0, \ldots, \beta_k = 0$.

При расчете *F*-статистики для уравнения в целом, возможно, было бы удобно разделить числитель и знаменатель уравнения (5.55) на *TSS* (общую сумму квадратов), заметив, что *ESS/TSS* равняется R^2, а *RSS/TSS* равняется $(1 - R^2)$. В результате можно записать:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}. \tag{5.57}$$

В табл. 5.6 показан анализ дисперсии, иллюстрирующий *F*-статистику для регрессионной зависимости дохода от образования, стажа работы и возраста, представленной уравнением (5.43).

Таблица 5.6

	Сумма квадратов отклонений (с.к.о.) (млн.)	*Число степеней свободы (с.с.)*	*С.к.о., деленная на с.с.*	*F-статистика*
Объяснено S, X и A	207,49	3	69,16	69,16/5,70=12,1
Остаток	91,18	16	5,70	

Критический уровень F с 3 и 16 степенями свободы при уровне значимости в 1% составляет 5,27; таким образом, F-статистика, равная 12,1, указывает на значимый уровень объяснения.

Дальнейший анализ дисперсии

Помимо проверки уравнения в целом F-тест можно использовать для определения значимости совместного предельного вклада группы переменных. Предположим, что вы сначала оцениваете регрессию с k независимыми переменными и объясненная сумма квадратов составляет ESS_k. Затем вы добавляете еще несколько переменных, доведя их общее число до m, и объясненная сумма квадратов возрастает до ESS_m. Таким образом, вы объяснили дополнительную величину $(ESS_m - ESS_k)$, использовав для этого дополнительные $(m - k)$ степеней свободы, и требуется выяснить, превышает ли данное увеличение то, которое может быть получено случайно.

Вновь используется F-тест, и соответствующая F-статистика может быть описана следующим образом:

$$F = \frac{\text{Улучшение качества уравнения / Число использованных степеней свободы}}{\text{Необъясненная сумма квадратов отклонений / Оставшееся число степеней свободы}}. \quad (5.58)$$

Поскольку RSS_m — необъясненная сумма квадратов отклонений в уравнении со всеми m переменными — равняется $(TSS - ESS_m)$ и RSS_k — необъясненная сумма квадратов отклонений в уравнении с k переменными — равняется $(TSS - ESS_k)$, улучшение качества уравнения при добавлении $(m - k)$ переменных, представленное как разность $(ESS_m - ESS_k)$, записывается в виде выражения $(RSS_k - RSS_m)$. Следовательно, соответствующая F-статистика равна:

$$F = \frac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m-1)}, \quad (5.59)$$

и в соответствии с нулевой гипотезой о том, что дополнительные переменные не увеличивают возможности объяснения уравнения, она распределена с $(m - k)$ и $(n - k - 1)$ степенями свободы. В табл. 5.7 дается анализ таблицы дисперсий для совместного предельного вклада новых переменных.

Например, вернемся к эксперименту по методу Монте-Карло, в котором доход зависит от продолжительности обучения, стажа работы и возраста. Оценка

Таблица 5.7

	Сумма квадратов отклонений (с.к.о.)	*Число степеней свободы (с.с.)*	*С.к.о., деленная на с.с.*	*F-статистика*
Объяснено исходным набором переменных	ESS_k	k	ESS_k/k	$\dfrac{ESS_k / k}{RSS_k / (n-k-1)}$
Остаток	$RSS_k=TSS-ESS_k$	$n-k-1$	$RSS_k/(n-k-1)$	
Объяснено новыми переменными	$ESS_m-ESS_k=RSS_k-RSS_m$	$m-k$	$\dfrac{RSS_k - RSS_m}{m-k}$	
Остаток	$RSS_m=TSS-ESS_m$	$n-m-1$	$RSS_m/(n-m-1)$	$\dfrac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m-1)}$

парной регрессионной зависимости дохода от продолжительности обучения дает *ESS*, равную 90 020 000, *TSS* составила 298 680 000, а *RSS* = 208 650 000 (табл. 5.8).

Критическое значение F с 1 и 18 степенями свободы при уровне значимости в 5% равно 4,41, а при уровне значимости в 1% составляет 8,29. Таким образом, модель, включающая только продолжительность обучения, обеспечивает значимое объяснение при уровне значимости в 5%, но не в 1%.

Если теперь рассмотреть регрессию, включающую также X и A, то можно проверить значимость их совместного предельного вклада. Мы имеем $k = 1$, $m = 3$, и $RSS_m = 91\ 180\ 000$ (см. табл. 5.8). Следовательно, $(RSS_k - RSS_m)$ составляет 117 470 000. Число степеней свободы после добавления X и A равняется 16.

Значение F-статистики равно 10,31, а критическое значение F с 2 и 16 степенями свободы при уровне значимости в 1% составляет 6,23. Таким образом, при добавлении X и A наблюдается значительное улучшение в объяснении дисперсии y.

Таблица 5.8

	Сумма квадратов отклонений (с.к.о.) (млн.)	*Число степеней свободы (с.с.)*	*С.к.о., деленная на с.с.*	*F-статистика*
Объяснено S	90,02	1	90,02	90,02/11,59=7,77
Остаток (кроме S)	208,65	18	11,59	
Объяснено X и A	117,47	2	58,74	58,74/5,70=10,31
Остаток (кроме S, X и A)	91,18	16	5,70	

Зависимость между F- и t-статистиками

Предположим, что вы оцениваете регрессию с несколькими объясняющими переменными, а затем повторяете расчет, отбросив одну из них. Используя разницу в объясненной сумме квадратов, можно выполнить *F*-тест для предельного вклада независимой переменной, которая была отброшена. Можно показать, что такой тест эквивалентен двустороннему *t*-тесту для гипотезы о том, что для этой переменной в первоначальной регрессии $\beta = 0$.

Другими словами, *t*-тесты обеспечивают эффективную проверку предельного вклада каждой переменной при допущении, что все другие переменные уже включены в уравнение.

Если объясняющие способности независимых переменных перекрываются, то предельный вклад в объяснение при добавлении каждой из них может оказаться совсем небольшим. Отсюда вполне возможно, что *t*-тест для каждой переменной окажется незначимым, в то время как *F*-тест для уравнения в целом вполне значим.

Например, рассмотрим вновь эксперимент по методу Монте-Карло (уравнение 5.44), где оценивается регрессионная зависимость дохода (y) от продолжительности обучения (S), стажа работы (X') и возраста (A):

$$\hat{y} = \underset{\text{(с.о.)}}{-7524} + \underset{(4202)}{781S} - \underset{(529)}{207X'} + \underset{(538)}{664A}; \qquad \underset{(476)}{R^2 = 0{,}84}. \tag{5.44}$$

При 16 степенях свободы *t*-тесты показывают, что ни один из коэффициентов не отличается значимо от нуля при уровне значимости в 5%. Тем не менее коэффициент R^2 равен 0,84, и соответствующий *F*-тест значим при уровне значимости в 1%. Результаты оценки регрессии показывают, что совместная объясняющая способность независимых переменных высока, несмотря на тот факт, что не представляется возможным выделить влияние каждой из них. Это неудивительно, поскольку в рассматриваемой модели наблюдалась высокая степень мультиколлинеарности, вызванной почти строгой линейной зависимостью между S, X' и A, а дисперсия случайного члена была большой.

Скорректированный коэффициент R^2

Если вы посмотрите на распечатку уравнений регрессии, то почти наверняка найдете рядом с коэффициентом R^2 показатель, который называют *скорректированным* коэффициентом R^2 (adjusted R^2). Иногда его также называют «исправленным» *коэффициентом* R^2, хотя это определение не означает, по мнению многих, что такой коэффициент улучшен по сравнению с обычным.

Как отмечалось в разделе 5.2, при добавлении объясняющей переменной к уравнению регрессии коэффициент R^2 никогда не уменьшается, а обычно увеличивается. Скорректированный коэффициент R^2, который обычно обозначают $\bar{R}^2$, обеспечивает компенсацию для такого автоматического сдвига вверх путем наложения «штрафа» за увеличение числа независимых переменных. Этот коэффициент определяется следующим образом:

$$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k-1} = \frac{n-1}{n-k-1}R^2 - \frac{k}{n-k-1} = R^2 - \frac{k}{n-k-1}(1-R^2), \qquad (5.60)$$

где k — число независимых переменных. По мере роста k увеличивается отношение $k/(n-k-1)$ и, следовательно, возрастает размер корректировки коэффициента R^2 в сторону уменьшения.

Можно показать, что добавление новой переменной к регрессии приведет к увеличению $\bar{R}^2$, если и только если соответствующая t-статистика больше единицы (или меньше -1). Следовательно, увеличение $\bar{R}^2$ при добавлении новой переменной необязательно означает, что ее коэффициент значимо отличается от нуля. Поэтому отнюдь не следует, как можно было бы предположить, что увеличение $\bar{R}$ означает улучшение спецификации уравнения.

Это является одной из причин того, почему $\bar{R}^2$ не стал широко использоваться в качестве диагностической величины. Другая причина состоит в уменьшении внимания к самому коэффициенту R^2. Ранее среди экономистов наблюдалась тенденция рассматривать коэффициент R^2 в качестве основного индикатора успеха в спецификации модели. Однако на практике, как будет показано в следующих главах, даже плохо определенная модель регрессии может дать высокий коэффициент R^2, и признание этого факта привело к снижению значимости R^2. Теперь он рассматривается в качестве одного из целого ряда диагностических показателей, которые должны быть проверены при построении модели регрессии, и, вероятно, как один из менее важных. Следовательно, и корректировка этого коэффициента мало что дает.

Упражнения

5.11. Величина коэффициента R^2 в логарифмической регрессии между расходами на продукты питания, располагаемым личным доходом и относительной ценой продовольствия (см. уравнение 5.26) составила 0,9867. Проверьте, что критерий F оказался приблизительно равным 820,1 и оцените его значимость (820,1 является фактическим значением критерия F; число, которое вы вычислите на основе коэффициента R^2, будет несколько отличаться от этой величины из-за ошибки округления).

5.12. Проверьте, что критерий F в соответствующей регрессии для выбранного вами товара (см. упражнение 5.6) был правильно вычислен на основе коэффициента R^2, и проверьте его значимость.

5.13. Сумма квадратов отклонений в регрессии в упражнении 5.6 оказалась меньше той, которая была получена в оценке регрессионной зависимости расходов на выбранный вами товар от располагаемого личного дохода в упражнении 4.2. Используйте F-тест для оценки значимости уменьшения указанной суммы. Этот тест эквивалентен некоторому тесту, который вы уже выполняли; объясните, о каком тесте идет речь, и проверьте идентичность сделанных выводов.

6

СПЕЦИФИКАЦИЯ ПЕРЕМЕННЫХ В УРАВНЕНИЯХ РЕГРЕССИИ: ПРЕДВАРИТЕЛЬНОЕ РАССМОТРЕНИЕ

К каким результатам приведет включение в уравнение регрессии переменной, которой там не должно быть? Каковы последствия невключения переменной, которая должна там присутствовать? Что произойдет, если при наличии трудностей в поиске исходных данных вы решите использовать вместо них «заменители»? В данной главе, представляющей собой предварительную попытку решения этих вопросов, основное внимание сосредоточено на последствиях неправильной спецификации переменной. Более сложный предмет — процедура выбора модели — будет затронут в последней главе книги.

В главе показано, каким образом могут быть проверены простейшие ограничения по параметрам. Глава завершается рассмотрением проблем введения переменных с запаздыванием (лагом) и описания фактора времени в моделях, основанных на данных временных рядов.

6.1. Моделирование

Построение экономической модели включает спецификацию составляющих ее соотношений, выбор переменных, входящих в каждое соотношение, а также определение математической функции, представляющей каждое соотношение. Последний элемент был рассмотрен в главе 4 и затем еще раз в главе 5. В данной главе мы рассмотрим второй из вышеперечисленных элементов и будем по-прежнему предполагать, что модель состоит только из одного уравнения. Вопрос о применении регрессионного анализа в моделях, состоящих из систем одновременных уравнений, будет рассмотрен в главе 11.

Если точно известно, какие объясняющие переменные должны быть включены в уравнение при проведении регрессионного анализа, то наша задача — ограничиться оцениванием их коэффициентов, определением доверительных интервалов для этих оценок и т. д. Однако на практике мы никогда не можем быть уверены, что уравнение специфицировано правильно. Экономическая теория должна указывать направление, но теория не может быть совершенной. Не будучи уверенными в ней, мы можем включить в уравнение переменные, которых там не должно быть, и в то же время мы можем не включить другие переменные, которые должны там присутствовать.

Свойства оценок коэффициентов регрессии в значительной мере зависят от правильности спецификации модели. Результаты неправильной спецификации переменных в уравнении могут быть в обобщенном виде выражены следующим образом.

1. Если опущена переменная, которая должна быть включена, то оценки коэффициентов регрессии, вообще говоря, хотя и не всегда, оказываются смещенными. Стандартные ошибки коэффициентов и соответствующие t-тесты в целом становятся некорректными.

2. Если включена переменная, которая не должна присутствовать в уравнении, то оценки коэффициентов регрессии будут несмещенными, однако, вообще говоря (хотя и не всегда), — неэффективными. Стандартные ошибки будут в целом корректны, но из-за неэффективности регрессионных оценок они будут излишне большими.

Мы начнем с рассмотрения этих двух случаев, а затем перейдем к более широким аспектам спецификации модели.

6.2. Влияние отсутствия в уравнении переменной, которая должна быть включена

Проблема смещения

Предположим, что переменная y зависит от двух переменных x_1 и x_2 в соответствии с соотношением:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u, \tag{6.1}$$

однако вы не уверены в значимости x_2. Считая, что модель должна выглядеть как

$$y = \alpha + \beta_1 x_1 + u_1, \tag{6.2}$$

вы оцениваете регрессию

$$\hat{y} = a + b_1 x_1 \tag{6.3}$$

и вычисляете b_1 по формуле $\mathrm{Cov}(x_1, y)/\mathrm{Var}(x_1)$ вместо правильного выражения, данного в уравнении (5.12). По определению, b_1 является несмещенной оценкой величины β_1, если $E(b_1)$ равняется β_1. Практически, если соотношение (6.1) верно, то

$$E\left\{\frac{\mathrm{Cov}(x_1, y)}{\mathrm{Var}(x_1)}\right\} = \beta_1 + \beta_2 \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)}. \tag{6.4}$$

Сначала мы дадим интуитивное объяснение этого, а затем — формальное доказательство.

В разделе 5.2 показано, что если опустить x_2 в регрессионном соотношении, то переменная x_1 будет играть двойную роль: отражать свое прямое влияние и

заменять переменную x_2 в описании ее влияния. Данное кажущееся опосредованное влияние величины x_1 на y будет зависеть от двух факторов: от видимой способности x_1 имитировать поведение x_2 и от влияния величины x_2 на y.

Кажущаяся способность переменной x_1 объяснять поведение x_2 определяется коэффициентом наклона h в псевдорегрессии:

$$\hat{x}_2 = g + hx_1. \tag{6.5}$$

Величина h, естественно, рассчитывается при помощи обычной формулы для парной регрессии, в данном случае $\text{Cov}(x_1, x_2)/\text{Var}(x_1)$. Влияние величины x_2 на y определяется коэффициентом β_2. Таким образом, эффект имитации посредством величины β_2 может быть записан как $\beta_2\text{Cov}(x_1, x_2)/\text{Var}(x_1)$. Прямое влияние величины x_1 на y описывается с помощью β_1. Таким образом, при оценивании регрессионной зависимости y от переменной x_1 (без включения в нее переменной x_2) коэффициент при x_1 определяется формулой:

$$\beta_1 + \beta_2\text{Cov}(x_1, x_2)/\text{Var}(x_1) + \text{Ошибка выборки}. \tag{6.6}$$

При условии, что величина x_1 не является стохастической, ожидаемым значением коэффициента будет сумма первых двух членов этой формулы. Присутствие второго слагаемого предполагает, что математическое ожидание коэффициента будет отличаться от истинной величины β_1, другими словами, оценка будет смещенной.

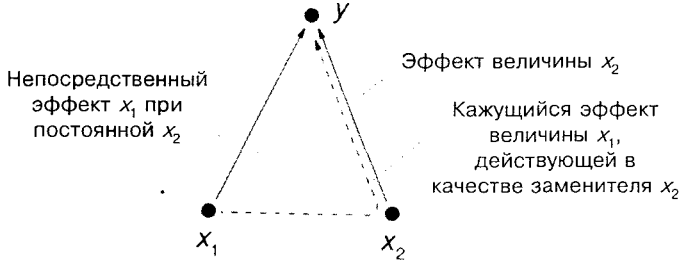


Рис. 6.1

Формальное доказательство соотношения (6.4) не представляет труда. Выполним ряд теоретических преобразований оценки b_1:

$$b_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} = \frac{\text{Cov}(x_1, [\alpha + \beta_1 x_1 + \beta_2 x_2 + u])}{\text{Var}(x_1)} =$$

$$= \frac{1}{\text{Var}(x_1)}[\text{Cov}(x_1, \alpha) + \text{Cov}(x_1, \beta_1 x_1) + \text{Cov}(x_1, \beta_2 x_2) + \text{Cov}(x_1, u) =$$

$$= \frac{1}{\text{Var}(x_1)}[0 + \beta_1\text{Var}(x_1) + \beta_2\text{Cov}(x_1, x_2) + \text{Cov}(x_1, u)] =$$

$$= \beta_1 + \beta_2\frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)} + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}. \tag{6.7}$$

Если величины x_1 и x_2 являются нестохастическими, то при вычислении математического ожидания величины b_1 первые два члена в уравнении (6.7) ос-

таются неизменными, а третий будет равен нулю. Отсюда мы получаем формулу (6.4).

Этим подтверждается наш интуитивный вывод, что b_1 смещена на величину, равную $\beta_2 \text{Cov}(x_1, x_1)/\text{Var}(x_1)$. Направление смещения будет зависеть от знака величин β_2 и $\text{Cov}(x_1, x_2)$. Например, если β_2 положительна, а также положительна ковариация, то смещение будет положительным, а b_1 будет в среднем давать завышенные оценки β_1. Самостоятельно вы можете рассмотреть и другие случаи.

Есть, однако, один исключительный случай, когда оценка b_1 остается несмещенной. Это случается, когда выборочная ковариация между x_1 и x_2 в точности равняется нулю. Если $\text{Cov}(x_1, x_2) = 0$, то смещение исчезает. Действительно, коэффициент, полученный с использованием парной регрессии, будет точно таким же, как если бы вы оценили правильно специфицированную множественную регрессию. Конечно, величина смещения здесь равнялась бы нулю и при $\beta_2 = 0$, но в этом случае неправильной спецификации не возникает.

Неприменимость статистических тестов

Другим серьезным следствием невключения переменной, которая на самом деле должна присутствовать в регрессии, является то, что формулы для стандартных ошибок коэффициентов и тестовые статистики, вообще говоря, становятся неприменимыми. Это, разумеется, означает, что, основываясь на полученных результатах оценки регрессии, в принципе нельзя заниматься проверкой каких-либо гипотез.

Иллюстрация, основанная на методе Монте-Карло

Проведенный нами анализ проиллюстрируем при помощи эксперимента, являющегося одной из вариаций метода Монте-Карло, рассмотренного в разделе 5.4. Предположим, что доход в какой-то стране определяется продолжительностью обучения (S), индексом интеллекта (IQ) и степенью удачи. К основному доходу, составляющему 10 000, добавляется по 1500 за каждый год обучения сверх минимальных 10 лет и по 250 за каждый балл IQ свыше 85. Кроме того, имеется еще фактор удачи (u):

$$y = 10\,000 + 1500\,(S - 10) + 250\,(IQ - 85) + u. \tag{6.8}$$

После упрощения это уравнение становится таким:

$$y = -26250 + 1500S + 250IQ + u. \tag{6.9}$$

Первые три колонки в табл. 6.1 представляют данные для воображаемой выборки из 20 человек. Значения S и IQ выбраны произвольно, но они оказываются положительно коррелированными. Положительная корреляция этих величин наблюдается во многих странах, и одним из объяснений (но ни в коем случае не единственным) является то, что студенты с большими способностями чаще выдерживают конкурсные экзамены, определяющие рейтинг для допуска к продолжению образования. Значения величины u были определены путем получе-

ния выборки из 20 наблюдений нормально распределенной случайной величины с нулевой средней и единичной дисперсией и умножения каждого наблюдения на 2000. В табл. 6.1 показаны также итоговые значения величины y, полученные по формуле (6.9).

Исследователь изучает факторы, определяющие доход в данной стране, не подозревая важности величины IQ, и оценивает парную регрессионную зависимость дохода от продолжительности обучения в годах:

$$\hat{y} = a + b_1 S. \tag{6.10}$$

Таблица 6.1

Индивид	*Эксперимент 1*				*Эксперимент 2*	
	S	*IQ*	*u*	*y*	*IQ'*	*y'*
1	10	95	1380	13880	100	15130
2	10	100	1560	15310	120	20310
3	10	100	-3280	10470	105	11720
4	11	105	780	17280	100	16030
5	11	85	980	12480	125	22480
6	11	115	-340	18660	100	14910
7	11	95	720	14720	115	19720
8	12	100	2640	19390	100	19390
9	12	100	-1240	15510	105	16760
10	12	110	340	19590	95	15840
11	13	90	20	15770	90	15770
12	13	120	-460	22790	105	19040
13	14	110	-1340	20910	100	18410
14	14	95	-1780	16720	95	16720
15	15	105	700	23200	95	20700
16	16	110	-560	24690	100	22190
17	16	100	380	23130	90	20630
18	17	125	4440	34940	105	29940
19	18	105	780	27780	85	22780
20	19	105	1880	30380	100	29130

Исследователь получает результат:

$$\hat{y} = -6418 + 1985S; \qquad R^2 = 0{,}78. \tag{6.11}$$
$$\text{(с. о.)} \quad (3349) \quad (248)$$

К несчастью для исследователя, величины S и IQ коррелированы. Для данной выборки выражение Cov (S, IQ)/Var (S) равно 1,29. Таким образом,

$$E(b_1) = \beta_1 + \beta_2 \frac{\text{Cov}(S, IQ)}{\text{Var}(S)} = 1500 + 250 \times 1{,}29 = 1823. \tag{6.12}$$

Поскольку исследователь не включил в уравнение величину IQ, то оценка коэффициента при S будет иметь положительное смещение на 323. Конечно, фактически полученная оценка может равняться 1823, но это будет просто совпадением, если только фактор удачи примет нулевое значение. Мы видим, что исследователь фактически получил несколько более высокую оценку, равную 1985. Различие объясняется влиянием остаточного члена в данной выборке.

Если бы исследователь включил в уравнение регрессии переменную IQ, то результат оценивания для той же выборки получился бы следующим:

$$\hat{y} = -29\,586 + 1640S + 268IQ; \qquad R^2 = 0{,}93. \tag{6.13}$$
$$\text{(с. о.)} \quad (4155) \quad (151) \quad (43)$$

Полученные исследователем оценки коэффициентов были бы несмещенными и, по крайней мере в данном случае, существенно более близкими к их истинным значениям.

Очевидно, что как полученное исследователем уравнение регрессии, так и уравнение, составленное с использованием правильной спецификации, зависят от фактических значений случайного члена в выборке, и было бы несправедливо придавать большой вес одному эксперименту, даже если он дает предсказуемые результаты. В соответствии с этим данный эксперимент был проведен еще 9 раз с использованием тех же значений величин S и IQ в каждом наблюдении и тех же значений величин α, β_1 и β_2, но с различными наборами случайных реализаций остаточного члена.

Результаты оценивания соответствующих регрессий даны в обобщенном виде в табл. 6.2. Из этой таблицы можно видеть, что полученные результаты подтверждают наши прежние выводы. Исследователь получает оценки коэффициентов при S, которые произвольно разбросаны около смещенного числа 1823 (их среднее значение равно 1854). При правильной спецификации оценки разбросаны вокруг истинного значения, равного 1500. Такие же замечания могут быть сделаны относительно постоянного члена уравнения.

А что бы произошло, если бы исследователь вместо величины IQ не включил в уравнение регрессии переменную S? В этом случае величина IQ частично действовала бы в качестве переменной сама по себе и отчасти в качестве заменителя отсутствующей переменной S. Повторением проведенного выше анализа, можно показать, что ее коэффициент был бы смещен на величину β_1Cov (S, IQ)/Var (IQ). Поскольку $\beta_1 = 1500$ и Cov (S, IQ)/Var (IQ) = 0,104, то коэффициент был бы смещен вверх на величину, равную 156, и его математическое ожидание составило бы 406. Такой вывод подкрепляется оценивани-

ем регрессии с использованием данных из первой части табл. 6.1. В результате получим:

$$\hat{y} = -25488 + 438IQ; \qquad R^2 = 0{,}47. \tag{6.14}$$
$$\text{(с.о.)} \quad (11362) \quad (109)$$

Таблица 6.2

Эксперимент	Спецификация исследователя				Правильная спецификация					
	Константа	с.о.	S	с.о.	Константа	с.о.	S	с.о.	IQ	с.о.
1	-6418	3349	1985	248	-29586	4155	1640	151	268	43
2	-6576	2718	1979	201	-19269	5067	1790	184	147	52
3	-1729	3880	1642	287	-27713	5150	1255	187	301	53
4	-1788	4070	1541	301	-31621	1097	1097	155	345	44
5	-3752	3796	1774	281	-30785	4371	1372	158	313	45
6	-7083	3463	2052	256	-32992	3352	1667	121	300	35
7	-7147	3038	1999	225	-22978	5330	1763	193	183	55
8	-6124	4075	1986	301	-34437	5005	1565	181	327	51
9	631	3556	1465	263	-21634	5254	1134	190	258	54
10	-8379	3263	2112	241	-30089	4386	1789	159	251	45

Направление смещения

Возвращаясь к общему случаю, мы видим, что если истинная модель выражается формулой (6.1), где y — функция переменных x_1 и x_2, и если в уравнении регрессии опустить x_2, то коэффициент при x_1 смещается на величину, равную $\beta_2 \text{Cov}(x_1, x_2)/\text{Var}(x_1)$. Поскольку величина $\text{Var}(x_1)$ не может быть отрицательной, то направление смещения определяется знаками величин β_2 и $\text{Cov}(x_1, x_2)$. В примере с экспериментом по методу Монте-Карло величина β_2 была положительной, а S и IQ имели положительную корреляцию, поэтому смещение оказалось положительным, а невключение переменной IQ привело к систематическому завышению коэффициента при S. Это, однако, не должно означать, что смещения обязательно являются положительными. Если β_2 отрицательна или же отрицательна ковариация между x_1 и x_2, то смещение будет отрицательным. Естественно, что если обе эти величины отрицательны, то смещение в результате будет положительным.

Проиллюстрируем это при помощи следующего эксперимента по методу Монте-Карло, являющегося модификацией предыдущего эксперимента. Мы используем ту же модель определения размера дохода [уравнение (6.9)], те же данные для S и случайных реализаций u, но другие данные для IQ. Предположим теперь, что мы находимся в стране, где индивиды с наиболее низкими

способностями, определяемыми величиной IQ, проходят самое длительное обучение для достижения ими одинакового со всеми уровня образования. Новые данные по величине IQ, обозначенные как IQ', даны во второй части табл. 6.1. Из таблицы видно, что величины S и IQ' отрицательно коррелированы.

Таблица 6.3

Объясняющие переменные	*Оценки (стандартные ошибки)*			
	Константа	b_1	b_2	R^2
S, IQ'	−35864	1776	312	0,89
	(5808)	(155)	(44)	
S	3309	1213	—	0,55
	(3483)	(258)		
IQ'	13875	—	54	0,01
	(11070)		(109)	

В табл. 6.3 приведены результаты оценивания регрессий: множественной регрессии (правильно специфицированной) и двух парных регрессий (неправильно специфицированных). Коэффициенты при S и IQ' в парных регрессиях значительно ниже соответствующих значений коэффициентов множественной регрессии. Если опустить переменную IQ', то смещение коэффициента регрессии при S будет равным $\beta_2 \text{Cov}(S, IQ')/\text{Var}(S) = 250 \times (-1{,}80) = -450$. При невключении переменной S смещение коэффициента при IQ' составит $\beta_1 \text{Cov}(S, IQ')/\text{Var}(IQ') = 1500 \times (-0{,}145) = -217$. Очевидно, что результаты оценивания регрессии подтверждают наши выводы.

Интуитивное объяснение говорит, что в соответствии с данной моделью индивиды с длительными сроками обучения имеют за счет этого относительно высокий доход, но они в то же время обычно обладают более низким IQ, а это уменьшает их доход. Следовательно, в парной регрессионной зависимости дохода от фактора продолжительности обучения эффект последнего недооценивается. Точно так же индивиды с высоким IQ получают вследствие этого относительно высокие доходы, но они в то же время, как правило, имеют относительно короткий срок обучения, что сокращает их преимущества. Отсюда в парной регрессионной зависимости дохода от величины IQ влияние IQ недооценивается.

Иногда это смещение бывает достаточно сильным для того, чтобы заставить коэффициент регрессии сменить знак. Допустим, что в рассматриваемой модели истинный коэффициент при IQ' был равен 25 вместо 250. Используя те же, что и раньше, данные по S и IQ', получим смещение коэффициента при IQ', равное −217, если переменная S опущена. Отсюда математическое ожидание коэффициента при IQ', равное $25 - 217 = -192$, вместо положительного станет отрицательным. В табл. 6.4 представлены результаты оценивания правильно и неправильно специфицированных регрессий. В третьей регрессии коэф-

фициент при IQ' действительно отрицателен (−171). (Это несколько больше, чем математическое ожидание, и расхождение здесь объясняется наличием в модели случайного члена.)

Поведение значения коэффициента R^2 при невключении объясняющей переменной

В разделе 5.6 указывалось, что при анализе множественной регрессии невозможно определить вклад каждой объясняющей переменной в величину коэффициента R^2, и сейчас мы поясним, почему это так.

Таблица 6.4

Объясняющие переменные	*Оценки (стандартные ошибки)*			
	Константа	b_1	b_2	R^2
S, IQ'	−35864	1776	87	0,90
	(5808)	(155)	(44)	
S	−24898	1618	—	0,88
	(1942)	(144)		
IQ'	13875	—	−177	0,12
	(11070)		(109)	

Сначала мы рассмотрим данную проблему, используя эксперимент по методу Монте-Карло «доход—обучение—интеллект», где продолжительность обучения и величина *IQ* положительно коррелированы. Мы видели, что при оценивании регрессионной зависимости величины *Y* только от величины *S*, значение коэффициента R^2 было равно 0,78; при оценивании регрессионной зависимости *Y* только от величины *IQ* значение коэффициента R^2 равнялось 0,47. Означает ли это, что величина *S* объясняет 78% дисперсии дохода, а величина *IQ* — 47%? Конечно, нет, так как это подразумевало бы, что вместе они объясняли бы 125% дисперсии, что невозможно. Практически их совместная объясняющая способность, выраженная коэффициентом R^2 во множественной регрессии [уравнение (6.13)], составляет 0,93.

Объяснение состоит в том, что в парной регрессии между доходом и продолжительностью обучения величина *S* играет собственную роль и отчасти роль заменителя отсутствующей переменной *IQ* (рис. 6.1). Следовательно, коэффициент R^2 для данной регрессии отражает общую объясняющую способность величины *S* в обеих этих ролях, а не непосредственную объясняющую способность переменной *S*. Отсюда число 0,78 является завышенной оценкой последней. Аналогично переменная *IQ* в парной регрессии между доходом и показателем уровня интеллекта *IQ* отчасти заменяет отсутствующую переменную *S*, и уровень коэффициента R^2 в этой регрессии отражает общую объясняющую способ-

ность величины IQ в обеих указанных ролях, а не просто объясняющую способность самой величины IQ.

В данном эксперименте по методу Монте-Карло уровни коэффициента R^2, наблюдавшиеся в простой регрессии, увеличиваются за счет эффекта замещения. В эксперименте по методу Монте-Карло, результаты которого приведены в табл. 6.3, происходит обратное. В этом эксперименте имела место отрицательная корреляция между величинами S и IQ'. В результате этого коэффициенты при переменных в парных регрессиях оказались смещенными в сторону занижения. Была подорвана также и их кажущаяся объясняющая способность. Переменная S объясняла только 55% дисперсии дохода, а IQ' — только 1%. Таким образом, вместе они объясняли только 56% дисперсии. В то же время коэффициент R^2 в правильно специфицированной множественной регрессии показывает, что их совместная объясняющая способность фактически составляла 89%.

В парной регрессии между доходом и величиной IQ' этот эффект был особенно резким. Разрушающий эффект отсутствия переменной S в функции почти уравновесил прямое влияние переменной IQ', в результате чего коэффициент регрессии составил лишь малую часть истинной величины, а кажущаяся объясняющая способность, составившая ничтожный 1%, здесь явно преуменьшена.

Таблица 6.4 иллюстрирует иной вариант исхода, который является вполне обычным. Здесь снова имеет место отрицательная корреляция между S и IQ', однако истинный коэффициент при переменной IQ' равнялся всего лишь 25 вместо 250 прежде. В действительности величина S явилась «ответственной» за большую часть дисперсии дохода, а уровень коэффициента R^2 в парной регрессии между величиной y и S почти так же высок, как и коэффициента R^2 во множественной регрессии. В парной регрессии между доходом и величиной IQ' разрушающий эффект отсутствия переменной S доминирует над прямым влиянием величины IQ'. Результатом этого является то, что последняя имеет отрицательный коэффициент при довольно высоком значении коэффициента R^2, но данное значение R^2 в основном можно объяснить тем, что переменная IQ' выполняет роль заменителя отсутствующей переменной S.

Упражнения

6.1. Предположив, что множественная регрессия [уравнение (5.3)] между расходами на питание (y), располагаемым личным доходом (x) и относительной ценой (p) правильно специфицирована, определите направление смещения коэффициента при другой переменной, если не включена: 1) переменная p и 2) переменная x. Воспользуйтесь тем, что относительная цена продовольствия в течение выборочного периода слегка возросла и, таким образом, p и x положительно коррелированы. В таблице приводятся результаты оценивания множественной и парной регрессий:

Объясняющие переменные	Оценки коэффициентов (стандартные ошибки)			R^2
	Константа	x	p	
x, p	116,7 (9,6)	0,112 (0,003)	-0,739 (0,114)	0,99
x	55,3 (2,4)	0,093 (0,003)	—	0,98
p	-125,9 (42,1)	—	2,462 (0,407)	0,62

Проверьте, подтверждают ли эти результаты ваши выводы, и дайте комментарии относительно уровня коэффициента R^2 в этих трех регрессиях.

6.2. В таблице приведены в обобщенном виде логарифмические аналоги указанных выше трех регрессий спроса:

Объясняющие переменные	Оценки коэффициентов (стандартные ошибки)			R^2
	Константа	log x	log p	
log x, log p	2,82 (0,42)	0,64 (0,03)	-0,48 (0,12)	0,99
log x	1,20 (0,11)	0,55 (0,02)	—	0,98
log p	-4,62 (1,52)	—	2,04 (0,33)	0,63

Прокомментируйте различия в коэффициентах трех указанных уравнений.

6.3. В упражнении 2.4 вы построили парную регрессию между расходами на выбранный вами вид благ (y) и располагаемым личным доходом (x), а в упражнении 5.3 — множественную регрессию между величинами y, x и p — ценой вашего блага относительно общего уровня инфляции. Постройте теперь парную регрессию только между величиной y и показателем p. Представьте результаты этих трех регрессий в форме, использованной в упражнении 6.1, и прокомментируйте вариации коэффициентов при x и p, а также уровней коэффициента R^2. Укажите, в частности, направление смещения, которое вы ожидали бы в этих же регрессиях при правильной множественной спецификации, принимая во внимание тенденцию (если таковая существует) в области относительных цен за тот же период, что и в упражнении 5.1, и используя уравнение (6.4).

6.4. Повторите упражнение 6.3, используя логарифмические функции спроса вместо линейных.

6.5. В таблице приведены данные (в млн. ф. ст. в постоянных ценах 1975 г.) по расходам на табак (y) и по располагаемым личным доходам (x) для Великобритании за период 1962–1981 гг. Переменная времени t определялась так: $t = 1$ в 1962 г., $t = 2$ в 1963 г. и т. д.

Год	t	y	x	*Год*	t	y	x
1962	1	2701	51484	1972	11	2747	70214
1963	2	2787	53684	1973	12	2918	75059
1964	3	2753	55754	1974	13	2885	74049
1965	4	2652	56970	1975	14	2748	74005
1966	5	2737	58278	1976	15	2653	73437
1967	6	2753	59226	1977	16	2523	72288
1968	7	2740	60367	1978	17	2746	78259
1969	8	2707	60576	1979	18	2731	83666
1970	9	2702	62485	1980	19	2685	84771
1971	10	2605	64544	1981	20	2492	82903

Следующие регрессии были построены с использованием данных из предыдущей таблицы:

Объясняющие переменные	*Оценки коэффициентов (стандартные ошибки)*			R^2
	Константа	x	t	
x, t	1257 (351)	0,031 (0,007)	−57,7 (12,6)	0,56
x	2794 (152)	−0,001 (0,002)	—	0,02
t	2763 (46)	—	−4,8 (3,8)	0,08

Прокомментируйте вариации коэффициентов при x и t, а также уровней коэффициента R^2. В какой степени данные из таблицы непосредственно подтверждают ваши выводы? (Тщательно рассмотрите отдельные наблюдения.)

6.6. В эксперименте по методу Монте-Карло, отраженном в табл. 6.4, истинное значение коэффициента при S было равным 1500. В парной регрессии между доходами и длительностью обучения коэффициент при переменной S дол-

жен быть смещенным вниз. Как бы вы отнеслись к фактической расчетной величине (1618), которая *превышает* истинное значение?

6.7. В эксперименте по методу Монте-Карло, описанном в разделе 5.4, доходы (Y) определялись длительностью обучения (S), стажем работы (X) и возрастом (A). Величины X и A были положительно коррелированными, а S была коррелирована отрицательно с каждой из них. Левая часть приведенной ниже таблицы показывает уровни коэффициента R^2, когда была оценена зависимость: 1) только от величины S; 2) от величин S и X; 3) от величин S, X и A. В правой части таблицы даны уровни коэффициента R^2, когда была оценена регрессионная зависимость: 1) только от A; 2) от A и X; 3) от A, X и S.

S	0,301	A	0,189
S, X	0,688	A, X	0,213
S, X, A	0,695	A, X, S	0,695

Объясните, почему видимая объясняющая способность переменной A, когда она в качестве дополнительной переменной была включена в уравнение последней, оказалась меньшей, чем при добавлении ее первой, в то время как соответствующая объясняющая способность переменной S больше, когда она включается в уравнение последней.

6.8. Допустим, что величина y определяется величинами x_1 и x_2 в соответствии с уравнением (6.1) и что $\mathrm{Cov}(x_1, x_2)$ равна нулю. Используйте это для упрощения формулы (5.12) вычисления коэффициента множественной регрессии b_1 и покажите, что она сводится к выражению для парной регрессии. Как бы вы выбрали здесь спецификацию уравнения регрессии и почему?

6.3. Влияние включения в модель переменной, которая не должна быть включена

Допустим, что истинная модель представляется в виде:

$$y = \alpha + \beta_1 x_1 + u, \tag{6.15}$$

а вы считаете, что ею является

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u, \tag{6.16}$$

и рассчитываете оценку величины b_1, используя формулу (5.12) вместо выражения $\mathrm{Cov}(x_1, y)/\mathrm{Var}(x_1)$.

В целом проблемы смещения здесь нет, даже если b_1 будет рассчитана неправильно. Величина $E(b_1)$ остается равной β_1, но в общем оценка будет неэффективной. Она будет более неустойчивой, в смысле наличия большей дисперсии относительно β_1, чем при правильном вычислении. Это проиллюстрировано на рис. 6.2.

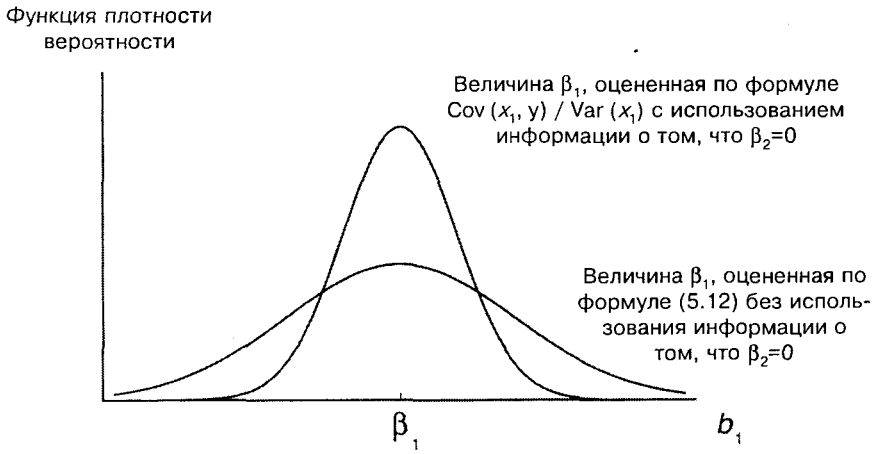


Рис. 6.2

Это можно легко объяснить интуитивно. Истинная модель может быть записана в виде:

$$y = \alpha + \beta_1 x_1 + 0x_2 + u. \tag{6.17}$$

Таким образом, если вы строите регрессионную зависимость y от x_1 и x_2, то b_1 будет являться несмещенной оценкой величины β_1, а b_2 будет несмещенной оценкой нуля (при выполнении условий Гаусса—Маркова). Практически вы обнаруживаете для себя, что β_2 равно нулю. Если бы вы заранее поняли, что β_2 равно нулю, то могли бы использовать эту информацию для исключения x_2 и применить парную регрессию, которая в данном случае является более эффективной.

Утрата эффективности в связи со включением x_2 в случае, когда она не должна была быть включена, зависит от корреляции между x_1 и x_2. Сравните дисперсии величины b_1 при построении парной и множественной регрессии (табл. 6.5).

Дисперсия в общем окажется большей при множественной регрессии, и разница будет тем большей, чем ближе коэффициент корреляции к единице или −1. Единственным исключением в связи с проблемой утраты эффективности является вариант, когда коэффициент корреляции точно равен нулю. В этом случае оценка b_1 для множественной регрессии совпадает с оценкой для парной регрессии. Доказательство этого опустим, поскольку оно довольно простое.

Таблица 6.5

Парная регрессия	*Множественная регрессия*
$\text{pop.var}(b_1) = \dfrac{\sigma_u^2}{n\text{Var}(x_1)}$	$\text{pop.var}(b_1) = \dfrac{\sigma_u^2}{n\text{Var}(x_1)} \times \dfrac{1}{1 - r_{x_1,x_2}^2}$

В выводе о несмещенности есть одно исключение, которое необходимо иметь в виду. Если величина x_2 коррелирует с u, то коэффициенты регрессии будут в конечном счете смещенными. Если модель записать как уравнение (6.17), то это будет означать, что четвертое условие Гаусса—Маркова применительно к величине x_2 не выполняется.

Иллюстрация, основанная на эксперименте по методу Монте-Карло

В эксперименте по методу Монте-Карло, описанном в разделе 6.2, исследователь переоценил влияние образования на доход из-за того, что он не учел зависимости дохода в данной стране от величины IQ и того обстоятельства, что величина S там отчасти играла роль замещающей переменной для IQ в неправильно специфицированном уравнении парной регрессии. Будем помнить об этом и предположим, что наш исследователь, являющийся уже экспертом в данном вопросе, приглашен в качестве консультанта для проведения аналогичного исследования в соседней стране.

Может оказаться, что в новой стране подход более формален, чем в первой, и доход здесь определяется только образованием (и удачей) без учета способностей как таковых. Пусть базовый доход здесь снова равен 10 000, с добавлением 2000 за каждый год учебы сверх минимальных 10 лет, плюс (или минус) некоторая величина, зависящая от фактора удачи. Истинным соотношением поэтому будет:

$$y = 10\,000 + 2000\,(S - 10) + u = -10\,000 + 2000S + u. \tag{6.18}$$

Исследователь снова делает выборку из 20 человек, и по удивительному совпадению все они имеют одинаковые характеристики, показанные в первой части табл. 5.2. В этом случае имеются также данные о величине IQ. Считая, что включение величины IQ в уравнение регрессии не причинит вреда, исследователь проводит эту операцию и получает следующее соотношение (стандартные ошибки указаны в скобках):

$$\hat{y} = \underset{(4155)}{-13\,336} + \underset{(151)}{2140S} + \underset{(43)}{18IQ}. \tag{6.19}$$

Результат действительно неплохой. 95-процентный доверительный интервал для константы включает в себя ее истинное значение −10 000, и аналогичный интервал для S включает значение 2000. Таким образом, полученные оценки незначимо отличаются от истинных величин с 5-процентным уровнем значимости. Точно так же коэффициент IQ незначимо отличается от нуля.

Если бы при этом была использована правильная спецификация, то результатом было бы:

$$\hat{y} = \underset{(\text{с.о.})}{-11\,782} + \underset{(137)}{2163S}. \tag{6.20}$$

Оценка константы здесь лучше, однако оценка коэффициента при переменной S недостаточно хороша (влияние фактора удачи оказалось относительно незначительным).

И вновь здесь нельзя слишком полагаться на результаты одного эксперимента.

Таблица 6.6

Эксперимент	Правильная спецификация				Спецификация исследователя					
	Константа	с.о.	S	с.о.	Константа	с.о.	S	с.о.	IQ	с.о.
1	-11781	1851	2163	137	-13336	4155	2140	151	18	43
2	-11940	2490	2157	184	-3019	5067	2290	184	-103	52
3	-7092	2342	1820	173	-11463	5150	1755	187	51	53
4	-7152	2138	1720	158	-15371	4273	1597	155	95	44
5	-9116	2044	1952	151	-14535	4371	1872	158	63	45
6	-12446	1573	2230	116	-16742	3352	2167	121	50	35
7	-12510	2462	2177	182	-6727	5329	2263	193	-67	55
8	-11487	2361	2164	175	-18187	5005	2065	181	77	52
9	-4733	2329	1644	172	-5384	5354	1634	190	8	54
10	-13742	1943	2290	144	-13839	4386	2289	159	1	45

В табл. 6.6 сведены вместе результаты повторения еще девяти таких же экспериментов с изменением в каждой выборке только значений случайной составляющей. Из табл. 6.6 можно сделать следующие выводы.

1. Результаты исследователя не выглядят смещенными, даже если спецификация является неправильной. Оценка константы колеблется около −10 000, а оценка коэффициента величины S — около 2000. (Естественно, что результаты оценивания правильной спецификации тоже будут несмещенными.)

2. Результаты оценивания правильной спецификации в целом более точны, поскольку эта спецификация оказывается более эффективной. Но данное утверждение не всегда верно, и в ряде случаев неправильная спецификация дает результат ближе к истине. Причиной этого является то, что относительная неэффективность спецификации исследователя зависит от корреляции между S и IQ, а корреляция оказывается не достаточно тесной (в выборке из табл. 6.1), чтобы вызвать большие расхождения с истинными значениями.

3. Более высокая эффективность правильной спецификации должна отражаться в меньших стандартных ошибках, и это в целом действительно подтверждается.

4. Оценки коэффициентов при IQ в спецификации исследователя в целом незначительно отличны от нуля. В эксперименте *4* имеется единственное отклонение, когда истинная гипотеза о нулевом значении отвергается при 5-процентном уровне значимости. Это является хорошим примером ошибки I рода (см. Обзор).

6.9. Социолог считает, что уровень активности в «теневой» экономике (Y_t) зависит либо положительно от уровня налогового бремени (X_t), либо отрицательно от активности государства в стремлении сделать невыгодной деятельность в сфере «теневой» экономики (Z_t). Величина Y_t может зависеть также от X_t и Z_t одновременно. Имеются годовые данные временного ряда за 20 лет, где величины Y_t, X_t и Z_t измерены в одних и тех же единицах. Социолог строит регрессионные зависимости: 1) Y_t только от величины X_t; 2) Y_t только от величины Z_t; 3) Y_t от обеих величин X_t и Z_t, применительно к каждому городу со следующими результатами (в скобках даны стандартные ошибки).

		Оценки коэффициентов		
	Константа	X_t	Z_t	R^2
Город *A*				
1	315,7 (18,5)	1,54 (0,97)	—	0,12
2	128,6 (50,9)	—	−0,96 (0,06)	0,94
3	218,0 (76,6)	2,85 (0,25)	−1,21 (0,03)	0,99
Город *B*				
1	197,6 (16,8)	2,86 (0,25)	—	0,88
2	512,2 (202,6)	—	−0,05 (0,08)	0,02
3	230,8 (82,5)	2,94 (0,27)	−0,01 (0,03)	0,88

Произведя соответствующую статистическую проверку, напишите краткий отчет с рекомендацией социологу о том, как интерпретировать эти результаты.

6.10. Проведите эксперимент по методу Монте-Карло, по аналогии с экспериментами данного и предыдущего разделов, исследовав сначала эффекты невключения переменной, которая должна быть включена в уравнение, а затем со включением переменной, которой там не должно быть. При желании используйте модель «доход—образование—*IQ*», изменяя при этом коэффициенты, но при достаточном воображении можно изменить и саму модель. (Данное упражнение, вероятно, потребует определенной помощи со стороны преподавателя.)

6.4. Замещающие переменные

Часто бывает, что вы не можете найти данных по переменной, которую хотелось бы включить в уравнение регрессии. Некоторые переменные, относящиеся к социально-экономическому положению или к качеству образования, имеют такое расплывчатое определение, что их в принципе даже невозможно измерить. Другие могут поддаваться измерению, но оно требует столько времени и энергии, что на практике их приходится отбрасывать. Иногда вы можете быть расстроены тем, что пользуетесь какими-то данными, собранными другим человеком, в которых (с вашей точки зрения) опущена важная переменная.

Независимо от причины обычно бывает полезно вместо отсутствующей переменной использовать некоторый ее *заменитель* (proxy), а не пренебрегать ею совершенно. В качестве показателя общего социально-экономического положения вы можете использовать его заменитель — показатель дохода, если данные о нем имеются. В качестве показателя качества образования можно использовать отношение числа преподавателей и сотрудников к числу студентов или расходы на одного студента. Вместо переменной, опущенной в каком-либо обзоре, вы можете обратиться к другим, уже фактически собранным данным, если в них имеется подходящая замещающая переменная.

Имеются две причины для поиска такой переменной. Во-первых, если вы просто опустите важную переменную, то регрессия может пострадать от смещения оценок, описанного в разделе 6.2, и статистическая проверка будет неполноценной. Во-вторых, результаты оценки регрессии с включением замещающей переменной могут дать косвенную информацию о той переменной, которая замещена данной переменной.

Пример 1. Время как замещающая переменная для показателя технического прогресса

Мы уже встречались с замещающей переменной в разделе 5.5, где время использовалось для описания роста выпуска вследствие технического прогресса. В рассматривавшемся там периоде рост производительности, связанный с техническим прогрессом, оказался относительно малозначительным фактором. В последующие годы технический прогресс стал значительно более важным фактором, и при полном его исключении из спецификации производственной функции очевидно, что результаты оценки регрессии оказались бы сильно искаженными.

Когда на основе совокупных данных по экономике США за период 1949–1978 гг., подготовленных Дж. Кендриком и Э. Гроссменом (Kendrick, Grossman, 1980), была построена производственная функция Кобба–Дугласа и получено следующее уравнение (в скобках даны стандартные ошибки):

$$\begin{array}{lllll} \widehat{\log} Y = -1{,}03 + 0{,}17 \log K + 0{,}93 \log L + 0{,}024t; & & R^2 = 0{,}99; & & (6.21) \\ \quad\quad (2{,}33) \quad (0{,}66) \quad\quad (0{,}17) \quad\quad (0{,}016) & & F = 1297, & & \end{array}$$

где Y — индекс объема выпуска внутреннего частного сектора; K — индекс затрат капитала; L — индекс затрат труда; t — время, равное единице в 1948 г.,

двум — в 1949 г., и т. д., все эти индексы были взяты в реальном выражении (1967 = 100).

Если не считать весьма высокой эластичности выпуска по труду, то полученный результат вполне правдоподобен. Правда, ни оценка эластичности выпуска по капиталу, ни оценка темпов роста, связанных с техническим прогрессом, не отличаются значимо от нуля, но это может быть отнесено на счет мультиколлинеарности.

Если бы время не было использовано в качестве замещающей переменной для показателя технического прогресса, то оцененное уравнение выглядело бы следующим образом:

$$\begin{array}{lllll} \widehat{\log} Y = & -4{,}50 + 1{,}19 \log K + 0{,}77 \log L; & & R^2 = 0{,}99; & (6.22) \\ \text{(с.о.)} & (0{,}57) \quad (0{,}10) & (0{,}15) & F = 2012. & \end{array}$$

В уравнении (6.22) видно, что роль замещающей переменной для показателя технического прогресса играет log K. Коэффициент при log K неправдоподобно велик с двух точек зрения. Во-первых, он указывает, что увеличение затрат капитала должно привести к еще большему (пропорционально) увеличению выпуска при сохранении затрат труда постоянными. На практике же при неизменности других факторов можно ожидать снижения отдачи данного фактора. Во-вторых, если предположить, что рынки имеют конкурентный характер, то полученный результат означал бы, что доля дохода, приходящегося на капитал, превышает единицу, что, естественно, является абсурдом.

При добавлении в уравнение переменной t коэффициент при log K уже больше не смещается под действием того, что log K играл роль замещающей переменной для показателя технического прогресса; так что этот коэффициент становится более обоснованным в обоих отношениях. Естественно, фактор времени может заключать в себе и другие факторы, относящиеся ко времени и влияющие на выпуск помимо технического прогресса. Но это только усиливает аргумент в пользу включения его в уравнение, хотя все это означает, что в интерпретации значения его коэффициента следует быть осторожным.

Пример 2. Замещающая переменная для показателя дохода в функции спроса

В качестве второго примера, который хотя и не контролируется подобно эксперименту по методу Монте-Карло, но тем не менее позволяет судить об успехе той или иной замещающей переменной, рассмотрим еще раз модель, связывающую расходы потребителя на питание (y) с располагаемым личным доходом (x) и с относительной ценой продовольствия (p):

$$\log y = \alpha + \beta_1 \log x + \beta_2 \log p + u, \tag{6.23}$$

и предположим, что по какой-то причине мы не имеем доступа к данным о располагаемом личном доходе. Допустим, что нам, тем не менее, хотелось бы получить оценку ценовой эластичности спроса.

Как мы видели в разделе 6.2, парная регрессия между log y и log p дает смещенную оценку величины β_2, при этом тестовые статистики оказываются не-

корректными. Пусть, однако, мы считаем (и считаем правильно), что $\log x$ имеет ярко выраженный временной тренд. В этом случае мы могли бы частично решить проблему путем использования времени в качестве замещающей переменной для x, построив регрессию:

$$\log \hat{y} = a + b_2 \log p + b_3 t. \tag{6.24}$$

Таблица 6.7

Объясняющие переменные	*Оценки коэффициентов (стандартные ошибки)*			
	b_1	b_2	b_3	R^2
$\log x$, $\log p$	0,64 (0,03)	−0,48 (0,12)	—	0,99
$\log p$	—	2,04 (0,33)	—	0,63
$\log p$, t	—	−0,47 (0,13)	0,023 (0,001)	0,98

В табл. 6.7 даны результаты, полученные: 1) для правильно специфицированной регрессии между $\log y$, $\log x$ и $\log p$; 2) для неправильно специфицированной парной регрессии только между $\log y$ и $\log p$; 3) для множественной регрессии при использовании t в качестве замещающей переменной для $\log x$ (с указанием стандартных ошибок в скобках).

Во второй регрессии при невключении в уравнение $\log x$ оценка ценовой эластичности спроса настолько сильно смещается вверх, что становится положительной, а уровень коэффициента R^2 значительно ниже, чем в первой регрессии. В третьей регрессии введение t явно устраняет смещение в оценке ценовой эластичности, а коэффициент R^2 восстанавливается до предшествующего высокого уровня. Устранение смещения вызывается тем, что t в этом случае берет на себя роль замещающей переменной для отсутствующего $\log x$, оставляя для $\log p$ выполнение только собственных функций. Почти полное восстановление коэффициента R^2 до предыдущего уровня можно объяснить тем, что величина t значительно лучше выполняет роль замещающей переменной для отсутствующего показателя $\log x$, чем $\log p$.

Обобщение

Теперь мы можем обобщить сделанные выводы. Предположим, что истинной моделью является

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \tag{6.25}$$

и допустим, что мы не имеем данных по переменной x_1, но другая переменная

(z) выступает идеальным заменителем для нее в том смысле, что имеется строгая линейная связь между величинами x_1 и z:

$$x_1 = \lambda + \mu z, \tag{6.26}$$

где λ и μ являются постоянными, но неизвестными величинами. (Заметьте, что если бы λ и μ были известными, то мы могли бы вычислить x_1 по величине z и тогда не было бы необходимости использовать z в качестве замещающей переменной для нее. Заметьте также, что мы не можем оценить величины λ и μ посредством регрессионного анализа, поскольку для этого потребовались бы данные по величине x_1.)

Если мы построим регрессию

$$\hat{y} = a + b_2 x_2 + \ldots + b_k x_k + cz, \tag{6.27}$$

то оценки величин b_2, ..., b_k, их стандартные ошибки и коэффициент R^2 будут такими же, какими они были бы при наличии возможности построения регрессии с использованием x_1. Единственным недостатком является то, что нет оценок коэффициента при самой величине x_1, а величина a не является оценкой α. Коэффициент c будет оценкой величины $\beta_1\mu$. Для того чтобы получить оценку β_1, нужно разделить величину c на μ. Зачастую вы можете не иметь представления о величине μ, и тогда на этом дело будет закончено. Но иногда вы сможете сделать о ней субъективное предположение на основе опыта, интуиции или логики.

Например, предположим, что вы исследуете вопрос об «утечке мозгов» из страны A в страну B и используете (весьма наивную) модель:

$$y = \alpha + \beta x + u, \tag{6.28}$$

где y — показатель относительного уровня миграции определенного вида трудовых ресурсов из страны A в страну B; x — показатель отношения уровня заработной платы в стране B к заработной плате в стране A. Вы полагаете, что при более высокой разнице в заработной плате будет более высокой и миграция. Однако предположим, что у вас есть данные только по валовому внутреннему продукту (ВВП) на душу населения, но не по заработной плате. В этом случае можно ввести замещающую переменную p, которая является отношением ВВП страны B к ВВП страны A.

В этом случае в качестве первого приближения было бы разумно предположить, что относительные уровни заработной платы пропорциональны относительным величинам ВВП. Если бы эта зависимость была строгой, то уравнение (6.26) можно было бы записать с величиной λ, равной нулю, и величиной μ, равной единице. Отсюда c — коэффициент при относительном ВВП дал бы непосредственную оценку величины β — коэффициента при относительной заработной плате. Поскольку переменные в регрессионном анализе зачастую определяются в относительной форме, то этот частный случай в действительности имеет широкое применение.

В данном рассуждении мы приняли, что z является идеальной замещающей переменной для x, и справедливость всех приведенных выше результатов зависит именно от этого условия. На практике обычно невозможно найти замещающую переменную, имеющую строгую линейную связь с недостающей переменной. Но если связь близка к линейной, то результаты будут приблизительно на

том же уровне. Основной проблемой является отсутствие средств для проверки того, удовлетворительно или нет выполняется указанное условие. Здесь приходится оправдывать использование замещающей переменной на основе субъективных критериев. Использование несовершенных замещающих переменных будет рассмотрено далее в главе 8.

Непреднамеренное использование замещающих переменных

Иногда случается, что вы используете замещающую переменную, не осознавая этого. Вы полагаете, что y зависит от z, а в действительности эта величина зависит от x.

Если корреляция между величинами z и x незначительна, то результаты будут плохими, и вы поймете, что тут что-то неладно. Но если корреляция тесная, то результаты окажутся удовлетворительными (коэффициент R^2 будет близок к желаемому уровню и т. п.), и вы можете даже не подозревать, что полученное соотношение неверно.

Имеет ли это какое-то значение? Это, во-первых, зависит от того, с какой целью вы строите данную регрессию. Если целью оценивания регрессии является предсказание будущих значений величины y, то использование замещающей переменной не будет иметь большого значения при условии, конечно, что корреляция тесная и не является в то же время статистической счастливой случайностью. Однако если вы намерены использовать объясняющую переменную в качестве инструмента экономической политики для оказания влияния на поведение зависимой переменной, то последствия могут оказаться катастрофическими. Если только не будет функциональной связи между замещающей переменной и истинной объясняющей переменной, манипулирование замещающей переменной не окажет никакого влияния на зависимую переменную. Если мотивом построения регрессии является чисто научное любопытство, то исход будет столь же неудовлетворительным.

Непреднамеренное использование замещающих переменных особенно распространено при анализе временных рядов, в частности в макроэкономических моделях. Если истинная объясняющая переменная имеет временной тренд, то вы, вероятно, получите хорошую оценку формулы, если замените (преднамеренно или нет) ее на любую другую переменную с временным трендом. Даже если вы связываете приращения зависимой переменной с приращениями объясняющей переменной, вы, вероятно, получите аналогичные результаты независимо от того, используется ли правильная объясняющая переменная или же замещающая переменная, поскольку макроэкономические переменные обычно изменяются взаимосвязанно, в соответствии с экономическим циклом.

Упражнения

6.11. В приведенной ниже таблице даны коэффициенты (с указанными в скобках стандартными ошибками) логарифмической регрессионной зависимости расходов на жилье: 1) от располагаемого личного дохода (*dpi*) и цены; 2) только

от цены; 3) от цены и времени. Показатель *dpi* и цена, вычисленные по данным выборочного периода, оказались отрицательно коррелированными. Прокомментируйте результаты.

	Оценки коэффициентов			
	dpi	*Цена*	*t*	R^2
(1)	1,18 (0,06)	−0,34 (0,31)	—	0,99
(2)	—	−6,72 (0,70)	—	0,80
(3)	—	−0,63 (0,13)	0,041 (0,001)	1,00

6.12. Используя данные по товару, выбранному вами в упражнении 2.4, постройте парную логарифмическую регрессионную зависимость спроса от относительной цены и множественную регрессионную зависимость спроса от относительной цены и времени (*не* включая *dpi*). Сопоставьте результаты со множественной регрессионной зависимостью спроса от располагаемого личного дохода и относительной цены, оцененной вами в упражнении 5.6. Сделали бы вы вывод, что время может служить удовлетворительной замещающей переменной для располагаемого личного дохода, если бы вам не удалось получить данные о последнем?

6.13. Исследователь считает, что соотношение между годовым доходом индивида (y), числом лет трудового стажа (x) и количеством лет обучения (S) выражается формулой:

$$y = \alpha + \beta_1 x + \beta_2 S + u,$$

где u — случайный член. Исследователь располагает данными перекрестной выборки по y и S для 1000 человек, но не имеет прямых данных по x. Имеются данные о возрасте каждого индивида, а отсюда z — число лет, прошедших с момента официального завершения обучения, может быть вычислено по формуле:

$$z = \text{Возраст} - S - 6,$$

исходя из предположения, что каждый индивид начал учиться в возрасте 6 лет. Подчеркните преимущества и недостатки построения регрессионной зависимости y: 1) только от величины S; 2) от S и z, используя z в качестве замещающей переменной для x. Обсудите в каждом случае, каким образом должны интерпретироваться результаты регрессии и какие могут быть выполнены статистические тесты.

6.5. Проверка линейного ограничения

В разделах 5.3 и 5.5 было показано, что число объясняющих переменных в уравнении регрессии можно уменьшить на единицу, если известно, что параметры этого уравнения линейно зависимы. Воспользовавшись этой зависимостью, вы сделаете оценки регрессии более эффективными. Если до этого имела место проблема мультиколлинеарности, то она может быть смягчена. Даже если эта проблема в исходной модели отсутствовала, то выигрыш в эффективности может дать улучшение точности оценок, что отражается их стандартными ошибками.

Например, в разделе 5.3 мы видели, что наиболее общая форма функции Кобба—Дугласа

$$Y = AK^{\alpha}L^{\beta}v \tag{5.31}$$

при наложении ограничения $\beta = 1 - \alpha$ могла бы быть преобразована к виду:

$$Y/L = A(K/L)^{\alpha}v. \tag{5.28}$$

Соответственно этому регрессии, построенные на основе функции Кобба—Дугласа, рассчитанной для производственного сектора США за 1899–1922 гг., выглядели так (в скобках указаны стандартные ошибки):

$$\begin{array}{lll} \widehat{\log Y} = -0{,}18 + 0{,}23 \log K + 0{,}81 \log L; & R^2 = 0{,}96; \\ \qquad\quad (0{,}43)\ (0{,}06) \qquad\ (0{,}15) & F = 236{,}1; \end{array} \tag{5.32}$$

и (с учетом линейного ограничения на параметры):

$$\begin{array}{lll} \widehat{\log Y/L} = 0{,}02 + 0{,}25 \log K/L; & R^2 = 0{,}63; \\ \quad (\text{с.о.})\ (0{,}02)\ (0{,}04) & F = 38{,}0. \end{array} \tag{5.30}$$

Оценки величин α и β в формуле (5.32) действительно в сумме дают примерно единицу, что может служить обоснованием для использования ограничения, учет которого, как видно, повышает эффективность, поскольку стандартная ошибка оценки величины α в версии с ограничением составляет всего 0,04 против 0,06 при отсутствии ограничения. Однако прежде, чем использовать версию с ограничением, мы должны провести формальную проверку гипотезы о наличии ограничения. Имеется несколько способов сделать это, но мы рассмотрим два наиболее распространенных, которые оказываются эквивалентными.

Проверка ограничения с помощью критерия F

Постройте обе формы регрессии — как с учетом ограничения, так и без него — и обозначьте сумму квадратов остатков (автоматически рассчитанных компьютером) через RSS_R — в варианте с ограничением и RSS_U — в варианте без ограничения. Поскольку ввод ограничения ведет к сужению возможностей подбора уравнения регрессии, обеспечивающего наибольшее соответствие с имеющимися данными, RSS_R не может быть меньше, чем RSS_U, а будет в общем случае больше. Нам хотелось бы проверить, является ли улучшение качества регрессии

при переходе от варианта с ограничением к варианту без ограничения статистически значимым. Если это так, то ограничение должно быть отброшено.

Для этой цели мы можем использовать критерий F, сконструированный так же, как и в разделе 5.6:

$$F = \frac{\textit{Улучшение качества уравнения / Число использованных степеней свободы}}{\textit{Оставшаяся сумма квадратов отклонений / Оставшееся число степеней свободы}}. \quad (5.58)$$

Здесь улучшение качества регрессии, получаемое при переходе от модели с ограничением к модели без ограничения, выражается величиной $(RSS_R - RSS_U)$, в модели без ограничения появляется одна дополнительная степень свободы (поскольку оценивается на один параметр больше), и сумма квадратов отклонений, остающаяся после перехода от ограниченного к неограниченному варианту, составляет RSS_U. Следовательно, F-статистика в данном случае равна:

$$F = \frac{RSS_R - RSS_U}{RSS_U / (n - k - 1)}, \quad (6.29)$$

где k — число объясняющих переменных в варианте без ограничения. Она распределена с одной и $(n - k - 1)$ степенями свободы при предположении, что ограничение верно.

В случае с производственной функцией Кобба—Дугласа сумма квадратов отклонений составила 0,0716 в модели с ограничением и 0,0710 — в модели без ограничения. Отсюда F-статистика равнялась:

$$F = \frac{0{,}0716 - 0{,}0710}{0{,}0710 / 21} = 0{,}18. \quad (6.30)$$

Критический уровень величины F с 1 и 21 степенью свободы при 5-процентном уровне значимости равен 4,32. Поскольку значение F-статистики оказалось ниже критического уровня, мы не отбрасываем ограничение. Другими словами, Ч. Кобб и П. Дуглас были правы, используя ограничение о постоянном эффекте от масштаба применительно к рассматриваемому периоду.

Проверка ограничения с помощью критерия t [1]

При проверке ограничения с помощью критерия t используется факт, что модель с ограничением может быть сведена к модели без ограничений путем добавления в уравнение соответствующего члена. Для удобства мы назовем эту формулировку модели «вариант *3*»[2]. Коэффициент дополнительного члена в варианте *3* будет равен нулю, если и только если ограничение выполняется. Поэтому вы можете проверить ограничение, оценив регрессию для варианта *3* и выяснив, значимо или нет отличается от нуля коэффициент дополнительного члена.

В случае функции Кобба—Дугласа добавление члена $(\beta + \alpha - 1) \log L$ в урав-

[1] Данный подраздел можно пропустить без потери целостности изложения.

[2] Данный термин не является стандартным.

нение преобразует модель с ограничением в модель без ограничений. Если расширить уравнение (5.29) для перехода к варианту *3*, то

$$\log Y/L = \log A + \alpha \log K/L + (\beta + \alpha - 1) \log \mathrm{L} + \log v. \qquad (6.31)$$

Отсюда

$$\log Y - \log L = \log A + \alpha[\log K - \log L] + \beta \log L + \alpha \log L - \log L + \log v. \qquad (6.32)$$

Путем упрощения можно вернуться вновь к модели без ограничений:

$$\log Y = \log A + \alpha \log K + \beta \log L + \log v. \qquad (6.33)$$

Таким образом, формула (6.31) является новым способом записи модели без ограничений. Если ограничение верно, то коэффициент при $\log L$ не должен значимо отличаться от нуля (если мы не совершим, к несчастью, ошибку I рода), и тогда мы имеем право исключить этот член, т. е. использовать модель с ограничением.

В рассматриваемом случае, оценивая регрессию для варианта *3*, мы получим (в скобках даны стандартные ошибки):

$$\widehat{\log Y/L} = \underset{(0,43)}{-0,18} + \underset{(0,06)}{0,23} \log K/L + \underset{(0,09)}{0,04} \log L; \qquad R^2 = 0,64. \qquad (6.34)$$

Коэффициент при $\log L$ не отличается значимо от нуля. Это подразумевает, что $(\alpha + \beta)$ не отличается значимо от единицы. Поэтому ограничение мы не отбрасываем.

Каким образом найти дополнительный член, который преобразует модель с ограничением обратно в модель без ограничений? Попрактиковавшись немного, вы сможете делать это путем изучения и проверки. Если вы предпочитаете формально строгий, но более механический путь его определения, то напишите сначала вариант модели без ограничений, а затем — с ограничением со всеми членами в левой стороне уравнения и проведите вычитание. Разность и будет тем выражением, которое вы ищете.

Почему этот способ эквивалентен использованию *F*-теста? Напомним, что *F*-тест проверяет улучшение качества регрессии при переходе от модели с ограничением к модели без ограничений. Это осуществляется путем включения в уравнение дополнительного члена, но, как нам известно, *F*-тест для проверки улучшения качества регрессии путем включения в уравнение дополнительного члена эквивалентен *t*-тесту для проверки значимости коэффициента этого члена (см. раздел 5.6).

Еще один пример

Допустим, вы предполагаете, что совокупный расход на продовольствие (*y*) зависит от совокупного личного дохода (*z*), совокупного личного налога (*tax*) и относительной цены продовольствия (*p*). Вы допускаете наличие зависимости:

$$y = \alpha + \beta_1 z + \beta_2 tax + \beta_3 p + u. \qquad (6.35)$$

Пользуясь данными по США за период 1959–1983 гг. из табл. Б.1 и Б.2, а также из упражнения 6.17 и вычисляя налог (*tax*) как разность между личным доходом и располагаемым личным доходом, мы получим регрессию:

$$\hat{y} = 116{,}7 + 0{,}113z - 0{,}115tax - 0{,}741p; \qquad R^2 = 0{,}99. \tag{6.36}$$
$$\text{(с.о.)}\ (9{,}8)\quad (0{,}009)\quad (0{,}040)\quad (0{,}120)$$

Заметив, что коэффициент при *tax* близок к коэффициенту при z по абсолютной величине, но противоположен по знаку, мы видим, что величина y в конечном счете может зависеть в большей степени не от z или *tax* по отдельности, а от располагаемого личного дохода, т. е. разности между ними, и поэтому мы имеем право ввести ограничение

$$\beta_2 = -\beta_1, \tag{6.37}$$

для того чтобы повысить эффективность оценок. Последнее уравнение может быть переписано в виде:

$$y = \alpha + \beta_1 x + \beta_3 p + u, \tag{6.38}$$

где x — располагаемый личный доход, а соответствующая регрессия выглядит как

$$\hat{y} = 116{,}7 + 0{,}112x - 0{,}739p; \qquad R^2 = 0{,}99. \tag{5.3}$$
$$\text{(с.о.)}\ (9{,}6)\quad (0{,}003)\quad (0{,}114)$$

Мы действительно видим улучшение эффективности, так как стандартная ошибка коэффициента при доходе сейчас составляет только 0,003 вместо 0,009.

Суммы квадратов отклонений в вариантах уравнений без ограничений и с ограничениями составляют 65,379 и 65,398 соответственно, и F-статистика для проверки ограничения равна:

$$F = \frac{0{,}019}{65{,}379 / 21} = 0{,}006. \tag{6.39}$$

Критический уровень F с 1 и 21 степенью свободы при 5-процентном уровне значимости составляет 4,32, и мы, таким образом, не отвергаем ограничения. Фактически это было, в сущности, почти предрешенным выводом, поскольку коэффициенты регрессии в уравнении без ограничений (6.36) очень близки к значениям, полученным при выполнении ограничения.

Можно, конечно, использовать подход с t-тестом. В данном случае вариант *3* представляется в виде:

$$y = \alpha + \beta_1 x + \beta_3 p + (\beta_1 + \beta_2)\, tax + u. \tag{6.40}$$

Соответствующей регрессией является (в скобках даны стандартные ошибки):

$$\hat{y} = 116{,}7 + 0{,}113x - 0{,}741p - 0{,}002tax; \qquad R^2 = 0{,}99; \tag{6.41}$$
$$(9{,}8)\quad (0{,}009)\quad (0{,}120)\quad (0{,}031)$$

и мы приходим к выводу, что оценка коэффициента $(\beta_1 + \beta_2)$ не отличается значимо от нуля, т. е. что β_2 не отличается значимо от $-\beta_1$.

6.14. В разделе 5.5 мы рассмотрели добавление временного тренда к производственной функции Кобба—Дугласа с целью учета технического прогресса. Мы обнаружили, что это вызвало мультиколлинеарность [уравнение (5.48)], и получили значительно лучшие результаты, когда ввели ограничение, определяющее постоянный эффект от масштаба [уравнение (5.49)]. Суммы квадратов отклонений в уравнениях без ограничений и с ограничениями были, соответственно, равны 0,056 и 0,068. Проведите проверку ограничения, предполагающего постоянный эффект от масштаба.

6.15. Построение регрессионной зависимости расходов на жилищные услуги от личного дохода, налога и относительной стоимости жилья дает следующие результаты (в скобках указаны стандартные ошибки):

$$\hat{y} = \underset{(50,0)}{-41,6} + \underset{(0,020)}{0,177}z - \underset{(0,094)}{0,160}tax + \underset{(0,432)}{0,131}p; \qquad R^2 = 0,99.$$

Сравните данное уравнение с регрессией между спросом, располагаемым личным доходом и относительной ценой, представленной в упражнении 5.2. Суммы квадратов отклонений в вариантах без ограничений и с ограничением были равны 382,4 и 383,3 соответственно. Проведите проверку ограничения, тщательно сформулировав нулевую гипотезу.

6.16. Регрессионная зависимость для «варианта 3» от располагаемого личного дохода, относительной цены и налогов дает следующий результат (стандартные ошибки указаны в скобках):

$$\hat{y} = \underset{(50,0)}{-41,6} + \underset{(0,020)}{0,177}x + \underset{(0,432)}{0,131} + \underset{(0,075)}{0,017}tax; \qquad R^2 = 0,99.$$

Проведите проверку ограничения и сравните это уравнение с уравнением из упражнения 6.15.

6.17. Постройте регрессионную зависимость расходов на выбранный вами вид благ от личного дохода, налогов и относительной цены и сравните результаты с результатами оценивания регрессии между расходами, располагаемым личным доходом и относительной ценой в упражнении 5.3. Укажите, какую из регрессий следует считать лучше специфицированной.

2. Регрессия в упражнении 5.3 может рассматриваться как модель новой регрессии с ограничением. Сформулируйте соответствующее ограничение и проведите его формальную проверку.

Новая регрессия соответствует той же модели, что и регрессия функции спроса на продовольствие в уравнении (6.35). При работе с программой регрессионного анализа вам нужно будет использовать данные о личном доходе, взятые из приведенной ниже таблицы (с. 193). Однако нет необходимости вводить отдельно данные по налогам: эти данные вы можете вычислить (или «поручить» это сделать за вас компьютеру) как разность между личным доходом и располагаемым личным доходом.

6.18. В своей классической статье М. Нерлов (Nerlove, 1963) вывел следующую формулу функции издержек для производства электроэнергии:

Личный доход в США в 1959–1983 гг. (млрд. долл. США, в постоянных ценах 1972 г.)

1959	544,9	1966	740,6	1973	1007,9	1980	1209,5
1960	559,7	1967	774,4	1974	1004,8	1981	1248,6
1961	575,4	1968	816,2	1975	1010,8	1982	1254,4
1962	602,0	1969	853,5	1976	1056,2	1983	1284,6
1963	622,9	1970	876,8	1977	1105,4		
1964	658,0	1971	900,0	1978	1162,3		
1965	700,4	1972	951,4	1979	1200,7		

$$C = \alpha Y^{\beta} P_1^{\gamma_1} P_2^{\gamma_2} P_3^{\gamma_3} \nu,$$

где C — полные издержки производства; Y — выпуск (измеренный в киловатт-часах); P_1 — стоимость затрат труда; P_2 — цена использования капитала, P_3 — стоимость топлива (все показатели измеряются в соответствующих единицах) и ν — случайный член. Теоретически сумма показателей ценовой эластичности должна равняться единице:

$$\gamma_1 + \gamma_2 + \gamma_3 = 1,$$

и, следовательно, формула функции издержек может быть переписана:

$$\frac{C}{P_3} = \alpha Y^{\beta} \left(\frac{P_1}{P_3}\right)^{\gamma_1} \left(\frac{P_2}{P_3}\right)^{\gamma_2}.$$

Эти два варианта формулы функции издержек оценены для 29 фирм среднего размера, в выборке Нерлова, со следующими результатами (стандартные ошибки даны в скобках; RSS — сумма квадратов отклонений):

$$\underset{\substack{(1,62)\ (0,11)\qquad (0,23)\qquad (0,29)\qquad (0,07)}}{\log \hat{C} = -4,93 + 0,94 \log Y + 0,31 \log P_1 - 0,26 \log P_2 + 0,44 \log P_3}; \quad RSS = 0,336;$$

$$\underset{\substack{(0,16)\ (0,11)\qquad (0,19)\qquad\quad (0,16)}}{\log \hat{C}/P_3 = -6,55 + 0,91 \log Y + 0,51 \log P_1/P_3 + 0,09 \log P_2/P_3}; \quad RSS = 0,364.$$

Сравните результаты оценивания регрессии по указанным двум уравнениям и проведите формальную проверку выполнения ограничения.

6.6. Как извлечь максимум информации из анализа остатков

Существует два пути рассмотрения остатков, полученных в результате оценивания уравнения регрессии по какому-то набору данных. Если вы по натуре пессимист или проявляете пассивность, то будете смотреть на них как на сви-

детельство своей неудачи. Чем больше остатки, тем хуже регрессия и тем меньше коэффициент R^2. Общей целью является такая оценка уравнения регрессии, чтобы свести до минимума сумму квадратов остатков. Однако при некоторой предприимчивости вы будете видеть в этих остатках потенциально неограниченный источник для зарождения новых идей, а возможно, и новых теорий. Они дают одновременно основу для постановки задач и конструктивной критики. Формулируемые задачи создают стимул для научных исследований: необходимость найти лучшее объяснение для наблюдаемых событий. А конструктивная критика вызывается тем, что остатки, взятые по отдельности, указывают, когда, где и в какой степени существующая модель не смогла объяснить наблюдаемые события. Извлечение пользы из такой конструктивной критики требует от исследователя большого терпения. Если выборка достаточно мала, то вам следует очень внимательно рассмотреть каждое наблюдение с большим положительным или отрицательным отклонением и попытаться сформулировать для них гипотетические объяснения. Некоторые из этих объяснений могут включать какие-то особые факторы, которые вряд ли повторятся в дальнейшем. Такие факторы не приносят теоретику большой пользы. Они могут помочь вам дать объяснение явлениям в прошлом, но не могут оказать большой помощи в прогнозировании будущего.

Предположим, что вы исследуете связь между продажей каких-то предметов длительного пользования и располагаемым личным доходом, пользуясь данными погодового временного ряда. Если вы находите, что отрицательный остаток в каком-то году может быть отнесен к длительной забастовке у ведущего поставщика, то этим вы сделаете вклад в историю, но не в теорию.

Другие факторы, однако, могут оказаться связанными с отклонениями, появляющимися в нескольких наблюдениях. Как только вы обнаруживаете закономерность такого характера, вы делаете шаг вперед. Следующим шагом должно быть нахождение разумного способа для количественного описания данного фактора и включения его в модель. Например, в своих исследованиях продажи предметов длительного пользования вы можете обнаружить, что имели место большие положительные остатки в годы большей инфляции. Тут было бы естественным выдвинуть гипотезу, что покупатели пытаются защитить себя от инфляции путем приобретения товаров вместо сбережения денег, и вы, разумеется, должны включить темп этой инфляции в уравнение в качестве объясняющей переменной.

Заметим, что данным примером иллюстрируется исходный момент. Определение величины остатков является только частью решения задачи. Вам необходимо также иметь базовые знания и воображение, чтобы оценить факторы, способные объяснить их. Уже по одной только этой причине эконометрическое моделирование представляет собой вид искусства.

Иллюстрация

Для более подробной иллюстрации рассмотренных моментов вернемся к эксперименту по методу Монте-Карло, описанному в разделе 6.2, где исследователь изучает соотношение между доходом (y) и продолжительностью обучения (S) в некоторой стране. Истинным соотношением было:

$$y = -26250 + 1500S + 250IQ + u, \tag{6.9}$$

однако исследователь не учитывает влияния фактора способностей и оценивает уравнение регрессии в следующем виде:

$$\hat{y} = -6418 + 1985S. \tag{6.11}$$
$$\text{(с.о.)} \quad (3349) \quad (248)$$

Фактические значения величин y, их расчетные значения и остатки показаны в табл. 6.8.

Таблица 6.8

Наблюдение	*y*	$\hat{y}$	*e*	*Наблюдение*	*y*	$\hat{y}$	*e*
1	13880	13430	450	11	15770	19380	-3610
2	15310	13430	1880	12	22790	19380	3410
3	10470	13430	-2960	13	20910	21370	-460
4	17280	15410	1870	14	16720	21370	-4650
5	12480	15410	-2930	15	23200	23350	-150
6	18660	15410	3250	16	24690	25340	-650
7	14720	15410	-690	17	23130	25340	-2210
8	19390	17400	1990	18	34940	27320	7620
9	15510	17400	-1890	19	27780	29310	-1530
10	19590	17400	2190	20	30380	31290	-910

Здесь, как вы видите, имеются большие положительные остатки у индивидов *6, 12* и *18*, а также большие отрицательные остатки у индивидов *11* и *14*. Если бы исследователь побеседовал с ними, то он выяснил бы, что индивид *6* происходит из рабочей семьи и что он рано оставил школу, как и все его товарищи, и, тем не менее, достиг положения руководителя в сфере мелкого бизнеса, где продвижение основывается на результатах работы. Индивиды *12* и *18*, которые имеют более высокое образование, также исключительно хорошо продвигались по службе, что не было удивительным для тех, кто знал о том, что они всегда были среди лучших в учебе. Если посмотреть на отрицательные остатки, то при опросе индивидов *11* и *14* исследователь установил бы, что оба они были весьма неспособны к учебе и с радостью оставили школу, когда это им позволили сделать их родители.

Если бы эти опросы были сделаны, то исследователь, пусть даже смутно, понял бы, что природные способности являются важным фактором в определении уровня дохода и это привело бы его к регрессии с правильной спецификацией, с предположением, что может быть измерен показатель IQ каждого индивида. (Согласимся, что это является упрощением: IQ отражает только один

вид способностей, не самый важный для успеха в бизнесе или деятельности такого же рода.)

Анализ остатков имеет также важное значение при выборе наиболее подходящей формулы уравнения регрессии. Как мы увидим ниже, в разделе 7.9, поведение остатков может указывать на математически неправильную спецификацию модели. И наконец, анализ остатков может быть полезным при проверке того, удовлетворены ли второе и третье условия Гаусса—Маркова. Условия Гаусса—Маркова относятся к случайному члену u. Измерение величины u в каждом отдельном наблюдении невозможно, но остаток в этом наблюдении может быть взят в качестве замещающей переменной для u. Отсюда если остатки подчиняются второму и третьему условиям Гаусса—Маркова, то будет разумным считать, что им подчиняется и случайный член. К этому вопросу мы вернемся в главе 7.

6.7. Лаговые переменные[1]

До сих пор мы считали, что на текущее значение зависимой переменной влияют только текущие значения объясняющих переменных. Такое предположение делается, когда мы пользуемся перекрестными данными (статистическими данными, относящимися к различным отраслям экономики, к различным фирмам и т. д.), где выборка берется из совокупности индивидов, фирм, стран и т. п. в соответствии с теми или иными условиями на какой-то один момент времени. Однако при пользовании данными временного ряда мы можем ослабить это условие и исследовать, в какой степени запаздывает рассматриваемое влияние. Технически это известно как *лаговая структура* зависимости (структура с запаздыванием).

Например, если нас интересует исследование соотношения между расходами на жилье (y), располагаемым личным доходом (x), и индексом реальных цен на жилье (p) и если мы допустим, что логарифмическая регрессия является более подходящей, чем линейная, то мы можем построить регрессию:

$$\log y_t = \alpha + \beta_1 \log x_t + \beta_2 \log p_t + u_t, \tag{6.42}$$

индекс t добавлен здесь к переменным для того, чтобы показать, что мы связываем текущие расходы на жилье с текущим доходом. Воспользовавшись данными за 1959—1983 гг. из табл. Б.1 и Б.2 приложения, мы получим:

$$\widehat{\log y_t} = -1{,}60 + 1{,}18 \log x_t - 0{,}34 \log p_t; \qquad R^2 = 0{,}992. \tag{6.43}$$
$$\text{(с.о.)} \quad (1{,}75) \quad (0{,}05) \qquad (0{,}31)$$

В соответствии с данной регрессией расходы на жилье имеют положительную эластичность по доходам, близкую к единице, и отрицательную эластичность по цене, что и следовало ожидать.

Другой исследователь может предположить, что люди более склонны соотносить свои расходы на жилье не с текущими доходами и ценами, а с предше-

[1] Лаговая переменная — это переменная, влияние которой характеризуется некоторым запаздыванием. (*Прим. ред.*)

ствующими, например с прошлогодними доходом и ценами, которые мы обозначим x_{t-1} и p_{t-1} соответственно:

$$\log y_t = \alpha + \beta_1 \log x_{t-1} + \beta_2 \log p_{t-1} + u_t. \qquad (6.44)$$

Можно утверждать, что расходы на жилье подвержены инерции и медленно согласуются с изменениями доходов и цен. В табл. 6.9 приведены данные по x_t и p_t за рассматриваемый период. Для того чтобы получить данные для x_{t-1} и p_{t-1}, нужно просто сдвинуть данные для x_t и p_t на один уровень ниже в таблице.

В 1980 г. доход составлял 1021,6. По отношению к 1981 г. эта сумма становится прошлогодним доходом, то есть x_{t-1}. То же происходит и по отношению к другим годам. Точно так же индекс цен в 1980 г. равнялся 93,0, а в 1981 г. он становится значением p_{t-1} и т. д. Мы сталкиваемся с проблемой наблюдения величин x_{t-1} и p_{t-1} применительно к 1959 г. Они равны величинам x_t и p_t 1958 г., которых нет в табл. Б.1 и Б.2. Мы должны либо отдельно найти значения этих величин, либо ограничить период выборки 1960–1983 гг. Здесь мы выбрали первый способ действий.

Оценив регрессионную зависимость $\log y_t$ от $\log x_{t-1}$ и $\log p_{t-1}$, получим:

$$\widehat{\log y_t} = 0{,}42 + 1{,}10 \log x_{t-1} - 0{,}66 \log p_{t-1}; \qquad R^2 = 0{,}995. \qquad (6.45)$$
$$\text{(с.о.)} \quad (1{,}74) \quad (0{,}05) \qquad\quad (0{,}31)$$

Первоначальной является регрессия между $\log y$ и текущими величинами $\log x$ и $\log p$, второй — регрессия между $\log y$ и величинами $\log x$ и $\log p$, взятыми с запаздыванием на один период.

Естественно, что ничто не может помешать вам оценить регрессию между $\log y$ и величинами $\log x$ и $\log p$, взятыми с запаздыванием на два периода. Величина x_{t-2} в 1982 г. — та же, что x_t в 1980 г., и т. д. Данные для x_{t-2} и p_{t-2} показаны в пятой и восьмой колонках табл. 6.9. Оценив регрессионную зависимость $\log y_t$ от $\log x_{t-2}$ и $\log p_{t-2}$, получим:

$$\widehat{\log y_t} = 0{,}95 + 1{,}08 \log x_{t-2} - 0{,}72 \log p_{t-2}; \qquad R^2 = 0{,}995; \qquad (6.46)$$
$$\text{(с.о.)} \quad (1{,}77) \quad (0{,}05) \qquad\quad (0{,}31)$$

В общем, если какая-то переменная появляется в модели с запаздыванием на s периодов, то она записывается с нижним индексом $(t-s)$. Как мы увидим в главе 10, одна и та же переменная может в данном уравнении появляться несколько раз с разными запаздываниями. Например, при анализе функции спроса на жилье было бы разумным выдвинуть гипотезу, что текущий доход, прошлогодний доход и, возможно, доход за несколько предыдущих лет, вместе взятые, влияют на текущие расходы на жилье. Спецификация запаздываний применительно к переменным в модели называется лаговой структурой (структурой с запаздыванием); она может являться важным аспектом модели и выступать в качестве спецификации самих переменных. Мы встретимся с одной специальной лаговой структурой при рассмотрении автокорреляции в следующей главе. Расширенный анализ этой проблемы будет дан в главе 10.

Таблица 6.9

Расходы на жилищные услуги, располагаемый личный доход и индекс реальных цен на жилье в США в 1959–1983 гг. (расходы на жилье (y) и доход (x) в млрд. долл. США в ценах 1972 г.; индекс цен (p) для 1972 г. = 100).

Год	y_t	x_t	x_{t-1}	x_{t-2}	p_t	p_{t-1}	p_{t-2}
1959	60,9	479,7	460,7	455,5	104,5	104,6	104,6
1960	64,0	489,7	479,7	460,7	104,5	104,5	104,6
1961	67,0	503,8	489,7	479,7	105,1	104,5	104,5
1962	70,7	524,9	503,8	489,7	105,0	105,1	104,5
1963	74,0	542,3	524,9	503,8	104,8	105,0	105,1
1964	77,4	580,8	542,3	524,9	104,5	104,8	105,0
1965	81,6	616,3	580,8	542,3	104,0	104,5	104,8
1966	85,3	646,8	616,3	580,8	102,6	104,0	104,5
1967	89,1	673,5	646,8	616,3	102,2	102,6	104,0
1968	93,5	701,3	673,5	646,8	100,9	102,2	102,6
1969	98,4	722,5	701,3	673,5	100,0	100,9	102,2
1970	102,0	751,6	722,5	701,3	99,6	100,0	100,9
1971	106,4	779,2	751,6	722,5	100,0	99,6	100,0
1972	112,5	810,3	779,2	751,6	100,0	100,0	99,6
1973	118,2	865,3	810,3	779,2	99,1	100,0	100,0
1974	124,2	858,4	865,3	810,3	95,1	99,1	100,0
1975	128,3	875,8	858,4	865,3	93,3	95,1	99,1
1976	134,9	906,8	875,8	858,4	93,7	93,3	95,1
1977	141,3	942,9	906,8	875,8	94,5	93,7	93,3
1978	148,5	988,8	942,9	906,8	94,7	94,5	93,7
1979	154,8	1015,5	988,8	942,9	93,8	94,7	94,5
1980	159,8	1021,6	1015,5	988,8	93,0	93,8	94,7
1981	164,8	1049,3	1021,6	1015,5	94,2	93,0	93,8
1982	167,5	1058,3	1049,3	1021,6	96,7	94,2	93,0
1983	171,3	1095,4	1058,3	1049,3	99,2	96,7	94,2

Источники: Таблицы Б.1 и Б.2 дают индексы реальных цен, вычисляемых путем деления дефлятора цен на жилье на дефлятор всех расходов и умножения результатов на 100. Данные по лаговым переменным в 1959 и 1960 гг. были взяты непосредственно из журнала «Survey of Current Business».

Упражнение

6.19. Постройте логарифмическую регрессию между расходами на выбранный вами товар и доходами (с запаздыванием на один год) или относительной ценой (с запаздыванием на один год). Заметьте, что для этого вам придется укоротить период выборки до 1960–1983 гг. Постройте такую же регрессию без запаздываний на этот же период и сравните результаты.

7

ГЕТЕРОСКЕДАСТИЧНОСТЬ И АВТОКОРРЕЛИРОВАННОСТЬ СЛУЧАЙНОГО ЧЛЕНА

Медицина традиционно подразделяется на три отрасли — анатомию, физиологию и патологию, соответственно изучающие структуру организма, принцип действия его систем и нарушения их функционирования. Аналогично, сейчас наступил момент для рассмотрения недостатков («патологии») регрессионного анализа, основанного на методе наименьших квадратов.

Свойства оценок коэффициентов регрессии зависят от свойств случайного члена в регрессионной модели. До сих пор мы предполагали, что значения случайного члена в выборочных наблюдениях независимы и одинаково распределены. Теперь рассмотрим, что происходит при нарушении этого предположения. Мы обнаружим, что обычный МНК в некоторых ситуациях будет давать плохие результаты и что можно получить лучшие результаты при использовании других методов.

7.1. Еще раз об условиях Гаусса—Маркова

До сих пор мы предполагали, что случайный член в регрессионной модели удовлетворяет всем четырем условиям Гаусса—Маркова, изложенным в разделе 3.3. Если регрессионное уравнение имеет вид:

$$y = \alpha + \beta x + u, \tag{7.1}$$

то эти условия состоят в следующем:

$E(u_i) = 0$ для всех наблюдений;
дисперсия pop. var (u_i) одинакова для всех наблюдений;
pop. cov $(u_i, u_j) = 0$, при $i \neq j$;
объясняющая переменная является неслучайной (так что pop. cov $(x_i, u_i) = 0$ для каждого наблюдения),

где u_i и x_i — значения u и x в i-м наблюдении. Если регрессия не парная, а множественная, то условия будут те же самые с тем различием, что последнему из них должна удовлетворять каждая объясняющая переменная. Как пояснялось в разделе 3.3, если не принимать во внимание особые случаи, первое условие по

сути является частью определения, если постоянный член включен в уравнение. В последующих двух главах будут рассматриваться последствия ситуаций, при которых не выполнены остальные условия. В этой главе мы рассмотрим второе и третье условия. В каждом случае рассмотрение будет осуществляться в такой последовательности: 1) почему рассматриваемое условие важно; 2) как оно может быть нарушено; 3) обзор возможных средств, исправляющих положение.

7.2. Гетероскедастичность и ее последствия

Во втором условии Гаусса—Маркова утверждается, что дисперсия случайного члена в каждом наблюдении должна быть постоянной. Такое утверждение может показаться странным, и здесь требуется пояснение. Случайный член в каждом наблюдении имеет только одно значение, и может возникнуть вопрос о том, что означает его «дисперсия».

Имеется в виду его *возможное* поведение *до того*, как сделана выборка. Когда мы записываем модель (7.1), первые два условия Гаусса—Маркова указывают, что случайные члены u_1, u_2, ..., u_n в n наблюдениях появляются на основе вероятностных распределений, имеющих нулевое математическое ожидание и одну и ту же дисперсию. Их *фактические* значения в выборке иногда будут положительными, иногда — отрицательными, иногда — относительно далекими от нуля, иногда — относительно близкими к нулю, но у нас нет причин a priori ожидать появления особенно больших отклонений в любом данном наблюдении. Другими словами, вероятность того, что величина u примет какое-то данное положительное (или отрицательное) значение, будет одинаковой для всех наблюдений. Это условие известно как *гомоскедастичность*, что означает «одинаковый разброс». Оно проиллюстрировано в левой части рис. 7.1.

Вместе с тем для некоторых выборок, возможно, более целесообразно предположить, что теоретическое распределение случайного члена является разным для различных наблюдений в выборке. В правой части рис. 7.1 дисперсия величины u_i увеличивается по мере продолжения выборочных наблюдений. Это не означает, что случайный член *обязательно* будет иметь особенно большие (положительные или отрицательные) значения в конце выборки, но это значит, что априорная *вероятность* получения сильно отклоненных величин будет относительно высока. Это пример *гетероскедастичности*, что означает «неодинаковый разброс». Математически гомоскедастичность и гетероскедастичность могут определяться следующим образом:

Гомоскедастичность: pop. var $(u_i) = \sigma^2$ постоянна для всех наблюдений;

Гетероскедастичность: pop. var $(u_i) = \sigma_i^2$, она не обязательно одинакова для всех i.

На рис. 7.2 показано, как будет выглядеть характерная диаграмма рассеяния, если y — возрастающая функция от x и имеется гетероскедастичность типа, показанного на рис. 7.1. Можно видеть, что, хотя наблюдения не обязательно все дальше отстоят от основной нестохастической составляющей линии регрессии $y = \alpha + \beta x$, по мере роста x все же имеется тенденция к увеличению их

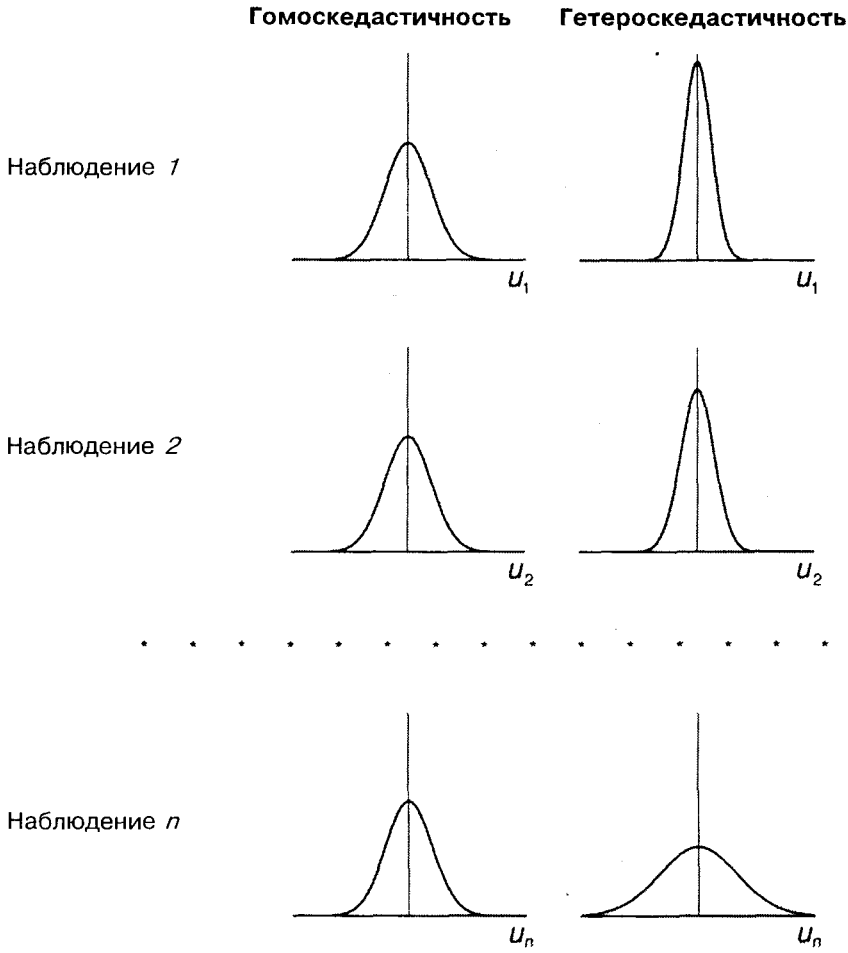


Рис. 7.1. Различия между гетероскедастичностью и гомоскедастичностью

разброса. (Следует иметь в виду, что гетероскедастичность не обязательно относится к типу, показанному на рис. 7.1. Данное понятие относится к любому случаю, в котором дисперсия вероятностного распределения случайного члена различна для разных наблюдений.)

Возникает вопрос, почему гетероскедастичность имеет существенное значение. В самом деле, соответствующее условие Гаусса—Маркова пока не использовалось в проводимом анализе, и оно может показаться практически не нужным. В частности, при рассмотрении простой модели (7.1) и оцененного уравнения

$$\hat{y} = a + bx, \tag{7.2}$$

в доказательстве того, что b является несмещенной оценкой β и a — несмещенной оценкой α, это условие не использовалось.

Это объясняется двумя причинами. Первая касается дисперсии оценок a и b. Желательно, чтобы она была как можно меньше, т.е. (в вероятностном смыс-

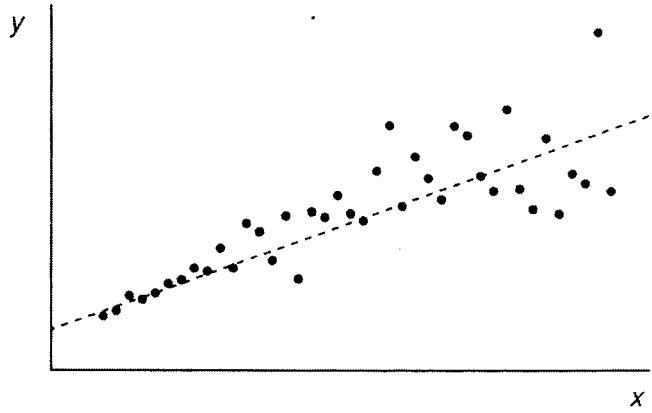


Рис. 7.2. Модель с гетероскедастичным случайным членом

ле) обеспечивала максимальную точность. При отсутствии гетероскедастичности обычные коэффициенты регрессии имеют наиболее низкую дисперсию среди всех несмещенных оценок, являющихся линейными функциями от наблюдений y. Если имеет место гетероскедастичность, то оценки МНК, которые мы до сих пор использовали, неэффективны. Можно, по меньшей мере в принципе, найти другие оценки, которые имеют меньшую дисперсию и, тем не менее, являются несмещенными.

Вторая, не менее важная причина заключается в том, что сделанные оценки стандартных ошибок коэффициентов регрессии будут неверны. Они вычисляются на основе предположения о том, что распределение случайного члена гомоскедастично; если это не так, то они неверны. Вполне вероятно, что стандартные ошибки будут занижены, а следовательно, t-статистика — завышена, и будет получено неправильное представление о точности оценки уравнения регрессии. Возможно, вы решите, что коэффициент значимо отличается от нуля при данном уровне значимости, тогда как в действительности это не так.

Свойство неэффективности можно легко объяснить интуитивно. Предположим, что имеется гетероскедастичность типа, показанного на рис. 7.1 и 7.2. Наблюдение, для которого теоретическое распределение случайного члена имеет малое стандартное отклонение (как в наблюдении *1* на рис. 7.1), будет обычно находиться близко к линии регрессии $y = \alpha + \beta x$ и, следовательно, может стать хорошим направляющим ориентиром, указывающим место этой линии. В противоположность этому наблюдение, где теоретическое распределение имеет большое стандартное отклонение (как в наблюдении *n* на рис. 7.1), не сможет существенно помочь в определении местоположения линии регрессии. Обычный МНК не делает различия между качеством наблюдений, придавая одинаковые «веса» каждому из них независимо от того, является ли наблюдение хорошим или плохим для определения местоположения этой линии. Из этого следует, что, если мы сможем найти способ придания большего «веса» наблюдениям высокого качества и меньшего — наблюдениям низкого качества, мы, вероятно, получим более точные оценки. Другими словами, оценки для α и β будут более эффективными. О том, как это делается, пойдет речь в разделе 7.4.

Гетероскедастичность становится проблемой, когда значения переменных в уравнении регрессии значительно различаются в разных наблюдениях. Если истинная зависимость описывается уравнением (7.1) и изменения значений невключенных переменных, и ошибки измерения, влияя на случайный член, делают его сравнительно малым при малых y и x и сравнительно большим — при больших y и x, то экономические переменные часто совместно меняют свой масштаб.

Предположим, например, что вы пользуетесь моделью парной регрессии (7.1) для рассмотрения зависимости между государственными расходами на образование (y) и валовым внутренним продуктом (x) в различных странах и вы сделали выборку наблюдений, представленных в табл. 7.1, включающую как малые страны, такие, как Сингапур, так и очень большие, такие, как США. Доля государственных расходов на образование в валовом внутреннем продукте обычно находится в диапазоне 3–9%; по-видимому, отдельные страны уделяют больше внимания частному образованию, чем другие, или правительства одних стран в большей степени, чем правительства других, осознают необходимость образования[1]. По социальным или иным причинам та или иная страна тратит на образование долю ВВП, которая может колебаться в пределах до 3% выше или ниже нормы. Очевидно, что при большом объеме ВВП изменение на 1% его абсолютной величины будет выражаться значительно большими цифрами, чем при малом.

Гетероскедастичность может также появляться при анализе временных рядов. Если наблюдения, используемые для построения регрессии вида (7.1), представляют собой данные временного ряда и если x и y увеличиваются со временем, то может случиться, что и дисперсия случайного члена со временем тоже будет расти.

7.3. Обнаружение гетероскедастичности

Очень часто появление проблемы гетероскедастичности можно предвидеть заранее, основываясь на знании характера данных. В таких случаях можно предпринять соответствующие действия по устранению этого эффекта на этапе спецификации модели регрессии, и это позволит уменьшить или, возможно, устранить необходимость формальной проверки. К настоящему времени для такой проверки предложено большое число тестов (и, соответственно, критериев для них). Мы рассмотрим три обычно используемых теста (критерия), в которых делаются различные предположения о зависимости между дисперсией случайного члена и величиной объясняющей переменной (или объясняющих переменных): тест ранговой корреляции Спирмена, тест Голдфелда—Квандта и тест Глейзера.

[1] В выборке имеются некоторые различия оценок расходов на образование, вызванные расхождениями в методике их определения.

Таблица 7.1

Государственные расходы на образование (EE), валовой внутренний продукт (GDP) и численность населения (P) в выборке стран, 1980 г.

Страна	*EE*	*GDP*	*EE/GDP*	*P*	*EE/P*	*GDP/P*
Люксембург	0,34	5,67	6,0	0,36	944	15750
Уругвай	0,22	10,13	2,1	2,90	76	3493
Сингапур	0,32	11,34	2,8	2,39	134	4745
Ирландия	1,23	18,88	6,5	3,44	358	5488
Израиль	1,81	20,94	8,6	3,87	468	5411
Венгрия	1,02	22,16	4,6	10,71	95	2069
Новая Зеландия	1,27	23,83	5,3	3,10	410	7687
Португалия	1,07	24,67	4,3	9,93	108	2484
Гонконг	0,67	27,56	2,4	5,07	132	5436
Чили	1,25	27,57	4,5	11,10	113	2484
Греция	0,75	40,15	1,9	9,60	78	4182
Финляндия	2,80	51,62	5,4	4,78	586	10799
Норвегия	4,90	57,71	8,5	4,09	1198	14110
Югославия	3,50	63,03	5,6	22,34	157	2821
Дания	4,45	66,32	6,7	5,12	869	12953
Турция	1,60	66,97	2,4	44,92	36	1491
Австрия	4,26	76,88	5,5	7,51	567	10237
Швейцария	5,31	101,65	5,2	6,37	834	15958
Саудовская Аравия	6,40	115,97	5,5	8,37	765	13855
Бельгия	7,15	119,49	6,0	9,86	725	12119
Швеция	11,22	124,15	9,0	8,31	1350	14940
Австралия	8,66	140,98	6,1	14,62	592	9643
Аргентина	5,56	153,85	6,5	27,06	205	5686
Нидерланды	13,41	169,38	7,9	14,14	948	11979
Мексика	5,46	186,33	2,9	67,40	81	2765
Испания	4,79	211,78	2,3	37,43	128	5658
Бразилия	8,92	249,72	3,6	123,03	73	2030
Канада	18,90	261,41	7,2	23,94	789	10919
Италия	15,95	395,52	4,0	57,04	280	6934
Великобритания	29,90	534,97	5,6	55,95	534	9562
Франция	33,59	655,29	5,1	53,71	625	12201
ФРГ	38,62	815,00	4,7	61,56	627	13239
Япония	61,61	1040,45	5,9	116,78	528	8909
США	181,30	2586,40	7,0	227,64	796	11362

Источники: Данные о государственных расходах на образование и о населении взяты из табл. 1.1, 4.1 и приложения С статистического ежегодника ЮНЕСКО «Statistical Yearbook» (1984). Данные о расходах на образование для Италии и Греции относятся к 1979 г. Данные о валовом внутреннем продукте взяты из источника Международного валютного фонда «International Financial Statistics», Supplement (1984).

Тест ранговой корреляции Спирмена

При выполнении теста ранговой корреляции Спирмена предполагается, что дисперсия случайного члена будет либо увеличиваться, либо уменьшаться по мере увеличения x, и поэтому в регрессии, оцениваемой с помощью МНК, абсолютные величины остатков и значения x будут коррелированы. Данные по x и остатки упорядочиваются, и коэффициент ранговой корреляции определяется как

$$r_{x,e} = 1 - \frac{6\sum D_i^2}{n(n^2-1)}, \tag{7.3}$$

где D_i — разность между рангом x и рангом e.

Если предположить, что коэффициент корреляции для генеральной совокупности равен нулю, то коэффициент ранговой корреляции имеет нормальное распределение с математическим ожиданием 0 и дисперсией $1/(n-1)$ в больших выборках. Следовательно, соответствующая тестовая статистика равна $r_{x,e}\sqrt{n-1}$, и при использовании двустороннего критерия нулевая гипотеза об отсутствии гетероскедастичности будет отклонена при уровне значимости в 5%, если она превысит 1,96, и при уровне значимости в 1%, если она превысит 2,58. Если в модели регрессии имеется более одной объясняющей переменной, то проверка гипотезы может выполняться с использованием любой из них.

Пример

По данным, приведенным в табл. 7.1, с помощью МНК оценена следующая регрессионная зависимость расходов на образование (*EE*) от валового внутреннего продукта (*GDP*) (стандартные ошибки указаны в скобках):

$$\begin{array}{ll} \hat{EE} = -2{,}32 + 0{,}067\,GDP; & R^2 = 0{,}98; \\ \quad (0{,}91)\ \ (0{,}002) & F = 1{,}524. \end{array} \tag{7.4}$$

Это предполагает, что при каждом увеличении ВВП в перекрестной выборке на 1 млрд. долл. на образование будет затрачиваться дополнительно 67 млн. долл. (другими словами, 6,7 цента на дополнительный доллар). Отклонения от линии регрессии, а также объемы ВВП приведены в упорядоченном по возрастанию виде в табл. 7.2, и на их основе вычислены показатели D_i и D_i^2. Сумма последних составила 2,676. Таким образом, коэффициент ранговой корреляции равен:

$$r_{x,e} = 1 - \frac{6(2{,}676)}{34(1{,}155)} = 0{,}59, \tag{7.5}$$

и тестовая статистика составляет $(0{,}59)(\sqrt{33}) = 3{,}39$. Это выше, чем 2,58, и, следовательно, нулевая гипотеза об отсутствии гетероскедастичности при однопроцентном уровне значимости отклоняется.

Таблица 7.2

x	Ранг	|e|	Ранг	D	D^2	x	Ранг	|e|	Ранг	D	D^2
5,67	1	2,28	17	−16	256	101,65	18	0,83	3	15	225
10,13	2	1,86	14	−12	144	115,97	19	0,96	4	15	225
11,34	3	1,88	15	−12	144	119,49	20	1,48	7	13	169
18,88	4	2,29	18	−14	196	124,15	21	5,24	27	−6	36
20,94	5	2,73	21	−16	256	140,98	22	1,55	8	14	196
22,16	6	1,86	13	−7	49	153,85	23	2,41	20	3	9
23,83	7	2,00	16	−9	81	169,38	24	4,40	25	−1	1
24,67	8	1,74	12	−4	16	186,33	25	4,68	26	−1	1
27,56	9	1,15	5	4	16	211,78	26	7,06	30	−4	16
27,57	10	1,73	11	−1	1	249,72	27	5,46	28	−1	1
40,15	11	0,38	1	10	100	261,41	28	3,73	24	4	16
51,62	12	1,67	10	2	4	395,52	29	8,19	32	−3	9
57,71	13	3,36	22	−9	81	534,97	30	3,56	23	7	49
63,03	14	1,60	9	5	25	655,29	31	7,92	31	0	0
66,32	15	2,33	19	−4	16	815,00	32	13,58	34	−2	4
66,97	16	0,56	2	14	196	1040,45	33	5,67	29	4	16
76,88	17	1,44	6	11	121	2586,40	34	10,61	33	1	1

Тест Голдфелда—Квандта

Вероятно, наиболее популярным формальным критерием является критерий, предложенный С. Голдфелдом и Р. Квандтом (Goldfeld, Quandt, 1956). При проведении проверки по этому критерию предполагается, что стандартное отклонение (σ_i) распределения вероятностей u_i пропорционально значению x в этом наблюдении. Предполагается также, что случайный член распределен нормально и не подвержен автокорреляции.

Все n наблюдений в выборке упорядочиваются по величине x, после чего оцениваются отдельные регрессии для первых n' и для последних n' наблюдений; средние $(n - 2n')$ наблюдений отбрасываются. Если предположение относительно природы гетероскедастичности верно, то дисперсия u в последних n' наблюдениях будет больше, чем в первых n', и это будет отражено в сумме квадратов остатков в двух указанных «частных» регрессиях. Обозначая суммы квадратов остатков в регрессиях для первых n' и последних n' наблюдений соответственно через RSS_1 и RSS_2, рассчитаем отношение RSS_2/RSS_1, которое имеет F-распределение с $(n' - k - 1)$ и $(n' - k - 1)$ степенями свободы, где k — число

объясняющих переменных в регрессионном уравнении. Мощность критерия зависит от выбора n' по отношению к n. Основываясь на результатах некоторых проведенных ими экспериментов, С. Голдфелд и Р. Квандт утверждают, что n' должно составлять порядка 11, когда $n = 30$, и порядка 22, когда $n = 60$. Если в модели имеется более одной объясняющей переменной, то наблюдения должны упорядочиваться по той из них, которая, как предполагается, связана с σ_i, и n' должно быть больше, чем $k + 1$ (где k — число объясняющих переменных).

Метод Голдфелда—Квандта может также использоваться для проверки на гетероскедастичность при предположении, что σ_i обратно пропорционально x_i. При этом используется та же процедура, что и описанная выше, но тестовой статистикой теперь является показатель RSS_1/RSS_2, который вновь имеет F-распределение с $(n' - k - 1)$ и $(n' - k - 1)$ степенями свободы.

Примеры

На основе данных табл. 7.1 с помощью обычного МНК были оценены регрессии сначала по наблюдениям для 12 стран с наименьшим валовым национальным продуктом (ВНП), а затем для 12 стран с наибольшим ВНП. Сумма квадратов отклонений в первой регрессии была равна 2,68, а во второй — 388,24. Соотношение RSS_2/RSS_1, следовательно, составило 144,9. Критическое значение $F(10,10)$ равно 4,85 при однопроцентном уровне значимости, и нулевая гипотеза об отсутствии гетероскедастичности снова отклоняется.

Тест Глейзера

Тест Глейзера позволяет несколько более тщательно рассмотреть характер гетероскедастичности. Мы снимаем предположение о том, что σ_i пропорционально x_i, и хотим проверить, может ли быть более подходящей какая-либо другая функциональная форма, например

$$\sigma_i = \alpha + \beta x_i^{\gamma}. \tag{7.6}$$

Чтобы использовать данный метод, следует оценить регрессионную зависимость y от x с помощью обычного МНК, а затем вычислить абсолютные величины остатков $|e_i|$ по функции (7.6) для данного значения γ. Можно построить несколько таких функций, изменяя значение γ. В каждом случае нулевая гипотеза об отсутствии гетероскедастичности будет отклонена, если оценка β значимо отличается от нуля. Если при оценивании более чем одной функции получается значимая оценка β, то ориентиром при определении характера гетероскедастичности может служить наилучшая из них.

Пример

На основе данных табл. 7.2 по x и $|e|$ с использованием значений γ от $-1,0$ до 1,5 были оценены уравнения (7.6). Результаты представлены в обобщенном виде в табл. 7.3.

Таблица 7.3

γ	a	$c.o.(a)$	b	$c.o.(b)$	R^2	F
−1,0	4,19	0,61	−28,0	14,0	0,11	4,0
−0,5	5,74	0,80	−17,1	5,0	0,27	11,7
0,5	0,58	0,51	0,24	0,03	0,62	52,7
1,0	2,37	0,42	0,0044	0,0008	0,49	31,1
1,5	2,90	0,44	0,000077	0,000019	0,35	17,5

Следует отметить, что различные оценки β несравнимы, так как определение объясняющей переменной (x^{γ}) в каждом случае разное. Статистически значимые оценки были получены для последних четырех значений γ. Уровни коэффициента R^2 сравнимы в том смысле, что зависимая переменная в каждом случае одна и та же. Наилучший результат соответствует значению $\gamma = 0{,}5$, и, следовательно, гетероскедастичность аппроксимируется уравнением:

$$s_i = 0{,}58 + 0{,}24\sqrt{x_i}. \qquad (7.7)$$

Другими словами, стандартное отклонение распределения величины u действительно увеличивается с ростом x, но не в такой же пропорции.

Упражнения

7. 1

Страна	*M*	*G*	*Страна*	*M*	*G*
Бельгия	849	2652	Люксембург	1368	3108
Канада	778	3888	Нидерланды	704	2429
Дания	853	3159	Норвегия	634	2881
Франция	1000	2777	Португалия	215	718
Германия	1331	3095	Испания	239	957
Греция	185	1091	Швеция	1025	4101
Ирландия	399	1331	Великобритания	609	2174
Италия	554	1731	США	1248	4799
Япония	679	1887			

Используя данные из приведенной выше таблицы, исследователь оценивает регрессионную зависимость выпуска продукции обрабатывающей промышленности на душу населения в 1970 г. (M) от валового внутреннего продукта на душу населения в том же году (G) (как M, так и G измеряются в долларах США) и получает формулу (в скобках приводятся стандартные ошибки):

$$\hat{M} = \underset{(128,1)}{74,2} + \underset{(0,05)}{0,27}G; \qquad R^2 = 0,69.$$

1. Изобразите диаграмму рассеяния, используя данные из таблицы, и объясните, почему исследователь может подозревать наличие гетероскедастичности.

2. Исследователь оценивает две «частные» регрессии для шести стран с наименьшими значениями показателя G и для шести стран с наибольшими значениями этого показателя. Сумма квадратов отклонений составляет 20,523 в первом случае и 313,842 — во втором. Выполните проверку на гетероскедастичность по критерию Голдфелда—Квандта.

3. Как гетероскедастичность будет влиять на свойства оцениваемых коэффициентов?

7.2. Что касается примера с государственными расходами на образование, то здесь можно высказать мнение о том, что гетероскедастичность в значительной степени обусловлена наблюдением для США, которые по сравнению с другими странами в выборке имеют значительно большие значения *EE* и *GDP*. Поэтому был повторно выполнен тест Голдфелда—Квандта с исключением из выборки этого наблюдения. Суммы квадратов отклонений в регрессиях с использованием первых 12 и последних 12 из 33 наблюдений соответственно составили 2,68 и 202,9. Какой вывод вы сделаете?

7.4. Что можно сделать в случае гетероскедастичности?

Пусть σ_i — стандартное отклонение случайного члена в наблюдении i. В том случае если бы было известно σ_i для каждого наблюдения, можно было бы устранить гетероскедастичность, разделив каждое наблюдение на соответствующее ему значение σ. Тогда случайный член в i-м наблюдении становится равным u_i/σ_i и его теоретическая дисперсия представляется в виде:

$$E\left\{\frac{u_i}{\sigma_i}\right\}^2,$$

что равняется:

$$\frac{1}{\sigma_i^2}E(u_i^2).$$

Это выражение переписывается как

$$\frac{1}{\sigma_i^2}(\sigma_i^2),$$

и, следовательно, оно равно единице. Таким образом, каждое наблюдение будет

иметь случайный член, полученный из генеральной совокупности с единичной дисперсией, и модель будет гомоскедастичной. Теперь модель имеет вид:

$$\frac{y_i}{\sigma_i} = \frac{\alpha}{\sigma_i} + \beta\frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i}, \tag{7.8.}$$

что может быть переписано как

$$y' = \alpha v + \beta x' + u', \tag{7.9}$$

где y_i' определяется как $\frac{y_i}{\sigma_i}$; x_i' представляет собой $\frac{x_i}{\sigma_i}$; v — новая переменная, i-е наблюдение которой равно $\frac{1}{\sigma_i}$; величина u_i' есть $\frac{u_i}{\sigma_i}$. Следует отметить, что в данном уравнении не должно быть постоянного члена. Оценивая регрессионную зависимость y' от v и x', мы получим эффективные оценки для α и β с несмещенными стандартными ошибками.

Математическое доказательство того, что уравнение (7.9) даст более эффективные оценки, чем уравнение (7.1), выходит за рамки данной книги, но здесь можно дать простое интуитивное объяснение. Наблюдения с наименьшими значениями σ_i будут наиболее полезными для определения истинной зависимости между y и x, поскольку величина случайного члена в них, как правило, наименьшая. Мы воспользуемся этим, оценивая так называемую *взвешенную регрессию*, придавая наибольшие веса наблюдениям самого «высокого качества», а наименьшие веса — соответственно, наблюдениям самого «низкого качества». Уравнение (7.9) можно рассматривать как «взвешенный» вариант уравнения (7.1), где значения y и x были умножены на величины $1/\sigma_i$, которые, конечно, тем больше, чем меньше σ_i.

Препятствием для этой процедуры является то, что вам почти наверняка будут неизвестны фактические значения σ_i. Однако процедура будет применимой, если мы сможем подобрать некоторую величину, пропорциональную, по нашему мнению, σ в каждом наблюдении, и разделим на нее обе части уравнения.

Допустим, есть основания предположить, что некоторая величина z пропорциональна σ_i и $z_i = \lambda\sigma_i$, где λ — некоторая константа. После деления на z уравнение принимает вид:

$$\frac{y_i}{z_i} = \frac{\alpha}{z_i} + \beta\frac{x_i}{z_i} + \frac{u_i}{z_i}, \tag{7.10}$$

Дисперсия случайного члена представлена как

$$E\left\{\frac{u_i}{z_i}\right\}^2 = E\left\{\frac{u_i}{\lambda\sigma_i}\right\}^2 = \frac{1}{\lambda^2}\frac{\sigma_i^2}{\sigma_i^2},$$

что равно $1/\lambda^2$. Следовательно, эта величина постоянна для всех наблюдений, и проблема устранена.

Например, может оказаться целесообразным предположить, что σ прибли-

зительно пропорционально x, как в критерии Голдфелда—Квандта. Если после этого мы разделим каждое наблюдение на соответствующее ему значение x, то уравнение (7.1) примет вид:

$$\frac{y}{x} = \alpha \frac{1}{x} + \beta + \frac{u}{x}, \tag{7.11}$$

и при этом, если повезет, новый случайный член u/x будет иметь постоянную дисперсию. Затем мы оцениваем регрессионную зависимость y/x от $1/x$, включив в уравнение постоянный член. Коэффициент при $1/x$ будет эффективной оценкой α, а постоянный член — эффективной оценкой β. В примере с расходами на образование, рассмотренном в предыдущем разделе, зависимой переменной будут государственные расходы на образование как доля ВВП, а объясняющей переменной — обратная к ВВП величина.

Иногда в нашем распоряжении может оказаться несколько переменных, каждую из которых можно использовать для масштабирования уравнения. В примере с расходами на образование альтернативной переменной является численность населения страны (P). Разделив обе части уравнения (7.1) на эту величину, получаем:

$$\frac{y}{P} = \alpha \frac{1}{P} + \beta \frac{x}{P} + \frac{u}{P}, \tag{7.12}$$

и мы снова надеемся на то, что случайный член u_i/P_i будет иметь постоянную дисперсию для всех наблюдений. Таким образом, теперь оценивается регрессионная зависимость государственных расходов на образование на душу населения от ВВП на душу населения и обратной величины от численности населения, причем на этот раз без постоянного члена.

На практике имеет смысл попробовать использовать несколько разных переменных для масштабирования наблюдений и сравнить затем результаты. Если каждый раз получаются сходные результаты и тесты не дают оснований отклонять нулевую гипотезу о гомоскедастичности, то проблему можно считать решенной.

Примеры

Регрессионная зависимость государственных расходов на образование от ВВП была построена в двух вариантах: 1) с делением на ВВП и 2) с делением на численность населения. Данные о численности населения, среднедушевых государственных расходах на образование и ВВП на душу населения приводятся в табл. 7.1. Результаты оказались следующими (в скобках приводятся стандартные ошибки):

$$\frac{\widehat{EE}}{GDP} = \underset{(0,094)}{-0,066} \frac{1}{GDP} + \underset{(0,004)}{0,053}; \qquad R^2 = 0,15; \quad F = 0,48; \tag{7.13}$$

$$\frac{\widehat{EE}}{P} = \underset{(0,057)}{-0,022} \frac{1}{P} + \underset{(0,003)}{0,062} \frac{GDP}{P}; \qquad R^2 = 0,83; \quad F = 160,9. \tag{7.14}$$

В каждом случае оценивались «частные» регрессии по первым 12 и последним 12 наблюдениям в выборке, которые определялись упорядочением по переменной GDP в первой регрессии и по GDP/P — во второй. В первом случае RSS_1 было больше, чем RSS_2, что показывает, что пересчет более чем компенсировал гетероскедастичность, но при этом отношение RSS_1/RSS_2 равнялось 1,37 и, следовательно, было недостаточно высоким, чтобы указывать на статистически значимую гетероскедастичность. Во втором случае RSS_2/RSS_1 было равно 4,60, что указывает на то, что нулевая гипотеза о гомоскедастичности должна быть отклонена при уровне значимости в 5% (критическое значение F составляет 2,98).

После преобразования этих уравнений обратно к форме (7.4) можно видеть, что оценки коэффициента при GDP — это такого же порядка величины, что и в данном уравнении, но они несколько ниже. На этой основе было бы целесообразно сделать вывод о том, что указанный коэффициент ближе к 0,06, чем к 0,07. Стандартные ошибки, как это видно, стали больше, но сравнение их со стандартной ошибкой коэффициента при GDP в уравнении (7.4) было бы некорректным по той причине, что последний почти наверняка был серьезно занижен. Оценки величины α незначимо отличаются от нуля как в уравнении (7.13), так и в уравнении (7.14).

Может вызывать беспокойство то, что уровень коэффициента детерминации R^2 стал ниже, чем в уравнении (7.4). Действительно, в уравнении (7.13) он настолько низок, что F-статистика не отличается значимо от нуля даже при уровне значимости в 5%. Следует, однако, помнить, что определение зависимой переменной в каждом уравнении свое, поэтому значения коэффициента R^2 в них несравнимы. В уравнении (7.13) коэффициент R^2 измеряет объясняющую способность переменной $1/GDP$, которая зависит от значимости в исходном уравнении постоянного члена α, а не β. Теперь же параметр β стал постоянным членом и поэтому не может внести какой-либо вклад в величину R^2.

Подход Глейзера

После выполнения теста Глейзера мы могли устранить гетероскедастичность за счет приравнивания z_i в уравнении (7.10) к σ_i в уравнении (7.6) и оценивания регрессионной зависимости y/s_i от $1/s_i$ и x_i/s_i, где s_i является оценкой σ_i. После расчета s_i на основе формулы (7.7) была оценена следующая регрессия (в скобках даны стандартные ошибки):

$$\frac{\hat{EE}}{s_i} = \underset{(0,30)}{-0,32}\frac{1}{s_i} + \underset{(0,003)}{0,059}\frac{GDP}{s_i}; \qquad R^2 = 0,89; \quad F = 263,8. \tag{7.15}$$

В сущности, данное уравнение схоже с уравнениями (7.13) и (7.14). Оценивание регрессионной зависимости его остатков от GDP^{γ}, где γ равно $-0,5$, 0,5 и 1,0, не привело к отклонению нулевой гипотезы о наличии гомоскедастичности.

Предположим, что истинная модель имеет нелинейную форму (7.16), рассмотренную в разделе 4.3, и что для определенности β положительна; таким образом, y является возрастающей функцией от x:

$$y = \alpha x^{\beta} v. \tag{7.16}$$

Мультипликативный случайный член v увеличивает или уменьшает y в соответствующей случайной пропорции. Распределение вероятностей v одинаково для всех наблюдений, что означает, например, что вероятность возрастания y на 2% под действием этой величины одинакова, независимо от того, является ли x малым или большим. Тем не менее абсолютная величина прироста y на 2% оказывается большей, когда x принимает высокие, а не низкие значения. Следовательно, будет проявляться тенденция к более сильному разбросу наблюдений вокруг истинной зависимости по мере увеличения x, и линейная регрессионная зависимость y от x может, следовательно, показывать гетероскедастичность.

Решением здесь, конечно, является переход к логарифмической регрессии. Это не только было бы более подходящей математической спецификацией, но и сделало бы модель регрессии гомоскедастичной:

$$\log y = \log \alpha + \beta \log x + \log v. \tag{7.17}$$

Случайный член $\log v$ теперь воздействует на зависимую переменную $\log y$ аддитивно, поэтому абсолютная величина его воздействия не зависит от величины $\log x$.

Пример

Оценка логарифмической регрессионной зависимости государственных расходов на образование от ВВП с использованием данных табл. 7.1 дает следующие результаты (в скобках приводятся стандартные ошибки):

$$\widehat{\log} EE = \underset{(0,24)}{-3,31} + \underset{(0,05)}{1,06} \log GDP; \qquad R^2 = 0,93; \quad F = 420,4, \tag{7.18}$$

откуда видно, что эластичность величины EE по объему ВВП приблизительно равна единице. Для повышения качества расчетов была также оценена логарифмическая регрессионная зависимость государственных расходов на образование на душу населения от ВВП на душу населения:

$$\widehat{\log} \frac{EE}{P} = \underset{\text{(с. о.)}}{-3,75} + \underset{(0,09)}{1,37} \log \frac{GDP}{P}; \qquad R^2 = 0,89; \quad F = 254,8. \tag{7.19}$$

Результат близок к предшествующему с несколько более высокой эластичностью. Для обоих вариантов были определены «частные» регрессии для первых 12 и последних 12 наблюдений, и в обоих случаях RSS_1 было больше, чем RSS_2.

а отношения RSS_1/RSS_2, соответственно, составили 1,92 и 2,78. Критическое значение F-статистики при 10 и 10 степенях свободы и уровне значимости в 5% составляет 2,98. Таким образом, в обоих случаях нулевая гипотеза о гомоскедастичности не будет отклонена.

Имеет ли в действительности значение гетероскедастичность?

Ответ на этот вопрос зависит от степени варьирования наблюдаемых значений объясняющих переменных (предполагается, что их разброс является ориентиром для определения величины стандартных ошибок случайного члена). Некоторые вычисления, выполненные Р. Гири (Geary, 1966), показывают, что если стандартное отклонение случайного члена пропорционально значениям объясняющей переменной в парной регрессии, то дисперсия оценки коэффициента наклона может быть в три раза больше при использовании обычного МНК по сравнению с тем случаем, когда делается поправка на гетероскедастичность.

Упражнения

7.3. Установив факт наличия гетероскедастичности, исследователь в упражнении 7.1 перешел к оцениванию следующих зависимостей:

$$\frac{\hat{M}}{G} = \underset{\text{(с. о.)}}{0{,}32} - \underset{(0{,}03)}{39{,}4}\, Z; \qquad R^2 = 0{,}03;$$

$$\widehat{\log} M = \underset{\text{(с. о.)}}{-1{,}66} + \underset{(0{,}92)}{1{,}05} \log G; \qquad \underset{(0{,}12)}{} \qquad R^2 = 0{,}84,$$

где Z является обратной величиной от G (логарифмы берутся по основанию e).

1. Почему проблема гетероскедастичности может в этих уравнениях стать менее существенной?

2. Сравните полученные результаты для этих уравнений и для уравнения в упражнении 7.1.

7.4. Студенту (назовем его A) дали 20 наблюдений двух переменных — y и x. Ему сообщили, что y линейно зависит от x и случайного члена u, и попросили получить оценку коэффициента при x. Истинная зависимость, неизвестная студенту, имеет вид:

$$y = 100 + 10x + u.$$

Используя обычный МНК, студент оценивает уравнение регрессии $y = a + bx$, получая (в скобках указаны стандартные ошибки):

$$\hat{y} = \underset{(12{,}2)}{92{,}0} + \underset{(1{,}0)}{11{,}4}\, x; \qquad R^2 = 0{,}87. \tag{1}$$

Затем студенту сообщают, что случайный член u_i в i-м наблюдении получен

путем умножения случайного числа ε_i на значение x в этом наблюдении, то есть $u_i = x_i \varepsilon_i$, где ε_i получено из нормального распределения с нулевым математическим ожиданием и единичной дисперсией. Студент определяет новые переменные y' и x', где $y' = y/x$, $x' = 1/x$, и оценивает уравнение $y' = c + dx'$, получая:

$$\hat{y}' = 10{,}2 + 101{,}7\, x'; \qquad R^2 = 0{,}99. \qquad (2)$$
$$\text{(с. о.)} \quad (0{,}7) \quad (2{,}3)$$

Девять других студентов (*B, C, ..., J*) выполняют такие же эксперименты с наблюдениями переменной y, полученными тем же способом и при таких же значениях x, но с другими наборами случайных чисел для ε. Полученные результаты обобщены в таблице.

Студент	Регрессия 1					Регрессия 2				
	a	*с.о.(a)*	*b*	*с.о.(b)*	R^2	*c*	*с.о.(c)*	*d*	*с.о.(d)*	R^2
A	92,0	12,2	11,4	1,0	0,87	10,2	0,7	101,7	2,3	0,99
B	93,6	7,5	11,3	0,6	0,95	10,7	0,5	98,1	1,7	0,99
C	99,0	11,1	10,2	0,9	0,87	10,1	0,6	99,9	2,0	0,99
D	99,4	8,3	9,6	0,7	0,91	9,3	0,4	102,5	1,4	0,99
E	114,9	19,1	7,8	1,6	0,57	9,7	0,8	97,7	2,8	0,99
F	88,4	12,6	12,2	1,1	0,88	11,2	0,6	97,9	2,3	0,99
G	99,6	9,8	10,0	0,8	0,89	9,9	0,6	99,9	2,0	0,99
H	100,3	6,1	9,8	0,5	0,95	9,9	0,5	100,2	1,7	0,99
I	104,7	9,4	9,9	0,8	0,78	10,7	0,5	97,3	1,8	0,99
J	91,1	11,3	11,0	0,9	0,88	10,0	0,6	100,2	2,1	0,99

1. Объясните, почему студенты оценили регрессию (2), когда им стал известен характер поведения случайного члена *u*.

2. Являются ли результаты, полученные для регрессии (2), лучшими, чем результаты, полученные для регрессии (1)?

7.5. Двум молодым исследователям поручено получить оценки (x_1 и x_2) математического ожидания μ случайной переменной x. Из прошлого опыта нам известно, что оба они получат несмещенные оценки. Однако один из них менее аккуратен, чем другой, и дисперсия x_2 будет в три раза больше, чем дисперсия x_1. Вы должны, используя результаты работы двух исследователей, представить единый результат. Возьмете ли вы среднее от x_1 и x_2, полностью проигнорируете x_2 или поступите как-нибудь иначе?

7.5. Автокорреляция и связанные с ней факторы

До сих пор предполагалось, что значение случайного члена u в любом наблюдении определяется независимо от его значений во всех других наблюдениях. Другими словами, мы предполагали, что удовлетворено третье условие Гаусса—Маркова, то есть pop. cov $(u_i, u_j) = 0$ при $i \neq j$[1].

Последствия *автокорреляции* в некоторой степени сходны с последствиями гетероскедастичности. Коэффициенты регрессии остаются несмещенными, но становятся неэффективными, и их стандартные ошибки оцениваются неправильно (вероятно, они смещаются вниз, т. е. занижаются).

Возможные причины автокорреляции

Автокорреляция обычно встречается только в регрессионном анализе при использовании данных временных рядов. Случайный член u в уравнении регрессии подвергается воздействию тех переменных, влияющих на зависимую переменную, которые не включены в уравнение регрессии. Если значение u в любом наблюдении должно быть независимым от его значения в предыдущем наблюдении, то и значение любой переменной, «скрытой» в u, должно быть некоррелированным с ее значением в предыдущем наблюдении.

Постоянная направленность воздействия не включенных в уравнение переменных является наиболее частой причиной *положительной автокорреляции* — ее обычного для экономического анализа типа. Предположим, что вы оцениваете уравнение спроса на мороженое по ежемесячным данным и что состояние погоды является единственным важным фактором, «скрытым» в u. Вероятно, у вас будет несколько последовательных наблюдений, когда теплая погода способствует увеличению спроса на мороженое и, таким образом, u положительно, и после этого — несколько последовательных наблюдений, когда ситуация складывается противоположным образом, после чего идет еще один ряд теплых месяцев и т. д.

Если доход постоянно возрастает со временем, схема наблюдений может быть такой, как показано на рис. 7.3. При обозначении объема продаж мороженого через y и дохода через x будет иметь место трендовая зависимость, отражающая рост объема продаж: $y = \alpha + \beta x$. Фактические наблюдения будут в основном сначала находиться выше линии регрессии, затем ниже ее и затем опять выше.

Изменения экономической конъюнктуры часто приводят к похожим результатам, особенно наглядным в макроэкономическом анализе, и в литературе о циклах деловой активности есть много таких примеров.

Здесь важно отметить, в частности, что автокорреляция в целом представляет тем более существенную проблему, чем меньше интервал между наблюдениями. Очевидно, что чем больше этот интервал, тем менее правдоподобно, что при переходе от одного наблюдения к другому характер влияния неучтенных переменных будет сохраняться.

---

[1] Так как мы предполагаем, что $E(u_i) = E(u_j) = 0$, это условие может быть также записано как $E(u_i, u_j) = 0$. Когда данное условие не выполняется, говорят, что случайный член подвержен *автокорреляции*, которую часто называют *сериальной корреляцией* (эти два термина взаимозаменяемы).

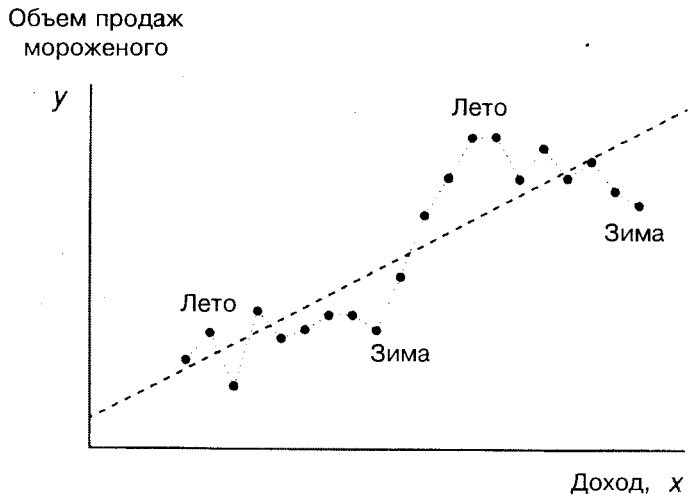


Рис. 7.3. Положительная автокорреляция

Если в примере с мороженым наблюдения проводятся не ежемесячно, а ежегодно, то автокорреляции, вероятно, вообще не будет. Мало вероятно, чтобы совокупное влияние погодных условий в одном году коррелировало с аналогичным влиянием в следующем году.

Пока мы рассматривали только положительную автокорреляцию. В принципе автокорреляция может также быть отрицательной. В нашем случае это означает, что корреляция между последовательными значениями случайного члена отрицательна. В этом случае, скорее всего, за положительным значением в одном наблюдении идет отрицательное значение в следующем, и наоборот; диаграмма рассеяния при этом выглядит так, как показано на рис. 7.4.

Здесь снова предполагается, что x со временем растет. Линия, соединяющая последовательные наблюдения друг с другом, будет пересекать линию, показывающую зависимость между y и x, чаще, чем можно было ожидать, если бы значения случайного члена не зависели друг от друга.

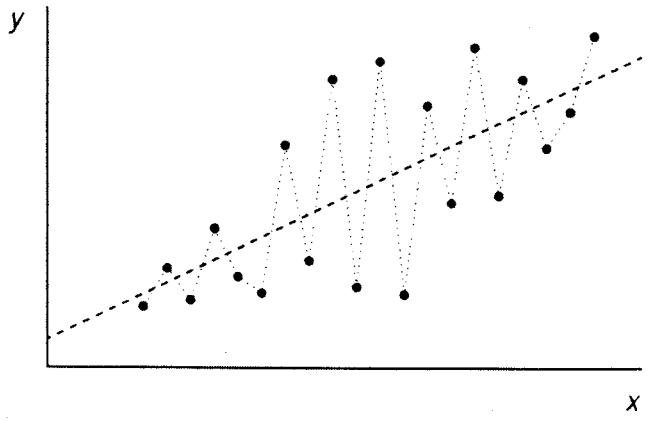


Рис. 7.4. Отрицательная автокорреляция

В экономике *отрицательная автокорреляция* встречается относительно редко. Но иногда она появляется при преобразовании первоначальной спецификации модели в форму, подходящую для регрессионного анализа. Мы встретим такой пример в разделе 10.2.

При рассмотрении автокорреляции мы будем предполагать, что имеем дело с данными временного ряда, и поэтому станем ссылаться на наблюдение t, а не i и обозначать размер выборки через T вместо n. Таким образом, базовая модель будет записана в виде:

$$y_t = \alpha + \beta x_t + u_t. \tag{7.20}$$

7.6. Обнаружение автокорреляции первого порядка: критерий Дарбина—Уотсона

Начнем с частного случая, в котором автокорреляция подчиняется авторегрессионной схеме первого порядка:

$$u_t = \rho u_{t-1} + \varepsilon_t. \tag{7.21}$$

Это означает, что величина случайного члена в любом наблюдении равна его значению в предшествующем наблюдении (т. е. его значению в период $t-1$), умноженному на ρ, плюс новый ε_t. Данная схема оказывается авторегрессионной, поскольку u определяется значениями этой же самой величины с запаздыванием, и схемой первого порядка, потому что в этом простом случае максимальное запаздывание равно единице. Предполагается, что значение ε в каждом наблюдении не зависит от его значений во всех других наблюдениях. Если ρ положительно, то автокорреляция положительная; если ρ отрицательно, то автокорреляция отрицательная. Если $\rho = 0$, то автокорреляции нет и третье условие Гаусса—Маркова удовлетворяется.

Конечно, мы не располагаем способом измерения значений случайного члена, поэтому мы не можем оценить регрессию (7.21) непосредственно. Тем не менее мы можем оценивать ρ путем оценивания регрессионной зависимости e_t от e_{t-1} с использованием обычного МНК. При этом оценка ρ равна $\dfrac{\mathrm{Cov}(e_{t-1}, e_t)}{\mathrm{Var}(e_{t-1})}$.

Так как среднее значение T остатков равно нулю, $\bar{e}_{T-1}$ (среднее значение остатков в наблюдениях от 1 до $T-1$) и $\bar{e}_T$ (среднее значение остатков в наблюдениях от 2 до T) будут близки к нулю, если выборка достаточно велика, и $\mathrm{Cov}(e_t, e_{t-1})$ и $\mathrm{Var}(e_{t-1})$ будут аппроксимироваться выражениями $\dfrac{1}{T-1}\sum e_{t-1}e_t$ и $\dfrac{1}{T-1}\sum e_{t-1}^2$, соответственно.

Кроме того, $\sum e_{t-1}^2$ будет приблизительно равно $\sum e_t^2$. Следовательно, $\dfrac{\mathrm{Cov}(e_{t-1}, e_t)}{\mathrm{Var}(e_{t-1})}$ аппроксимируется выражением $\dfrac{\sum e_{t-1}e_t}{\sum e_t^2}$.

Широко известная статистика Дарбина—Уотсона (d) определяется следующим образом [1]:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

Можно показать (см. приложение 7.3), что в больших выборках

$$d \to 2 - 2\rho. \tag{7.22}$$

Если автокорреляция отсутствует, то $\rho = 0$, и поэтому величина d должна быть близкой к двум. При наличии положительной автокорреляции величина d, вообще говоря, будет меньше двух; при отрицательной автокорреляции она, вообще говоря, будет превышать 2. Так как ρ должно находиться между значениями 1 и -1, то d должно лежать между 0 и 4.

Критическое значение d при любом данном уровне значимости зависит, как можно предполагать, от числа объясняющих переменных в уравнении регрессии и от количества наблюдений в выборке. К сожалению, оно также зависит от конкретных значений, принимаемых объясняющими переменными. Поэтому невозможно составить таблицу с указанием точных критических значений для всех возможных выборок, как это можно сделать для t- и F-статистик; но можно вычислить верхнюю и нижнюю *границы* для критического значения d. Для положительной автокорреляции они обычно обозначаются как d_U и d_L.

На рис. 7.5 данная ситуация представлена в виде схемы; стрелка указывает критический уровень d, который обозначается как $d_{крит}$. Если бы мы знали значение $d_{крит}$, то могли бы сравнить с ним значение d, рассчитанное для нашей регрессии. Если бы оказалось, что $d \geq d_{крит}$, то мы не смогли бы отклонить нулевую гипотезу от отсутствии автокорреляции. В случае $d \leq d_{крит}$ мы бы отклонили нулевую гипотезу и сделали вывод о наличии положительной автокорреляции.

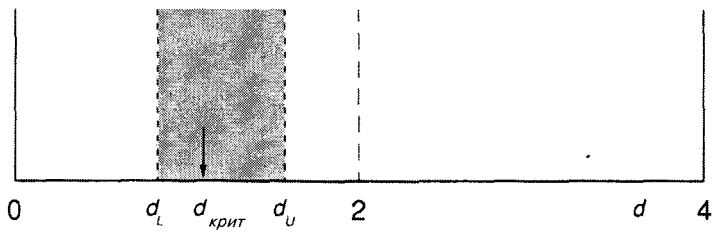


Рис. 7.5. Тест Дарбина—Уотсона на автокорреляцию (показана зона неопределенности в случае предполагаемой положительной автокорреляции)

Вместе с тем мы знаем только, что $d_{крит}$ находится где-то между d_L и d_U. Это предполагает наличие трех возможностей:

[1] В русскоязычной литературе эта статистика чаще обозначается как *DW*. (*Прим. ред.*)

1. Величина d меньше, чем d_L. В этом случае она будет также меньше, чем $d_{крит}$, и поэтому мы сделаем вывод о наличии положительной автокорреляции.

2. Величина d больше, чем d_U. В этом случае она также больше критического уровня, и поэтому мы не сможем отклонить нулевую гипотезу.

3. Величина d находится между d_L и d_U. В этом случае она может быть больше или меньше критического уровня. Поскольку нельзя определить, которая из двух возможностей налицо, мы не можем ни отклонить, ни принять нулевую гипотезу.

В случаях 1 и 2 тест Дарбина—Уотсона дает определенный ответ, но случай 3 относится к зоне невозможности принятия решения, и изменить создавшееся положение нельзя.

В табл. А.5 в конце книги даны значения (d_L и d_U), стоящие на пересечении строк и столбцов, соответствующих количеству наблюдений и числу объясняющих переменных для уровней значимости в 5 и 1%. В таблице показаны критические значения в случае положительной автокорреляции, наиболее часто встречающемся в экономических моделях. Можно видеть, что чем больше число наблюдений, тем уже зона неопределенности, представленная отрезком между d_L и d_U.

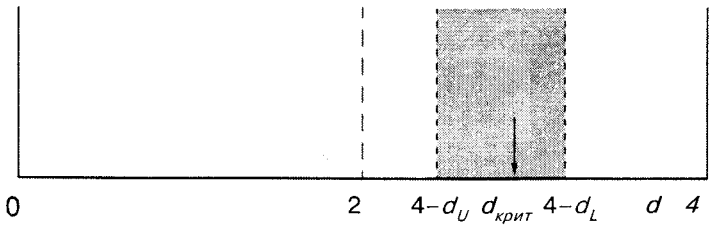


Рис. 7.6. Тест Дарбина—Уотсона на автокорреляцию
(показана зона неопределенности в случае предполагаемой отрицательной автокорреляции)

Проверка на отрицательную автокорреляцию проводится по аналогичной схеме, причем зона, содержащая критический уровень, расположена симметрично справа от 2. Так как отрицательная автокорреляция встречается относительно редко, предполагается, что при необходимости вы сами вычислите границы зоны на основе соответствующих значений для положительной автокорреляции при данном числе наблюдений и объясняющих переменных. Это достаточно легко сделать. Как показано на рис. 7.6, величина $(4 - d_U)$ есть нижний предел, ниже которого признается отсутствие автокорреляции, а $(4 - d_L)$ — верхний предел, выше которого делается вывод о наличии отрицательной автокорреляции.

7.6. Рассмотрите статистику Дарбина—Уотсона для логарифмической функции спроса, которую вы построили в упражнении 5.6. Есть ли там автокорреляция? Если да, то как это повлияло на результаты выполненных вами статистических тестов?

7.7. Если ваш регрессионный пакет позволяет распечатать остатки для уравнения регрессии, сделайте это для оцененных вами по МНК уравнений функции спроса. Подтверждает ли рассмотрение остатков наличие (или отсутствие) автокорреляции, на которую указывает статистика Дарбина—Уотсона? Можете ли вы дать экономическое объяснение поведения остатков?

7.7. Что можно сделать в отношении автокорреляции?

Возможно, вам удастся устранить автокорреляцию путем определения ответственного за нее фактора или факторов и соответствующего расширения уравнения регрессии. Когда такое возможно, это может оказаться наилучшим решением. Пример приводится в упражнении 10.4.

В других случаях процедура, которую следует принять, будет зависеть от характера зависимости между значениями случайного члена. В литературе наибольшее внимание уделяется так называемой авторегрессионной схеме первого порядка (7.21), так как она интуитивно правдоподобна, но для того, чтобы было целесообразным ее использование в более сложных моделях, оснований обычно не хватает. Вместе с тем если наблюдения проводятся ежеквартально или ежемесячно, могут оказаться более подходящими другие модели, но мы не будем их здесь рассматривать.

Если бы уравнение (7.21) было правильной спецификацией для измерения величины случайного члена, то вы могли бы полностью устранить автокорреляцию, если бы знали величину ρ. Это будет продемонстрировано на примере уравнения регрессии, включающего только одну объясняющую переменную, однако при большем их числе действует тот же принцип.

Предположим, что истинная модель задается выражением (7.20), так что наблюдения t и $t-1$ формируются как

$$y_t = \alpha + \beta x_t + u_t; \tag{7.23}$$

$$y_{t-1} = \alpha + \beta x_{t-1} + u_{t-1}. \tag{7.24}$$

Теперь вычтем из обеих частей уравнения (7.23) умноженное на ρ соотношение (7.24) и получим:

$$y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(x_t - \rho x_{t-1}) + u_t - \rho u_{t-1}. \tag{7.25}$$

Обозначим $\tilde{y}_t = y_t - \rho y_{t-1}$, $\tilde{x}_t = x_t - \rho x_{t-1}$ и $\tilde{q}_t = 1-\rho$. Тогда формулу (7.25) можно переписать как

$$\tilde{y}_t = \alpha \tilde{q}_t + \beta \tilde{x}_t + u_t - \rho u_{t-1}. \tag{7.26}$$

Вместе с тем из уравнения (7.21) имеем $u_t - \rho u_{t-1} = \varepsilon_t$. Таким образом, формула (7.26) принимает вид:

$$\tilde{y}_t = \alpha \tilde{q}_t + \beta \tilde{x}_t + \varepsilon_t. \qquad (7.27)$$

Мы предположили, что ρ известно. Тогда можно вычислить величины $\tilde{y}_t$, $\tilde{x}_t$ и $\tilde{q}_t$ (последняя одинакова для всех наблюдений) для наблюдений, включающих от 2 до T исходных данных. Если теперь оценить регрессию между $\tilde{y}_t$, $\tilde{x}_t$ и $\tilde{q}_t$ (заметим, что в уравнение не должна включаться постоянная), то будут получены оценки α и β, не связанные с проблемой автокорреляции, поскольку, согласно предположению, значения ε не зависят друг от друга.

Остается, однако, небольшая проблема. Если в выборке нет данных, предшествующих первому наблюдению, то мы не сможем вычислить $\tilde{y}_1$ и $\tilde{x}_1$ и потеряем первое наблюдение. Число степеней свободы уменьшается на единицу, и это вызовет потерю эффективности, которая может в небольших выборках перевесить повышение эффективности от устранения автокорреляции.

Эту проблему, к счастью, можно довольно легко обойти, пользуясь так называемой *поправкой Прайса—Уинстена* (Prais, Winsten, 1954). Случайный член ε, согласно определению, не зависит от значения u в любом предшествующем наблюдении. В частности, все величины $\varepsilon_2, \ldots, \varepsilon_T$ не зависят от u_1. Следовательно, если при устранении автокорреляции все другие наблюдения преобразуются, то не требуется преобразовывать первое наблюдение. Можно сохранить его, включив в новую схему, полагая, что $\tilde{y}_1 = y_1$, $\tilde{q}_1 = 1$, $\tilde{x}_1 = x_1$.

Мы можем таким способом спасти первое наблюдение, но здесь есть небольшая проблема, которую требуется решить. Если ρ велико, то первое наблюдение будет оказывать непропорционально большое воздействие на оценки, исчисленные по уравнению регрессии. Чтобы нейтрализовать этот эффект, уменьшим вес данного наблюдения умножением его на величину $\sqrt{1-\rho^2}$, полагая $\tilde{y}_1 = \sqrt{1-\rho^2}\, y_1$, $\tilde{x}_1 = \sqrt{1-\rho^2}\, x_1$ и $\tilde{q}_1 = \sqrt{1-\rho^2}$. Причина выбора такого столь необычного веса объясняется в приложении 7.4.

Конечно, на практике величина ρ неизвестна, его оценка получается одновременно с оценками α и β. Имеется несколько стандартных способов такого оценивания, и, вероятно, один или нескольких таких способов могут быть реализованы в имеющемся у вас регрессионном пакете.

Метод Кокрана—Оркатта представляет собой итеративный процесс, включающий следующие этапы.

1. Оценивается регрессия (7.20) с исходными непреобразованными данными.
2. Вычисляются остатки.
3. Оценивается регрессионная зависимость e_t от e_{t-1}, соответствующая формуле (7.21), и коэффициент при e_{t-1} представляет собой оценку ρ.
4. С этой оценкой ρ уравнение (7.20) преобразуется в (7.27), оценивание которого позволяет получить пересмотренные оценки α и β.
5. Повторно вычисляются остатки, и процесс возвращается к этапу 3.

Чередование этапов пересмотра оценок α и β и оценки ρ продолжается до тех пор, пока не будет получена требуемая точность сходимости, т. е. до тех пор, пока оценки на последнем и предпоследнем циклах не совпадут с заданной степенью точности.

Метод Хилдрета—Лу, также широко применяемый в регрессионных пакетах, основан на тех же самых принципах, но использует другой алгоритм вычислений. Здесь регрессия (7.27) оценивается для каждого значения ρ из определенного диапазона с заданным шагом внутри его. (Например, исследователь может задать диапазон от $\rho = -1{,}00$ до $\rho = 1{,}00$ с шагом 0,01.) Значение, которое дает минимальную стандартную ошибку для преобразованного уравнения, принимается в качестве оценки ρ, а коэффициенты регрессии определяются при оценивании уравнения (7.27) с использованием этого значения.

Когда статистика Дарбина—Уотсона указывает на очень тесную положительную автокорреляцию, можно применить упрощенную процедуру, заключающуюся в предположении, что $\rho = 1$. Тогда уравнение (7.25) принимает вид:

$$y_t - y_{t-1} = \beta (x_t - x_{t-1}) + u_t - u_{t-1}. \tag{7.28}$$

Другими словами, оценивается регрессионная зависимость разности значений y в последовательных наблюдениях от разности значений x. Она известна как уравнение регрессии первых разностей и часто записывается в виде:

$$\Delta y_t = \beta \Delta x_t + u_t - u_{t-1}. \tag{7.29}$$

Так как фактическое (неизвестное) значение ρ, вероятно, будет меньшим единицы, эта процедура, по-видимому, компенсирует автокорреляцию с некоторым избытком. Можно показать, что теоретическая корреляция между последовательными значениями $(u_t - u_{t-1})$ равна $-(1 - \rho)/2$. Чем ближе ρ к единице, тем эта корреляция меньше и, следовательно, тем более вероятно улучшение результатов.

Примеры использования метода первых разностей можно найти в литературе до конца 1970-х гг., но в современных исследованиях он обычно не применяется. До относительно недавнего времени привлекала присущая ему простота вычислений, но теперь, когда итеративные процессы благодаря разработке более мощных и быстродействующих компьютеров связаны с меньшими затратами времени и стали менее дорогостоящими, этот метод считается устаревшим.

Таблица 7.4

Метод	*Доход*		*Цена*					
	Эластичность	*с.о.*	*Эластичность*	*с.о.*	$\hat{\rho}$	*с.о.*	*d*	R^2
МНК	0,637	0,026	-0,476	0,121	—	—	0,63	0,987
CO—PW	0,651	0,034	-0,578	0,130	0,67	0,16	1,74	0,999

В табл. 7.4 представлены результаты построения логарифмических регрессий между расходами на питание (y), личным доходом (x) и ценой (p) с использованием данных для США (1959–1983 гг.), приведенных в табл. Б.1 и Б.2; применялись обычный метод наименьших квадратов и метод Кокрана—Оркатта с по-

правкой Прайса—Уинстена (CO—PW). Так как регрессии являются логарифмическими, коэффициенты при y и p следует интерпретировать как показатели эластичности.

Для регрессии по МНК d-статистика указывает на положительную автокорреляцию, статистически значимую при уровне значимости в 1% и выше; эта гипотеза подтверждается значимостью оценки ρ в регрессии Кокрана—Оркатта. Приведем пример ситуации, в которой часто возникает недоразумение. Напомним, что стандартная ошибка представляет собой оценку стандартного отклонения рассматриваемого коэффициента (это анализируется в разделах 3.1 и 3.5). Предположим, что мы оценили регрессию и статистика Дарбина—Уотсона (d) указывает на тесную положительную автокорреляцию. Предположим также, что рассчитанная стандартная ошибка данного коэффициента составляет 0,40. По причине автокорреляции стандартная ошибка представляет собой смещенную оценку истинного стандартного отклонения. Последнее могло быть значительно выше, например 0,90. При оценивании регрессии с использованием CO—PW или другого подобного метода истинное стандартное отклонение с повышением эффективности должно снизиться. Для определенности предположим, что оно снижается с 0,90 до 0,70. Стандартная ошибка в этой регрессии должна представлять собой приблизительно несмещенную оценку стандартного отклонения. Предположим, что она составляет 0,68. Часто бывает так, что студент делает вывод о том, что регрессия CO—PW *менее* эффективна, чем первоначальная, потому что рассчитанная стандартная ошибка возросла с 0,40 до 0,68. Это, конечно, неверно, поскольку оценка 0,40 неверна, и сравнение не имеет смысла. Такая ситуация отражена в табл. 7.5.

Таблица 7.5

	Первоначальная регрессия	*Регрессия CO—PW*
Стандартное отклонение (истинное)	0,90	0,70
Стандартная ошибка (рассчитанная)	0,40	0,68

В рассматриваемом случае представляется, что стандартные ошибки эластичности спроса по доходу и цене в регрессии по обычному МНК ниже, чем в регрессии по CO—PW; может показаться, что МНК более эффективен. Однако по указанным выше причинам это, вероятно, является не более чем иллюзией. Этот вопрос более подробно рассматривается в приложении 7.1, где дается также общая сравнительная оценка МНК и CO—PW.

Упражнения

7.8. Ваш регрессионный пакет, скорее всего, включает один или несколько методов, применяемых в случае автокорреляции. Используйте такой метод для оценивания измененного варианта функции спроса, если статистика Дарбина—Уотсона в упражнении 7.6 указала на наличие автокорреляции. Сравните

пересмотренные оценки коэффициентов и стандартные ошибки с предшествующими и дайте соответствующие комментарии.

7.9. Студент (назовем его *A*) получил 30 наблюдений по двум переменным (*y* и *x*). Ему сообщили, что *y* линейно зависит от *x* и от случайного члена *u*:

$$y_t = \alpha + \beta x_t + u_t,$$

и ему поручено получить оценку β. Истинное значение β, не известное студенту, составляет 5. Студент предпринимает следующее:

1. Использовав обычный метод наименьших квадратов, он получает оценку β, равную 4,64. Стандартная ошибка составляет 1,30; статистика Дарбина—Уотсона (*d*) равна 0,70.

2. Затем студент получает информацию о том, что случайный член подвергается воздействию автокорреляции первого порядка, определяемой выражением:

$$u_t = 0{,}70u_{t-1} + \varepsilon_t,$$

где ε_t удовлетворяет обычным условиям Гаусса—Маркова и нормально распределено. Студент определяет, что $Y_t = y_t - 0{,}70y_{t-1}$ и $X_t = x_t - 0{,}70x_{t-1}$ оценивает регрессию между Y_t и X_t, получая оценку 5,14 для β со стандартной ошибкой 0,75 и статистику Дарбина—Уотсона 1,91.

Девять других студентов (*B, C, ..., J*) выполняют такие же эксперименты с наблюдениями величины *y*, полученными по такой же модели и при тех же значениях *x*, но с другими случайными числами для ε_t. Результаты приводятся в таблице.

Студент	*Эксперимент 1*			*Эксперимент 2*		
	Оценка β	*с.о.*	*d*	*Оценка β*	*с.о.*	*d*
A	4,64	1,30	0,70	5,14	0,75	1,91
B	4,56	1,57	0,60	4,96	0,87	1,20
C	6,54	1,77	0,64	5,57	0,98	2,11
D	5,19	0,90	1,08	5,49	0,63	2,35
E	5,81	1,30	0,76	5,37	0,77	2,49
F	5,24	1,22	0,62	4,92	0,65	2,25
G	4,27	1,10	0,63	4,19	0,61	2,09
H	5,26	0,97	1,20	4,70	0,71	2,23
I	6,80	1,55	0,46	6,17	0,76	1,57
J	3,83	1,72	0,48	5,33	0,82	1,54

1. Объясните, почему студенты, должно быть, не удовлетворены результатами эксперимента *1*.

2. Объясните, почему студенты провели эксперимент *2*, когда стал известен характер автокорреляции.

3. Сравните результаты экспериментов. (Сначала посмотрите на коэффициенты и затем — на стандартные ошибки.)

7.8. Автокорреляция с лаговой зависимой переменной

Предположим, что имеется модель, в которой зависимая переменная, взятая с лагом в один период, используется в качестве одной из объясняющих переменных (мы встретим такие примеры в главе 10). В этом случае влияние автокорреляции, по-видимому, сделает оценки по обычному МНК несостоятельными.

Например, предположим, что модель имеет вид:

$$y_t = \alpha + \beta_1 x_t + \beta_2 y_{t-1} + u_t, \tag{7.30}$$

и допустим, что случайный член u_t подвержен воздействию автокорреляции первого порядка:

$$u_t = \rho u_{t-1} + \varepsilon_t. \tag{7.21}$$

Тогда уравнение (7.30) может быть переписано как

$$y_t = \alpha + \beta_1 x_t + \beta_2 y_{t-1} + \rho u_{t-1} + \varepsilon_t. \tag{7.31}$$

Вместе с тем y_{t-1} зависит от u_{t-1}, так как если соотношение (7.30) верно для t, то оно справедливо и для $(t-1)$:

$$y_{t-1} = \alpha + \beta_1 x_{t-1} + \beta_2 y_{t-2} + u_{t-1}. \tag{7.32}$$

Следовательно, имеется систематическая связь между одной из объясняющих переменных в уравнении (7.31) и первым компонентом случайного члена. Четвертое условие Гаусса—Маркова не удовлетворено, и оценки будут смещенными даже в больших выборках (см. разделы 3.3 и 3.4).

Обнаружение автокорреляции в модели с лаговой зависимой переменной

Как отметили в своей первоначальной статье Дж. Дарбин и Дж. Уотсон, d-статистика Дарбина—Уотсона неприменима в случае, когда уравнение регрессии включает лаговую зависимую переменную. В таком случае можно использовать h-статистику Дарбина (Durbin, 1970), которая также вычисляется на основе остатков. Она определяется как

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n\mathrm{Var}(b)}}, \tag{7.33}$$

где $\hat{\rho}$ — оценка ρ в автокорреляции первого порядка (7.21); $\mathrm{Var}(b)$ — оцененная дисперсия коэффициента при лаговой зависимой переменной; n — число наблюдений в выборке. Приблизительная оценка ρ получается из выражения $(1 - 0{,}5d)$, где d — обычная статистика Дарбина—Уотсона и $\mathrm{Var}(b)$ — квадрат

стандартной ошибки b. Поэтому h можно вычислить на основе обычных результатов оценивания регрессии.

В больших выборках h распределяется как $N(0,1)$, т. е. как нормальная переменная со средним значением 0 и дисперсией, равной единице по нулевой гипотезе отсутствия автокорреляции. Следовательно, гипотеза отсутствия автокорреляции может быть отклонена при уровне значимости в 5%, если абсолютное значение h больше, чем 1,96, и при уровне в 1%, если оно больше, чем 2,58, при применении двустороннего критерия и большой выборке.

Основная проблема, связанная с использованием этого теста, заключается в невозможности вычисления h в том случае, если $n\,\mathrm{Var}\,(b)$ больше единицы. Альтернативная процедура, состоящая в применении теста с множителем Лагранжа, описана в приложении 7.2, где использование лаговой зависимой переменной в качестве объясняющей переменной не влияет на результат. Как и h-тест, эта процедура применима только для больших выборок.

Если в число объясняющих переменных включена лаговая зависимая переменная, то использование метода Кокрана—Оркатта может привести к локальному, а не к общему минимуму, что указали Р. Бетанкур и Х. Келейан (Betancourt, Kelejian, 1981) и Л. Оксли и К. Робертс (Oxley, Roberts, 1982). По этой причине в данном случае при построении модели рекомендуется использовать решетчатый поиск Хилдрета—Лу или подобный ему метод.

Упражнения

7.10. В эксперименте по методу Монте-Карло модель

$$y_t = \alpha + \beta y_{t-1} + u_t$$

оценивалась: 1) с использованием МНК; 2) с использованием метода Кокрана—Оркатта (CO); при этом истинные значения α и β равнялись соответственно 10 и 0,8. Случайный член u подвергался воздействию автокорреляции первого порядка:

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

где ρ было равно 0,7, а значения ε_t определялись умножением на число 5 независимых значений нормально распределенной случайной переменной с нулевым математическим ожиданием и единичной дисперсией.

Этот эксперимент был проведен 10 раз с выборками из 30 наблюдений; результаты представлены в таблице. Величина a представляет собой оценку α; b — оценка β; с. о. (b) — стандартная ошибка b; d есть d-статистика Дарбина—Уотсона; h — это h-статистика Дарбина.

1. Объясните, каким образом регрессии, построенные с помощью обычного МНК, указывают на наличие автокорреляции.

2. Объясните последствия автокорреляции для МНК-оценок в этой модели.

3. Объясните, подтверждают ли результаты оценивания регрессии по МНК ваш ответ на вопрос (2) или противоречат ему и дают ли какое-то улучшение оценки, полученные с помощью метода Кокрана—Оркатта.

Эксперимент	МНК					СО				
	a	*b*	*c.o.(b)*	*d*	*h*	*a*	*b*	*c.o.(b)*	*d*	*h*
1	4,7	0,95	0,07	1,14	2,50	17,3	0,79	0,10	2,05	–0,16
2	5,0	0,92	0,07	0,95	3,09	19,1	0,70	0,14	1,94	0,26
3	0,9	0,94	0,05	0,40	4,47	4,0	0,84	0,11	2,06	–0,20
4	6,1	0,84	0,11	0,89	3,68	14,2	0,65	0,15	1,50	2,32
5	5,1	0,83	0,07	1,11	2,60	8,5	0,75	0,13	2,09	–0,34
6	–1,7	1,01	0,05	1,04	2,70	5,0	0,91	0,09	2,01	0,03
7	5,3	0,90	0,08	0,78	3,65	17,4	0,71	0,13	1,93	0,27
8	–1,3	0,96	0,04	0,83	3,22	2,4	0,80	0,11	1,62	1,26
9	–0,6	0,98	0,04	0,55	4,01	3,8	0,83	0,10	1,83	0,54
10	–0,9	1,00	0,07	1,03	2,81	11,8	0,80	0,12	1,70	1,08

7.9. Автокорреляция как следствие неправильной спецификации модели

Автокорреляция в модели регрессии формально вызывается зависимостью между значениями случайного члена в выборке. Но этот вопрос можно рассмотреть и более глубоко. Причиной наличия случайного члена может быть какая-либо неточность в спецификации модели, например пропуск какой-либо важной объясняющей переменной или использование неподходящей математической функции (см. раздел 2.1). Следовательно, автокорреляция нередко может объясняться неправильной спецификацией модели; в этом случае, по-видимому, лучше вместо использования механической процедуры «исправления» непосредственно попытаться устранить ошибки в спецификации. Конечно, обычно лучше устранять причину, чем симптом.

Автокорреляция, вызванная неправильной спецификацией переменных

Явная автокорреляция может быть вызвана пропуском важной объясняющей переменной, и положение можно исправить, если эта переменная будет определена и включена. (Пример дан в упражнении 10.4.) Другая ее причина может заключаться в том, что не принята во внимание структура модели, включающая запаздывание. Метод Кокрана—Оркатта является эффективным способом отражения структуры запаздывания в модели, которая ранее была статической. Возможно, будет признана предпочтительной более общая спецификация. Мы видели, что при наличии автокорреляции в модели (7.21) ее можно устранить

в случае парной регрессии путем преобразования модели к виду (7.27). Это можно переписать таким образом:

$$y_t = \alpha(1-\rho) + \rho y_{t-1} + \beta x_t - \beta\rho x_{t-1} + \varepsilon_t. \tag{7.34}$$

Фактически мы оцениваем регрессионную зависимость y_t от y_{t-1}, x_t и x_{t-1}, налагая ограничение, заключающееся в требовании равенства коэффициента при x_{t-1} произведению коэффициентов при других двух переменных в правой части уравнения. Так как уравнение является нелинейным по параметрам, мы не можем для его оценивания использовать МНК. Вместо этого мы применяем метод Кокрана—Оркатта или какой-либо другой подобный ему метод оценивания, в сущности, нелинейной регрессии.

В целом мы не имеем права заранее утверждать, что указанное ограничение обосновано. Кроме того, мы должны проверять все ограничения, где это возможно, и в данном случае сделать это несложно. Мы вводим другую, не включающую ограничения модель:

$$y_t = \lambda_0 + \lambda_1 y_{t-1} + \lambda_2 x_t + \lambda_3 x_{t-1} + \varepsilon_t \tag{7.35}$$

и проверяем, равно ли λ_3 величине $-\lambda_1\lambda_2$. Если это ограничение не отклонено, мы принимаем предположение, что модель адекватно представлена выражениями (7.20) и (7.21), и продолжаем оценивать ее параметры, используя метод Кокрана—Оркатта или другой подобный ему метод. Если ограничение отклонено, то непосредственно оценивается регрессия (7.35) с использованием обычного МНК.

Следует отметить, что если лучшей спецификацией модели окажется (7.35), то из этого следует, что мы отказываемся от гипотезы, что случайный член формируется авторегрессионным процессом (7.21) и тест Дарбина—Уотсона перестает быть применимым при оценивании регрессии (7.20). Тем не менее он может быть полезен в диагностических целях, и часто первым указанием на наличие какой-либо проблемы в исходной регрессии служит d-статистика, недостаточно близкая к двум.

Теоретические положения, обосновывающие рассматриваемую процедуру проверки, здесь не представлены (они кратко излагаются в работе Д. Хендри и Г. Майзона [Hendry, Mizon, 1978]). Для данного случая подходит тестовая статистика

$$T \log (RSS_R / RSS_U), \tag{7.36}$$

где RSS_R и RSS_U — необъясненные суммы квадратов отклонений соответственно в вариантах с ограничением и без ограничений; логарифмы вычисляются по основанию e и T — количество наблюдений в выборке. В больших выборках статистика, лежащая в основе критерия, имеет распределение χ^2 с числом степеней свободы, равным количеству налагаемых ограничений.

Может возникнуть вопрос о количестве налагаемых ограничений. До сих пор анализировалась исходная модель с одной объясняющей переменной. В этом случае ограничение было только одно: λ_3 равно $-\lambda_1\lambda_2$. При наличии k объясняющих переменных количество ограничений также было бы равно k. Если исходная модель имеет вид:

$$y_t = \alpha + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + u_t, \tag{7.37}$$

где u_t формируется на основе соотношения (7.20), то преобразованная модель будет представлена выражением:

$$y_t = \alpha(1-\rho) + \rho y_{t-1} + \beta_1(x_{1t} - \rho x_{1t-1}) + \ldots + \beta_k(x_{kt} - \rho x_{kt-1}) + \varepsilon_t, \qquad (7.38)$$

и, таким образом, каждой объясняющей переменной соответствует ограничение, состоящее в том, что коэффициент при лаговом значении объясняющей переменной должен равняться произведению со знаком «минус» коэффициента при текущем значении этой переменной и коэффициента при y_{t-1}.

Пример

Преобразованная по методу Кокрана—Оркатта логарифмическая регрессия между расходами на жилье, располагаемым личным доходом и относительной ценой имеет следующий вид (в скобках приведены стандартные ошибки):

$$\widehat{\log y_t} = \underset{(1,05)}{4,47} + \underset{(0,11)}{0,40}\log x_t - \underset{(0,14)}{0,26}\log p_t; \qquad (7.39)$$

$$R^2 = 0,9994;\ RSS = 0,0014; \qquad \hat{\rho} = 0,98; \qquad d = 1,93.$$

Результаты оценивания регрессии (7.34) по МНК без учета ограничений могут быть представлены в виде:

$$\widehat{\log y_t} = \underset{(\text{с.о.})}{0,73} + \underset{(0,48)}{0,87}\log y_{t-1} + \underset{(0,06)}{0,22}\log x_t -$$

$$- \underset{(0,11)}{0,11}\log x_{t-1} - \underset{(0,14)}{0,19}\log p_t + \underset{(0,17)}{0,01}\log p_{t-1}; \qquad (7.40)$$

$$R^2 = 0,9997;\ RSS = 0,0008; \qquad d = 2,27; \qquad h = -0,67.$$

Рассмотрим это уравнение, прежде чем применить *тест на общий фактор*. Мы получаем оценку ρ из коэффициента при $\log y_{t-1}$. Верно ли, что коэффициент при $\log x_{t-1}$ приблизительно равен умноженному на −0,87 коэффициенту при $\log x_t$ и что коэффициент при $\log p_{t-1}$ приблизительно равен умноженному на −0,87 коэффициенту при $\log p_t$? Очевидно, нет, по меньшей мере на первый взгляд. Статистика, лежащая в основе критерия, рассчитывается как $24\log(0,0014/0,0008)$, что равняется 13,4. Критическое значение χ^2 с двумя степенями свободы при уровне значимости в 1% составляет 9,2 (см. табл. А.4). Следовательно, ограничение подлежит обоснованному отклонению (но при этом не нужно забывать, что данный тест следует использовать только для больших выборок). Еще одно свидетельство в пользу уравнения (7.40) обеспечивается тем, что *h*-тест показывает отсутствие статистически значимой автокорреляции.

Если мы выполним *t*-тест применительно к коэффициентам уравнения без ограничений, то увидим, что только одна лаговая переменная ($\log y_{t-1}$) имеет

значимый коэффициент. Это означает, что мы можем опустить два других лаговых члена. Если мы сделаем это и повторно оценим регрессию (снова используя обычный МНК), то получим:

$$\begin{array}{l}\widehat{\log y_t} = 0{,}49 + 0{,}85 \log y_{t-1} + 0{,}15 \log x_t - 0{,}16 \log p_t; \\ \text{(с.о.)}\ (0{,}38)\ (0{,}04)\qquad (0{,}05)\qquad (0{,}07)\end{array} \tag{7.41}$$

$$R^2 = 0{,}9996;\ RSS = 0{,}0008;\quad d = 1{,}94;\quad h = 0{,}16.$$

Здесь нет статистически значимой автокорреляции. *Вывод*: Ярко выраженная автокорреляция в первоначальной регрессии между расходами на жилье, доходом и ценой фактически объясняется пропуском лаговой зависимой переменной.

Резюме

В связи с проведенным анализом следует отметить, что если при оценивании регрессии мы получаем d-статистику, которая явно указывает на автокорреляцию, то в первую очередь следует выполнить общий факторный тест, используя как преобразование по методу Кокрана—Оркатта, так и вариант без ограничений. Если ограничение не отклоняется, следует придерживаться результата, полученного по методу Кокрана—Оркатта. Если оно отклоняется, следует сосредоточиться на варианте без ограничений и попробовать внести новые усовершенствования. Например, не всегда необходимо сохранять все лаговые переменные.

Автокорреляция, вызываемая ошибочной функциональной спецификацией

Автокорреляция остатков в регрессии может иметь место при ошибочной функциональной спецификации уравнения регрессии. Например, в разделе 4.1 мы видели, что если истинная модель имеет вид:

$$y = \alpha + \frac{\beta}{x} + u\,, \tag{7.42}$$

и мы оцениваем линейную регрессию, то получается результат, представленный на рис. 4.1 и в табл. 4.2: отрицательный остаток в первом наблюдении, положительные остатки в следующих шести и отрицательные остатки в последних трех. Другими словами, обнаруживается очень сильная положительная автокорреляция. Однако когда регрессия имеет форму

$$\hat{y} = \alpha + bx', \tag{7.43}$$

где x' определяется как $1/x$, то не только достигается гораздо лучшее качество оценок, но и исчезает автокорреляция.

Самый прямой способ обнаружения автокорреляции, вызванной ошибочной

функциональной спецификацией, заключается в непосредственном рассмотрении остатков. Это может дать определенное представление о правильной спецификации. d-статистика Дарбина—Уотсона также может сигнализировать об ошибочной функциональной спецификации, хотя, конечно, выполненная на ее основе проверка была бы необоснованной, так как случайный член не соответствует процессу, описанному формулой (7.21), и использование метода типа Кокрана—Оркатта было бы нецелесообразным. В описанном выше примере d-статистика Дарбина—Уотсона составляла 0,86, что указывает на наличие ошибки.

Упражнения

7.11. В упражнении 6.9 d-статистики для шести уравнений были следующими:

Город A (1) 1,18; (2) 1,42; (3) 1,98;

Город B (1) 2,28; (2) 0,76; (3) 2,13.

Рассмотрите поведение d-статистики в свете вашего ответа на задание, сформулированное в упражнении 6.9.

7.12. Функция спроса на продукты питания с преобразованием по методу Кокрана—Оркатта имеет следующий вид (в скобках указаны стандартные ошибки):

$$\log \hat{y}_t = \underset{}{3{,}11} + \underset{(0,55)}{0{,}69} \log x_t - \underset{(0,14)}{0{,}61} \log p_t;$$
$$(0,04)$$

$$R^2 = 0{,}9930; \; RSS = 0{,}0033; \qquad d = 1{,}93.$$

Вариант без ограничений выглядит следующим образом:

$$\log \hat{y}_t = \underset{(\text{с.о.})}{0{,}94} + \underset{(0,15)}{0{,}54} \log y_{t-1} + \underset{(0,17)}{0{,}56} \log x_t -$$
$$(0,59)$$

$$- \underset{(0,20)}{0{,}28} \log x_{t-1} - \underset{(0,13)}{0{,}68} \log p_t + \underset{(0,13)}{0{,}55} \log p_{t-1};$$

$$R^2 = 0{,}9949; \; RSS = 0{,}0024; \qquad d = 2{,}20.$$

Проанализируйте полученные значения коэффициентов в двух уравнениях и выполните общий факторный тест.

7.13. Предположим, что модель подвержена воздействию автокорреляции первого порядка и, следовательно, может быть представлена выражением (7.34). Почему при построении регрессионного уравнения не следует использовать МНК?

Приложение 7.1

Исследование, проведенное Р.Э. Парком и Б. Митчеллом на основе метода Монте-Карло

Почему в последнем предложении раздела 7.7 мы употребляем слово «вероятно», а не «определенно»? Причина этого заключается в том, что оценки, использующие преобразованное уравнение (7.27), будут иметь желательные для МНК свойства, только если в преобразовании используется истинное значение ρ (и если сохранено первое наблюдение). Если таким способом точно устранена автокорреляция, то полученная оценка является оценкой по обобщенному методу наименьших квадратов (ОМНК). Однако величину ρ мы должны были оценить. Это означает, что метод CO—PW будет работать определенно лучше, чем обычный МНК, только для больших выборок.

За последние годы усилия многих исследователей были направлены на поиск других способов оценки ρ и на исследование свойств преобразования по методу Кокрана—Оркатта и получаемых с его помощью вариантов моделей для малых выборок. Существенный вклад внесла статья Р.Э. Парка и Б. Митчелла (Park, Mitchell, 1980), обзор которой и приводится здесь, так как результаты, полученные авторами, имеют большое значение, а также по той причине, что она хорошо иллюстрирует возможность применения метода Монте-Карло для выявления свойств оценок в малых выборках.

Основой анализа, выполненного Р.Э. Парком и Б. Митчеллом, является модель парной регрессии (7.20) с коррелированными остатками, подчиняющимися авторегрессионному процессу первого порядка (7.21). Было принято, что как α, так и β в уравнении (7.20) равны единице. Для переменной x использовались три различных ряда данных (и эксперименты были повторены отдельно для каждого из них): простой временной тренд; фактические ежегодные данные о валовом национальном продукте США начиная с 1950 г.; фактические ежегодные данные о коэффициенте использования производственных мощностей (CAP) в США также начиная с 1950 г. Случайный член ε генерировался с использованием независимых нормально распределенных случайных чисел с нулевым математическим ожиданием и единичной дисперсией. Для большинства экспериментов размер выборки равнялся 20, и каждый эксперимент был повторен для 1000 выборок.

Причина, по которой эксперимент проводился с тремя различными рядами данных для x, заключалась в том, что, как показали более ранние исследования, качество получаемых оценок зависит (в числе других факторов) от того, содержат ли данные временной тренд. Первый ряд допускает сравнение свойств оценок в экстремальном случае при наличии чистого тренда. Второй ряд, где рассматривается валовой национальный продукт (GNP), допускает сравнение в неэкстремальном случае, где объясняющая переменная имеет определенную тенденцию, часто встречающуюся в экономических данных. Что касается третьего ряда, то здесь допускается сравнение, когда объясняющая переменная вообще не имеет какого-либо тренда.

В каждом эксперименте Р.Э. Парк и Б. Митчелл вычислили среднеквадратичную ошибку оценок в 1000 выборках. Потом они определили относительную эф-

фективность оценки как обратную величину отношения ее среднеквадратичной ошибки к соответствующей ошибке по обычному МНК для тех же рядов данных и значениях ρ. Результаты приводятся в табл. 7.6. Аббревиатурами CO и CO—PW обозначается метод Кокрана—Оркатта соответственно без поправки Прайса—Уинстена и с поправкой. В табл. 7.6 приводятся результаты для оценок коэффициента наклона в регрессии (7.20). Таблица включает результаты оценивания по обобщенному МНК, которые автоматически имеют наибольшую относительную эффективность, которая используется как критерий. Относительная эффективность устанавливает предел выигрыша, который может быть достигнут за счет замены обычного МНК другим методом оценивания.

Таблица 7.6

Эффективность методов ОМНК , CO и CO—PW в сравнении с обычным МНК

	ρ			
	0,4	*0,8*	*0,9*	*0,98*
Временной тренд				
ОМНК	1,02	1,09	1,10	1,08
CO	0,85	0,69	0,56	0,64
CO—PW	1,01	1,07	1,08	1,05
GNP				
ОМНК	1,02	1,14	1,20	1,26
CO	0,85	0,80	0,83	0,88
CO—PW	1,01	1,09	1,13	1,12
CAP				
ОМНК	1,14	1,86	2,21	2,52
CO	1,03	1,65	2,03	2,27
CO—PW	1,05	1,61	2,00	2,15

На основе табл. 7.6 можно сделать следующие выводы:

1. Выигрыш в эффективности, обеспечиваемый заменой обычного МНК на метод CO или CO—PW, может быть значительным при наличии неярко выраженного тренда и высоком значении ρ.

2. В условиях сильного тренда обычный МНК может быть не очень эффективным даже для высоких значений ρ. Наихудший результат МНК дает для показателя ВНП при значении ρ, близком к единице. В этом случае его эффективность на 26% ниже по сравнению с обобщенным МНК. Разница в эффективно-

сти меньше для случая с другими рассматриваемыми оценками: здесь ее наибольшее значение равно 13%. Другие исследования показали, что обычный МНК действительно более эффективен, чем любой другой метод, если данные подвержены сильному тренду и значение ρ мало. Следовательно, основной причиной, по которой в этих условиях не используется обычный МНК, является не недостаточно высокая эффективность, а смещение оценок стандартных ошибок — проблема, которая вскоре будет рассмотрена.

3. Использование метода Кокрана—Оркатта в чистом виде значительно менее эффективно, чем CO—PW, когда данные подвержены сильному тренду. В этом случае данный метод неэффективен даже в сравнении с обычным МНК. Когда же данные не подвержены тренду, CO работает так же, как и CO—PW.

Вывод: Следует всегда использовать CO—PW или какой-либо другой метод, позволяющий сохранить первое наблюдение.

Р.Э. Парк и Б. Митчелл рассмотрели также эффективность применения различных оценок при проверке гипотез, определяя, сколько раз в каждой совокупности из 1000 проб проверка по *t*-критерию при уровне значимости в 5% приведет к отклонению верной гипотезы о том, что коэффициент наклона β в регрессии (7.20) равен единице, и к ошибке I рода. Следует помнить, что даже при идеальных условиях тест при уровне значимости в 5% будет приводить к отклонению верной нулевой гипотезы в 5% случаев и, таким образом, даже использование оценки ОМНК вызовет появление приблизительно 50 ошибок I рода.

Таблица 7.7

Количество ошибок I рода при проведении 1000 проб при уровне значимости в 5%

	ρ			
	0,4	*0,8*	*0,9*	*0,98*
Временной тренд				
МНК	197	490	571	709
CO—PW	131	285	336	474
GNP				
МНК	185	449	596	666
CO—PW	136	246	322	397
CAP				
МНК	143	294	323	322
CO—PW	113	101	86	86

Р. Э. Парк и Д. Митчелл не приводят результатов оценивания для CO, потому что CO—PW оказывается более предпочтительным методом, имеющим более высокую эффективность. В табл. 7.7 прослеживается очень сильная тенден-

ция к недооценке стандартных ошибок для обычного МНК, что приводит к появлению ошибок I рода. Эта таблица также показывает, что тот же недостаток может быть свойственен методу CO—PW, хотя здесь в этом отношении наблюдается улучшение. Особенно ярко эта проблема проявляется, когда данные подвержены сильному тренду и значение ρ высоко. *Вывод*: В этих условиях при проверке гипотез должны использоваться более высокие уровни значимости, чем обычно.

Р. Э. Парк и Д. Митчелл сравнили также качество различных оценок при более крупной выборке, увеличив число наблюдений до 50 в каждой серии (испытании). Они сообщают только о результатах использования рядов данных по ВНП в качестве объясняющей переменной (причем не ежегодных, а ежеквартальных данных). Основные выводы здесь следующие: CO—PW остается более предпочтительным по сравнению с CO; эффективность CO—PW по отношению к обычному МНК возрастает, особенно для высоких значений ρ; использование обычного МНК приводит к появлению ошибок I рода даже чаще, чем для меньших по объему выборок; при применении метода CO—PW ошибки I рода возникают реже, но этот метод может по-прежнему в значительной степени вводить в заблуждение при проверке гипотез.

Приложение 7.2

Автокорреляция более высокого порядка: обнаружение и оценивание

До сих пор мы рассматривали очень простой вид автокорреляции — авторегрессионный процесс первого порядка, описанный выражением (7.21). Причина этого в том, что он обычно рассматривается как приемлемый способ аппроксимации общего вида автокорреляции, и в большинстве регрессионных пакетов *d*-статистика вычисляется автоматически.

В некоторых пакетах в настоящее время также используется тест Томаса—Уоллиса (Thomas, Wallis, 1971) на автокорреляцию четвертого порядка, когда связь случайных членов описывается процессом AR(4):

$$u_t = \rho u_{t-4} + \varepsilon_t. \tag{7.44}$$

Автокорреляция такого типа может иметь место при использовании в качестве данных ежеквартальных наблюдений, и сезонные колебания переходят из года в год. Статистика, лежащая в основе данного теста, по своей структуре сходна с *d*-статистикой Дарбина—Уотсона, но имеет другое распределение.

Вместе с тем встречаются случаи, когда целесообразно предположить общую зависимость более высокого порядка:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_s u_{t-s} + \varepsilon_t, \tag{7.45}$$

или когда случайные члены связываются не авторегрессионным процессом, а процессом скользящих средних, который может быть описан уравнением:

$$u_t = \alpha_1\varepsilon_t + \alpha_2\varepsilon_{t-1} + \ldots + \alpha_s\varepsilon_{t-s}, \tag{7.46}$$

или даже какой-либо комбинацией этих двух выражений.

Для обнаружения автокорреляции более высокого порядка разработано несколько тестов. Здесь будет представлен только один пример — тест с множителем Лагранжа, выбранный потому, что выполнять его относительно легко, даже когда это специально не предусматривается в регрессионном пакете. По этой причине возрастает популярность данного теста. Теория, обосновывающая его применение, не входит в круг вопросов, рассматриваемых в данной книге. Относительно простой обзор этой теории, выполненный ее авторами, содержится в работе Т. Бройша и Л. Годфри (Breusch, Godfrey, 1982).

Вначале следует решить, насколько глубоко вас интересует вопрос о порядке обнаруживаемой автокорреляции. Для процессов, представляемых выражениями (7.45) и (7.46), этот порядок обозначается с помощью нижнего индекса s. Затем вы обычным образом оцениваете регрессию с помощью МНК и «запоминаете» остатки. В конечном итоге вы оцениваете следующую регрессию:

$$\hat{e}_t = c_0 + c_1e_{t-1} + c_2e_{t-2} + \ldots + c_se_{t-s} + d_1x_{1t} + d_2x_{2t} + \ldots + d_kx_{kt}, \tag{7.47}$$

где e_{t-p} — остаток в наблюдении $(t-p)$; переменные x — объясняющие переменные первоначальной регрессии. Регрессия оценивается по данным для периодов от $(s+1)$ до T, так как величины e_{t-s} не определены для первых s периодов. Затем для этой регрессии вычисляется $(T \times R^2)$, и при нулевой гипотезе об отсутствии автокорреляции эта статистика имеет распределение χ^2 с s степенями свободы.

Необходимо сделать два предупреждения. Во-первых, тест с множителем Лагранжа, как и альтернативные ему варианты, рассчитан на работу с большими выборками. Поэтому нужно проявлять осторожность при интерпретации результатов, полученных на малых выборках. Во-вторых, в отличие от теста Дарбина—Уотсона он обнаруживает не только авторегрессионную корреляцию остатков типа (7.21), но и автокорреляцию, описываемую скользящими средними (7.46). Здесь также необходима осторожность при толковании результатов.

Пример

Тест с множителем Лагранжа был применен по отношению к логарифмической функции спроса на продукты питания; полученные результаты представлены в уравнении (5.26), где значения s меняются от 1 до 4. Результаты обобщены в табл. 7.8, где показано, что нулевая гипотеза отклоняется по меньшей мере на уровне 5% для всех значений s, но тестовая статистика увеличивается очень медленно для значений s, больших единицы, в предположении о том, что автокорреляция хорошо аппроксимируется процессом первого порядка. (Возможно, у вас вызывает удивление то, что коэффициент R^2 в действительности уменьшается, когда в уравнение добавляется e_{t-2}; объясняется это изменением периода выборки. При $s = 1$ период выборки охватывал 1960—1983 гг.; при $s = 2$ период выборки включал 1961—1983 гг. Следовательно, здесь не действует обычное правило, согласно которому коэффициент R^2 не может уменьшаться при добавлении к уравнению новых переменных.)

Оценивание регрессии с автокорреляцией более высокого уровня

Предположим, что мы обнаружили автокорреляцию более высокого порядка. Как в этом случае оценить регрессию? Если у нас есть причины предполагать, что автокорреляция представляет собой авторегрессионный процесс типа (7.45), то можно использовать обобщенный вариант оценки по методу Кокрана—Оркатта. Предположим, что у нас имеется модель парной регрессии (7.20). Если определить

$$\tilde{y}_t = y_t - \rho_1 y_{t-1} - \ldots - \rho_s y_{t-s}; \tag{7.48}$$

$$\tilde{x}_t = x_t - \rho_1 x_{t-1} - \ldots - \rho_s x_{t-s}; \tag{7.49}$$

$$\tilde{\alpha} = \alpha(1 - \rho_1 - \ldots - \rho_s), \tag{7.50}$$

то можно легко показать, что

$$\tilde{y}_t = \tilde{\alpha} + \beta\tilde{x}_t + \varepsilon_t \quad (t = s+1, \ldots, T). \tag{7.51}$$

Таблица 7.8

Проверка на автокорреляцию Бройша—Годфри с множителем Лагранжа: функция спроса на продукты питания

Порядок	$T \times R^2$	*Критическое значение*	
		5%	*1%*
1	11,9	3,8	6,6
2	11,4	6,0	9,2
3	11,7	7,8	11,3
4	11,9	9,5	13,3

Если бы мы знали величину ρ, то по данным наблюдений y и x можно было бы рассчитать $\tilde{y}$ и $\tilde{x}$ и оценить данное уравнение, устранив проблему автокорреляции. Вообще говоря, мы не знаем величины ρ, и ее приходится оценивать наряду с α и β. Обобщенный вариант оценки по методу Кокрана—Оркатта следует той же схеме, что и первоначальный, в том смысле, что сначала оценивается уравнение регрессии с помощью обычного МНК, а затем оценивается регрессия (7.45) с заменой случайных членов на рассчитанные остатки. Далее вычисляются $\tilde{y}$ и $\tilde{x}$ и оценивается регрессия (7.51), дающая пересмотренные оценки α и β. Получив новую совокупность остатков, вновь оцениваем уравнение (7.45) и т. д. до тех пор, пока не будет достигнута сходимость процесса.

Здесь нужно сделать такое же предостережение, как и в случае авторегрессии первого порядка. Так как требуется оценить ρ, обобщенная оценка по методу Кокрана—Оркатта будет иметь требуемые свойства только на больших

выборках. Имеется также проблема относительно применения поправки Прайса—Уинстена, которая становится все менее надежной по мере возрастания порядка авторегрессии. В принципе эта проблема преодолима [см. работу Э. Харви (Harvey, 1981, p. 206)], но требуемая корректировка нечасто встречается в компьютерных регрессионных пакетах.

Уравнение (7.52) представляет результаты оценивания регрессии по методу Кокрана—Оркатта четвертого порядка для функции спроса на продукты питания (без поправки Прайса—Уинстена):

$$\begin{matrix}\widehat{\log} y_t = 3{,}12 + 0{,}73 \log x_t - 0{,}68 \log p_t; & R^2 = 0{,}991; \\ \text{(с.о.)} \quad (0{,}63) \quad (0{,}06) \qquad (0{,}16)\end{matrix} \tag{7.52}$$

$$\begin{matrix}\hat{u}_t = 0{,}61u_{t-1} - 0{,}03u_{t-2} - 0{,}28u_{t-3} + 0{,}23u_{t-4}. \\ \quad (0{,}26) \quad (0{,}29) \quad (0{,}30) \quad (0{,}27)\end{matrix} \tag{7.53}$$

Уравнение (7.53) подтверждает, что если автокорреляция остатков является авторегрессионным процессом, то она должным образом аппроксимируется моделью первого порядка.

Упражнение

Выполните аналогичный тест с множителем Лагранжа для логарифмической регрессии функции спроса на выбранный вами товар[1].

Приложение 7.3

Иллюстрация того, что d-статистика Дарбина—Уотсона приближается к (2 – 2ρ)

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = \frac{\sum_{t=2}^{T}(e_t^2 - 2e_{t-1}e_t + e_{t-1}^2)}{\sum_{t=1}^{T} e_t^2} =$$

$$= \frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2} + \frac{\sum_{t=2}^{T} e_{t-1}^2}{\sum_{t=1}^{T} e_t^2} - 2\frac{\sum_{t=2}^{T} e_{t-1}e_t}{\sum_{t=1}^{T} e_t^2} \cong 2 - 2\frac{\sum_{t=2}^{T} e_{t-1}e_t}{\sum_{t=1}^{T} e_t^2}, \tag{7.54}$$

[1] При выполнении теста вам может потребоваться техническая помощь со стороны преподавателя. Если в вашем регрессионном пакете реализован метод Кокрана—Оркатта более высокого порядка, то следует выполнить оценивание уравнения с его использованием.

так как $\frac{\sum_{t=2}^{T} e_t^2}{\sum_{t=1}^{T} e_t^2}$ и $\frac{\sum_{t=2}^{T} e_{t-1}^2}{\sum_{t=1}^{T} e_t^2}$ будут близкими к единице, если выборка достаточно большая. Поскольку $\frac{\sum_{t=2}^{T} e_{t-1}e_t}{\sum_{t=1}^{T} e_t^2}$ является оценкой ρ, то d является оценкой для $(2 - 2\rho)$.

Приложение 7.4

Взвешивание первого наблюдения при использовании метода Кокрана—Оркатта

Мы видели в разделе 7.7, что если y_t связано с x_t соотношением (7.20) и u_t выражается через u_{t-1} и ε_t в форме авторегрессионного процесса первого порядка (7.21), то можно устранить автокорреляцию путем оценивания регрессии $\tilde{y}_t$ от $\tilde{x}_t$, где

$\tilde{y}_t = y_1$ для наблюдения *1*;

$\tilde{y}_t = y_t - \rho y_{t-1}$ для наблюдений *2*, ..., *T*;

$\tilde{q}_t = 1$ для наблюдения *1*;

$\tilde{q}_t = 1 - \rho$ для наблюдений *2*, ..., *T*;

$\tilde{x}_t = x_1$ для наблюдения *1*;

$\tilde{x}_t = x_t - \rho x_{t-1}$ для наблюдений *2*, ..., *T*.

Тогда случайные члены в T наблюдениях (u_1 в первом наблюдении и ε_2, ..., ε_T в остальных) будут распределены независимо. Остается одна проблема: теоретическая дисперсия u в первом наблюдении (σ_u^2) отличается от теоретической дисперсии в остальных наблюдениях (σ_ε^2). Таким образом, мы решили проблему автокорреляции за счет введения особого случая гетероскедастичности. Используя формулу (7.21), мы можем выразить σ_u^2 через σ_ε^2:

$$\sigma_u^2 = \text{pop. var}\,(u_t) = \text{pop. var}\,(\rho u_{t-1} + \varepsilon_t) =$$

$$= \rho^2 \text{pop. var}\,(u_{t-1}) + \text{pop. var}\,(\varepsilon_t) + 2\rho \text{ pop. cov}\,(u_{t-1}, \varepsilon_t) = \rho^2\sigma_u^2 + \sigma_\varepsilon^2, \qquad (7.55)$$

так как u_{t-1} и ε_t независимы. (Мы предполагаем, что $|\rho| < 1$, и это неравенство, как можно показать, является условием, чтобы дисперсия u была независимой от t, что позволяет нам записать pop. var (u_t) = pop. var $(u_{t-1}) = \sigma_u^2$.) Таким образом,

$$\sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}. \qquad (7.56)$$

Если ρ близко к единице, то σ_u^2 будет значительно больше, чем σ_ε^2, и первое наблюдение станет оказывать, вообще говоря, непропорционально сильное воздействие на результаты оценивания регрессии. Вместе с тем использование соотношения (7.56) позволяет устранить гетероскедастичность. Если мы учтем первое наблюдение с «весом» $\sqrt{1-\rho^2}$, то случайный член в этом наблюдении примет вид $\sqrt{1-\rho^2}u_1$. Дисперсия этой величины равна $(1-\rho^2)\sigma_u^2$, что, конечно, равно дисперсии σ_ε^2 случайного члена в остальных наблюдениях.

8

СТОХАСТИЧЕСКИЕ ОБЪЯСНЯЮЩИЕ ПЕРЕМЕННЫЕ И ОШИБКИ ИЗМЕРЕНИЯ

В базовой регрессионной модели, построенной на основе метода наименьших квадратов, предполагается, что объясняющие переменные являются нестохастическими. Часто это предположение оказывается нереалистичным, и поэтому важно знать, к каким последствиям приведут более слабые модельные ограничения. Мы увидим, что в некоторых ситуациях сможем продолжать использовать МНК, но в других, например, когда объясняющая переменная (или переменные) подвержена воздействию ошибок измерения, МНК приводит к смещенным и несостоятельным оценкам. Глава заканчивается введением другого метода оценивания, основанного на инструментальных переменных, который позволяет получать оценки с более приемлемыми свойствами.

8.1. Стохастические объясняющие переменные

До сих пор мы считали, что объясняющие переменные в регрессионной модели являются нестохастическими. Это означает, что если бы нам пришлось повторить регрессионный анализ с новой выборкой, то значения объясняющих переменных остались бы неизменными. При этом значения зависимой переменной изменились бы, потому что новая выборка содержала бы новую совокупность значений случайного члена.

Такое допущение может показаться странным. На практике в эконометрике мы оцениваем параметры модели регрессии только один раз. Редко бывает возможность повторить расчет с теми же или с другими значениями объясняющих переменных. Единственным общим исключением являются эксперименты лабораторного типа, основанные на использовании метода Монте-Карло.

Причина выдвижения данного предположения была технической и заключалась в упрощении анализа свойств оценок регрессии. Например, мы видели, что в модели парной регрессии

$$y = \alpha + \beta x + u \tag{8.1}$$

оценка МНК коэффициента наклона может быть представлена в виде разложения:

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \beta + \frac{\mathrm{Cov}(x, u)}{\mathrm{Var}(x)}. \tag{8.2}$$

Теперь если x — нестохастическая переменная, то Var (x) также является нестохастической величиной, и математическое ожидание ошибки может быть записано как $E[\text{Cov}(x, u)]/\text{Var}(x)$. Кроме того, если переменная x неслучайна, то $E[\text{Cov}(x, u)] = 0$. Поэтому доказательство того, что b — несмещенная оценка β, не вызвало затруднений. Хотя доказательство не было представлено, предположение о нестохастичности использовалось и при получении выражения для стандартной ошибки коэффициента. Кроме того, оно используется в теореме Гаусса—Маркова, доказывающей, что если удовлетворены условия Гаусса—Маркова, то оценки МНК эффективны.

В экономической практике предположение о нестохастичности часто оказывается весьма нереалистичным. Обычно обнаруживается, что объясняющие переменные модели сами были определены из других экономических зависимостей. Как мы увидим в главе 11, часто желательно рассматривать не одну зависимость изолированно, а целую систему зависимостей, действующих одновременно.

Мы изучим три типа моделей со стохастическими объясняющими переменными, классифицируемых в соответствии с тем, какова связь между распределениями этих переменных и распределением случайного члена. Все они имеют важное практическое значение.

1. В моделях первого типа объясняющие переменные распределены независимо от случайного члена.

2. В моделях второго типа объясняющие переменные и случайный член не являются независимыми, но их значения в каждый момент времени некоррелированы (т. е. текущие значения объясняющих переменных не коррелируют с текущим значением случайного члена).

3. В моделях третьего типа значения объясняющих переменных и случайного члена коррелируют в каждый момент времени.

Прежде чем начать рассмотрение указанных типов моделей, необходимо прояснить вопрос, о котором нередко забывают, — о дисперсиях и ковариациях объясняющих переменных в больших выборках. Обычно предполагается, что они стремятся к конечным пределам. Для упрощения мы примем здесь сильную форму этого допущения, заключающуюся в том, что объясняющие переменные могут рассматриваться как особый вид случайных переменных, для которых выборочные значения извлекаются (не обязательно независимо) из генеральных совокупностей с конечными средними, дисперсиями и ковариациями.

Справедливости ради следует отметить, что, по-видимому, мотивация для этого предположения — прежде всего практическая, так как это делает несложным определение поведения оценки в больших выборках. Действительно, это предположение является целесообразным, если вы имеете дело со статистическими данными, относящимися к различным отраслям экономики, и наблюдения берутся (случайно или в рамках схемы расслоенной выборки) из данной генеральной совокупности. Это предположение может быть также обоснованным в том случае, когда вы имеете дело с данными временного ряда, сформированными стационарным процессом, т. е. таким, в котором распределение x не зависит от времени. Уравнение (7.21) является примером стационарного процесса при выполнении условия устойчивости $-1 < \rho < 1$.

Тем не менее во многих моделях, особенно в тех, где используются данные временных рядов, это предположение не является целесообразным. Весьма очевидно, что когда модель включает переменные с трендом, имеет смысл считать, что Var (x) неограниченно увеличивается по мере расширения периода выборки. Примером являются функции спроса, которым в этой книге уделяется особое внимание. Поэтому мы будем рассматривать также и эту альтернативу. Анализ ограничим рассмотрением модели парной регрессии (8.1), но результаты легко распространяются и на случай множественной регрессии.

Случай, когда распределение x имеет конечное математическое ожидание и конечную дисперсию

Сначала рассмотрим случай, когда x извлекается из генеральной совокупности с конечными математическим ожиданием и дисперсией, обозначаемой σ_x^2.

а) x *и* u *независимо распределены*

Если x и u распределяются независимо друг от друга, то обычный МНК сохраняет все свои важные свойства. Сюда относятся несмещенность, эффективность и состоятельность. Кроме того, критическая статистика может использоваться, как обычно, при условии, что распределение x не зависит от параметров α, β или σ_u. Мы покажем, что выполняются условия несмещенности и состоятельности, а соблюдение требования эффективности примем на веру.

Несмещенность

Если x — стохастическая переменная, то Var (x) не может рассматриваться как скаляр, поэтому мы не можем переписать E [Cov (x, u)/Var (x)] как E [Cov (x, u)]/Var (x). Следовательно, обычное доказательство несмещенности здесь не проходит. Однако мы можем найти другой способ разложения ошибки:

$$\frac{\mathrm{Cov}(x,u)}{\mathrm{Var}(x)} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(u_i - \bar{u})}{\mathrm{Var}(x)} =$$

$$= \frac{1}{n}\sum \left(\frac{x_i - \bar{x}}{\mathrm{Var}(x)} \right)(u_i - \bar{u}) = \frac{1}{n}\sum f(x_i)(u_i - \bar{u}), \qquad (8.3)$$

где $f(x_i) = (x_i - \bar{x}) / \mathrm{Var}(x)$.

Далее, если x и u распределены независимо, то также независимо будут распределены $f(x)$ и u. Следовательно, используя одно из свойств независимости (см. Обзор), получаем:

$$E[f(x_i)(u_i - \bar{u})] = E[f(x_i)]E[u_i - \bar{u}] = E[f(x_i)] \times 0, \qquad (8.4)$$

так как $E(u_i)$, согласно предположению, равно нулю в каждом наблюдении. Следовательно, если мы берем математическое ожидание обеих частей уравнения (8.3), то правая часть приводится к виду: $(1/n)$, умноженное на сумму n членов, каждый из которых равен нулю. Следовательно, математическое ожидание ошибки равно нулю.

Состоятельность

Показать состоятельность также легко, если x имеет конечную теоретическую дисперсию σ_x^2. Мы знаем, что в общем случае plim (A/B) равен plim (A)/ plim (B), где A и B — произвольные случайные величины, у которых plim (A) и plim (B) существуют и plim (B) не равен нулю (см. Обзор; plim означает предельное значение при увеличении объема выборки). Мы также знаем, что plim Cov (x, u) равен pop. cov (x, u), которая равна нулю, если x и u независимо распределены. Следовательно,

$$\text{plim}b = \beta + \frac{\text{plim}\,\text{Cov}(x,u)}{\text{plim}\,\text{Var}(x)} = \beta + \frac{0}{\sigma_x^2} = \beta. \tag{8.5}$$

б) x *и* u *одномоментно некоррелированы*

Классический пример (единственный, который мы здесь рассмотрим), когда объясняющая переменная и случайный член одномоментно некоррелированы, заключается в использовании лаговой зависимой переменной в качестве одной из объясняющих переменных. Если мы имеем модель

$$y_t = \alpha + \beta_1 x_t + \beta_2 y_{t-1} + u_t, \tag{8.6}$$

то y_{t-1} находится непосредственно под воздействием u_{t-1} и косвенно — под влиянием всех предшествующих значений случайного члена. Следовательно, одна из объясняющих переменных в этой модели не имеет независимого от случайного члена распределения, и МНК не дает несмещенных оценок. Тем не менее несмотря на то, что приведенное выше доказательство несмещенности становится некорректным, доказательство состоятельности остается справедливым; если y_{t-1} и u_t некоррелированы, то можно показать, что plim Cov (y_{t-1}, u_t) равен нулю. Таким образом, МНК сохраняет желательные свойства в больших выборках, хотя в малых это не обязательно так.

в) x и u одномоментно коррелированы

Если x и u одномоментно коррелированы, то Cov (x, u) не будет стремиться к нулю даже в больших выборках, и оценка, полученная обычным МНК, является как смещенной, так и несостоятельной. Смещение в большой выборке равно пределу по вероятности Cov $(x, u)/\sigma_x^2$.

При разумных предпосылках представленные выше выводы, за одним исключением, остаются неизменными и в случае, когда взамен применявшегося ранее делается предположение о том, что по мере увеличения объема выборки Var (x) неограниченно растет. Основное отличие состоит в том, что, даже когда значения x и u коррелированы в каждый данный момент времени, МНК может обеспечивать состоятельность получаемых оценок. Если Cov (x, u) имеет конечный предел, а Var (x) увеличивается неограниченно, то ошибка в оценке β будет в больших выборках исчезать. С другой стороны, если Cov (x, u) не имеет конечного предела, то, очевидно, мало что можно сказать при отсутствии дополнительной информации. В этом случае возникает также проблема интерпретации проверок статистической значимости. Наиболее легкое решение состоит в том, чтобы трактовать их как условные, взятые при фактических выборочных значениях объясняющих переменных.

Мы не будем рассматривать здесь отдельно модель с независимо распределенной стохастической объясняющей переменной. Достаточно сказать, что большинство предыдущих примеров и упражнений в книге, вероятно, в большей мере соответствуют этой модели, чем первоначальной модели с нестохастической объясняющей переменной. Модели с лаговыми зависимыми переменными будут рассматриваться в главе 10. Оставшаяся часть главы 8 и вся глава 11 будут посвящены двум важным случаям, когда x и u в каждый отдельный момент коррелированы: когда данные подвержены воздействию ошибок измерения (глава 8) и когда осуществляется оценка параметров уравнения, входящего в состав системы одновременных зависимостей (глава 11). В обоих этих случаях если мы хотим получить состоятельные оценки, то должны найти какую-либо альтернативу методу наименьших квадратов.

8.2. Последствия ошибок измерения

В экономике при исследовании какой-либо зависимости используемые переменные часто оказываются неправильно измеренными. Например, в обследованиях часто имеются ошибки, сделанные по вине опрашиваемого, неправильно понимающего вопрос (а в некоторых случаях и по вине опрашивающего). Вместе с тем сообщение неправильных сведений является не единственным источником неточностей. Иногда случается, что вы каким-то образом определили переменную в модели, но имеющиеся данные свидетельствуют о несколько другом определении. Широко известным примером такого случая является рассматриваемый в разделе 8.3 критический анализ М. Фридменом стандартной функции потребления.

Первоначально мы рассмотрим классический случай, в котором дисперсия объясняющей переменной стремится в больших выборках к конечной теоретической дисперсии. В конце этого раздела мы проанализируем последствия принятия альтернативного допущения о том, что дисперсия неограниченно увеличивается.

Допустим, переменная y зависит от переменной z, что задано следующим соотношением:

$$y = \alpha + \beta z + v, \tag{8.7}$$

где v — случайный член с нулевым средним и дисперсией σ_v^2.

Предположим, что z невозможно измерить абсолютно точно, и мы будем использовать x для обозначения его измеренного значения. В i-м наблюдении x_i равно истинному значению z_i плюс ошибка измерения w_i:

$$x_i = z_i + w_i. \tag{8.8}$$

Допустим, что w имеет нулевое среднее и дисперсию σ_w^2, что Var (z) в больших выборках стремится к конечному пределу σ_z^2 и что z и v распределены независимо.

Подставляя формулу (8.8) в уравнение (8.7), получим:

$$y = \alpha + \beta x + v - \beta w. \tag{8.9}$$

Это уравнение имеет две случайные составляющие — первоначальный случайный член v и ошибку измерения w (умноженную на $-\beta$). Вместе они образуют составную случайную переменную, которую мы назовем u:

$$u = v - \beta w. \tag{8.10}$$

Соотношение (8.9) можно теперь записать как

$$y = \alpha + \beta x + u. \tag{8.11}$$

Имея значения переменных y (временно будем предполагать, что они измерены точно) и x, мы, несомненно, можем оценить регрессионную зависимость y от x.

Коэффициент регрессии b, как обычно, представляется выражением (8.2). Анализируя ошибку, можно заметить, что она, вероятно, поведет себя не так, как требуется. Переменная x зависит от w (8.8), от этой величины зависит также и u (8.10). Когда ошибка измерения в наблюдении оказывается положительной, происходят две вещи: x_i имеет положительную составляющую w_i, а u_i имеет отрицательную составляющую $-\beta w_i$. Аналогично, если ошибка измерения отрицательна, она вносит отрицательный вклад в величину x_i и положительный вклад в величину u_i. Следовательно, корреляция между x и u отрицательна. Величина pop. cov (x, u) не равна нулю, а из соотношения (8.2) следует, что b является несостоятельной оценкой β.

Даже если бы у нас была очень большая выборка, оценка оказалась бы неточной. Она бы занижала β на величину

$$\frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}\beta. \tag{8.12}$$

Доказательство этого дается ниже. Сначала мы отметим его очевидные следствия. Чем больше теоретическая дисперсия ошибки измерения по отношению к теоретической дисперсии z, тем больше будет отрицательное смещение. Напри-

мер, если бы σ_w^2 было равно $0{,}25\sigma_z^2$, то отрицательное смещение составило бы:

$$-\frac{0{,}25\sigma_z^2}{1{,}25\sigma_z^2}\beta,$$

что равняется $0{,}2\beta$. Даже если бы выборка была очень большой, оценка оказалась бы на 20% ниже истинного значения при положительном β и на 20% выше его при отрицательном β.

Рисунок 8.1 показывает, как ошибка измерения приводит к появлению смещенных коэффициентов регрессии, если использовать модель, представленную выражениями (8.7) и (8.8). На рис. 8.1А мы предполагаем, что ошибка измерения отсутствует и что отклонения от линии регрессии вызываются только случайным членом v. На рис. 8.1Б предполагается, что переменная x подвержена воздействию существенной ошибки измерения, которая сдвигает наблюдения вправо при их положительном значении и влево — при отрицательном. По причине горизонтального рассеяния множество точек наблюдений здесь кажется более пологим, чем на рис. 8.1А, и оцененная линия регрессии будет иметь тенденцию к занижению угла наклона истинной линии зависимости. Чем больше дисперсия ошибки измерения по отношению к дисперсии x, тем больше окажется эффект уменьшения угла наклона и тем сильнее будет смещение.

Несовершенные замещающие переменные

В главе 6 было показано, что если мы не можем получить данные об одной из объясняющих переменных в регрессионной модели и оцениваем регрессию без нее, то коэффициенты при других переменных, вообще говоря, будут смещенными, их стандартные ошибки — некорректными, а коэффициент R^2 — трудно интерпретируемым. Однако в разделе 6.4 мы видели, что если можно найти полноценную замену для отсутствующей переменной, т. е. другую переменную, связанную с ней строгой линейной зависимостью, и использовать ее в регрессии вместо отсутствующей переменной, то основная часть результатов оценивания регрессии будет сохранена. Таким образом, коэффициенты при других переменных не будут смещенными, их стандартные ошибки и соответствующие t-тесты будут обоснованными, и коэффициент R^2 будет таким же, как если бы мы могли непосредственно включить переменную, которую невозможно измерить. Мы не сможем получить оценку коэффициентов последней, но ее t-статистика будет такой же, как t-статистика для замещающей переменной.

К сожалению, крайне редко удается найти идеальную замещающую переменную. Обычно самое большее, на что мы можем рассчитывать, — это замещающая переменная, связанная нестрогой линейной зависимостью с отсутствующей переменной. Последствия использования несовершенной замещающей переменной (взамен совершенной) близки к последствиям использования переменной, подверженной воздействию ошибки измерения (вместо переменной, когда такие ошибки отсутствуют). Они заключаются в том, что коэффициенты регрессии оказываются смещенными, оцененные стандартные ошибки некорректны и т. д.

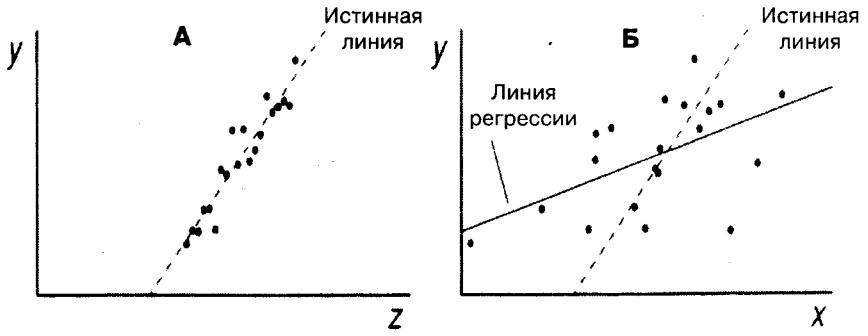


Рис. 8.1. Влияние ошибок измерения объясняющей переменной

Вместе с тем можно признать оправданным использование замещающей переменной, если есть основания полагать, что степень ее несовершенства не настолько велика, чтобы смещение было серьезным, а стандартные ошибки вводили в заблуждение. Так как обычно нет способа проверить, насколько велика или мала степень несовершенства, решение об использовании замещающей переменной или отказе от него приходится принимать, основываясь на субъективных соображениях и учитывая конкретные условия, связанные с моделью.

Доказательство несостоятельности

Доказательство справедливости выражения (8.12) не содержит ничего такого, что было бы нам неизвестно, и оно не особенно длинно, поэтому и приводится в данной книге. Если, однако, оно покажется трудным, можно пропустить его и перейти к следующему разделу.

Так как x и u не являются независимо распределенными величинами, не существует простого способа описать результирующее поведение отношения $\mathrm{Cov}\,(x, u)/\mathrm{Var}\,(x)$ в малых выборках. Нельзя даже получить выражение для его математического ожидания. Самое большее, что можно сделать, — это предсказать его поведение в том случае, если бы выборка была очень большой. Это отношение стремилось бы к теоретической ковариации между x и u, деленной на теоретическую дисперсию x. Мы рассмотрим этот вопрос отдельно.

Пользуясь определениями x и u, а также правилами вычисления ковариаций, можно получить разложение их выборочной ковариации:

$$\mathrm{Cov}\,(x, u) = \mathrm{Cov}\,\{(z + w),(v - \beta w)\} =$$

$$= \mathrm{Cov}\,(z, v) + \mathrm{Cov}\,(w, v) - \mathrm{Cov}\,(z, \beta w) - \mathrm{Cov}\,(w, \beta w). \qquad (8.13)$$

Выборочные дисперсии и ковариации с ростом объема выборки стремятся к своим теоретическим аналогам, если последние существуют. В нашем случае как $\mathrm{pop.\,cov}\,(z, v)$, так и $\mathrm{pop.\,cov}\,(z, \beta w)$ равны нулю. Предположим, что связь между v и w отсутствует, так что $\mathrm{pop.\,cov}\,(w, v)$ равна нулю. Тогда остается член $-\mathrm{pop.\,cov}\,(w, \beta w)$, представляющий собой $-\beta\ \mathrm{pop.\,cov}\,(w, w)$ или $-\beta\ \mathrm{pop.\,var}\,(w)$. Следовательно, теоретическая ковариация между x и u равна $-\beta\sigma_w^2$.

Теперь мы рассмотрим Var (x): она равна Var $(z + w)$. Поэтому, используя правила вычисления дисперсий, имеем:

$$\text{Var}(x) = \text{Var}(z + w) = \text{Var}(z) + \text{Var}(w) + 2\,\text{Cov}(z, w). \tag{8.14}$$

В предположении, что z и w распределены независимо, получим, что pop. cov (z, w) равна нулю, а Var (x) будет в больших выборках стремиться к $[\sigma_z^2 + \sigma_w^2]$.

Сопоставляя эти два результата, можно показать, что Cov (x, u)/Var (x) в больших выборках стремится к $-\beta\sigma_w^2/(\sigma_z^2 + \sigma_w^2)$; поэтому ввиду (8.2) b стремится к

$$\beta - \left(\frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}\right)\beta. \tag{8.15}$$

Если выразить это, используя показатели сходимости по вероятности, то можно записать:

$$\text{plim}(b) = \beta + \frac{\text{plim}\{\text{Cov}(x,u)\}}{\text{plim}\{\text{Var}(x)\}} = \beta + \frac{-\beta\sigma_w^2}{\sigma_z^2 + \sigma_w^2}. \tag{8.16}$$

Ошибки измерения зависимой переменной

Ошибки измерения зависимой переменной не имеют столь большого значения. На практике их можно считать составляющими случайного члена. Они нежелательны, так как все, что увеличивает «уровень шума» в модели, приводит к уменьшению точности оценок коэффициентов регрессии; тем не менее, они не вызывают смещения этих оценок.

Предположим, что истинное значение зависимой переменной равно q и истинная зависимость имеет вид:

$$q = \alpha + \beta x + v, \tag{8.17}$$

где v — случайный член. Если y_i — это измеренное значение зависимой переменной в i-м наблюдении и r_i — ошибка измерения, то

$$y_i = q_i + r_i. \tag{8.18}$$

Следовательно, зависимость между наблюдаемым значением зависимой переменной и x представляется выражением:

$$y - r = \alpha + \beta x + v, \tag{8.19}$$

которое может быть переписано как

$$y = \alpha + \beta x + u, \tag{8.20}$$

где u — составная случайная переменная $(v + r)$.

Единственное отличие этой модели от обычной заключается в том, что случайный член в уравнении (8.20) имеет две составляющие: первоначальный случайный член и ошибку измерения y. Важно, что здесь нет воздействия на объясняющую переменную x. Следовательно, если переменная x является неслучай-

ной или если она распределяется независимо от u, то МНК по-прежнему будет давать несмещенные оценки.

Дисперсия Var (x), не стремящаяся к конечному пределу при увеличении объема выборки

Если с ростом объема выборки Var (x) неограниченно увеличивается, то в обсуждение последствий включения в объясняющую переменную ошибок измерения требуется внести поправку. Мы видели, что для любой конечной выборки

$$b = \beta + \frac{\mathrm{Cov}(z, v) + \mathrm{Cov}(w, v) - \mathrm{Cov}(z, \beta w) - \mathrm{Cov}(w, \beta w)}{\mathrm{Var}(z) + \mathrm{Var}(w) + 2\mathrm{Cov}(z, w)}. \qquad (8.21)$$

Можно показать, что при разумных предположениях, когда Var (z) увеличивается, все другие составляющие ошибки становятся пренебрежимо малыми по сравнению с Var (z), и, следовательно, при росте объема выборки ошибка будет стремиться к нулю. Другими словами, влияние ошибок измерения становится пренебрежимо малым в больших выборках, в результате чего оказывается, что МНК приводит к состоятельным оценкам. Тем не менее в малых выборках они будут смещенными.

Более важное предположение состоит в том, что переменная w действительно гомоскедастична. Это значит, что σ_w^2 постоянна; следовательно, мы предполагаем, что дисперсия ошибки измерения не увеличивается по мере роста x. Если же это не так, то наши рассуждения и выкладки становятся некорректными.

Упражнения

8.1. В некоторой отрасли промышленности фирмы определяют соотношение между запасами готовой продукции (Y) и ожидаемыми годовыми объемами продаж (X^e) в соответствии с линейной зависимостью:

$$Y = \alpha + \beta X^e.$$

Фактические объемы продаж X отличаются от ожидаемых на случайную величину u, которая распределена с нулевым математическим ожиданием и постоянной дисперсией:

$$X = X^e + u.$$

При этом распределение u независимо от X^e.

В распоряжении исследователя имеются данные об Y и X (но не об X^e), полученные по результатам перекрестной выборки для фирм в стране. Опишите проблемы, с которыми придется иметь дело в случае использования обычного МНК при построении регрессионной зависимости Y от X и оценивании α и β.

8.2. В аналогичной отрасли промышленности фирмы связывают *предполагаемые* запасы готовой продукции (Y^*) с ожидаемыми годовыми объемами продаж (X^e), используя линейную зависимость:

$$Y^* + \alpha + \beta X^e.$$

Фактические объемы продаж X отличаются от ожидаемых на случайную величину u, которая распределена с нулевым математическим ожиданием и постоянной дисперсией:

$$X = X^e + u.$$

Величина u распределена независимо от X^e. Так как непредусмотренные объемы продаж приводят к уменьшению запасов, фактические запасы Y выражаются в виде:

$$Y = Y^* - u.$$

В распоряжении исследователя имеются данные по Y и X перекрестной выборки фирм в масштабе страны (но нет данных по Y^* и X^e). Опишите проблемы, с которыми придется столкнуться в этом случае, если для оценивания α и β при построении регрессионной зависимости Y от X используется обычный МНК.

8.3. Критика М. Фридменом стандартной функции потребления

Теперь мы приступим к рассмотрению наиболее известного применения анализа ошибок измерения в экономической теории — к изложению критических взглядов М. Фридмена на использование МНК для оценивания функции потребления (Friedman, 1957). Здесь рассматривается предпринятый М. Фридменом анализ данной проблемы, а в главе 10 представлено предложенное им решение.

В модели Фридмена потребление i-го индивида в период t связывается не с фактическим текущим доходом, а с постоянным доходом, который будет обозначаться как Y_{it}^P. Постоянный доход следует рассматривать как долговременное понятие дохода — сумму, на которую человек может в большей или меньшей степени рассчитывать, принимая во внимание ее возможные колебания. Постоянный доход субъективно определяется на основе полученного в последнее время опыта и ожиданий на будущее, и поскольку это понятие субъективное, он не может быть измерен непосредственно. Фактический доход в том или ином году может быть выше или ниже постоянного дохода в зависимости от конкретных условий в данном году. Разность между фактическим и постоянным доходом рассматривается как переменный доход Y_{it}^T:

$$Y_{it} = Y_{it}^P + Y_{it}^T. \tag{8.22}$$

Таким же образом М. Фридмен проводит различие между фактическим потреблением C_{it} и постоянным потреблением C_{it}^P. Постоянное потребление представляет собой уровень потребления, обусловленный уровнем постоянного дохода. Фактическое потребление может отличаться от него в случае возникновения особых непредусмотренных обстоятельств (например непредусмотренных расходов на медицинское обслуживание) или в случае непредвиденных покупок. Разность этих величин описывается как переменное потребление C_{it}^T:

$$C_{it} = C_{it}^P + C_{it}^T. \tag{8.23}$$

Предполагается, что Y_{it}^T, C_{it}^T — случайные переменные с нулевым математическим ожиданием и постоянной дисперсией, не коррелированные с Y_{it}^P и C_{it}^P и друг с другом. Далее М. Фридмен выдвинул гипотезу о том, что постоянное потребление прямо пропорционально постоянному доходу:

$$C_{it}^P = \beta Y_{it}^P. \tag{8.24}$$

Если модель Фридмена правильна, то возникает вопрос, что случится, если вы по незнанию попытаетесь оценить обычную функцию потребления, описанную регрессионной зависимостью измеренного потребления от измеренного дохода, и оцените регрессию в виде:

$$\hat{C} = a + bY. \tag{8.25}$$

В регрессии как зависимая, так и объясняющая переменные были измерены неточно; и ошибки измерения равны C^T и Y^T. В соответствии с результатами анализа, проведенного в разделе 8.2,

$$z = Y^P; \quad w = Y^T; \quad q = C^P; \quad r = C^T. \tag{8.26}$$

Как мы видели в этом разделе, ошибки измерения зависимой переменной ведут лишь к увеличению дисперсии случайного члена. Использование неправильной концепции дохода ведет к более серьезным последствиям. В результате этого оценка β становится несостоятельной. Из выражения (8.12) можно видеть, что в больших выборках β будет недооцениваться примерно на величину:

$$\frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}\beta = \frac{\sigma_{Y^T}^2}{\text{plimVar}(Y^P) + \sigma_{Y^T}^2}\beta, \tag{8.27}$$

где $\sigma_{Y^T}^2$ — теоретическая дисперсия Y^T, если существует plim Var (Y^P). Мы рассмотрим два случая: первый, в котором Y^P берется из генеральной совокупности с конечной дисперсией $\sigma_{Y^P}^2$; и второй, в котором Var (Y^P) неограниченно возрастает.

Предел по вероятности plim Var $(Y^P) = \sigma_{Y^P}^2$

К первому случаю, очевидно, относится задача построения регрессионной функции потребления по данным перекрестной выборки. Предельное значение коэффициента наклона в больших выборках представляется выражением:

$$\text{plim}\, b = \beta - \frac{\sigma_{Y^T}^2}{\sigma_{Y^P}^2 + \sigma_{Y^T}^2}\beta. \tag{8.28}$$

Это означает, что даже в больших выборках выявленная предельная склонность к потреблению (полученная оценка b) будет ниже, чем истинное значение β в зависимости (8.24). Величина смещения зависит от соотношения

между дисперсиями переменного и постоянного доходов. Смещение будет наибольшим для тех профессий, где доход в наибольшей степени подвержен колебаниям. Очевидным примером может служить ведение фермерского хозяйства. Модель Фридмена предсказывает, что даже если для фермеров было бы такое же значение β, как и для остальной части населения, их предельная склонность к потреблению по отношению к измеренному доходу была бы относительно низкой, и это согласуется с фактами (Friedman, 1957, p. 57 и далее).

Иллюстрация

Анализ, проделанный М. Фридменом, будет проиллюстрирован с использованием метода Монте-Карло. Предположим, что 20 человек, включенные в выборку, имеют постоянный доход 2000, 2100, 2200, ..., 3900. Допустим также, что переменный доход каждого из указанных индивидов равен случайному числу, извлеченному из нормальной генеральной совокупности с нулевым средним и единичной дисперсией, умноженному на 200 (как обычно, случайные числа берутся из таблицы нормальных случайных чисел). Измеренный доход для каждого из 20 человек представляет собой сумму постоянного и переменного доходов. Предположим, что истинное значение β равно 0,9; таким образом, постоянное потребление составляет 0,9 от соответствующего постоянного дохода. Переменная составляющая потребления здесь не рассматривается, и измеренное потребление равно постоянному потреблению. Результат оценивания регрессионной зависимости измеренного потребления от измеренного дохода имеет вид:

$$\underset{\text{(с. о.)}}{\hat{C} = \underset{(179)}{443} + \underset{(0,06)}{0,75}Y}; \qquad R^2 = 0,89. \tag{8.29}$$

Как и предполагалось, оцененная предельная склонность к потреблению оказалась ниже истинного значения. В самом деле, если построить 95-процентный доверительный интервал, используя результаты оценки регрессии, то истинное значение оказалось бы за его границами и, следовательно, было бы отклонено при 5-процентном уровне значимости. При 18 степенях свободы критический уровень t составляет 2,10; таким образом, доверительный интервал вычислялся бы как

$$0,75 - 2,10 \times 0,06 \le \beta \le 0,75 + 2,10 \times 0,06, \tag{8.30}$$

или

$$0,62 \le \beta \le 0,88. \tag{8.31}$$

Следовательно, вы допустили бы ошибку I рода. Фактически наличие ошибок измерения делает некорректной стандартную ошибку Y, а значит, и доверительный интервал. Еще один побочный эффект заключается в том, что постоянный член, который должен быть равным нулю, так как он отсутствовал при расчете значений C, имеет значимо отличное от нуля (на 5-процентном уровне значимости) положительное значение. Данный эксперимент был повторен еще 9 раз, и результаты приводятся в табл. 8.1, серия A.

Оценка b показывает явно отрицательно смещенную предельную склонность

к потреблению. В девяти из десяти экспериментов она ниже, чем истинное значение 0,90. Проверим, согласуются ли эти результаты с выводами теоретического анализа, на основе которого получено уравнение (8.28). В нашем примере σ^2_{YT} равно 40 000, так как Y^T имеет стандартное отклонение 200. Предположим, что в больших выборках Y^P принимает значения 2000, 2100, ..., 3000 с равной вероятностью и, следовательно, что величина σ^2_{YP} — конечна и равна дисперсии этого набора чисел, составляющей 332 500. Таким образом, в больших выборках оценка коэффициента β будет заниженной на величину:

$$\frac{40000}{332500+40000} \times 0{,}90 = 0{,}11 \times 0{,}90 = 0{,}10. \qquad (8.32)$$

Следует подчеркнуть, что такой вывод справедлив только для больших выборок и что ничего нельзя сказать о поведении оценки в выборках небольшого объема. Однако в данном случае можно видеть, что на самом деле это значение представляет собой хороший ориентир. Проанализировав оценки коэффициента β в 10 указанных экспериментах, мы видим, что они, по-видимому, случайно распределены вокруг 0,80 (а не 0,90) и что таким образом имеется отрицательное смещение приблизительно на 0,10.

Таблица 8.1

№	*Серия экспериментов А*				*Серия экспериментов Б*			
	a	*с.о.(a)*	*b*	*с.о.(b)*	*a*	*с.о.(a)*	*b*	*с.о.(b)*
1	443	179	0,75	0,06	1001	251	0,56	0,08
2	152	222	0,83	0,07	755	357	0,62	0,11
3	101	222	0,89	0,08	756	376	0,68	0,13
4	195	179	0,83	0,06	668	290	0,66	0,09
5	319	116	0,78	0,04	675	179	0,64	0,06
6	371	200	0,78	0,07	982	289	0,57	0,10
7	426	161	0,74	0,05	918	229	0,56	0,07
8	−146	275	0,93	0,09	625	504	0,66	0,16
9	467	128	0,74	0,04	918	181	0,58	0,06
10	258	153	0,80	0,05	679	243	0,65	0,08

Последствием занижения оценки *b* является завышение оценки *a*, которое оказывается положительным, несмотря на то что истинное значение α равно нулю. Действительно, в четырех случаях *t*-тест показывает, что эта величина значимо отличается от нуля при 5-процентном уровне значимости. Вместе с тем в этих условиях *t*-тесты являются некорректными, потому что невыполнение четвер-

того условия Гаусса—Маркова делает некорректным расчет стандартных ошибок, а значит, и t-статистик.

Что произойдет, если мы увеличим дисперсию Y^T, оставив все остальное без изменения? В данных серии *Б* из табл. 8.1 первоначальные случайные числа умножались на 400 вместо 200, поэтому величина $\sigma^2_{Y^T}$ составила 160 000 вместо 40 000. Значение ошибки в выражении (8.28) теперь равно 160 000/(332 500 + + 160 000), что составляет 0,32, поэтому можно предполагать, что в выборках увеличивающегося объема b будет стремиться к (0,9 − 0,32 × 0,9), то есть к 0,61. Мы снова видим, что это хороший ориентир, позволяющий судить о подлинном поведении b, несмотря на то что в каждой выборке содержится всего лишь 20 наблюдений. Как и следовало предполагать, значения оценки a здесь даже больше, чем в серии *А*.

Неограниченный рост Var (Y^p)

Если Var (Y^p) неограниченно увеличивается, а $\sigma^2_{Y^T}$ конечна, то в принципе смещение исчезнет по мере роста числа наблюдений в выборке. Тем не менее в малых выборках оно может быть значительным, и могут потребоваться поправки — либо по методу, который использовался М. Фридменом, либо по методу, рассматриваемому в следующем разделе.

Выводы для экономической политики

Имеются два отдельных и противоположных вывода относительно мультипликатора. Во-первых, если М. Фридмен прав, то регрессионная зависимость фактического потребления от фактического дохода в результате даст заниженную величину предельной склонности к потреблению и, следовательно, заниженную оценку мультипликатора. В примере из предыдущего раздела истинное значение β было равно 0,90 и, таким образом, истинное значение мультипликатора равнялось 10. Однако в серии *А* оценка β стремилась к 0,80, что означало, что мультипликатор равен всего лишь 5. В серии *Б* его значение было еще ниже. Оценка β стремилась к 0,61, что дает значение мультипликатора 2,6.

Если правительство пользуется заниженной оценкой мультипликатора, то в результате будут недооцениваться последствия бюджетно-налоговой политики. Например, повышение государственных расходов в целях снижения безработицы может фактически привести к избыточному повышению действительного спроса и к усилению инфляции.

Второй вывод заключается в том, что мультипликатор относится только к той части изменения дохода, которая воспринимается как постоянная, так как (согласно М. Фридмену) потребление зависит только от постоянного дохода. Таким образом, если предполагается, что увеличение государственных расходов является временным, то оно (в первом приближении) вообще не будет оказывать влияния на потребление и связанный с ним мультипликатор будет равен единице.

Эти замечания должны быть связаны с рассмотрением формы, в которой люди хранят сбережения. Мы пока неявно предполагали, что они держат их в форме финансовых активов (банковские депозиты, облигации и т. д.). Вместе с тем в модели Фридмена в качестве одной из форм сбережения рассматриваются расходы на потребительские товары длительного пользования. Дополнительные средства, образовавшиеся в результате увеличения переменного дохода, не будут потрачены на обычные предметы потребления, но возможно частичное их сбережение в форме закупок потребительских товаров длительного пользования, и повышение спроса на них приведет к эффекту мультипликатора. Суммарный краткосрочный эффект мультипликатора может быть не таким малым.

Первое замечание говорит о том, что если М. Фридмен прав, то регрессионная зависимость C_t от Y_t является неудачной с эконометрической точки зрения. Второе замечание означает, что предположение о зависимости C_t только от Y_t неудачно с точки зрения экономической теории (подчеркнем, если М. Фридмен прав). Вместе эти замечания означают, что мультипликатор, вычисленный по оценкам регрессии между C^t и Y^t, вероятно, будет неточным как для краткосрочного, так и для долгосрочного периода.

В данном случае рассматривалась только теория потребления — первоначальная область применения введенного М. Фридменом понятия постоянного дохода; но это понятие может применяться и применяется и в других областях. В частности, в денежной теории можно показать, что спрос на наличные деньги для сделок должен соотноситься не с фактическим, а с постоянным доходом; в теории инвестиций можно утверждать, что акселератор должен быть связан с изменениями не в фактическом, а в постоянном доходе. Первоначальный вклад в решение этих вопросов был сделан М. Фридменом (Friedman, 1959) и Р. Эйснером (Eisner, 1967).

Упражнения

8.3. В некоторой экономике дисперсия переменного дохода составляет 0,5 от дисперсии постоянного дохода, склонность к потреблению товаров кратковременного пользования за счет постоянного дохода равна 0,6 и нет расходов на товары длительного пользования. Каким будет значение мультипликатора, полученного на основе построения простейшей регрессионной зависимости потребления от дохода, и каково его истинное значение?

8.4. В определение постоянного потребления М. Фридмен включает потребление услуг, обеспечиваемых товарами длительного пользования. Закупки товаров длительного пользования характеризуются как одна из форм сбережений. В экономике, подобной той, которая была рассмотрена в упражнении 8.3, дисперсия переменного дохода составляет 0,5 от дисперсии постоянного дохода, склонность к потреблению товаров кратковременного пользования за счет постоянного дохода равна 0,6 и половина текущих сбережений (фактический доход минус расходы на товары кратковременного пользования) принимает форму расходов на товары длительного пользования. Каким будет значение мультипликатора, полученного на основе простой регрессии между потреблением и доходом, и каково его истинное значение?

8.4. Инструментальные переменные

Что следует делать при наличии ошибок измерения? Если их причиной является неточность при подготовке данных, то единственное, что можно сделать, — это обрабатывать данные более тщательно. Если же их причина заключается в том, что измеряемая переменная принципиально отличается от истинной объясняющей переменной в зависимости, то можно попытаться получить более подходящие данные. Часто это бывает трудно осуществить на практике. Если требуется получить временной ряд по совокупному измеренному доходу, то его можно найти в национальных счетах, но не существует прямого способа получения данных по совокупному постоянному доходу. М. Фридмен решил эту проблему, предложив оригинальный косвенный метод, рассматриваемый в главе 10.

Здесь мы объясним использование *метода инструментальных переменных* (ИП) — наиболее важной разновидности метода наименьших квадратов — для решения данной задачи. Это также будет иметь большое значение, когда мы приступим к оцениванию параметров моделей, состоящих из нескольких уравнений.

В сущности, метод инструментальных переменных заключается в частичной замене непригодной объясняющей переменной такой переменной, которая не коррелирована со случайным членом. Ограничимся случаем парной регрессии:

$$y = \alpha + \beta x + u \tag{8.33}$$

и допустим, что по какой-либо причине x имеет случайную составляющую, зависящую от u. Будем также предполагать, что в больших выборках $\mathrm{Var}(x)$ стремится к конечному пределу σ_x^2. В этих условиях непосредственное применение МНК для построения регрессионной зависимости y от x привело бы к несостоятельным оценкам параметров.

Теперь предположим, что можно найти другую переменную z, которая коррелирована с x, но не коррелирована с u. Покажем, что основанная на использовании инструментальных переменных оценка параметра β, определяемая как

$$b_{\text{ИП}} = \frac{\mathrm{Cov}(z, y)}{\mathrm{Cov}(z, x)}, \tag{8.34}$$

является состоятельной при условии, что при увеличивающемся числе наблюдений $\mathrm{Cov}(z, x)$ стремится к конечному, отличному от нуля пределу, который мы обозначим как σ_{xz}. Это означает, что в больших выборках $b_{\text{ип}}$ стремится к истинному значению β. Перед этим полезно сравнить $b_{\text{ип}}$ с оценкой МНК, которую обозначим как $b_{\text{мнк}}$:

$$b_{\text{МНК}} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Cov}(x, x)}. \tag{8.35}$$

так как $\mathrm{Cov}(x, x)$ и $\mathrm{Var}(x)$ — одно и то же. Оценка ИП в парном регрессионном анализе получается путем подстановки инструментальной переменной z вместо x в числителе и вместо одного x (но не обоих) в знаменателе.

Используя уравнение (8.33), мы можем записать выражение для $b_{\text{ип}}$ следующим образом:

$$b_{\text{ип}} = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)} = \frac{\text{Cov}(z, \{\alpha + \beta x + u\})}{\text{Cov}(z, x)} =$$

$$= \frac{\text{Cov}(z, \alpha)}{\text{Cov}(z, x)} + \frac{\text{Cov}(z, \beta x)}{\text{Cov}(z, x)} + \frac{\text{Cov}(z, u)}{\text{Cov}(z, x)} = \beta + \frac{\text{Cov}(z, u)}{\text{Cov}(z, x)}, \tag{8.36}$$

так как Cov (z, α) равна нулю (α является постоянной) и Cov $(z, \beta x)$ равна β Cov (z, x). Таким образом, можно заметить, что оценка по методу инструментальных переменных равна истинному значению плюс ошибка, равная Cov (z, u)/Cov (z, x). В больших выборках ошибка исчезает, так как

$$\text{plim} b_{\text{ип}} = \beta + \frac{\text{plimCov}(z, u)}{\text{plimCov}(z, x)} = \beta + \frac{0}{\sigma_{xz}} = \beta \tag{8.37}$$

при условии, что переменная z действительно распределена независимо от u. Следовательно, на больших выборках $b_{\text{ип}}$ будет стремиться к истинному значению β.

Почти ничего нельзя сказать о распределении оценки $b_{\text{ип}}$ на малых выборках, но при увеличении n ее распределение будет стремиться к нормальному с математическим ожиданием β и дисперсией:

$$\text{pop.var}(b_{\text{ип}}) \to \frac{\sigma_u^2}{n \, \text{pop.var}(x)} \times \frac{1}{r_{x,z}^2}, \tag{8.38}$$

где $r_{x,z}$ — выборочный коэффициент корреляции между x и z.

Сравним полученное выражение с дисперсией оценки МНК:

$$\text{pop.var}(b_{\text{МНК}}) = \frac{\sigma_u^2}{n \, \text{Var}(x)}. \tag{8.39}$$

Основное различие заключается в том, что дисперсия $b_{\text{ип}}$ умножается на $1/r_{x,z}^2$. Чем теснее корреляция между x и z, тем меньше будет этот коэффициент и, следовательно, тем меньше будет дисперсия $b_{\text{ип}}$. Следовательно, если мы стоим перед выбором между несколькими возможными инструментальными переменными, то следует выбрать наиболее тесно коррелированную с x, потому что при прочих равных условиях она даст наиболее эффективные оценки. Вместе с тем было бы нежелательно использовать инструментальную переменную, полностью коррелированную с x, даже если бы ее удалось найти, потому что тогда она автоматически оказалась бы коррелированной также и с u, и мы по-прежнему получили бы несостоятельные оценки. Нам нужна инструментальная переменная, наиболее тесно коррелированная с x, но без корреляции с u.

Что следует делать при невозможности найти инструментальную переменную, достаточно тесно коррелированную с x? Тогда можно вновь вернуться к методу наименьших квадратов. Если, например, критерием выбора оценки является ее стандартная ошибка, то вы можете предпочесть оценку МНК любой оценке, полученной по методу инструментальных переменных, несмотря на смещение, потому что здесь дисперсия меньше.

Использование инструментальных переменных для оценивания функции потребления Фридмена

В контексте гипотезы Фридмена о постоянном доходе впервые инструментальные переменные использовались Н. Ливиатаном (Liviatan, 1963). В распоряжении Н. Ливиатана были данные о потреблении и доходе в 883 домашних хозяйствах для двух последовательных лет. Обозначим потребление и доход в первом году через C_1 и Y_1, а во втором году — через C_2 и Y_2.

Н. Ливиатан обнаружил, что если теория Фридмена правильна, то Y_2 может выступать в качестве инструментальной переменной для Y_1. Очевидно, что она, скорее всего, тесно коррелирована с Y_1, так что одно из двух условий для хорошей инструментальной переменной выполнено. Во-вторых, если переменная составляющая измеренного дохода за соседние годы некоррелирована, как это предполагал М. Фридмен, то Y_2 будет некоррелирована со случайным членом зависимости между C_1 и Y_1; таким образом, удовлетворено и другое условие.

Можно также попробовать использовать C_2 как инструментальную переменную для Y_1. Она будет тесно коррелирована с Y_2 и, следовательно, с Y_1, а также не будет коррелирована со случайным членом зависимости между C_1 и Y_1, если в соответствии с гипотезой Фридмена переменные составляющие потребления не коррелированы друг с другом.

По аналогии с этим можно оценивать регрессии по данным за второй год, используя Y_1 и C_1 в качестве инструментальных переменных для Y_2. Н. Ливиатан опробовал здесь все четыре комбинации, разделив выборку на лиц наемного труда и ведущих собственное дело. Он обнаружил, что в четырех случаях оценки предельной склонности к потреблению были значимо большими при однопроцентном уровне значимости, чем те, которые были получены непосредственно методом наименьших квадратов. В одном случае оценка была значимо большей, чем оценка МНК при уровне значимости в 5%, а в других трех случаях различие было незначимым; в целом полученные результаты подтверждают гипотезу постоянного дохода. Однако предельная склонность к потреблению в общем была не такой высокой, как средняя склонность к потреблению. Следовательно, полученные результаты не подтверждают гипотезу о единичной эластичности потребления по постоянному доходу, не явно предполагаемую соотношением (8.24).

Упражнения

8.5. В упражнении 8.1 количество труда L, применяемого фирмами, является линейной функцией от объема ожидаемых продаж:

$$L = \gamma + \delta X^e.$$

Объясните, как эта зависимость может использоваться исследователем для решения проблемы смещения, вызванного ошибками измерения.

8.6. В чем разница между инструментальной переменной и замещающей переменной (см. раздел 6.4)? Когда целесообразно воспользоваться одной из этих переменных и когда — другой?

9

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

Часто случается так, что отдельные факторы, которые вы хотели бы ввести в регрессионную модель, являются качественными по своей природе и, следовательно, не измеряются в числовой шкале. Приведем несколько примеров.

1. Исследуется зависимость между продолжительностью полученного образования и доходом, и в выборке представлены лица как мужского, так и женского пола. Нужно выяснить, обусловливает ли пол различие в результатах.

2. Исследуется зависимость между доходом и потреблением в Бельгии, и выборка включает как франкоговорящие семьи, так и семьи, говорящие по-фламандски. Нужно выяснить, имеет ли существенное значение это этническое различие.

3. Исследуются факторы, определяющие инфляцию, и в некоторые годы периода наблюдений правительство проводило политику регулирования доходов. Нужно проверить, оказало ли это какое-либо влияние на исследуемую зависимость.

В каждом из этих примеров одним из возможных решений было бы оценивание отдельных регрессий для двух указанных категорий с последующим выяснением, различаются ли полученные коэффициенты. Другой возможный подход к решению состоит в оценивании единой регрессии с использованием всей совокупности наблюдений и измерением степени влияния качественного фактора посредством введения так называемой *фиктивной переменной*. Второй подход обладает двумя важными преимуществами: во-первых, имеется простой способ проверки, является ли воздействие качественного фактора значимым; во-вторых, при условии выполнения определенных предположений регрессионные оценки оказываются более эффективными.

9.1. Иллюстрация использования фиктивной переменной

Мы проиллюстрируем метод использования фиктивных переменных на примере регрессионного анализа основных факторов, влияющих на вес новорожденных младенцев. Если вы думаете, что эта тема не представляет достаточного интереса для экономиста, то ошибаетесь. Типичный экономист, работающий в прикладной сфере, не тратит все свое время на создание макроэкономических моделей. Значительно вероятнее, что он участвует в работе, имеющей более

непосредственное отношение к практике, например, занимается анализом использования ресурсов в какой-либо конкретной сфере. Поскольку в большинстве стран на медицинское обслуживание направляется достаточно большая часть личных и общественных ресурсов, вполне оправдан интерес к нему со стороны экономистов; на эту тему имеется обширная экономическая литература, причем объем ее постоянно растет. Несмотря на то что наибольшим вниманием средств массовой информации обычно пользуется неотложная медицинская помощь, самыми важными с точки зрения затрат являются акушерская помощь, уход за пожилыми и забота о лицах, имеющих психические заболевания.

Так как стоимостное выражение результатов медицинской помощи обычно является весьма спорным предметом, невозможно провести удовлетворительный сравнительный анализ затрат и результатов по большинству видов расходов на медицину. Вместо этого широко используется следующий подход: берется какой-либо показатель успеха (или неудачи), определяются основные факторы, его обусловливающие, и затем делается попытка найти наиболее эффективный путь достижения заданной цели, выраженной этим показателем.

В принципе задача выявления основных факторов, формирующих значения целевого показателя, должна решаться специалистами в области медицинской статистики, а задача наилучшего распределения ресурсов — экономистами; но в здравоохранении, как и в других отраслях прикладной экономики, высокопрофессиональным экономистам часто приходится выходить за непосредственные рамки своей дисциплины и проводить такой статистический анализ, прежде чем перейти к своей основной работе. В области акушерской помощи двумя главными показателями являются младенческая смертность и вес новорожденных. Так как коэффициент смертности новорожденных для большинства стран очень низок, для исследования определяющих его факторов требуются выборки большого объема, и поэтому вес новорожденных во многих случаях является более практичным альтернативным показателем.

Регрессионные зависимости веса новорожденного, о которых в этой главе идет речь, являются побочным продуктом исследования, основная цель которого состоит в выяснении того, оказывает ли предродовая подготовка ощутимое воздействие на результат родов. Во время беременности будущая мать посещает врача для консультаций, а в некоторых странах ей также предлагается посещать занятия по предродовой подготовке, где она может получить знания о немедицинских аспектах беременности и родов.

Чтобы определить в ходе исследования, оказывает ли посещение занятий по предродовой подготовке положительное воздействие на вес младенца при рождении, что рассматривается как показатель результата родов, недостаточно было оценить регрессионную зависимость веса новорожденного от посещения занятий, так как парная регрессия такого типа определенно даст смещенные результаты.

Например, (1) у матерей, которые рожают не в первый раз, появляются обычно младенцы с большим весом, чем у женщин, рожающих впервые; и (2) они не склонны посещать занятия по предродовой подготовке, так как уже прошли через это. Если эта взаимосвязь не принимается во внимание, то результаты исследования могут показать, что посещение занятий по предродовой подготовке оказывает неблагоприятное воздействие на вес новорожденного. По аналогии с этим женщины, которые курят во время беременности, в меньь-

шей степени склонны посещать занятия, чем те, которые не курят. Курение оказывает неблагоприятное воздействие на вес новорожденного ребенка. Если данная сторона не принимается во внимание, это приведет к смещению в сторону завышения оценки влияния предродовой подготовки.

Соответственно было необходимо провести полное исследование вопроса, включающее всесторонний учет социально-экономических, медицинских и физических факторов, влияющих на вес новорожденного, чтобы получить несмещенную оценку воздействия любого из факторов (подробная информация о выборке, использовавшейся в рассматриваемых регрессиях, приводится в работе К. Доугерти и А. Д. Джонса [Dougherty, Jones, 1982]).

Наибольшая часть дисперсии веса новорожденного обусловлена генетической наследственностью ребенка и продолжительностью беременности; таким образом, коэффициент R^2 в регрессиях веса новорожденного младенца всегда является очень низким. Те, кто не имеет достаточного опыта в области регрессионного анализа, стремятся задать желаемый уровень R^2 и считают, что если коэффициент R^2 высок, то уравнение является точным, а если он низок, то данная регрессия оценивалась впустую. Оба вывода неправильны. В рассматриваемом случае курение во время беременности объясняет только очень малую долю всей дисперсии, но тем не менее является значимым фактором. Если предположить, что воздействие всех остальных факторов постоянно, то курение 10 сигарет в день во время беременности снижает вес новорожденного в среднем приблизительно на 80 граммов. Хотя само по себе это, видимо, не столь серьезно, тот факт, что курение оказывает неблагоприятное воздействие на вес новорожденного, вероятно, означает, что оно также оказывает неблагоприятное воздействие на умственное развитие плода, и это имеет большое значение. Зависимость между весом при рождении и курением — тема, вызывающая много дискуссий, которой по понятным причинам уделялось большое внимание в медицинской литературе.

В качестве отправной точки возьмем модель:

$$y = \alpha + \beta x + u, \tag{9.1}$$

где y — вес новорожденного в граммах и x — количество сигарет, выкуриваемых в день будущей матерью во время беременности. Оценив регрессию по выборке, включающей данные о 964 родах (описанной в указанной выше статье), получаем:

$$\hat{y} = 3418 - 7{,}2x; \qquad R^2 = 0{,}012. \tag{9.2}$$
$$\text{(с.о.)} \quad (14) \quad (2{,}1)$$

Это означает, что ребенок, рожденный некурящей матерью, будет иметь при рождении средний вес около 3400 г и что уменьшение веса новорожденного по причине курения составит несколько больше 7 г на каждую сигарету, выкуриваемую в день будущей матерью.

Это только отправная точка. Далее мы будем исследовать воздействие качественного фактора: рожала ли женщина до этого или нет. Это можно смоделировать с помощью двух уравнений:

$$y = \alpha + \beta x + u \tag{9.3}$$

и

$$y = \alpha' + \beta x + u, \tag{9.4}$$

где первое уравнение относится к детям, родившимся у своих матерей первыми (первенцам), а второе — ко всем остальным.

Заметим, что эти два уравнения записаны с одним и тем же коэффициентом при x, но с разными свободными членами. Мы предполагаем, что тот факт, является ли ребенок первенцем или нет, влияет на основной вес, но не на вес, теряемый при каждой выкуриваемой матерью сигарете.

Эквивалентным способом записи модели было бы сохранить уравнение (9.1) для первенцев и записать другое уравнение в виде:

$$y = \alpha + \delta + \beta x + u. \tag{9.5}$$

Основной вес ребенка, не являющегося первенцем (α'), разделен здесь на две составляющие: основной вес ребенка-первенца (α) и дополнительный вес, обусловленный тем, что ребенок родился не первым (δ). Эта модель иллюстрируется на рис. 9.1. Две прямые линии показывают зависимость между весом новорожденного и курением без учета случайного фактора. Они изображены с наклоном вниз, так как на практике коэффициент β отрицателен.

Линия регрессии для ребенка, который родился не первым, такая же, как для первенца, с тем различием, что она сдвинута вверх на величину δ. Нашей целью является оценка этого неизвестного параметра сдвига, и мы получим ее с помощью введения так называемой фиктивной переменной. Перепишем модель в виде:

$$y = \alpha + \delta D + \beta x + u, \tag{9.6}$$

где D — фиктивная переменная, т. е. искусственно введенная переменная, которая принимает значение 0, если наблюдение относится к первенцу, и значение 1, если оно относится к ребенку, родившемуся не первым.

Мы видим, что ситуация определяется тем, что происходит при D, равном нулю или единице. Если ребенок — первенец, то D берется равным нулю и уравнение упрощается до вида (9.3). Если ребенок родился не первым, то D принимается равным единице и уравнение записывается в виде (9.5). Набор данных для иллюстрации сказанного представлен в табл. 9.1.

Данные загружаются в компьютерную программу регрессионного анализа,

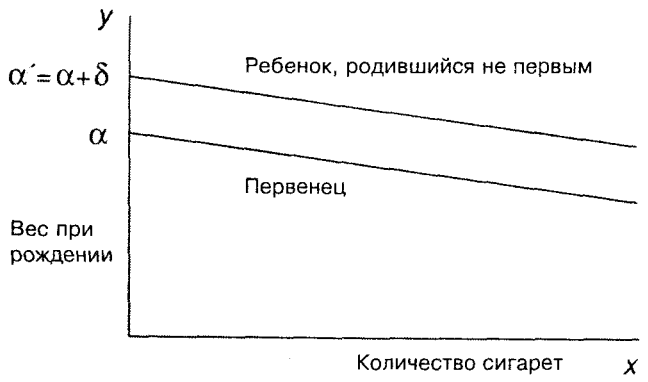


Рис. 9.1. Зависимость веса новорожденного от степени пристрастия будущей матери к курению и от того, является ли ее ребенок первенцем или нет

и для оценивания зависимости y от x и D используется множественная регрессия; D рассматривается точно так же, как обычная переменная, хотя набор ее значений состоит только из нулей и единиц.

Таблица 9.1

Наблюдение	*Первенец?*	*y*	*x*	*D*	*Наблюдение*	*Первенец?*	*y*	*x*	*D*
1	Нет	3520	10	1	11	Нет	3210	29	1
2	Нет	3460	19	1	12	Нет	3290	15	1
3	Нет	3000	16	1	13	Да	3190	3	0
4	Нет	3320	26	1	14	Да	3060	12	0
5	Нет	3540	4	1	15	Да	3270	17	0
6	Нет	3310	14	1	16	Да	3170	14	0
7	Нет	3360	21	1	17	Да	3230	18	0
8	Нет	3650	10	1	18	Да	3700	11	0
9	Нет	3150	22	1	19	Да	3300	14	0
10	Нет	3440	8	1	20	Да	3460	9	0

Результаты оценивания регрессии для наблюдений, представленных в табл. 9.1, таковы:

$$\hat{y} = 3444 + 103D - 11{,}9x; \qquad R^2 = 0{,}19. \tag{9.7}$$
$$\text{(с.о.)} \quad (99) \quad (84) \quad (6{,}3)$$

Параметр сдвига составляет 103 грамма (или приблизительно 4 унции).

Уравнение (9.7) можно переписать в соответствии с (9.3) и (9.4):

$$\hat{y} = 3444 - 11{,}9x \qquad \text{(для первенца);} \tag{9.8}$$

$$\hat{y} = 3547 - 11{,}9x \qquad \text{(для непервенца).} \tag{9.9}$$

Эти линии вместе с точками наблюдений в выборке показаны на рис. 9.2. Оценивание регрессии по реальным данным о 964 родах дало результат:

$$\hat{y} = 3373 + 119D - 7{,}8x; \qquad R^2 = 0{,}032. \tag{9.10}$$
$$\text{(с.о.)} \quad (17) \quad (26) \quad (2{,}1)$$

Стандартные ошибки и проверка гипотез

Стандартные ошибки коэффициентов при фиктивных переменных, рассчитанные с помощью компьютера, так же как и стандартные ошибки других коэффициентов, используются для проверки гипотез и построения доверительных интервалов.

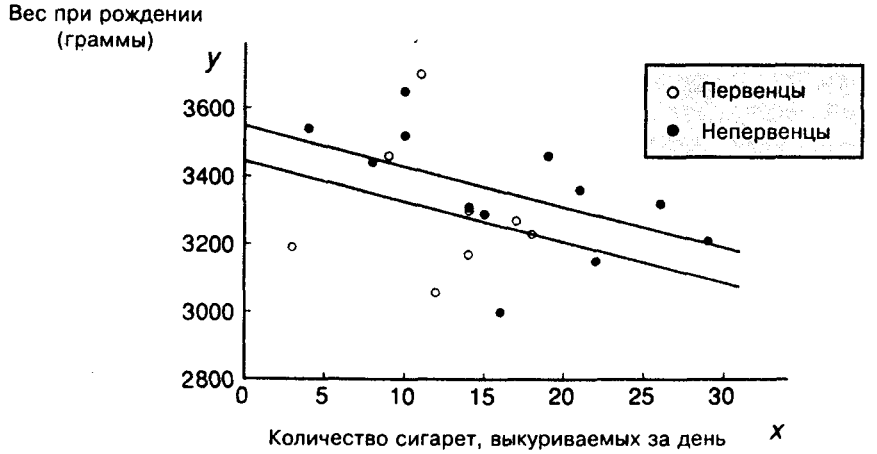


Рис. 9.2. Регрессия, представляющая зависимость веса новорожденного от степени пристрастия будущей матери к курению

Наиболее распространенное их применение состоит в проверке значимости отличия коэффициента от нуля. Она выполняется обычным способом — делением коэффициента на стандартную ошибку для получения t-статистики, которая сравнивается с критическим значением t при заданном уровне значимости. Если t-статистика значима, то из этого следует, что свободные члены для двух категорий наблюдений значимо различаются.

Например, в уравнении (9.7) t-статистика для коэффициента при фиктивной переменной составляет 1,23. Таким образом, коэффициент незначимо отличается от нуля, что означает, что сдвиг линий регрессии для первенцев и детей, родившихся не первыми, не является значимым. Это можно объяснить малым размером выборки. Эффект, вызываемый тем, что ребенок — первенец (или непервенец), проявляется только как тенденция, и он слишком невелик, чтобы можно было выявить его значимость по выборке, содержащей только 20 наблюдений. Если мы рассмотрим регрессию на реальных данных, то увидим, что t-статистика составляет 4,58, а это указывает, что в действительности сдвиг линии регрессии весьма значим.

Пример с временным рядом

В табл. 9.2 можно видеть, что в 1974 г. наблюдалось резкое снижение расходов на автомобили. Имел место нефтяной кризис, и такое снижение было одним из его результатов. Однако впоследствии расходы на автомобили начали снова расти. Следовательно, мы можем выдвинуть гипотезу, что функция спроса в 1974 г. сдвинулась вниз, как показано на рис. 9.3, где y — расходы на автомобили и x — располагаемый личный доход.

Мы можем выразить этот сдвиг математически, введя в уравнение фиктивную переменную D, принимая ее значения равными нулю для 1963–1973 гг. и единице для 1974–1982 гг.:

$$y = \alpha + \delta D + \beta x + u. \tag{9.11}$$

Таблица 9.2

Расходы на автомобили в 1963–1982 гг. (млрд. долл., в постоянных ценах 1972 г.)

1963	18,5	1968	26,5	1973	33,9	1978	34,8
1964	19,7	1969	26,7	1974	25,5	1979	32,9
1965	23,5	1970	22,7	1975	25,4	1980	28,7
1966	23,6	1971	28,0	1976	31,1	1981	29,6
1967	22,2	1972	31,6	1977	34,4	1982	29,8

Источник: тот же, что и в табл. Б.1.

Для периода 1963–1973 гг. при $D = 0$ уравнение принимает вид:

$$y = \alpha + \beta x + u, \tag{9.12}$$

а для периода 1974–1982 гг. при $D = 1$:

$$y = (\alpha + \delta) + \beta x + u. \tag{9.13}$$

Коэффициент δ при фиктивной переменной, конечно, отрицателен. В случае оценивания функции спроса по данным для y и x из табл. Б.1 и значений D, которые представляют собой ряд из 11 нулей, за которыми идут 9 единиц, получаем:

$$\hat{y} = 0{,}57 - 4{,}40D + 0{,}035x; \qquad R^2 = 0{,}69. \tag{9.14}$$
$$\text{(с.о.)}\ (5{,}34)\ \ (2{,}40)\ \ (0{,}008)$$

Это означает, что величина свободного члена в уравнении регрессии для периода 1963–1973 гг., показанная на рис. 9.3, составляет 0,57, а для периода 1974–1982 гг. она равна −3,83.

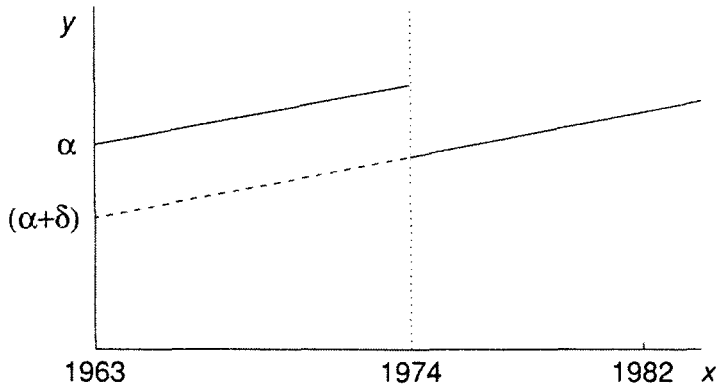


Рис. 9.3. Пример использования фиктивной переменной для описания сдвига в модели с временным рядом

Проверка значимости с помощью t-теста для коэффициента при фиктивной переменной с использованием одностороннего критерия (поскольку мы предвидим, что коэффициент будет отрицательным), показывает, что сдвиг является значимым при уровне значимости в 5%.

Если включить в регрессию также и относительную цену автомобилей, то мы получим:

$$\hat{y} = 18{,}64 - 4{,}47D + 0{,}027x - 11{,}5p; \qquad R^2 = 0{,}69. \tag{9.15}$$
$$\text{(с.о.)}\ (32{,}41)\ \ (2{,}45)\ \ (0{,}016)\ \ (20{,}3)$$

Сдвиг остается значимым на 5-процентном уровне.

В этом частном случае фактически нет реальной потребности в фиктивной переменной. Мы знаем причину сдвига функции спроса после 1973 г., когда увеличилась относительная цена бензина. Если теперь включить в уравнение регрессии также и относительную цену на бензин, обозначив ее *pgas*, то получится уравнение:

$$\hat{y} = -38{,}57 - 1{,}66D + 0{,}065x + 29{,}70p - 13{,}18pgas; \qquad R^2 = 0{,}87. \tag{9.16}$$
$$\text{(с.о.)}\ (25{,}04)\ \ (1{,}75)\ \ (0{,}013)\ \ (16{,}28)\ \ (2{,}89)$$

Коэффициент при фиктивной переменной уже не отличается значимо от нуля. Спецификация в целом улучшилась. Оба коэффициента при x и при *pgas* являются значимыми при уровне значимости в 0,1%, имеют ожидаемые знаки [заметим, что коэффициент при x не был значимым в (9.15) даже при 5-процентном уровне]. Значение коэффициента R^2 выросло с 0,69 до 0,87. Вместе с тем теперь обнаруживается одна странность, заслуживающая дальнейшего внимания: оценка коэффициента при p является положительной, хотя и незначимой при 5-процентном уровне значимости.

Упражнения

9.1. Дайте полную интерпретацию уравнения (9.10).

9.2. Существует закономерность, согласно которой младенцы мужского пола имеют в среднем больший вес при рождении по сравнению с младенцами женского пола. Определяя фиктивную переменную $M = 1$ для мальчиков и $M = 0$ для девочек и используя выборку из 964 родов, получим следующую оценку регрессионной зависимости веса новорожденного от показателя курения и фиктивной переменной M:

$$\hat{y} = 3354 - 119M + 7{,}0x; \qquad R^2 = 0{,}033.$$
$$\text{(с.о.)}\ \ (20)\ \ \ (26)\ \ \ (2{,}1)$$

Дайте полную интерпретацию регрессии и выполните соответствующие статистические проверки.

9.3. Вы исследуете зависимость между расходами на зарубежные поездки и располагаемым личным доходом для Франции, используя ежегодные данные за период 1966–1985 гг. В течение 1982–1983 гг. правительство Франции значительно ограничило нормы использования иностранной валюты для этой цели с тем, чтобы уменьшить дефицит платежного баланса. Объясните, как бы вы использовали фиктивную переменную для оценки эффективности введения этих ограничений.

9.2. Общий случай

В предыдущем примере были только две категории качественной переменной: дети, родившиеся первыми, и дети, родившиеся не первыми. Ввиду высокой значимости коэффициента при фиктивной переменной у нас может появиться желание развить модель и выяснить, влияет ли на вес новорожденного число родов, имевшихся у его матери в прошлом.

Одним из путей такого исследования, конечно, было бы использование модели:

$$y = \alpha + \beta_1 x + \beta_2 z + u, \tag{9.17}$$

где z — число предшествующих родов. Однако эта модель внутренне исходит из того, что вес новорожденного возрастает как линейная функция от z, т. е. с постоянным приращением на каждые дополнительные предшествующие роды. А это в общем-то само по себе неочевидно. По физиологическим причинам было бы естественным предполагать, что вторые или последующие роды будут иметь относительно небольшой дополнительный эффект.

В этой ситуации, возможно, было бы лучше использовать систему фиктивных переменных для более точного изучения влияния количества родов, применяя, например, следующую классификацию состояний: отсутствие родов в прошлом (которое мы впредь будем отмечать как категорию *0*); одни роды в прошлом (категория *1*); двое родов в прошлом (категория *2*); трое или более родов в прошлом (категория *3*). (Выборка из 964 родов не включала достаточного количества примеров с четырьмя или более предшествующими родами, которое могло бы оправдать дальнейшее выделение отдельных категорий.)

Затем мы выбираем одну из этих категорий как эталонную и определяем фиктивные переменные для остальных. Способ выбора эталонной категории будет рассмотрен ниже, но в данном контексте для этого было бы естественно использовать категорию *0*. Мы определяем фиктивные переменные $D1$, $D2$ и $D3$ для других категорий следующим образом:

Категория *0*	$D1 = D2 = D3 = 0;$
Категория *1*	$D1 = 1;\ D2 = D3 = 0;$
Категория *2*	$D2 = 1;\ D1 = D3 = 0;$
Категория *3*	$D3 = 1;\ D1 = D2 = 0.$

Запишем модель в следующем виде:

$$y = \alpha + \beta x + \delta_1 D1 + \delta_2 D2 + \delta_3 D3 + u, \tag{9.18}$$

где δ_1, δ_2 и δ_3 — коэффициенты при фиктивных переменных. Причем δ_1 — разность между весом новорожденных в категориях *1* и *0* при сохранении воздействия x на постоянном уровне, δ_2 — разность между весом новорожденных в категориях *2* и *0*, и δ_3 — разность в весе в категориях *3* и *0*.

В табл. 9.3 показано число предшествующих родов и соответствующие значения фиктивных переменных для первых 20 из рассматриваемых 964 родов.

Оценивая регрессионную зависимость веса новорожденного от степени пристрастия будущей матери к курению и от этих фиктивных переменных (для выборки из всех 964 случаев), получаем:

Таблица 9.3

Случай	Пред. родов*	D1	D2	D3	Случай	Пред. родов	D1	D2	D3
1	1	1	0	0	11	0	0	0	0
2	2	0	1	0	12	0	0	0	0
3	0	0	0	0	13	0	0	0	0
4	1	1	0	0	14	0	0	0	0
5	3	0	0	1	15	0	0	0	0
6	2	0	1	0	16	1	1	0	0
7	0	0	0	0	17	0	0	0	0
8	1	1	0	0	18	0	0	0	0
9	0	0	0	0	19	0	0	0	0
10	1	1	0	0	20	1	1	0	0

*Пред. родов — число предшествующих родов.

$$\hat{y} = 3373 - 7{,}8x + 127D1 + 102D2 + 105D3; \qquad R^2 = 0{,}033. \tag{9.19}$$
(с.о.) (17) (2,1) (30) (49) (61)

Коэффициент при каждой фиктивной переменной представляет собой оценку разницы в весе новорожденного между соответствующей и эталонной категориями при фиксированном уровне воздействия эффекта курения. Отсюда мы заключаем, что новорожденные из категории *1* в среднем имеют вес на 127 г больше по сравнению с новорожденными из категории *0*, новорожденные из категории *2* — на 102 г больше по сравнению с новорожденными из категории *0* и новорожденные из категории *3* — на 105 г больше в сравнении с новорожденными из категории *0*. Результаты подтверждают гипотезу о том, что важным фактором является не число предшествующих родов, а то, рожала мать в прошлом или нет.

Используя определения фиктивных переменных, мы могли бы при желании получить из уравнения (9.19) четыре соотношения, по одному для каждой категории. Например, в случае категории *0* все фиктивные переменные берутся равными нулю, и получается уравнение:

$$\hat{y} = 3373 - 7{,}8x. \tag{9.20}$$

Для категории *1*, где $D1 = 1$, $D2 = D3 = 0$, получаем:

$$\hat{y} = 3373 - 7{,}8x + 127 = 3500 - 7{,}8x. \tag{9.21}$$

Аналогично уравнения для категорий *2* и *3* имеют вид:

$$\hat{y} = 3475 - 7{,}8x; \tag{9.22}$$

$$\hat{y} = 3478 - 7{,}8x. \tag{9.23}$$

Проверка гипотез

Проверка гипотез с помощью t-критерия показывает, что все коэффициенты при фиктивных переменных значимо отличаются от нуля, другими словами, что средний вес новорожденного в каждой из остальных категорий значимо выше, чем в случае, когда женщина рожала впервые. Также может быть интересно рассмотреть, привело ли включение группы фиктивных переменных к значимому повышению объясняющей способности уравнения регрессии. Сумма квадратов остатков без включения фиктивных переменных составила 158,6 млн., а с их включением — 155,3 млн. Как отмечалось в разделе 5.6, соответствующая F-статистика имеет вид:

$$F = \frac{\textit{Улучшение качества уравнения / Использованные степени свободы}}{\textit{Необъясненная дисперсия / Число остающихся степеней свободы}} =$$

$$= \frac{(3{,}3 \times 10^6) / 3}{(155{,}3 \times 10^6) / 959} = 6{,}79. \qquad (9.24)$$

Она распределена с 3 и 959 степенями свободы и превышает критическое значение F, равное 5,42 при уровне значимости в 0,1%.

Выбор эталонной категории

Выбор эталонной категории не оказывает воздействия на сущность уравнений регрессии; но от этого выбора зависит, какие тесты вы сможете выполнить, и это соображение, как правило, должно служить ориентиром. Хотя сам выбор определяет форму представления коэффициентов регрессии, он отражает лишь внешнюю сторону вопроса. Это не оказывает влияния на уравнения, соответствующие отдельным категориям, когда они выводятся из основного уравнения.

Это можно доказать формально, но мы ограничимся иллюстрацией. Предположим, что в примере с весом новорожденных мы выбрали в качестве эталонной категорию *1*, означающую, что ранее мать рожала ровно один раз, и вновь оценим регрессию. Теперь нам надо ввести новую фиктивную переменную ($D0$), которая равна единице для категории *0* и нулю — в остальных случаях. Мы опускаем $D1$, так как фиктивная переменная для эталонной категории не включается. Переменные $D2$ и $D3$ включаются в уравнение с теми же определениями, что и раньше. Результатом построения регрессии является:

$$\hat{y} = 3500 - 7{,}8x + 127D0 - 25D2 - 22D3; \qquad R^2 = 0{,}033. \qquad (9.25)$$
$$\text{(с.о.)} \quad (26) \quad (2{,}1) \quad (30) \quad (52) \quad (64)$$

Так как теперь эталонной является категория *1*, коэффициенты при фиктивных переменных дают оценки добавочного веса младенцев, относящихся к другим категориям, *по сравнению с новорожденными из категории 1.* Коэффициент $D0$ является, конечно, отрицательным, потому что новорожденные в категории *0* обычно имеют меньший вес, чем новорожденные в категории *1*. Коэффициенты при $D2$ и $D3$ невелики и отрицательны, что показывает, что вес

новорожденного в действительности уменьшается при более высоком числе предшествующих родов, но несущественно.

Чтобы получить вариант уравнения для категории *0*, устанавливаем $D0 = 1$, $D2 = D3 = 0$. Для категории *1* все фиктивные переменные принимают значение 0. Для категории *2* переменная $D2 = 1$, $D0 = D3 = 0$. Для категории *3* переменная $D3 = 1$, $D0 = D2 = 0$. Можно проверить, что мы получаем здесь уравнения (9.20)—(9.23), как и раньше.

Интерпретация проверки гипотез для коэффициентов при фиктивных переменных будет, однако, теперь другой. Например, коэффициент при $D2$ уже оценивает разность между весом новорожденных в категориях *2* и *1*, а не между весом младенцев в категориях *2* и *0*.

Таким образом, выбор эталонной категории будет определяться набором проверок гипотез, которые вы хотите провести. В данном случае если вы хотите проверить, был ли вес новорожденных в категории *0* значительно ниже, чем в других категориях, то следует использовать в качестве эталонной категории первоначальный вариант с категорией *0*. Если вы уже знаете, что результат для категории *0* значительно ниже, то, возможно, будете в большей степени заинтересованы в проверке, которая показала бы, увеличился ли (или уменьшился) значимо вес новорожденных в категориях более высоких, чем категория *1*. В этом случае следует использовать в качестве эталонной категории второй вариант с категорией *1*. В уравнении (9.25) t-статистики коэффициентов при $D2$ и $D3$ показывают, что они не отличаются значимо от нуля при уровне значимости в 5%. Отсюда мы делаем вывод, что между весом новорожденных в категории *1* и новорожденных в более высоких категориях значимого различия нет.

Ловушка при применении фиктивных переменных

Что произойдет, если включить в уравнение фиктивную переменную для эталонной категории? Произойдут два явления. Во-первых, если бы было возможно вычислить коэффициенты регрессии, то вы не смогли бы дать им интерпретацию. Коэффициент a является оценкой базового значения постоянного члена в уравнении регрессии, а коэффициенты при фиктивных переменных служат оценками приращения постоянного члена по сравнению с этим базовым уровнем. Теперь, однако, отсутствует то, что является «базой», поэтому интерпретация оказывается несостоятельной. Фактически станет невыполнимой процедура вычисления коэффициентов регрессии. Компьютер просто выдаст сообщение об ошибке или, возможно (в более совершенных регрессионных пакетах), отбросит одну из фиктивных переменных.

Использование сезонных фиктивных переменных

Исследователи, использующие данные временных рядов, в целом предпочитают погодовым данным сведения по кварталам по той простой причине, что за счет этого они получают в 4 раза больше наблюдений в рассматриваемый период. Вместе с тем иногда заметное воздействие на зависимость оказывает сезонный фактор. В этом случае желательно непосредственно принять его во вни-

мание. Если не учитывать это воздействие, то оно вносит свой вклад в случайный член и «шум» в уравнении, в результате чего происходит ненужное снижение эффективности оценок других коэффициентов.

Таблица 9.4

Расходы потребителей на газ и электричество в США (млрд. долл., в постоянных ценах 1972 г.; без сезонной поправки)

1977	I	7,33	1979	I	7,96	1981	I	8,04
	II	4,70		II	5,01		II	5,27
	III	5,10		III	5,05		III	5,51
	IV	5,46		IV	5,59		IV	6,04
1978	I	7,65	1980	I	7,74	1982	I	8,26
	II	4,92		II	5,10		II	5,51
	III	5,15		III	5,67		III	5,41
	IV	5,55		IV	5,92		IV	5,83

Источник: Вычислено на основе табл. 7.11 и 9.2 обзора «Survey of Current Business», July 1982, July 1983.

В табл. 9.4 представлены расходы потребителей на газ и электричество в США в постоянных ценах с I квартала 1977 г. по IV квартал 1982 г. Следует обратить внимание, что для обозначения кварталов года используются римские цифры I–IV. Ряд характеризуется небольшой тенденцией к повышению и сильными сезонными колебаниями. Как и следовало предполагать, расходы такого рода всегда значительно выше зимой, чем летом.

Произвольно возьмем I квартал года в качестве эталонной категории и будем использовать фиктивные переменные для оценки разницы между ним и другими кварталами. Запишем модель как

$$y = \alpha + \beta t + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4 + u, \qquad (9.26)$$

где D_2, D_3 и D_4 — фиктивные переменные, определяемые следующим образом: D_2 равно единице, когда наблюдение относится ко II кварталу, и нулю в остальных случаях; D_3 равно единице в III квартале и нулю в остальных случаях; D_4 равно единице в IV квартале и нулю в остальных случаях.

Величины δ_2, δ_3 и δ_4 — их коэффициенты; они дают численную величину эффекта, вызываемого сменой сезонов. Коэффициент δ_2 показывает дополнительное потребление газа и электричества во II квартале относительно I квартала, связанное со сменой времени года. По аналогии с этим δ_3 и δ_4 показывают соответствующие дополнительные количества в III и IV кварталах относительно I квартала. Все эти «сдвиги» даются относительно I квартала, потому что он выбран в качестве эталонной категории.

Полная совокупность наблюдений за расходами на газ и электричество, данные о времени и фиктивные переменные приведены в табл. 9.5. Оценив регрессионную зависимость расходов от времени и фиктивных переменных, получаем:

$$\hat{y} = 7{,}50 + 0{,}030t - 2{,}78D_2 - 2{,}58D_3 - 2{,}19D_4; \qquad R^2 = 0{,}98. \tag{9.27}$$
$$\text{(с.о.)}\ (0{,}09)\ (0{,}005)\ (0{,}09)\ (0{,}10)\ (0{,}10)$$

Из этого результата мы выводим отдельные уравнения для каждого квартала:

$$\begin{aligned} \hat{y} &= 7{,}50 + 0{,}030t \quad &&\text{(I квартал)} \\ \hat{y} &= 4{,}72 + 0{,}030t \quad &&\text{(II квартал)} \\ \hat{y} &= 4{,}92 + 0{,}030t \quad &&\text{(III квартал)} \\ \hat{y} &= 5{,}31 + 0{,}030t \quad &&\text{(IV квартал)} \end{aligned} \tag{9.28}$$

Таблица 9.5

y	t	D_2	D_3	D_4	y	t	D_2	D_3	D_4
7,33	1	0	0	0	7,74	13	0	0	0
4,70	2	1	0	0	5,10	14	1	0	0
5,10	3	0	1	0	5,67	15	0	1	0
5,46	4	0	0	1	5,92	16	0	0	1
7,65	5	0	0	0	8,04	17	0	0	0
4,92	6	1	0	0	5,27	18	1	0	0
5,15	7	0	1	0	5,51	19	0	1	0
5,55	8	0	0	1	6,04	20	0	0	1
7,96	9	0	0	0	8,26	21	0	0	0
5,01	10	1	0	0	5,51	22	1	0	0
5,05	11	0	1	0	5,41	23	0	1	0
5,59	12	0	0	1	5,83	24	0	0	1

Уравнения (9.28) можно графически проиллюстрировать (рис. 9.4). (Следует отметить, что в этом конкретном случае временной тренд настолько незначителен, что линии оказываются почти горизонтальными.)

При желании можно использовать оцененную регрессию для получения оценки сезонных колебаний в каждом квартале. Выражение (9.28) дает четыре отдельные линии регрессии. Усредняя их, получаем:

$$\hat{y} = 5{,}61 + 0{,}030t. \tag{9.29}$$

Расстояние между отдельной линией регрессии для любого квартала и усредненной линией, которое представлено разностью значений постоянного члена

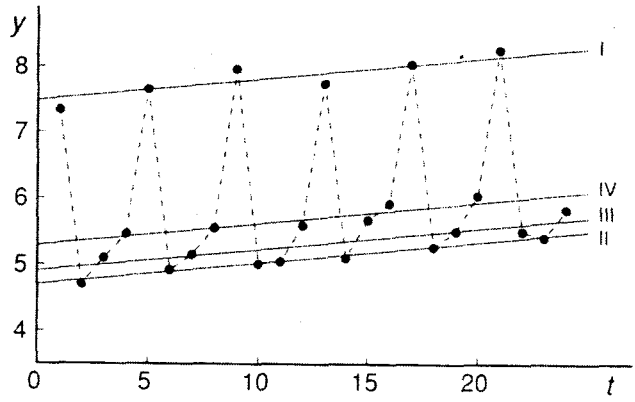


Рис. 9.4. Сезонные колебания, смоделированные при помощи фиктивных переменных

в уравнении регрессии, дает оценку сезонных отклонений в этом квартале. Она составляет:

для I квартала:	$7{,}50 - 5{,}61 = 1{,}89;$
для II квартала:	$4{,}72 - 5{,}61 = -0{,}89;$
для III квартала:	$4{,}92 - 5{,}61 = -0{,}69;$
для IV квартала:	$5{,}31 - 5{,}61 = -0{,}30.$

(*Проверка*: Сумма сезонных отклонений должна равняться нулю, и в данном случае это действительно так.)

Все t-тесты, относящиеся к коэффициентам при фиктивных переменных, показывают высокую значимость, как и F-тест для их совместной объясняющей способности. Суммы квадратов остатков в регрессиях с фиктивными переменными и без них равны соответственно 0,51 и 29,76; таким образом, F-статистика равна $(29{,}25/3)/(0{,}51/19) = 363{,}2$. Критический уровень F с 3 и 19 степенями свободы составляет 5,01 при однопроцентном уровне значимости.

Упражнения

9.4. В нижеследующей таблице приведены поквартальные данные о жилищном строительстве (кроме сельской местности) в США в течение периода 1977–1982 гг. (в миллиардах долларов, в ценах 1972 г.). Оценка регрессионной зависимости этого показателя от временного тренда и сезонных фиктивных переменных, определенных для II, III и IV кварталов, дала следующий результат (в скобках указаны стандартные ошибки):

$$\hat{y} = 13{,}69 + 3{,}02D2 + 4{,}08D3 + 3{,}00D4 - 0{,}31t; \qquad R^2 = 0{,}83.$$
$$\text{(с.о.)}\quad (0{,}65)\quad (0{,}73)\quad (0{,}73)\quad (0{,}73)\quad (0{,}04)$$

Дайте полную интерпретацию регрессии.

9.5. Оцените уравнения регрессии аналогично тому, как это сделано в упражнении 9.4, используя данные для одного из видов потребительских расходов

Жилищное строительство (кроме сельской местности) в США (млрд. долл., в ценах 1972 г.; без сезонной поправки)

1977	I	10,7	1979	I	11,8	1981	I	9,7
	II	15,4		II	14,8		II	11,8
	III	17,2		III	15,8		III	11,2
	IV	14,5		IV	13,7		IV	9,3
1978	I	11,7	1980	I	10,3	1982	I	7,1
	II	15,9		II	10,4		II	9,3
	III	17,1		III	11,6		III	9,3
	IV	14,7		IV	11,8		IV	9,6

Источник: Рассчитано по данным табл. 7.1 и 9.1 обзора «Survey of Current Business», July 1982, July 1983.

(табл. Б.3) и компьютер. Дайте интерпретацию полученных результатов и выполните соответствующие статистические тесты. Оцените регрессию еще раз без фиктивных переменных и выполните F-тест для проверки их совместной значимости.

9.6. Предположим, что вы оцениваете регрессионную зависимость расходов на мороженое от располагаемого личного дохода, используя наблюдения по месяцам. Объясните, как вы введете совокупность фиктивных переменных для оценки сезонных колебаний.

9.3. Множественные совокупности фиктивных переменных

Может потребоваться включение в уравнение регрессии более одной совокупности фиктивных переменных. Это особенно часто встречается при работе со статистическими данными перекрестных выборок, когда могут быть собраны данные по ряду как качественных, так и количественных характеристик. При этом если четко определены рамки работы, то расширение использования в данном случае фиктивных переменных не представляет проблемы.

Мы поясним эту процедуру, используя пример с весом новорожденных. Предположим, что вы желаете исследовать воздействие семейного положения матери на вес при рождении, а также влияние того, рожала ли она раньше. Одинокие матери (матери, живущие на собственные средства независимо от того, состоят ли они формально в браке) как группа подвержены экономическим и социальным лишениям, что вызывает неблагоприятные последствия для течения беременности и развития ребенка. В странах с достаточно развитой системой социального обеспечения обычно прилагаются усилия, направленные на снижение такого неблагоприятного воздействия; в первую очередь, конечно, из соображений гуманности, но также и потому, что такая забота может снизить потребность в лечении после рождения ребенка и, следовательно, приве-

сти к экономии ресурсов. Указанием на успех или неудачу работы социальной системы в этом отношении может служить сравнение веса новорожденных, родившихся у одиноких матерей, с весом новорожденных, родившихся у замужних матерей, предполагая, что воздействие других характеристик является постоянным.

В данном контексте мы введем фиктивную переменную *UM*, которая по определению должна принимать значение 1 для одиноких матерей и 0 — для всех остальных. Мы определим также фиктивную переменную числа родов в прошлом (*D*), равную, как и раньше, единице для матерей, которые рожали в прошлом, и нулю для матерей, которые ранее не рожали.

При этой двойной классификации мы имеем четыре возможных случая с соответствующими комбинациями значений фиктивных переменных:

1. Замужняя мать, первые роды $UM = 0;\ D = 0;$
2. Одинокая мать, первые роды $UM = 1;\ D = 0;$
3. Замужняя мать, не первые роды $UM = 0;\ D = 1;$
4. Одинокая мать, не первые роды $UM = 1;\ D = 1.$

Первый случай по смыслу является основной совместной эталонной категорией. Коэффициент при *UM* будет представлять собой оценку разности веса новорожденных, если мать одинока (мы ожидаем получить отрицательную величину). Коэффициент при *D* будет представлять собой оценку дополнительного веса при рождении, если ребенок не является первенцем. Ребенок в четвертой категории будет подвержен одновременно обоим воздействиям.

Оценивание регрессии с использованием данных о 964 родах дает результат:

$$\hat{y} = 3386 + 109D - 132UM - 7{,}2x; \qquad R^2 = 0{,}040. \tag{9.30}$$
$$\text{(с.о.)}\quad (18)\quad (27)\quad (47)\quad (2{,}1)$$

Используя четыре рассмотренные комбинации значений для *D* и *UM*, можно получить следующие подуравнения:

$$\begin{aligned} &1.\ \hat{y} = 3386 - 7{,}2x; \\ &2.\ \hat{y} = 3254 - 7{,}2x; \\ &3.\ \hat{y} = 3495 - 7{,}2x; \\ &4.\ \hat{y} = 3363 - 7{,}2x. \end{aligned} \tag{9.31}$$

Графическая иллюстрация этих уравнений представлена на рис. 9.5. Мы делаем вывод, что в совокупности, из которой была взята выборка, имеется значимая тенденция, согласно которой вес детей, рождающихся у одиноких матерей, меньше среднего веса.

Эту процедуру можно обобщить. В рассмотренном случае обе классификации качественной переменной имеют только две категории. На практике каждая классификация может иметь несколько категорий, и в этом случае в каждой классификации фиктивная переменная определяется для каждой категории, кроме эталонной категории. Например, если бы мы более подробно рассмотрели по категориям число родов у матери в прошлом в соответствии с классификацией, используемой в разделе 9.2, то аналог регрессии (9.19) после добавления фиктивной переменной для одиноких матерей имел бы вид:

$$\hat{y} = 3386 + 118D1 + 90D2 + 92D3 - 132UM - 7{,}2x; \qquad R^2 = 0{,}041. \qquad (9.32)$$
(с.о.) (18) (30) (49) (61) (47) (2,1)

Число различных классификаций качественной переменной, которая может быть включена в уравнение, не ограничено. Можно еще дальше развертывать уравнение (9.32), вводя дополнительные совокупности фиктивных переменных, относящихся к профессии матери, полу ребенка и т. д., если требуется исследовать потенциальные воздействия этих характеристик.

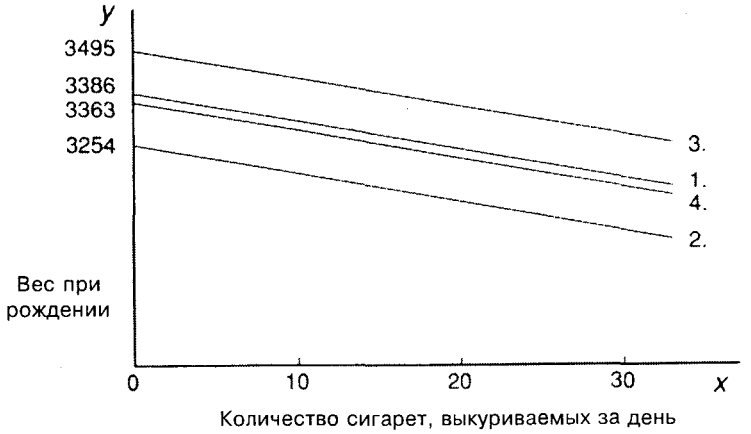


Рис. 9.5. Модель веса при рождении с двумя совокупностями фиктивных переменных

Упражнения

9.7. Дайте полную интерпретацию уравнения (9.32).

9.8. В свете данных, представленных в таблице, прокомментируйте результат оценивания регрессии, включающей только одну фиктивную переменную *UM*, определяемую в тексте:

$$\hat{y} = 3412 - 169UM; \qquad R^2 = 0{,}014.$$
(с.о.) (14) (47)

В частности, сравните этот результат с регрессией, показанной в уравнении (9.30).

	Количество в выборке	*Процент матерей, которые рожают впервые*	*Процент матерей, курящих в период беременности*
Замужние матери	881	58,7	21,6
Одинокие матери	83	80,7	42,2

9.4. Фиктивные переменные для коэффициента наклона

Мы пока предположили, что качественные переменные, введенные в уравнение регрессии, отвечают только за сдвиги в значении постоянного члена в уравнении регрессии. Мы неявно предположили, что наклон линии регрессии одинаков для каждой категории качественных переменных. Это предположение не обязательно верно, и теперь мы рассмотрим, как сделать его менее строгим и проверить, воспользовавшись инструментом, известным как *фиктивная переменная для коэффициента наклона* (иногда называемая также фиктивной переменной взаимодействия).

Для объяснения его использования вернемся к примеру с оцениванием регрессионной зависимости веса при рождении (y) от интенсивности курения матери (x) и фиктивной переменной числа родов в прошлом ($D = 1$, если мать рожала раньше; $D = 0$, если мать раньше не рожала):

$$y = \alpha + \delta D + \beta x + u. \tag{9.6}$$

В этой формулировке модели мы предполагаем, что воздействие курения матери на вес новорожденного одинаково, независимо от того, рожала ли мать раньше.

Предположим, что теперь мы добавим в уравнение член γDx — произведение D и x с коэффициентом γ:

$$y = \alpha + \delta D + \beta x + \gamma Dx + u. \tag{9.33}$$

Это можно переписать как

$$y = \alpha + \delta D + (\beta + \gamma D)x + u. \tag{9.34}$$

Если $D = 0$, то коэффициент при x, как и раньше, равен β.
Если $D = 1$, то коэффициент приобретает вид $(\beta + \gamma)$. Поэтому величина γ может рассматриваться как разность между коэффициентом при показателе интенсивности курения для матерей, которые рожали раньше, и коэффициентом при показателе интенсивности курения для матерей, которые раньше не рожали.

Коэффициент γ можно оценить, используя уравнение (9.33), где y связан регрессионной зависимостью с D, x и Dx; показатель Dx, представляющий собой фиктивную переменную для коэффициента наклона, рассматривается как третья и отдельная объясняющая переменная. В табл. 9.6 показано, как вычисляется переменная Dx по 20 наблюдениям, приведенным в табл. 9.1.

Оценивание регрессии по данным выборки о 964 родах дает результат:

$$\hat{y} = 3363 + 143D - 4{,}0x - 8{,}1Dx; \qquad R^2 = 0{,}036. \tag{9.35}$$
$$\text{(с.о.)}\ (18)\quad (29)\quad (2{,}8)\quad (4{,}1)$$

Положив $D = 0$ или $D = 1$, можно вывести два частных соотношения:

$$\hat{y} = 3363 - 4{,}0x \text{ (для первенцев);} \tag{9.36}$$

$$\hat{y} = 3506 - 12{,}1x \text{ (для детей, рожденных не первыми).} \tag{9.37}$$

Результат оценивания регрессии показывает, что снижение веса новорожденного, связанное с курением матери в период беременности, значительно

Таблица 9.6

Наблюдение	*Первенец?*	*y*	*x*	*D*	*Dx*	*Наблюдение*	*Первенец?*	*y*	*x*	*D*	*Dx*
1	Нет	3520	10	1	10	11	Нет	3210	29	1	29
2	Нет	3460	19	1	19	12	Нет	3290	15	1	15
3	Нет	3000	16	1	16	13	Да	3190	3	0	0
4	Нет	3320	26	1	26	14	Да	3060	12	0	0
5	Нет	3540	4	1	4	15	Да	3270	17	0	0
6	Нет	3310	14	1	14	16	Да	3170	14	0	0
7	Нет	3360	21	1	21	17	Да	3230	18	0	0
8	Нет	3650	10	1	10	18	Да	3700	11	0	0
9	Нет	3150	22	1	22	19	Да	3300	14	0	0
10	Нет	3440	8	1	8	20	Да	3460	9	0	0

больше для матерей, которые рожали раньше, чем для матерей, которые раньше не рожали (12,1 г на каждую сигарету в день против 4,0 г), и что различие значимо при уровне значимости в 5%.

Взаимодействие между фиктивными переменными

Мы до сих пор предполагали, что воздействия качественных характеристик на зависимую переменную являются независимыми друг от друга. Например, в регрессии (9.30) предполагалось, что различие в весе при рождении для детей, родившихся у замужних и одиноких матерей, не зависит от того, рожала ли мать раньше, и наоборот. Мы можем сделать это предположение менее строгим за счет ввода *фиктивных переменных взаимодействия*, которые определяются по аналогии с фиктивными переменными для коэффициента наклона и имеют такое же назначение.

В рассматриваемом случае мы могли бы ввести фиктивную переменную взаимодействия (*UMD*), которая определяется как произведение *UM* и *D* и которая, следовательно, равна единице для одиноких матерей, рожавших раньше, и равна нулю для трех других комбинаций. Модель регрессии имеет вид:

$$y = \alpha + \delta D + \gamma UM + \lambda UMD + \beta x + u, \tag{9.38}$$

и ее можно переписать либо как

$$y = \alpha + (\delta + \lambda UM)D + \gamma UM + \beta x + u, \tag{9.39}$$

либо как

$$y = \alpha + \delta D + (\gamma + \lambda D)UM + \beta x + u. \tag{9.40}$$

Поэтому коэффициент λ можно по выбору (оба альтернативных варианта эк-

вивалентны) рассматривать либо как прирост коэффициента при фиктивной переменной числа предшествующих родов, если мать является одинокой, либо как прирост коэффициента для одиноких матерей, если мать рожала раньше.

Оценивание регрессии с использованием данных о 964 родах дает следующий результат:

$$\hat{y} = 3{,}385 + 113D - 117UM - 72UMD - 7{,}3x; \qquad R^2 = 0{,}041. \qquad (9.41)$$
$$\text{(с.о.)} \quad (18) \quad (28) \quad (52) \quad (115) \quad (2{,}1)$$

Мы видим, что коэффициент при фиктивной переменной взаимодействия значимо не отличается от нуля при уровне значимости в 5%, и делаем вывод, что может не быть взаимодействия между переменной числа родов в прошлом и переменной для одиноких матерей. Однако следует отметить, что в выборке было только 16 одиноких матерей, которые рожали не в первый раз, и, следовательно, коэффициент при UMD имеет очень большую стандартную ошибку. Этот пример дает предупреждение о том, что даже если выборка большая, но имеется несколько совокупностей фиктивных переменных, то число наблюдений в отдельных подкатегориях может легко оказаться очень малым, и, следовательно, проведение удовлетворительных проверок гипотез может быть затруднено.

Упражнения

9.9. При использовании выборки, включающей данные о 964 родах, оценена регрессионная зависимость веса новорожденных (y) от интенсивности курения матери (x), фиктивной переменной (D), характеризующейся числом предыдущих родов, фиктивной переменной (M) пола ребенка (определенной как в упражнении 9.2) и фиктивной переменной для коэффициента наклона (Mx), определяемой как произведение M и x (в скобках указаны стандартные ошибки):

$$\hat{y} = 3312 + 124D + 108M - 10{,}5x + 5{,}7Mx; \qquad R^2 = 0{,}057.$$
$$(23) \quad (26) \quad (28) \quad (2{,}9) \quad (4{,}1)$$

Прокомментируйте этот результат.

9.10. Та же самая регрессия повторно оценена с включением фиктивной переменной взаимодействия (DM), определяемой как произведение D и M (в скобках указаны стандартные ошибки):

$$\hat{y} = 3304 + 144D + 123M - 39DM - 10{,}6x + 5{,}9x; \qquad R^2 = 0{,}058.$$
$$(26) \quad (38) \quad (35) \quad (53) \quad (2{,}9) \quad (4{,}1)$$

Прокомментируйте этот результат.

9.5. Тест Чоу

Иногда выборка наблюдений состоит из двух или более подвыборок, и трудно установить, следует ли оценивать одну объединенную регрессию или отдельные регрессии для каждой подвыборки. На практике проблема выбора стоит

обычно не столь жестко, поскольку могут быть некоторые возможности объединения подвыборок при использовании соответствующих фиктивных переменных и фиктивных переменных для коэффициента наклона, чтобы сделать менее строгим предположение о том, что все коэффициенты должны быть одинаковыми для каждой подвыборки. К этому вопросу мы еще вернемся.

Предположим, что имеется выборка, состоящая из двух подвыборок, и что возникает вопрос, следует ли объединить их для оценивания общей регрессии P или оценить отдельные регрессии A и B. Обозначим суммы квадратов остатков для регрессий подвыборок U_A и U_B. Пусть U_A^P и U_B^P — суммы квадратов остатков в объединенной регрессии для наблюдений, относящихся к двум рассматриваемым подвыборкам. Так как отдельные регрессии для подвыборок должны соответствовать наблюдениям по меньшей мере так же хорошо, если не лучше, чем объединенная регрессия, то $U_A \leq U_A^P$ и $U_B \leq U_B^P$. Следовательно, $(U_A + U_B) \leq U_P$, где общая сумма квадратов остатков в объединенной регрессии U_P равна сумме U_A^P и U_B^P.

Это поясняется на рис. 9.6. Предположим, что имеются данные временного ряда по двум переменным и что в период выборки произошло структурное изменение, разделяющее наблюдения на подвыборки A и B. На рис. 9.6Б регрессии для подвыборок обеспечивают вполне адекватное соответствие данным, обусловливая низкие значения U_A и U_B. Если бы требовалось оценить объединенную регрессию, как на рис. 9.6А, то остатки в обеих подвыборках в целом были бы значительно больше.

Равенство между U_P и $(U_A + U_B)$ будет иметь место только при совпадении коэффициентов регрессии для объединенной регрессии и регрессий подвыборок. В общем случае при разделении выборки будет наблюдаться улучшение качества уравнения, что можно представить как $(U_P - U_A - U_B)$. Это имеет свою цену: используются $(k + 1)$ дополнительных степеней свободы, так как вместо $(k + 1)$ параметров для одной объединенной регрессии мы теперь должны оценить в сумме $(2k + 2)$ параметров (k — число объясняющих переменных, единица соответствует постоянному члену). После разделения выборки, однако, остается необъясненная сумма квадратов остатков $(U_A + U_B)$ и, кроме того, $(n - 2k - 2)$ степеней свободы.

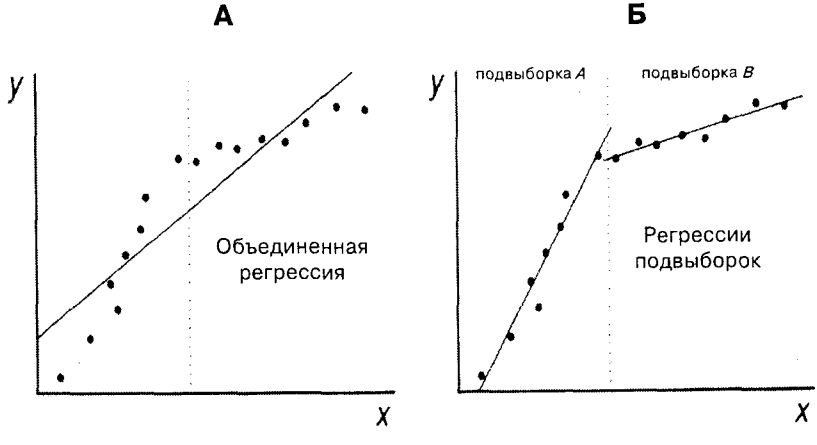


Рис. 9.6. Регрессии, оцениваемые для теста Чоу

Теперь мы можем определить, является ли значимым улучшение качества уравнения после разделения выборки. Для этого используется F-статистика:

$$\frac{\textit{Улучшение качества уравнения / Использованные степени свободы}}{\textit{Необъясненная дисперсия / Число остающихся степеней свободы}} = \frac{(U_P - U_A - U_B)/(k+1)}{(U_A + U_B)/(n-2k-2)}, \tag{9.42}$$

которая распределена с $(k+1)$ и $(n-2k-2)$ степенями свободы.

Теперь, например, давайте вернемся к случаю парной регрессионной зависимости веса новорожденных от интенсивности курения их матерей, и пусть мы еще не решили, следует ли объединять подвыборки, включающие 584 матери, которые ранее не рожали, и 380 матерей, которые ранее рожали. Оценивание объединенной регрессии и регрессий для подвыборок дает результаты, показанные в таблице.

Выборка	*Оцененное уравнение*	R^2	*Сумма квадратов остатков*	
Объединенная выборка	$\hat{y} = 3418 - 7{,}2x$ (с.о.) (143) (2,1)	0,012	$158{,}6 \times 10^6$	(9.43)
Первенцы	$\hat{y} = 3363 - 4{,}0x$ (с.о.) (18) (2,8)	0,004	$91{,}2 \times 10^6$	(9.44)
Непервенцы	$\hat{y} = 3506 - 12{,}1x$ (с.о.) (23) (3,1)	0,039	$63{,}5 \times 10^6$	(9.45)

Соответствующая F-статистика, следовательно, равна:

$$F = \frac{(158{,}6 - 91{,}2 - 63{,}5)/2}{(91{,}2 + 63{,}5)/960} = 12{,}1. \tag{9.46}$$

Критическое значение F с 2 и 960 степенями свободы составляет 6,91 (при уровне значимости в 0,1%), поэтому мы делаем вывод, что не следует оценивать объединенную регрессию.

Регрессии для подвыборок идентичны регрессиям, представленным соотношениями (9.36) и (9.37), и это не простое совпадение. В основной регрессии (9.35) составляющая, не связанная с фиктивной переменной, включает постоянный член и показатель зависимости от интенсивности курения. К этому добавляются фиктивная переменная, позволяющая различать значения постоянного члена для первенцев и детей, родившихся не первыми, и фиктивная переменная для коэффициента наклона, также позволяющая различать коэффициенты при показателе интенсивности курения для двух рассматриваемых подвыборок. Следовательно, в (9.35) мы не задаем заранее какой-либо коэффициент одинаковым для обеих подвыборок и, таким образом, получаем такие же оценки коэффициентов, как и в отдельных регрессиях для подвыборок.

Рассматривая лишь соотношение (9.35), мы можем проверить, оправдана ли

эта гибкость, выяснив, вносят ли указанные фиктивные переменные как группа значимый вклад в объясняющую способность уравнения. Сумма квадратов остатков, если фиктивные переменные не включены в уравнение, составляет $158{,}6 \times 10^6$, а когда они включены в уравнение, эта сумма равна $154{,}7 \times 10^6$. Следовательно, F-статистика для проверки объясняющей способности фиктивных переменных как группы имеет вид:

$$F = \frac{(158{,}6 - 154{,}7)/2}{154{,}7/960} = 12{,}1, \tag{9.47}$$

т. е. она в точности такая же, как в тесте Чоу.

Можно показать, что это общий результат. Выбор между использованием рассмотренной процедуры теста Чоу или оцениванием сложной регрессии с фиктивными переменными на основе соотношения (9.35) будет зависеть от целей, которые ставит перед собой исследователь. Тест Чоу выполняется быстрее, и он достаточен, если требуется только установить, что зависимости в подвыборках в некоторой степени различаются. Оценивание регрессии с фиктивными переменными более информативно в том отношении, что оно позволяет выполнять t-тесты с рассмотрением вклада каждой фиктивной переменной, а также всей группы в целом и может привести к компромиссу, в котором исследователь предполагает, что некоторые коэффициенты одинаковы в обеих подвыборках, и использует фиктивные переменные для дифференциации значений остальных коэффициентов.

Упражнения

9.11. Выполните тест Чоу, чтобы определить, имел ли место структурный разрыв зависимости расходов на автомобили от располагаемого личного дохода в 1973 г., используя данные табл. 9.2.

9.12. Исследователь, интересующийся воздействием особенностей национальной культуры на структуру потребления, предлагает 20 малайским, 20 китайским, 20 индийским и 20 другим семьям, живущим в Куала-Лумпуре, подробно записывать свои расходы на продукты питания в течение одного года. Кратко опишите преимущества и недостатки оценивания одной функции спроса, описанной уравнением с фиктивными переменными, для всех 80 семей в сравнении с оцениванием четырех отдельных уравнений для различных этнических групп.

Приложение 9.1

Качественные зависимые переменные

Может случиться так, что переменная, детерминанты которой требуется исследовать, является качественной по своему характеру. Например, в нашем исследовании в области акушерства можно рассмотреть вопрос оценки факторов, приводящих в критических обстоятельствах к необходимости родоразреше-

ния путем кесарева сечения. Наша цель заключается в уменьшении частоты проведения такой операции, что важно само по себе и для снижения расходов на специальное оборудование для ухода за младенцами и т. д.

Упрощенный способ рассмотрения этой проблемы заключается в определении зависимой переменной *emsec* как фиктивной переменной и в оценивании регрессии обычным способом. Например, мы можем считать *emsec* равной единице, если родоразрешение проводилось путем кесарева сечения, и равной нулю, когда роды были нормальными. Используя данные о 964 родах, мы получаем следующий результат:

$$\widehat{emsec} = 0{,}08 - 0{,}08D + 0{,}01old + 0{,}07short + 0{,}05heavy - \\ - 0{,}02class + 0{,}01UM + 0{,}0018\,x, \qquad (9.48)$$

где D — фиктивная переменная числа родов в прошлом (значение 1 — если мать рожала раньше, значение 0 — в других случаях); *old* — фиктивная переменная возраста (1 — когда матери 36 или более лет, 0 — в других случаях); *short* — фиктивная переменная роста матери (1 — если мать находится в низшем квинтиле по росту, т. е. имеет рост 157 см или меньше, 0 — в других случаях); *heavy* — фиктивная переменная веса матери (1 — если мать относится к верхнему квинтилю по весу, т. е. имеет вес 68 кг или больше, 0 — в других случаях); *class* — фиктивная переменная посещения занятий по предродовой подготовке (1 — если мать регулярно посещала эти занятия, 0 — в других случаях); *UM* — фиктивная переменная семейного положения (1 — если мать является одинокой, 0 — в противном случае); x — количество сигарет, выкуриваемых в день в период беременности.

Последние три переменные представляют интерес для социальной политики; остальные включены в уравнение, потому что, как известно, они имеют отношение к частоте проведения операции кесарева сечения, и если они не будут включены, это может привести к смещению оценок коэффициентов регрессии.

Прогнозируемое значение *emsec* для любого наблюдения показывает вероятность родоразрешения путем кесарева сечения, если даны значения параметров в правой части уравнения. Коэффициент при каждой переменной увеличивает вероятность кесарева сечения для матери с соответствующим параметром. Например, эта вероятность на 8% ниже для матерей, которые рожали ранее, по сравнению с матерями, которые ранее не рожали.

Недостатки линейной вероятностной модели, как известно, связаны с тем, что ее случайный член не удовлетворяет обычным предположениям. В частности, он не распределен нормально, поэтому нельзя выполнить обычную проверку значимости. Кроме того, он может привести к появлению прогнозируемых значений зависимой переменной больше единицы или меньше нуля, что невозможно.

Для преодоления этих трудностей разработано несколько статистических методов, аналогичных методам построения линейной вероятностной модели, но основанных на других принципах. Возможно, наиболее широко известным из них является логит-анализ, основанный на методе максимального правдоподобия. Рассмотрение этого метода выходит за рамки данной книги, и дос-

таточно отметить, что его возможное использование на практике во многом совпадает с практическим применением регрессионного анализа.

В рассматриваемом примере логит-анализ дает следующий результат (в скобках приведены t-статистики):

$$\begin{array}{l} \widehat{emsec} = \text{Константа} - \underset{(t)}{0{,}11}D + \underset{(-4{,}61)}{0{,}11}old + \underset{(3{,}46)}{0{,}05}short + \underset{(2{,}45)}{0{,}05}heavy - \underset{(2{,}27)}{} \\ - \underset{(-1{,}09)}{0{,}02}class + \underset{(0{,}33)}{0{,}01}UM + \underset{(1{,}26)}{0{,}0025}x \end{array} \qquad (9.49)$$

(постоянный член не был вычислен при помощи использованного алгоритма). Уравнение показывает значимые воздействия первых четырех переменных, как это и ожидалось, но не показывает значимого влияния социальных переменных.

10

МОДЕЛИРОВАНИЕ ДИНАМИЧЕСКИХ ПРОЦЕССОВ

Многие экономические процессы имеют долговременный характер, поэтому в эконометрическом моделировании необходимо учитывать временное измерение. В данной главе рассматриваются некоторые принятые подходы к решению этой проблемы. В ней также показано, как регрессионные модели могут использоваться для построения прогнозов и как могут быть оценены их прогнозные свойства.

10.1. Введение

В разделе 6.7 нами был сделан первый шаг к анализу динамического аспекта эконометрической модели, когда мы ввели понятие лаговой переменной и рассматривали вероятность, например, того, что объем затрат на некоторый товар определяется не текущим доходом и ценой этого товара, а доходом и ценой в предыдущий период времени или доходом и ценой за два прошедших периода. Что касается расходов на жилье, то результаты построения регрессионной зависимости этих расходов от текущих доходов и цены, от доходов и цены в прошедшем периоде и от значений этих переменных за два прошедших периода оказались следующими:

$$\underset{\text{(с.о.)}}{\widehat{\log}\, y_t} = \underset{(1{,}75)}{-1{,}60} + \underset{(0{,}05)}{1{,}18} \log x_t - \underset{(0{,}31)}{0{,}34} \log p_t; \qquad R^2 = 0{,}992; \tag{10.1}$$

$$\underset{\text{(с.о.)}}{\widehat{\log}\, y_t} = \underset{(1{,}75)}{0{,}42} + \underset{(0{,}05)}{1{,}10} \log x_{t-1} - \underset{(0{,}31)}{0{,}66} \log p_{t-1}; \qquad R^2 = 0{,}995; \tag{10.2}$$

$$\underset{\text{(с.о.)}}{\widehat{\log}\, y_t} = \underset{(1{,}77)}{0{,}95} + \underset{(0{,}05)}{1{,}08} \log x_{t-2} - \underset{(0{,}31)}{0{,}72} \log p_{t-2}; \qquad R^2 = 0{,}995. \tag{10.3}$$

Продолжая начатое исследование, можно выдвинуть предположение, что расходы на жилье частично зависят от текущих значений дохода и цены, а частично — от их значений в прошлом году, и построить уравнение регрессионной зависимости $\log y_t$ от $\log x_t$ и $\log x_{t-1}$, а также от $\log p_t$ и $\log p_{t-1}$:

$$\underset{\text{(с.о.)}}{\widehat{\log}\, y_t} = \underset{(1{,}55)}{0{,}27} + \underset{(0{,}29)}{0{,}22} \log x_t + \underset{(0{,}30)}{0{,}90} \log x_{t-1} +$$

$$+ \underset{(0,36)}{0,98} \log p_t - \underset{(0,39)}{1,51} \log p_{t-1}; \qquad R^2 = 0,995. \tag{10.4}$$

Для полной уверенности можно учесть также $\log x_{t-2}$ и $\log p_{t-2}$:

$$\widehat{\log y_t} = \underset{(\text{с. о.})(1,88)}{1,00} + \underset{(0,29)}{0,28} \log x_t + \underset{(0,47)}{0,53} \log x_{t-1} + \underset{(0,34)}{0,27} \log x_{t-2} +$$

$$+ \underset{(0,56)}{0,24} \log p_t - \underset{(0,97)}{0,01} \log p_{t-1} - \underset{(0,57)}{0,98} \log p_{t-2}; \qquad R^2 = 0,997. \tag{10.5}$$

Анализируя полученные результаты, можно заметить два обстоятельства, вызывающих беспокойство. Во-первых, между уравнениями (10.1), (10.2) и (10.3) нет особого выбора. Значения эластичности затрат по доходу почти одинаковы в каждом случае, значения эластичности затрат по цене выше в лаговых уравнениях и значимо отличаются от нуля при 5-процентном уровне значимости при односторонней проверке; стандартные отклонения и коэффициенты R^2 почти одинаковы во всех трех уравнениях. За исключением более легко интерпретируемых значений эластичности затрат по цене в лаговых уравнениях, у нас нет никаких оснований предпочесть какое-либо одно уравнение двум другим.

Во-вторых, уравнения (10.4) и (10.5) уступают всем трем предыдущим. Коэффициенты в этих уравнениях нестабильны в том смысле, что их значения существенно различаются при изменении спецификации, и их стандартные отклонения значительно выше, чем в предыдущих уравнениях. Полученные результаты иллюстрируют проблему мультиколлинеарности. Очевидно, значения $\log x_t$, $\log x_{t-1}$ и $\log x_{t-2}$ тесно коррелированы, поскольку они представляют один и тот же набор наблюдений с лагом в один или два периода (см. табл. 6.9). Значения $\log p_t$, $\log p_{t-1}$ и $\log p_{t-2}$ также тесно коррелированы. Если вы используете текущие и лаговые значения в качестве объясняющих переменных, то неудивительно, что коэффициенты при них выглядят несколько странно.

Для оценки лаговой структуры зависимостей было разработано несколько подходов, позволяющих ограничить число объясняющих переменных в уравнении регрессии с целью избежать появления проблемы мультиколлинеарности или по крайней мере минимизировать ее эффект. Мы рассмотрим два широко известных подхода: распределение Койка и лаги Алмон.

Упражнение

10.1. Дайте экономическую интерпретацию коэффициентов при $\log x_t$, $\log x_{t-1}$ и $\log x_{t-2}$ в уравнении (10.5).

10.2. Распределение Койка

В распределении Койка (Koyck, 1954) делается простое предположение, что коэффициенты (известные также как «веса») при лаговых значениях объясняющей переменной убывают в геометрической прогрессии. Если имеется единственная объясняющая переменная, то модель принимает вид:

$$y_t = \alpha + \beta x_t + \beta\delta x_{t-1} + \beta\delta^2 x_{t-2} + \beta\delta^3 x_{t-3} + \ldots + u_t, \qquad (10.6)$$

где значение δ находится в границах от −1 до 1. Во многих приложениях предполагается, что оно лежит между 0 и 1.

В данной зависимости имеются всего три параметра: α, β и δ. Для их оценки вам *не нужно* оценивать уравнение регрессионной зависимости y_t от x_t, x_{t-1}, x_{t-2} и т. д. В этом случае, во-первых, наверняка возникла бы серьезная проблема мультиколлинеарности. Во-вторых, из полученных оценок не удалось бы вывести значения β и δ. Здесь можно получить одно значение β с помощью коэффициента при x_t и другое, совершенно иное, возведя в квадрат коэффициент при x_{t-1} и разделив его на коэффициент при x_{t-2} или возведя в квадрат коэффициент при x_{t-2} и разделив его на коэффициент при x_{t-4}. Точно так же существует много различных и противоречивых способов получения оценки δ.

Однако можно довольно легко избежать как этой проблемы, так и проблемы мультиколлинеарности. Один эффективный способ заключается в применении нелинейного метода наименьших квадратов. Вы начинаете с задания границ возможных значений δ и рассматриваете все возможные значения внутри этих пределов с достаточно малым шагом. Например, пределы изменения могут быть от 0 до 1, и вы рассматриваете все значения 0,00, 0,01, 0,02 и т. д., увеличивая их каждый раз на 0,01. Чем меньше шаг, тем более точными будут полученные результаты, но тем больше времени займут расчеты. Теперь, когда компьютеры стали такими мощными и дешевыми, вы, как правило, сможете достичь любой желаемой точности. Для каждого значения δ рассчитывается

$$z_t = x_t + \delta x_{t-1} + \delta^2 x_{t-2} + \delta^3 x_{t-3} + \ldots + \delta^p x_{t-p}, \qquad (10.7)$$

с таким значением p, при котором дальнейшие лаговые значения x не оказывают существенного воздействия на z. Затем оценивается уравнение регрессии

$$y_t = \alpha + \beta z_t + u_t. \qquad (10.8)$$

Вы проделываете эти расчеты для всех значений δ и выбираете такое значение δ, которое обеспечивает наибольший коэффициент R^2 при оценке уравнения (10.8). В качестве оценок α и β выбираются их оценки в этом уравнении. Уравнения (10.7) и (10.8) в совокупности, конечно же, эквивалентны уравнению (10.6).

Другой метод использует так называемое преобразование Койка. Если выражение (10.6) выполняется для периода t, то оно также выполняется для периода $t-1$:

$$y_{t-1} = \alpha + \beta x_{t-1} + \beta\delta x_{t-2} + \beta\delta^2 x_{t-3} + \ldots + u_{t-1}. \qquad (10.9)$$

Умножив обе части этого уравнения на δ и вычтя их из уравнения (10.6), вы получите:

$$y_t - \delta y_{t-1} = \alpha(1-\delta) + \beta x_t + u_t - \delta u_{t-1}, \qquad (10.10)$$

где уже отсутствуют лаговые значения x. Как следствие имеем:

$$y_t = \alpha(1-\delta) + \beta x_t + \delta y_{t-1} + u_t - \delta u_{t-1}. \qquad (10.11)$$

Эта форма позволяет анализировать кратко- и долгосрочные динамические свойства модели. В краткосрочном аспекте (в текущем периоде) значение y_{t-1} нужно рассматривать как фиксированное, и воздействие x на y отражается ко-

эффициентом β. В долгосрочном периоде (не учитывая случайный член), если x_t стремится к некоторому своему равновесному значению $\bar{x}$, y_t и y_{t-1} также будут стремиться к равновесному уровню $\bar{y}$, определяемому как

$$\bar{y} = \alpha(1-\delta) + \beta\bar{x} + \delta\bar{y}, \tag{10.12}$$

из которого следует:

$$\bar{y} = \alpha + \frac{\beta}{(1-\delta)}\bar{x}. \tag{10.13}$$

Итак, долгосрочное воздействие x на y отражается коэффициентом $\beta/(1-\delta)$. Если δ находится в границах от 0 до 1, то этот коэффициент превысит β, т. е. долгосрочное воздействие оказывается сильнее краткосрочного.

Модель с преобразованием Койка привлекательна с практической точки зрения, поскольку оценивание парной регрессии с помощью МНК позволяет получить оценки α, β и δ (оценка α получается делением свободного члена на разность единицы и коэффициента при y_{t-1}). Разумеется, это требует гораздо меньших усилий, чем описанный ранее поиск с помощью перебора, но здесь, к сожалению, возникает серьезная эконометрическая проблема, которая делает этот метод менее привлекательным, — нарушение четвертого условия Гаусса—Маркова. Одна из объясняющих переменных y_{t-1} в уравнении (10.11) частично зависит от u_{t-1}. Поэтому она коррелирует с одной из составляющих случайного члена $-\delta u_{t-1}$ в уравнении (10.11). В итоге оценки, полученные с помощью МНК, оказываются смещенными и несостоятельными. В таком случае не остается иного выбора, кроме использования первого из подходов — нелинейного метода на базе уравнения (10.7) и (10.8). Теперь мы рассмотрим две хорошо известные динамические модели, обе относящиеся к семейству моделей Койка, хотя на первый взгляд это может показаться неочевидным.

10.3. Частичная корректировка

В модели частичной корректировки предполагается, что поведенческое уравнение определяет не фактическое значение зависимой переменной y_t, а ее *желаемый* (или «целевой») уровень y_t^*:

$$y_t^* = \alpha + \beta x_t + u_t. \tag{10.14}$$

Предполагается также, что фактическое приращение зависимой переменной $y_t - y_{t-1}$ пропорционально разнице между ее желаемым уровнем и значением в предыдущий период, то есть $y_t^* - y_{t-1}$:

$$y_t - y_{t-1} = \lambda(y_t^* - y_{t-1}) + v_t \qquad (0 \le \lambda \le 1), \tag{10.15}$$

где v_t — случайный член. Это выражение может быть переписано как

$$y_t = \lambda y_t^* + (1-\lambda) y_{t-1} + v_t \qquad (0 \le \lambda \le 1), \tag{10.16}$$

откуда видно, что y_t получается как взвешенное среднее желаемого уровня и фактического значения этой переменной в предыдущем периоде. Чем больше

значение λ, тем быстрее происходит процесс корректировки. Если значение $\lambda = 1$, то y_t равно y_t^*, и полная корректировка происходит за один период. В другом крайнем случае, когда значение $\lambda = 0$, корректировка y_t не происходит совсем.

Подставив выражение (10.14) в формулу (10.16), можно получить:

$$y_t = \alpha\lambda + \beta\lambda x_t + (1 - \lambda)\, y_{t-1} + v_t + \lambda u_t. \qquad (10.17)$$

Как следствие, параметры α, β и λ поведенческой модели (10.14) и (10.15) могут быть оценены с помощью построения уравнения регрессионной зависимости y_t от x_t и y_{t-1}. Коэффициент при y_{t-1} дает оценку $(1 - \lambda)$, а следовательно, и λ; коэффициент при x_t, деленный на оценку λ, дает оценку β; а постоянный член, деленный на оценку λ, дает оценку α.

Как и в уравнении (10.11), эта модель включает стохастическую объясняющую переменную, которой снова является y_{t-1}. Но теперь эта переменная по крайней мере коррелирует не с текущим значением совокупного случайного члена уравнения, поскольку как v_t, так и u_t рассчитываются после того, как определилось значение y_{t-1}. Как мы видели в разделе 8.1, при таких условиях МНК позволяет получать асимптотически несмещенные и эффективные оценки (*«асимптотически»* здесь означает «с ростом объема выборки»), однако оценки не будут обладать этими свойствами на малых выборках.

Хотя модель частичной корректировки на первый взгляд и не относится к моделям Койка, мы покажем, что на самом деле это именно так. Если уравнение (10.17) выполняется для y_t, оно должно выполняться и для y_{t-1}:

$$y_{t-1} = \alpha\lambda + \beta\lambda x_{t-1} + (1 - \lambda)\, y_{t-2} + v_{t-1} + \lambda u_{t-1}. \qquad (10.18)$$

Подставив выражение для y_{t-1} в уравнение (10.17), получаем (опустив для простоты случайный член):

$$y_t = \alpha\lambda\,(1 + [1 - \lambda]) + \beta\lambda x_t + (1 - \lambda)\,\beta\lambda x_{t-1} + (1 - \lambda)^2 y_{t-2}. \qquad (10.19)$$

Точно так же выбрав позапрошлый период в уравнении (10.18), можно вывести выражение для y_{t-2}, которое в свою очередь может быть подставлено в уравнение (10.19), и так до бесконечности. В итоге мы получим выражение для y_t через текущие и лаговые значения x с геометрически убывающими весами в виде модели Койка. Заменив $(1 - \lambda)$ на δ, а $\beta\lambda$ — на β', мы имеем:

$$y_t = \alpha + \beta'(x_t + \delta x_{t-1} + \delta^2 x_{t-2} + \delta^3 x_{t-3} + \ldots), \qquad (10.20)$$

что по форме совпадает с уравнением (10.6).

Пример: модель корректировки размера дивидендов (модель Линтнера)

Пионерские исследования политики компаний относительно распределения дивидендов, проведенные Дж. Линтнером (Lintner, 1956), — широко известный пример использования модели частичной корректировки. Обычно производственные компании распределяют прибыль, остающуюся после уплаты налогов, частично на выплату доходов акционерам в форме дивидендов, а оставшиеся средства направляют на финансирование инвестиций. Когда прибыль растет, диви-

денды тоже увеличиваются, но, как правило, не в той же пропорции. Причиной этого в основном является осторожность руководства компаний. Возрастание прибыли может оказаться временным, и если дивиденды будут увеличиваться слишком быстро, вполне вероятно, что впоследствии их придется сокращать. По мнению руководства компании, ничто не наносит такой сильный удар по репутации фирмы, как сокращение дивидендов, поэтому оно проявляет большую осторожность с целью избежать риска. (Это, конечно, самоусиливающийся процесс. Именно по причине того, что многие компании так неохотно уменьшают дивиденды, если какая-нибудь из них вынуждена так поступить, это служит признаком наличия серьезных проблем.) Вторым доводом против немедленного увеличения дивидендов в той же пропорции, что и рост прибыли, является соображение о том, что рост прибыли может свидетельствовать об улучшении инвестиционных возможностей, требующих финансирования.

Моделируя описанное поведение, Дж. Линтнер предположил, что у фирм имеется целевая долгосрочная доля выплат γ и что желаемый объем дивидендов D_t^* соотносится с текущей прибылью Π_t как

$$D_t^* = \gamma \Pi_t. \tag{10.21}$$

Однако реальный объем дивидендов подвержен процессу частичной корректировки:

$$\Delta D_t = \lambda (D_t^* - D_{t-1}) + u_t, \tag{10.22}$$

где u_t — случайный член. Как следствие

$$D_t - D_{t-1} = \gamma (D_t^* - D_{t-1}) + u_t = \gamma\lambda \Pi_t - \lambda D_{t-1} + u_t, \tag{10.23}$$

или

$$D_t = \gamma\lambda \Pi_t + (1 - \lambda) D_{t-1} + u_t. \tag{10.24}$$

Используя данные о деятельности корпоративного сектора США за период 1918–1941 гг., Дж. Линтнер построил следующее уравнение регрессии:

$$\hat{D}_t = 352{,}3 + 0{,}15\Pi_t + 0{,}70 D_{t-1}, \tag{10.25}$$

где все коэффициенты значимо отличаются от нуля при уровне значимости в 1%.

Коэффициент при D_{t-1} позволяет оценить $(1 - \lambda)$ как 0,70 и, следовательно, коэффициент скорости корректировки — как 0,3. Поскольку коэффициент при Π_t служит оценкой $\gamma\lambda$, то, разделив его на 0,3, мы получим оценку для доли выплат, равную 0,5.

Упражнения

10.2. Линейное и логарифмическое уравнения регрессионной зависимости и объема расходов на жилье (y_t) от располагаемых доходов (x_t), индекса цен на жилье (p_t) и объемов этих расходов с лагом в один период, построенные на данных за 1959–1983 гг. из табл. 6.9, выглядят следующим образом (в скобках указаны значения стандартных ошибок):

$$\hat{y}_t = \underset{(9,54)}{21,86} + \underset{(0,007)}{0,022}x_t - \underset{(0,081)}{0,210}p_t + \underset{(0,036)}{0,871}y_{t-1}; \qquad R^2 = 0,999;$$

$$\widehat{\log} y_t = \underset{(0,37)}{0,44} + \underset{(0,05)}{0,15}\log x_t - \underset{(0,06)}{0,15}\log p_t + \underset{(0,037)}{0,845}\log y_{t-1}; \quad R^2 = 0,999.$$

Дайте интерпретацию обоих уравнений и проанализируйте их динамические свойства.

10.3. В логарифмическом уравнении регрессии в упражнении 10.2 d-статистика Дарбина—Уотсона равна 1,89. Рассчитайте h-статистику и проведите тест на автокорреляцию, приняв предположение, что выборка достаточно велика для того, чтобы эта статистика имела распределение, близкое к $N(0,1)$, при принятии гипотезы об отсутствии автокорреляции. (Об h-тесте см. раздел 7.8.)

10.4. Если в предыдущем упражнении убрать лаговое значение зависимой переменной из уравнения регрессии, т. е. если оценить зависимость $\log y_t$ от $\log x_t$ и $\log p_t$, как это показано в уравнении (10.1), d-статистика Дарбина—Уотсона оказывается равной 0,35. Объясните, почему во вновь построенном уравнении возникает автокорреляция, в то время как в предыдущем ее нет.

10.5. Оцените линейное и логарифмическое уравнения регрессии, как и в упражнении 10.2, для товара, выбранного для вас в упражнении 2.4. Дайте интерпретацию полученных результатов.

10.6. Проведите тест на автокорреляцию для уравнений регрессии, построенных в упражнении 10.5. Если вы обнаружите существенную автокорреляцию, проведите переоценку регрессии с помощью метода Кокрана—Оркатта или другого подобного метода.

10.7. В одном из первых исследований динамики совокупного потребления Т.М. Браун (Brown, 1952), используя годичные данные для Канады за период 1926—1949 гг. и исключив военное время 1942—1945 гг., построил следующее уравнение регрессии с помощью метода оценки одновременных уравнений (в скобках указаны значения t-статистики):

$$\hat{C}_t = \underset{(4,8)}{0,90} + \underset{(7,4)}{0,61}W_t + \underset{(4,2)}{0,28}NW_t + \underset{(2,8)}{0,22}C_{t-1} + \underset{(4,8)}{0,69}A,$$

где C_t — совокупное потребление; W_t — совокупный фонд заработной платы; NW_t — совокупный доход за вычетом фонда заработной платы; A — фиктивная переменная, равная нулю для довоенного периода и единице — для послевоенного. Переменные доходов и потребления измерены в миллиардах канадских долларов в постоянных ценах 1935—1939 гг. Разделение дохода на заработную плату и остальную часть сделано в соответствии с предположением Лоуренса Клейна о том, что предельная склонность к потреблению за счет заработной платы превышает предельную склонность к потреблению за счет прочих доходов и поэтому две данные составляющие должны фигурировать отдельно. Покажите, что эта регрессионная модель формально может рассматриваться как модель частичной корректировки, и предложите соответствующую интерпретацию для ее коэффициентов.

10.8. Как бы вы проверили предположение Л. Клейна с помощью данных, использованных Т.М. Брауном?

10.9. В своей классической работе «Распределенные лаги и анализ инвестиционных процессов» (1954) Л.М. Койк исследовал связь между инвестициями

на приобретение железнодорожных вагонов и объемом перевозок на железных дорогах США на данных за период 1894–1939 гг. Предположив, что желаемый парк вагонов в году t зависит от объема перевозок в годы $t-1$ и $t-2$ и от временного тренда, а также что затраты на приобретение вагонов подлежат частичной корректировке, он с помощью МНК получил следующее уравнение регрессии (стандартные отклонения и постоянный член не приводятся):

$$\hat{I}_t = 0{,}077F_{t-1} + 0{,}017F_{t-2} - 0{,}0033t - 0{,}110K_{t-1}; \qquad R^2 = 0{,}85,$$

где $I_t = K_t - K_{t-1}$ — число приобретенных вагонов в году t (тысяч); K_t — парк вагонов на конец года t (тысяч); F_t — объем перевозок в году t (млн. тонно-миль).

Предложите интерпретацию приведенному уравнению и опишите динамический процесс, который оно представляет [1].

10.4. Адаптивные ожидания

Моделирование ожиданий часто становится наиболее ответственной и сложной задачей в прикладной экономике. Это особенно верно для макроэкономики, где инвестиции, сбережения и спрос на активы оказываются чувствительными к ожиданиям относительно будущего. Вводные учебники по макроэкономике, анализируя базовую модель определения доходов (модель IS-LM), рассматривают валовые инвестиции как заданные или по крайней мере как строго убывающую функцию от нормы процента. В итоге остается такая проблема, как исследование воздействия роста государственных расходов на валовой объем производства в рамках предположения о том, что валовые инвестиции реагируют только на норму процента. Однако последнее неверно. Если государство проводит стимулирующую политику, то это оказывает воздействие на ожидания бизнесменов как относительно общего состояния экономики в будущем, так и относительно уровня прибыльности, которые определяют их планы независимо от того, что происходит с нормой процента.

Так, например, если в стране наблюдается существенная безработица, то действия правительства могут рассматриваться как позитивные, и это стимулирует инвестиции. С другой стороны, если экономика близка к состоянию полной занятости, то та же самая государственная политика может рассматриваться как ведущая к росту уровня инфляции и это вызовет снижение доверия бизнесменов и падение инвестиционной активности.

Все это создает непростую проблему, что признавал и Дж. М. Кейнс. Перечитайте главы «Общей теории», посвященные инвестициям. Конечно, Дж. М. Кейнс отводил много времени рассмотрению предельной эффективности капитальных вложений, связи инвестиций с нормой процента, но он также делал акцент на зависимости инвестиций от ожиданий, и это не оставляет сомнений в том, что и сам он считал IS-кривую (или то, что мы под ней сейчас понимаем) чрезвычайно подвижной.

К сожалению, в настоящее время отсутствуют удовлетворительные методы

[1] Может оказаться удобным заменить в уравнении I_t на $K_t - K_{t-1}$ и рассматривать это уравнение как динамическое соотношение, определяющее значение K_t

измерения ожиданий для решения макроэкономических задач. Как следствие макроэкономические модели не позволяют получать достаточно точные прогнозы, что затрудняет управление экономикой.

В качестве паллиатива решения описанной проблемы в некоторых моделях используется косвенный метод, известный как «процесс адаптивных ожиданий». Этот процесс заключается в простой процедуре корректировки ожиданий, когда в каждый период времени реальное значение переменной сравнивается с ее ожидаемым значением. Если реальное значение оказывается больше, то значение, ожидаемое в следующем периоде, корректируется в сторону его повышения; если меньше — то в сторону уменьшения. Предполагается, что размер корректировки пропорционален разности между реальным и ожидаемым значениями переменной.

Таким образом, если рассматривается переменная x, а x_t^e — ее значение, ожидаемое в период t, то

$$x_{t+1}^e - x_t^e = \lambda(x_t - x_t^e) \qquad (0 \le \lambda \le 1). \tag{10.26}$$

Это выражение может быть переписано в виде:

$$x_{t+1}^e = \lambda x_t + (1 - \lambda)x_t^e \qquad (0 \le \lambda \le 1). \tag{10.27}$$

Выражение (10.27) служит утверждением, что значение переменной, ожидаемое в следующий период времени, формируется как взвешенное среднее ее реального и ожидаемого значений в текущем периоде. Чем больше величина λ, тем быстрее ожидаемое значение адаптируется к предыдущим реальным значениям переменной.

Сходство моделей адаптивных ожиданий и частичной корректировки очевидно. Однако следует заметить два различия между ними. Во-первых, процесс адаптивных ожиданий направлен в будущее, в то время как процесс частичной корректировки базируется в основном на инерции и прошлой динамике показателей. Во-вторых, выведение выражения, которое включает только наблюдаемые значения переменной в модели, более гибко, чем в случае модели частичной корректировки.

Предположим, например, что зависимая переменная y_t связана с ожидаемым значением объясняющей переменной x в году $t + 1$:

$$y_t = \alpha + \beta x_{t+1}^e + u_t. \tag{10.28}$$

В уравнении (10.28) y выражена через величину x_{t+1}^e, которая не наблюдаема и которую необходимо так или иначе заменить наблюдаемыми переменными, т. е. реальными текущим и (или) прошлыми значениями переменной x и, может быть, прошлыми значениями переменной y. Процесс адаптивных ожиданий, описываемый уравнением (10.26), не позволяет это сделать прямо, поскольку он ставит x_{t+1}^e в зависимость частично от наблюдаемых переменных, но частично — и от ненаблюдаемых (x_t^e).

Тем не менее если (10.27) выполняется для периода t, то оно также должно выполняться и для периода $t - 1$:

$$x_t^e = \lambda x_{t-1} + (1 - \lambda)x_{t-1}^e. \tag{10.29}$$

Величину x_t^e в уравнении (10.27) можно заметить, но вместо нее появляется x_{t-1}^e:

$$x_{t+1}^e = \lambda x_t + \lambda(1-\lambda)x_{t-1} + (1-\lambda)^2 x_{t-1}^e. \tag{10.30}$$

В выражении (10.29) можно выбрать позапрошлый период и использовать полученный результат для исключения x_{t-1}^e ценой введения x_{t-2}^e. Повторив эту процедуру бесконечное число раз, мы получим:

$$x_{t+1}^e = \lambda[x_t + (1-\lambda)x_{t-1} + (1-\lambda)^2 x_{t-2} + \ldots]. \tag{10.31}$$

В итоге модель адаптивных ожиданий сводится к утверждению, что ожидаемое значение переменной является взвешенным средним ее прошлых значений с геометрически убывающими весами.

Подставив полученное выражение в (10.28) и заменив $(1-\lambda)$ на δ, мы имеем:

$$y_t = \alpha + \beta\lambda[x_t + \delta x_{t-1} + \delta^2 x_{t-2} + \ldots] + u_t, \tag{10.32}$$

откуда видно, что значение y определяется текущим и прошлыми значениями x с лагами, подчиняющимися распределению Койка. Параметры уравнения можно оценить с помощью метода нелинейного оценивания, описанного в разделе 10.2. (Преобразование Койка позволяет упростить уравнение математически, но оно неприменимо в качестве модели регрессии по причинам, уже обсуждавшимся в этом разделе.)

Пример: модель гиперинфляции Кейгана

Возможно, впервые модель адаптивных ожиданий была применена в исследовании, проведенном Ф. Кейганом, соотношения между спросом на реальные денежные остатки и ожидаемым изменением уровня цен (Cagan, 1956). Одним из факторов, определяющих спрос на денежные остатки, являются издержки их хранения, вызываемые обесцениванием наличности в реальном выражении. Предположив, что этот фактор будет главным при высоком уровне инфляции, Ф. Кейган исследовал эту зависимость для семи периодов гиперинфляции, имевших место между 1921 и 1956 гг., с помощью модели:

$$\log (M/P)_t = -\alpha E_{t+1} - \gamma + u_t, \tag{10.33}$$

где M — индекс изменения объема денег в обращении; P — индекс цен; $\log (P/M)$ — логарифм спроса на реальные денежные остатки; E — ожидаемый уровень инфляции; α и γ — неизвестные параметры. Поскольку переменная E ненаблюдаема, Ф. Кейган дополнил модель выражением для адаптивных ожиданий:

$$\Delta E_{t+1} = \beta(C_t - E_t), \tag{10.34}$$

которое определяет ожидаемое в период t изменение уровня инфляции ΔE_{t+1} как долю от величины разности между реальным текущим уровнем инфляции C_t и его предсказанным значением E_t.

С помощью формулы (10.34) величина E_{t+1} может быть выражена через текущее и прошлые значения C аналогично тому, как уравнение (10.26) было преобразовано в (10.31):

$$E_{t+1} = \beta C_t + (1-\beta)E_t = \beta[C_t + (1-\beta)C_{t-1} + (1-\beta)^2 C_{t-2} + \ldots]. \qquad (10.35)$$

Подставив это выражение в (10.33), мы получим следующую регрессионную модель:

$$\log (M/P)_t = -\alpha\beta[C_t + (1-\beta)C_{t-1} + (1-\beta)^2 C_{t-2} + \ldots] - \gamma + u_t. \qquad (10.36)$$

(Хотя уравнение, выведенное Ф. Кейганом, казалось бы, несколько отличается от представленного здесь, в принципе оно такое же. Разница в том, что оно было построено для непрерывной переменной времени, а не для дискретной, как в данном случае.)

Ф. Кейган оценил зависимости как отдельно для каждого из рассмотренных им семи случаев гиперинфляции, так и совместно для всех этих случаев, используя метод оценки нелинейной регрессии, описанный в разделе 10.2. Мы приведем здесь лишь последнюю версию:

$$\log (M/P)_t = -4{,}68E_{t+1} + \text{Константа}; \qquad (10.37)$$

$$\Delta E_{t+1} = 0{,}20(E_t - C_t). \qquad (10.38)$$

(Доверительные интервалы были приведены лишь для отдельных зависимостей, для совместного уравнения не были указаны ни доверительные интервалы, ни стандартные ошибки.) Полученные результаты означают что: 1) спрос на реальные денежные остатки сокращается в пропорции, равной 4,68 прироста ожидаемого уровня инфляции; 2) текущие ожидания корректируются каждый месяц только на $^1/_5$ от величины разности между реальным и ожидаемым уровнем инфляции. Например, если ожидаемый месячный уровень инфляции был равен 10 процентным пунктам, то спрос на реальные денежные остатки будет на долю 0,468, т. е. на 47%, ниже, чем он был бы в случае стабильных цен[1].

10.5. Гипотеза Фридмена о постоянном доходе

Хотя первым модель адаптивных ожиданий предложил Ф. Кейган, самым известным ее приложением, без сомнения, является модель потребления, основанная на гипотезе Фридмена о постоянном доходе (Friedman, 1957). Как было показано в разделе 8.3, в этой модели постоянное потребление индивида i в период t, обозначаемое C_{it}^P, предполагается пропорциональным его постоянному доходу Y_{it}^P:

$$C_{it}^P = \beta Y_{it}^P. \qquad (10.39)$$

Далее предполагается, что фактический объем потребления C_{it} и фактичес-

[1] Не совсем точный числовой пример: линейная формула расчета процентных изменений величины (M/P) для данной зависимости применима лишь при малых значениях E_{t+1}: например, если $E_{t+1} = 1\%$, или 0,01, то процентное изменение величины $(M/P)_t$ составит $e^{-0{,}0468} - 1$, то есть –0,046, или –4,6%. Если же $E_{t+1} = 10\%$, то есть 0,1, то процентное изменение величины $(M/P)_t$ составит $e^{-0{,}0468} - 1$, то есть –0,374, или –37,4%, а не –47%. (*Прим. ред.*)

кий уровень дохода Y_{it} включают временные составляющие C_{it}^{T} и Y_{it}^{T} соответственно, которые зависят от ситуации в году t:

$$C_{it} = C_{it}^{P} + C_{it}^{T}; \tag{10.40}$$

$$Y_{it} = Y_{it}^{P} + Y_{it}^{T}. \tag{10.41}$$

Предполагается, что временная составляющая потребления и временная составляющая дохода являются случайными переменными со средним значением 0 и постоянными значениями дисперсии, распределенными независимо от величины постоянного дохода, постоянного потребления и друг от друга.

Величина постоянного дохода в уравнении (10.39) ненаблюдаема. Для решения этой проблемы М. Фридмен расширил свою модель, предположив, что изменение постоянного дохода подчиняется процессу адаптивных ожиданий. Если фактический текущий доход индивида выше (или ниже) величины его постоянного дохода в предыдущем периоде, то индивид увеличивает (или уменьшает) значение последнего путем умножения λ на соответствующую разность:

$$\Delta Y_{it}^{P} = \lambda(Y_{it} - Y_{it-1}^{P}). \tag{10.42}$$

В общем случае предполагается, что величина λ лежит в границах между 0 и 1. Индивиды корректируют свое представление о постоянном доходе с ростом фактического дохода, но не на полное значение прироста, сознавая, что изменения фактического дохода частично объясняются вариацией временной составляющей дохода.

Выражение (10.42) может быть переписано как

$$Y_{it}^{P} - Y_{it-1}^{P} = \lambda(Y_{it} - Y_{it-1}^{P}), \tag{10.43}$$

или

$$Y_{it}^{P} = \lambda Y_{it} + (1 - \lambda)Y_{it-1}^{P}. \tag{10.44}$$

Это уравнение имеет простую интерпретацию. Оно говорит, что оценка индивидом величины постоянного дохода в году t равна средневзвешенной величине текущего фактического дохода и предыдущей оценки постоянного дохода. Если величина λ близка к единице, то индивид придает больший вес фактическому доходу, и значение Y^{P} быстро приближается к Y. Если величина λ, наоборот, близка к нулю, то фактическому доходу придается относительно меньший вес и процесс корректировки происходит медленно.

Подставив величину C_{it}^{P} из формулы (10.40) в (10.39), мы имеем:

$$C_{it} - C_{it}^{T} = \beta Y_{it}^{P}, \tag{10.45}$$

или

$$C_{it} = \beta Y_{it}^{P} + C_{it}^{T}. \tag{10.46}$$

В итоге мы получили соотношение между фактическим потреблением и постоянным доходом, где C_{it}^{T} играет роль случайного члена, который до этого отсутствовал в модели.

Использование фактического текущего значения дохода в качестве «замени-

теля» для показателя постоянного дохода в случае принятия гипотезы о постоянном доходе неприемлемо, поскольку это дает, как было показано в разделе 8.3, смещенные и несостоятельные оценки параметров. Вместо этого М. Фридмен использовал уравнение (10.44) для оценки связи постоянного дохода с текущим и прошлыми фактическими значениями дохода. Конечно, уравнение (10.44) не может использоваться напрямую для измерения постоянного дохода в году t по двум причинам: мы не знаем значения λ и нет метода измерения Y^P_{t-1}. Вторую причину можно устранить, заметив, что если выражение (10.44) выполняется для периода t, то оно выполняется также и для периода $(t-1)$:

$$Y^P_{it-1} = \lambda Y_{it-1} + (1-\lambda)Y^P_{it-2}. \tag{10.47}$$

Подставив это выражение в (10.44), мы получим:

$$Y^P_{it} = \lambda Y_{it} + \lambda(1-\lambda)Y_{it-1} + (1-\lambda)^2 Y^P_{it-2}. \tag{10.48}$$

Конечно, это уравнение включает ненаблюдаемую составляющую Y^P_{it-2}, но можно устранить ее, сдвинув выражение (10.44) на два периода назад и подставив его в (10.48), получить таким образом зависимость Y^P_{it} от Y_{it}, Y_{it-1}, Y_{it-2} и Y^P_{it-3}. Повторяя эту процедуру до бесконечности, можно выразить Y^P_{it} как взвешенную сумму текущего и прошлых фактических значений дохода:

$$Y^P_{it} = \lambda Y_{it} + \lambda(1-\lambda)Y_{it-1} + \lambda(1-\lambda)^2 Y_{it-2} + \lambda(1-\lambda)^3 Y_{it-3} + \ldots. \tag{10.49}$$

Опираясь на обоснованное предположение о том, что значение λ лежит в границах от 0 до 1, можно сделать вывод, что $(1-\lambda)$ лежит в тех же границах, а следовательно, величина $(1-\lambda)^s$ убывает с ростом s. Это свидетельствует о том, что текущее значение дохода имеет самый большой вес, значение дохода в предыдущем периоде имеет более низкий вес и значение этого веса постепенно убывает по мере продвижения назад к более отдаленным прошлым периодам. В конце концов оно становится настолько малым, что все предшествующие значения можно не принимать во внимание.

Тем не менее остается проблема оценки величины λ. Решение М. Фридмена схоже с решением, предложенным Ф. Кейганом в его исследовании гиперинфляции. Он испытал большое число различных значений λ между 0 и 1, рассчитал соответствующие ряды постоянного дохода для каждого из них, построил уравнения зависимости потребления для каждого ряда данных о постоянном доходе, используя коэффициент R^2 для измерения качества оценки. Затем он выбрал то значение λ, которое позволяло получить ряд Y^P, дающий наилучшую оценку.

Динамические свойства модели

Динамические свойства модели Фридмена удобнее анализировать после проведения преобразования Койка. Предположим, что мы используем агрегированные данные, и поэтому индекс i можно опустить. Подставив выражение (10.44) в (10.46), мы получим:

$$C_t = \beta\lambda Y_t + \beta(1-\lambda)Y_{t-1}^P + C_t^T. \qquad (10.50)$$

Сдвигая выражение (10.45) на один период назад, имеем:

$$\beta Y_{t-1}^P = C_{t-1} - C_{t-1}^T. \qquad (10.51)$$

Подставив это выражение в (10.50), получим:

$$C_t = \beta\lambda Y_t + (1-\lambda)C_{t-1} + C_t^T - (1-\lambda)C_{t-1}^T. \qquad (10.52)$$

Это уравнение позволяет одновременно оценить кратко- и долгосрочную предельную склонность к потреблению. Краткосрочная предельная склонность к потреблению $\partial C_t/\partial Y_t$ равна коэффициенту при Y_t, то есть $\beta\lambda$. Слагаемое $(1-\lambda)C_{t-1}$ в краткосрочном аспекте выступает как константа, поскольку изменение Y_t не может сказываться на значении C_{t-1}.

Что происходит в случае, когда величина дохода изменяется во времени — скажем, постепенно возрастает? Увеличение дохода в этом году прямо воздействует на объем потребления в этом году и косвенно — на объем потребления в следующем году, поскольку величина $(1-\lambda)C_t$ в следующем году будет больше, чем в текущем. Другими словами, график функции потребления сдвинется вверх. Если доход продолжит рост в будущем, график функции будет сдвигаться и дальше, а сложившаяся зависимость между объемом потребления и доходом, которой соответствует пунктирная линия на рис. 10.1, будет иметь более крутой наклон, чем краткосрочная зависимость.

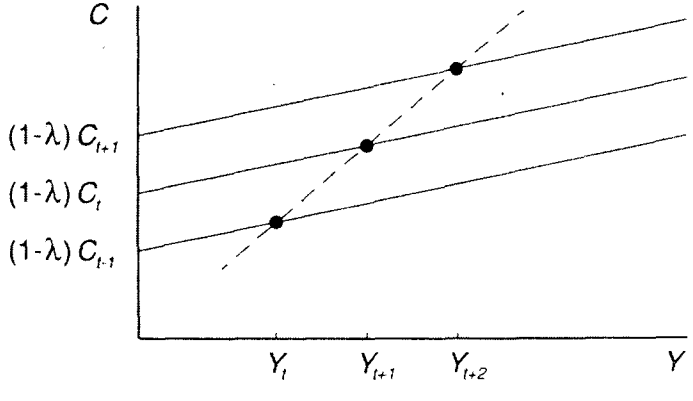


Рис. 10.1. Динамические свойства модели постоянного дохода Фридмена

Зависимость для состояния долгосрочного равновесия, без учета случайного члена, может быть получена подстановкой $C_t = C_{t-1} = \overline{C}$ и $Y_t = \overline{Y}$ в формулу (10.52):

$$\overline{C} = \beta\lambda\overline{Y} + (1-\lambda)\overline{C}, \qquad (10.53)$$

что может быть упрощено до вида:

$$\overline{C} = \beta\overline{Y}. \tag{10.54}$$

Модель сводится к фридменовской зависимости постоянного потребления от дохода с равным нулю свободным членом и коэффициентом наклона, равным β. Угол наклона краткосрочной функции $\beta\lambda$ оказывается меньше, поскольку значение λ находится в границах между 0 и 1. Следовательно, модель Фридмена объясняет сосуществование краткосрочной предельной склонности к потреблению, меньшей единицы, и приблизительно постоянной средней склонности к потреблению, которое в послевоенные годы было загадкой для эконометристов. Заметим, однако, что модель Брауна, представленная в упражнении 10.7, приводит к похожему уравнению. Здесь мы имеем пример двух радикально различных экономических моделей, приводящих к одинаковой зависимости между наблюдаемыми переменными. Верно также, что если случайный член в поведенческом уравнении удовлетворяет условиям Гаусса—Маркова, то он будет также удовлетворять им в преобразованной по Койку модели Брауна, но не в модели Фридмена, где он будет отрицательно коррелировать со своим значением в следующий период, и поэтому МНК окажется неприменимым. В принципе этот вывод может быть положен в основу для выбора модели, но если есть причины полагать, что случайный член в поведенческом уравнении может не удовлетворять требуемым условиям, то данный мотив для выбора модели становится неприемлемым.

Пример

Для сравнения своего варианта функции потребления с другими функциями М. Фридмен оценил ее на годовых рядах данных о реальном потреблении на душу населения и о реальном располагаемом доходе на душу населения в США в период 1905—1951 гг., за исключением военных лет (Friedman, 1957, pp. 142—152). В пошаговом поиске он рассчитал значения постоянного дохода как взвешенную сумму текущего и 16 предшествующих значений дохода, и оптимальное значение λ оказалось равным 0,37. В уравнении функции потребления он получил значение $\beta = 0{,}88$. Как следствие, краткосрочная предельная склонность к потреблению была равна 0,33, а краткосрочный мультипликатор — 1,5. Долгосрочные показатели составили 0,88 и 8,5 соответственно.

Упражнение

10.10. Исследователь полагает, что расходы на одежду определяются следующей моделью:

$$\log y_t = \alpha + \beta_1 \log z_t + \beta_2 \log p_t + u, \tag{1}$$

где y_t — расходы на одежду (млрд. долл., в постоянных ценах); z_t — постоянный доход; p_t — реальный индекс цен на одежду (отнесенный к уровню инфляции). Исследователь также полагает, что постоянный доход зависит от фактического дохода x_t, согласно следующему уравнению:

$$\log z_t = \gamma [\log x_t + \delta \log x_{t-1} + \delta^2 \log x_{t-2} + ...], \quad (2)$$

где γ рассчитывается так, чтобы сделать сумму весов равной единице (значение γ обратно $[1 + \delta + \delta^2 + ...]$). Показатель $\log z_t$ рассчитывается на годичных данных для США за период 1959–1983 гг. с лагом в четыре периода для следующих значений δ: 0; 0,1; 0,2; ... 0,9. Для каждого значения δ оценивается уравнение (1), результаты оценивания приведены в таблице (СКО — сумма квадратов отклонений, с. о. — стандартные ошибки).

1) Прокомментируйте полученные результаты.

2) Как текущее значение цены в уравнении (1) можно было бы заменить на значение «постоянной» цены?

	b_1		b_2			
δ	*Коэффициент*	*с. о.*	*Коэффициент*	*с. о.*	R^2	*СКО*
0,0	0,71	0,03	–0,63	0,11	0,9860	0,0042
0,1	0,70	0,03	–0,64	0,11	0,9862	0,0041
0,2	0,70	0,03	–0,65	0,11	0,9863	0,0040
0,3	0,70	0,03	–0,65	0,11	0,9863	0,0040
0,4	0,69	0,03	–0,65	0,11	0,9862	0,0041
0,5	0,69	0,03	–0,65	0,11	0,9859	0,0042
0,6	0,68	0,03	–0,65	0,11	0,9854	0,0043
0,7	0,67	0,03	–0,64	0,12	0,9847	0,0045
0,8	0,66	0,03	–0,63	0,12	0,9840	0,0047
0,9	0,66	0,03	–0,62	0,12	0,9832	0,0050

10.6. Полиномиально распределенные лаги Алмон[1]

Распределение Койка основывается на ограничивающем предположении, что коэффициенты при лаговых объясняющих переменных убывают в геометрической прогрессии. Для многих исследований это предположение вполне удовлетворительно, но для других оно малореалистично. Например, в некоторых случаях более уместно предположить, что изменение зависимой переменной в ответ на изменение объясняющей переменной сначала невелико, затем возрастает со временем, а потом снова уменьшается. Распределенные лаги Алмон (Almon, 1965) обладают достаточной гибкостью для моделирования поведения такого рода, используя при этом минимальное число параметров.

В основе модели лежит предположение о том, что если y зависит от теку-

[1] Раздел содержит относительно сложный материал, который можно опустить без потери в последовательности изложения.

щих и лаговых значений x, то веса в этой зависимости подчиняются полиномиальному распределению. По этой причине лаги Алмон также часто описываются как полиномиально распределенные лаги. Приведем простые примеры, когда значения весов подчиняются квадратичной зависимости (как на рис. 10.2А и 10.2Б) кубической функции (как на рис. 10.3) или полиному более высокой степени. Выбор функции остается за исследователем, и он, конечно, может быть сделан на основе экспериментов.

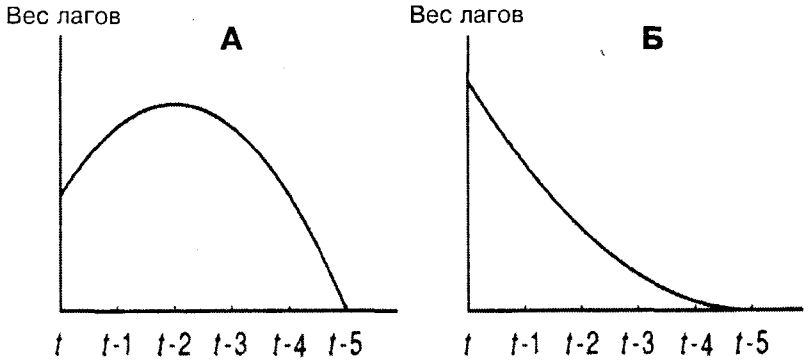


Рис. 10.2

В общем случае модель регрессии может быть записана как

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_n x_{t-n} + u_t, \qquad (10.55)$$

где

$$\beta_s = \gamma_0 + \gamma_1 s + \gamma_2 s^2 + \ldots + \gamma_m s^m. \qquad (10.56)$$

Рисунки 10.2А и 10.2Б, следовательно, соответствуют случаю, когда величина $m = 2$, рис. 10.3 — случаю, когда $m = 3$.

Далее исследователь должен выбрать число лаговых значений объясняющей переменной n, которое будет использоваться в модели. И снова это число может быть определено в результате экспериментов, направленных на получение хорошего описания имеющихся данных.

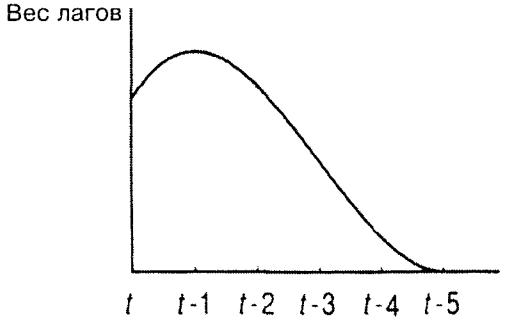


Рис. 10.3

Для того чтобы сделать наш анализ проще, предположим, что была выбрана квадратичная функция и число лагов равно трем. Подставив выражение (10.56) в (10.55), мы получим:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + u_t =$$
$$= \alpha + \gamma_0 x_t + (\gamma_0 + \gamma_1 + \gamma_2)\, x_{t-1} + (\gamma_0 + 2\gamma_1 + 4\gamma_2)\, x_{t-2} + (\gamma_0 + 3\gamma_1 + 9\gamma_2)\, x_{t-3} + u_t =$$
$$= \alpha + \gamma_0 (x_t + x_{t-1} + x_{t-2} + x_{t-3}) + \gamma_1 (x_{t-1} + 2x_{t-2} + 3x_{t-3}) + \gamma_2 (x_{t-1} + 4x_{t-2} + 9x_{t-3}) + u_t =$$
$$= \alpha + \gamma_0 z_0 + \gamma_1 z_1 + \gamma_2 z_2 + u_t, \qquad (10.57)$$

где $z_0 = x_t + x_{t-1} + x_{t-2} + x_{t-3}$, $z_1 = x_{t-1} + 2x_{t-2} + 3x_{t-3}$ и $z_2 = x_{t-1} + 4x_{t-2} + 9x_{t-3}$.

Для оценки параметров модели следует построить уравнение зависимости y не от текущего и лаговых значений x, а от сконструированных из них переменных z. Большинство статистических программ позволяют легко сделать это. С их помощью можно получить коэффициенты при сконструированных переменных, а на их основе — рассчитать коэффициенты в исходной модели.

Как видно из рис. 10.2А, сначала возрастающее, а затем убывающее распределение лагов в принципе может быть представлено квадратичной функцией. Однако у этого распределения имеется нежелательное свойство: выпуклый вверх график функции не позволяет весам плавно приближаться к нулю. С этой точки зрения кубическая функция на рис. 10.3, у которой две экстремальные точки, более привлекательна: в окрестности второй экстремальной точки происходит плавное убывание весов.

На практике распределение лагов объясняющей переменной может не соответствовать простой функции, и попытки их применения могут привести к нежелательным результатам: получение весов с неверными знаками, резкое уменьшение весов на краю распределения и т. д. Все эти проблемы в принципе можно преодолеть, используя полиномы более высокой степени: сама Ш. Алмон в своей статье использовала полином четвертой степени, получив вполне удовлетворительные результаты. Однако с ростом степени полиномов вновь возникает риск появления неучтенной мультиколлинеарности. Число переменных z равно числу слагаемых в полиноме, и переменные z коррелируют друг с другом, поскольку каждая из них является линейной комбинацией текущего и лаговых значений x.

Остается большой соблазн испытать все кажущиеся возможными комбинации степени полинома и числа лаговых значений и выбрать ту из них, которая дает результаты, наиболее близкие к априорным представлениям или по крайней мере наименее противоречащие им. В итоге вместо того, чтобы служить независимым тестом для гипотез, эксперимент превращается в поиск подтверждений для априорных представлений.

Пример

При построении уравнения регрессионной зависимости расходов на жилье от личных доходов и реального индекса цен на жилье по данным за период 1959–1983 гг. из табл. 6.9 было решено использовать описанную выше схему Алмон (квадратичная функция, три лаговых периода) для переменных дохода и цен. В уравнении регрессии, совпадающем с (10.57), постоянный член оказался равен −2,33, коэффициенты при z_0, z_1 и z_2 для переменной дохода были равны

0,417, −0,159 и 0,025 соответственно, откуда следовало, что полиномиальная функция имела вид:

$$\beta_s = 0{,}417 - 0{,}159s + 0{,}025s^2. \tag{10.58}$$

Подставляя в этой формуле для s значения 0, 1, 2 и 3, можно получить коэффициенты при текущем и лаговых значениях x:

$$\widehat{\log}\ y_t = -2{,}33 + 0{,}417 \log x_t + 0{,}283 \log x_{t-1} + 0{,}199 \log x_{t-2} + \\ + 0{,}165 \log x_{t-3} + \text{Слагаемые для цен.} \tag{10.59}$$

В данном случае график квадратичной функции оказался вогнутым вниз, как на рис. 10.2Б, и распределение лагов имеет такой же вид, как и укороченное распределение Койка.

Упражнение

10.11. Коэффициенты при z_0, z_1 и z_2 для переменной цен в регрессионном уравнении расходов на жилье были равны −0,035, 0,249 и −0,165, соответственно. Рассчитайте коэффициенты при $\log p_t$, $\log p_{t-1}$, $\log p_{t-2}$ и $\log p_{t-3}$ и прокомментируйте полученный результат.

10.7. Рациональные ожидания

Одним из потенциальных дефектов процесса адаптивных ожиданий и других похожих способов учета ожиданий является то, что получаемые с их помощью прогнозы в общем случае отличаются от прогнозов, получаемых с помощью модели в целом. Разработчик модели может встать на защиту этих методов, сказав, что субъекты, представленные в модели, обладают ограниченной информацией и не знают о других закономерностях и т. д., и как следствие их прогнозы будут уступать прогнозам, принимающим во внимание всю сложность данной модели.

В некоторых ситуациях это может быть обоснованной предпосылкой, но в большинстве случаев субъекты наверняка обладают не меньшей информацией, чем разработчик модели. Они далеко не наивны и в состоянии получать выводы, близкие к выводам разработчика модели, хотя и полагаются целиком на свои собственные представления и интуицию. В таких случаях механический подход к формированию ожиданий, как в методе адаптивных ожиданий, неадекватен. Наоборот, лучшей стартовой позицией будет принятие предположения, что все субъекты имеют доступ к модели и к получаемым с ее помощью прогнозам, и учет этого предположения *внутренне* присущ самой модели. Этот подход известен как подход с позиции рациональных ожиданий.

Для того чтобы сделать наши рассуждения более конкретными, рассмотрим модель спроса и предложения некоторого товара, производители которого определяют объем выпуска за один период до того, как поставить произведенный товар на рынок. Мы предположим также, что нельзя делать запасы товара, и рынок всегда приходит в равновесие. В результате имеем следующую модель:

$$y_t^d = \alpha + \beta p_t + u_t^d; \tag{10.60}$$

$$y_t^s = \delta + \varepsilon p_t^e + u_t^s, \tag{10.61}$$

где y_t^d и y_t^s — объемы спроса и предложения в период t, соответственно; p_t — цена рыночного равновесия в период t; p_t^e — ожидаемое значение p_t, сформированное в период $(t-1)$; u_t^d и u_t^s — случайные члены. Когда рынок находится в равновесии и $y_t^d = y_t^s$, модель дает следующее соотношение между реальной и ожидаемой ценой в период t:

$$p_t = \frac{\delta - \alpha}{\beta} + \frac{\varepsilon}{\beta} p_t^e + \frac{u_t^s - u_t^d}{\beta}. \tag{10.62}$$

В простейшей модели такого рода производитель предполагает, что цены периода $(t-1)$ будут действовать и в период t:

$$p_t^e = p_{t-1}. \tag{10.63}$$

Это соотношение порождает так называемый «цикл поставок свинины», названный так по товару, рынок которого, как предполагается, ведет себя подобным образом. Пренебрегая на время воздействием случайных членов, из уравнений (10.62) и (10.63) получаем, что равновесие будет поддерживаться при условии:

$$p_t = p_{t-1} = \frac{\alpha - \delta}{\varepsilon - \beta}. \tag{10.64}$$

Если первоначально рынок находился в состоянии неравновесия, то поведение цен и выпуска будет таким, как показано на рис. 10.4, из которого видно, почему эта модель также называется *паутинообразной* (или *паутинообразным циклом*). (Первый формальный анализ свойств этой модели можно найти в работе М. Езекиела [Ezekiel, 1938].) В период $t = 0$ производители принимают решения о том, сколько товара предложить в следующем периоде по текущей цене p_0. Этот объем предложения (y_1) представлен точкой A. Он меньше равновесного объема, и, следовательно, цена равновесия в период *1* (p_1) будет относительно высокой (точка B). Следуя предположению, что эта цена будет иметь место в периоде *2*, производители значительно увеличивают свой выпуск (точка C), что приводит к относительно низкой равновесной рыночной цене (точка D). Процесс будет сходиться, если функция спроса более эластична (крута), чем функция предложения, как на рис. 10.4.

Если функция спроса оказывается менее эластичной, то рынок с каждым циклом будет удаляться все дальше от точки равновесия. Случайные члены лишь смещают действительные значения p и y в каждый период времени, но не меняют общий характер процесса.

Подобная модель формируется производителями, которые не понимают, что их собственные решения воздействуют на цену рыночного равновесия. Если же

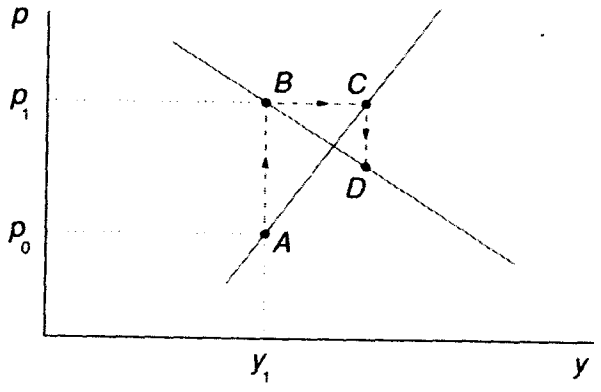


Рис. 10.4. Паутинообразная модель

производители осознали связь между спросом и предложением (а их представления об этом часто гораздо лучше, чем у рядового эконометриста), то они интуитивно будут использовать обобщенную модель для генерации своих ожиданий.

В таком случае ключевым становится уравнение (10.62), связывающее действительную цену с ожидаемой. Поскольку значение $p_t^e = E(p_t)$, т. е. ожидаемая цена определяется как математическое ожидание цены в период t, полученное в период $(t-1)$, то мы имеем:

$$p_t^e = E(p_t) = E\left(\frac{\delta-\alpha}{\beta} + \frac{\varepsilon}{\beta}p_t^e + \frac{u_t^s - u_t^d}{\beta}\right) = \frac{\delta-\alpha}{\beta} + \frac{\varepsilon}{\beta}p_t^e. \qquad (10.65)$$

Слагаемое $(\delta-\alpha)/\beta$ является константой и не изменяется под воздействием ожиданий. Значение $E(p_t^e) = p_t^e$, поскольку оба ожидания формируются в период $(t-1)$. Случайные члены исчезают, поскольку их значение не может быть предсказано в период $(t-1)$. Решив уравнение, мы получим:

$$p_t^e = \frac{\alpha-\delta}{\varepsilon-\beta}. \qquad (10.66)$$

Как следствие, объем предложения в период t равен:

$$y_t = \frac{\alpha\varepsilon - \beta\delta}{\varepsilon-\beta} + u_t^s, \qquad (10.67)$$

а цена рыночного равновесия составит:

$$p_t = \frac{\alpha-\delta}{\varepsilon-\beta} + \frac{u_t^s - u_t^d}{\beta}. \qquad (10.68)$$

Если вся информация используется в модели подобным образом, паутинообразный цикл исчезает. Производители выпускают одинаковое количество товара в каждый период, не считая случайной составляющей, а цена всегда

является ценой равновесия плюс случайная составляющая, которая зависит от обоих случайных членов. (Более развернутый анализ использования принципа рациональных ожиданий в этом контексте см. в работе Дж. Муса [Muth, 1961]. Общий обзор предложен в работе С. Шеффрина [Sheffrin, 1983].)

Упражнение

10.12. Предположим, что в модели спроса и предложения функция спроса (10.60) заменена на

$$y_t^d = \alpha + \beta p_t + \gamma x_{t-1} + u_t^d,$$

где x_{t-1} — располагаемый личный доход в период $(t-1)$, и он устойчиво возрастает в наблюдаемый период.

1. Какое воздействие это окажет на паутинообразный цикл?
2. Как будут определяться значения y и p, если ожидания формируются рационально?

10.8. Предсказание[1]

Предположим, что вы оценили модель

$$y_t = \alpha + \beta x_t + u_t \tag{10.69}$$

на наблюдениях периода $(t = 1, ..., T)$:

$$\hat{y}_t = a + bx_t. \tag{10.70}$$

Имея некоторое послевыборочное значение переменной x, скажем, x_{T+p}, вы можете предсказать соответствующее значение y:

$$\hat{y}_{T+p} = a + bx_{T+p}. \tag{10.71}$$

Такие предсказания могут быть важными по двум причинам. Во-первых, вы можете быть одним из тех эконометристов, чья работа — заглядывать в экономическое будущее. Некоторые эконометристы изучают экономические закономерности с целью улучшить понимание того, как работает экономика, но для других это является лишь средством достижения более практичной цели — предвидеть, что может случиться. Во многих странах макроэкономическое прогнозирование имеет высокую репутацию, и коллективы эконометристов поддерживаются министерством финансов или другими правительственными органами, частными финансовыми учреждениями, университетами и исследовательскими институтами, и их предсказания активно используются для формирования и толкования государственной политики или в деловых целях. Когда подобные предсказания публикуются в печати, они, как правило, привлекают гораздо больше внимания, чем большинство других видов экономического анализа, в основном благодаря своей сути и тому, что в отличие от большинства других

[1] Мы используем для перевода термины «предсказание» и «прогноз» в соответствии с терминологией автора, которую он объясняет чуть дальше. (*Прим. ред.*)

видов экономического анализа они легко могут быть поняты средним гражданином. Даже человек с совершенно нематематическим и нетехническим складом ума в состоянии понять, что подразумевается под оценками будущего уровня безработицы, инфляции и т. д.

Есть, однако, и другое применение эконометрического предсказания, которое делает его предметом заботы большинства эконометристов независимо от того, заняты они прогнозированием или нет. Оно дает метод оценки устойчивости регрессионной модели, который имеет большую исследовательскую направленность, чем диагностические статистики, использовавшиеся до сих пор.

Прежде чем продвигаться дальше, необходимо уточнить, что мы понимаем под *предсказанием*. К сожалению, в эконометрической литературе этот термин может иметь несколько различных значений в соответствии с пониманием x_{T+p} в уравнении (10.71). Мы будем различать предсказания и прогнозы. Это разделение сделано в соответствии с обычным использованием терминов (например, у Э. Харвея [Harvey, 1981]), но тем не менее используемая здесь терминология не вполне стандартна.

Предсказания

Мы опишем $\hat{y}_{T+p}$ как предсказание, если значение x_{T+p} известно. Как это возможно? В общем случае эконометристы хотят включить все имеющиеся данные в выборку для максимизации ее размера и, как следствие, для минимизации дисперсии оценок, поэтому x_T является последним зафиксированным значением x на момент оценки регрессии. Тем не менее возможны две ситуации, когда x_{T+p} известны: когда вы ждете p или больше периодов после оценки регрессии и когда вы заранее ограничили период выборки так, чтобы у вас остались несколько последних наблюдений. Как мы увидим в следующем разделе, весомой причиной так поступать может стать возможность без задержки оценить прогнозную точность модели.

Так, например, обращаясь снова к уравнению (3.34) модели связи общей инфляции и инфляции заработной платы, предположим, что для всего периода выборки мы оценили уравнение

$$\dot{p} = 1{,}0 + 0{,}8\dot{w}, \tag{10.72}$$

где $\dot{p}$ и $\dot{w}$ — годовой уровень общей инфляции и инфляции заработной платы (в процентах) соответственно, и что мы знаем, что в один послевыборочный год уровень инфляции заработной платы составлял 6%. Тогда мы можем утверждать, что предсказанный уровень общей инфляции равен 5,8%. Мы, конечно, должны иметь возможность сразу сравнить его с действительным уровнем инфляции в этом году и рассчитать *ошибку предсказания*, которая равна разности между предсказанным и действительным значениями. В общем случае если $\hat{y}_{T+p}$ — предсказываемое значение, а y_{T+p} — действительное, то ошибка предсказания f_{T+p} определяется как

$$f_{T+p} = \hat{y}_{T+p} - y_{T+p}. \tag{10.73}$$

Почему появляется ошибка предсказания? Это происходит по двум причинам. Во-первых, значение $\hat{y}_{T+p}$ было рассчитано с помощью оценок парамет-

ров a и b вместо их реальных значений. Во-вторых, $\hat{y}_{T+p}$ не учитывает воздействие случайного члена u_{T+p}, являющегося составной частью y_{T+p}. В дальнейшем мы будем предполагать, что данные включают $(T+m)$ наблюдений переменных, из них первые T наблюдений (период выборки) используются для построения регрессии, а последние m (период, или интервал предсказания) применяются для анализа точности предсказания.

Пример

Предположим, что когда мы оценивали функцию спроса на продукты питания с помощью данных из табл. Б.1 и Б.2, мы использовали лишь первые 21 наблюдение из выборки, т.е. данные за 1959–1979 гг., оставив последние 4 наблюдения для анализа предсказаний. Полученное на выборке 1959–1979 гг. уравнение выглядит следующим образом (в скобках приведены стандартные ошибки):

$$\widehat{\log} y = \underset{(0,42)}{2,78} + \underset{(0,03)}{0,61} \log x - \underset{(0,12)}{0,42} \log p; \qquad R^2 = 0,98. \tag{10.74}$$

Значения $\hat{y}$ для периода 1980–1983 гг., предсказанные с помощью этого уравнения, при использовании действительных значений личного располагаемого дохода и относительной цены на продукты питания в эти годы, показаны в табл. 10.1 вместе с фактическими значениями этой переменной и ошибками предсказания. Предсказания, как и исходные данные, приведены в логарифмической шкале. Для удобства в табл. 10.1 показаны также абсолютные значения, выраженные в миллиардах долларов (в ценах 1972 г.), которые могут быть рассчитаны на основе значений логарифмов.

Таблица 10.1

Предсказанные и действительные значения спроса на продукты питания, 1980–1983 гг.

Год	*Логарифмы*			*Абсолютные значения*		
	log ŷ	*log y*	*Ошибка*	*ŷ*	*y*	*Ошибка*
1980	4,995	5,031	−0,037	147,7	153,2	−5,5
1981	5,012	5,030	−0,019	150,2	153,0	−2,8
1982	5,024	5,041	−0,017	152,0	154,6	−2,6
1983	5,052	5,083	−0,031	156,4	161,2	−4,8

Как мы видим, предсказанные значения расходов на продукты питания примерно на 2–3 процентных пункта ниже фактических значений. Может ли такое предсказание считаться удовлетворительным? Мы обсудим это в следующем разделе.

Если вы хотите предсказать конкретное значение y_{T+p}, не зная действительное значение x_{T+p}, то говорится, что вы делаете прогноз (по крайней мере, если использовать терминологию этого текста). Макроэкономические предвидения, публикуемые в прессе, обычно являются прогнозами в таком смысле. Политиков, а в особенности широкую публику мало интересуют «двусторонние» экономисты, рассуждения которых имеют вид «с одной стороны... но если нет, то с другой стороны...». Обычно все желают точных однозначных оценок, дополненных, может быть, границами возможной ошибки, но часто даже и без этого. Прогнозы менее точны, чем предсказания, поскольку они подвержены воздействию дополнительного источника ошибки — предсказания значения x_{T+p}. Очевидно, что делающий прогноз эконометрист пытается, как правило, минимизировать эту дополнительную ошибку, моделируя как можно более точно поведение переменной x. Иногда для нее строят отдельную модель, иногда совмещают в одну модель уравнение для y и уравнение для x, дополняя их множеством других соотношений и оценивая так называемую систему одновременных уравнений (что рассматривается в главе 11).

Свойства предсказаний, полученных с помощью МНК

В последующих рассуждениях мы сосредоточимся в основном на предсказаниях, а не на прогнозах и рассмотрим свойства коэффициентов уравнения регрессии и свойства случайного члена, а не переменной x в случае, когда ее значения неизвестны. И в этом есть положительные моменты. Если значение y_{T+p} порождается тем же процессом, что и выборочные значения переменной y [то есть в соответствии с уравнением (10.69), где u_{T+p} удовлетворяет условиям Гаусса—Маркова], и если мы строим предсказание $\hat{y}_{T+p}$ с помощью уравнения (10.71), то ошибка предсказания f_{T+p} будет иметь нулевое среднее значение и минимальную дисперсию.

Первое свойство можно продемонстрировать довольно просто:

$$\begin{aligned} E(f_{T+p}) &= E(\hat{y}_{T+p}) - E(y_{T+p}) = \\ &= E(a + bx_{T+p}) - E(\alpha + \beta x_{T+p} + u_{T+p}) = \\ &= E(a) + x_{T+p}E(b) - \alpha - \beta x_{T+p} - E(u_{T+p}) = \\ &= \alpha + \beta x_{T+p} - \alpha - \beta x_{T+p} = 0, \end{aligned} \tag{10.75}$$

поскольку $E(a) = \alpha$, $E(b) = \beta$ и $E(u_{T+p}) = 0$. Мы не будем доказывать свойство минимума дисперсии (доказательство можно найти у Дж. Джонстона [Johnston, 1984] или Дж. Томаса [Thomas, 1983]). Оба эти свойства сохраняются и для общего случая множественного регрессионного анализа.

В случае уравнения парной регрессии выборочная дисперсия f_{T+p} определяется как

$$\text{pop.var}(f_{T+p}) = \left(1 + \frac{1}{n} + \frac{(x_{T+p} - \bar{x})^2}{\sum (x_t - \bar{x})^2}\right)\sigma_u^2 = \left(1 + \frac{1}{n} + \frac{(x_{T+p} - \bar{x})^2}{n\text{Var}(x)}\right)\sigma_u^2, \tag{10.76}$$

где $\bar{x}$ и Var (x) — выборочное среднее значение и дисперсия переменной x. Из формулы следует, и это неудивительно, что чем больше значение x отклоняется от выборочного среднего, тем больше дисперсия ошибки предсказания. Из формулы также следует, и это вновь неудивительно, что чем больше объем выборки, тем меньше дисперсия ошибки предсказания с нижним пределом, равным σ_u^2. С ростом объема выборки оценки a и b стремятся к истинным значениям соответствующих коэффициентов (в случае выполнения условий Гаусса—Маркова), и единственным источником ошибки при предсказании будет случайный член u_{T+p}, а он по определению имеет дисперсию σ_u^2.

Доверительные интервалы для предсказаний

Мы можем получить значение ***стандартной ошибки предсказания***, если заменим σ_u^2 в уравнении (10.76) на s_u^2 и извлечем квадратный корень. Тогда отношение величины $(\hat{y}_{T+p} - y_{T+p})$ к стандартной ошибке при оценивании уравнения для периода выборки будет подчиняться t-распределению с соответствующим числом степеней свободы. Отсюда можно получить доверительный интервал для действительного значения y_{T+p}:

$$\hat{y}_{T+p} - t_{крит} \times \text{с. о.} < y_{T+p} < \hat{y}_{T+p} + t_{крит} \times \text{с. о,} \qquad (10.77)$$

где $t_{крит}$ — критическое значение t при заданных уровне значимости и числе степеней свободы; с. о. — стандартная ошибка предсказания. На рис. 10.5 в общем виде показано соотношение между доверительным интервалом для предсказания и значением объясняющей переменной.

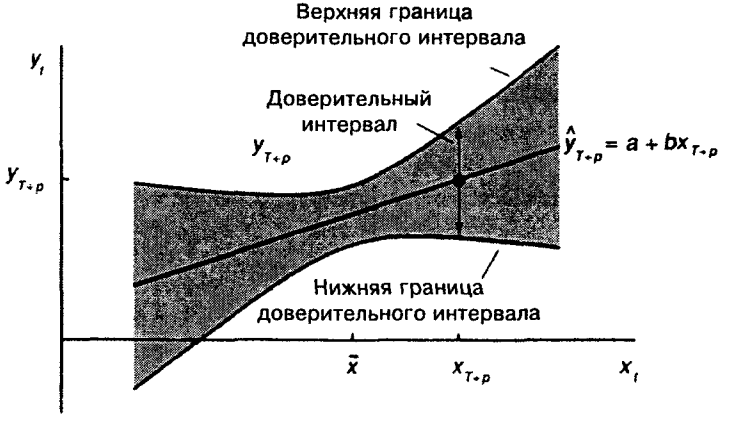


Рис. 10.5. Доверительный интервал для предсказания

В уравнении множественной регрессии выражение, соответствующее (10.76), имеет гораздо более сложный вид, и оно лучше может быть представлено с

помощью аппарата матричной алгебры. Однако имеется простой прием, который можно использовать для расчета значений стандартных ошибок с помощью компьютера. Вы оцениваете уравнение регрессии на выборке, совмещающей выборочный и прогнозный периоды, добавив (различные) фиктивные переменные для каждого из наблюдений периода предсказания. Это означает включение в модель набора фиктивных переменных D_{T+1}, D_{T+2}, ..., D_{T+m}, где значение $D_{T+p} = 0$ для всех наблюдений, кроме наблюдения $T + p$, для которого оно равно единице. Как может быть показано, оценки коэффициентов при нефиктивных переменных и их стандартные отклонения будут в точности такими же, как и в уравнении регрессии, оцененном только на периоде выборки (см. работы Д. Салкевера [Salkever, 1976] и Ж.-М. Дюфора [Dufour, 1980]). Компьютер использует фиктивные переменные для получения точного значения каждого наблюдения в период предсказания и делает это, приравнивая коэффициент при фиктивной переменной к значению ошибки предсказания, как она была определена выше. Стандартная ошибка этого коэффициента равна стандартной ошибке предсказания.

Пример

Стандартная ошибка предсказания в уравнении функции спроса на продукты питания для 1980 г. равна 0,019. При числе степеней свободы, равном 18, и уровне значимости в 5% критический уровень t-статистики равен 2,10, откуда можно получить следующий 95-процентный доверительный интервал для предсказания в этом году:

$$4{,}995 - 2{,}10 \times 0{,}019 < \log y < 4{,}995 + 2{,}10 \times 0{,}019, \qquad (10.78)$$

т. е.

$$4{,}955 < \log y < 5{,}035. \qquad (10.79)$$

Как мы видим, действительное значение переменной попадает в этот доверительный интервал, поэтому предсказание, по крайней мере в данном году, можно считать удовлетворительным. Это верно и для оставшихся лет периода предсказания.

Упражнение

10.13. Используйте косвенный метод Салкевера для расчета прогнозов и их стандартных ошибок для логарифмической функции спроса на выбранный вами товар. Добавьте фиктивные переменные для последних четырех наблюдений и рассчитайте ошибки предсказания для этих лет, базируясь на уравнении регрессии, полученном на первых 21 наблюдении. Добавьте это к реальным значениям для получения прогноза. Рассчитайте доверительный интервал для прогноза по крайней мере на год вперед.

10.9. Тесты на устойчивость

Тесты на устойчивость для регрессионной модели предназначены для оценки того, насколько поведение модели в послевыборочном периоде сравнимо с ее поведением в период выборки, на которой она была получена. В основе организации тестов на устойчивость могут лежать два принципа. Один подход — сосредоточиться на предсказательной способности модели; другой подход — оценить, происходит ли сдвиг параметров в период предсказания.

Тест Чоу на неудачу предсказания

Как мы видели в предыдущем разделе, ошибку предсказания можно рассчитать, добавив набор фиктивных переменных для наблюдений периода предсказания. Теперь вполне естественно определить, существенно ли ошибка предсказания отличается от нуля, и мы можем сделать это с помощью F-теста на совместную объясняющую способность фиктивных переменных. Совместив период выборки и период предсказания, мы оценим уравнение регрессии сначала без набора фиктивных переменных, а затем — вместе с этим набором. Обозначим полученные суммы квадратов отклонений как RSS_{T+m} и RSS^{D}_{T+m}, где нижний индекс показывает число наблюдений в регрессии, а верхний индекс «D» означает включение в уравнение фиктивных переменных. С помощью F-теста, описанного в разделе 5.6, мы можем определить, было ли существенным улучшение качества уравнения после добавления набора фиктивных переменных. Данное улучшение можно представить в виде $(RSS_{T+m} - RSS^{D}_{T+m})$; число фиктивных переменных равно m; сумма квадратов отклонений после включения фиктивных переменных составляет RSS^{D}_{T+m}; остающееся число степеней свободы равно числу наблюдений в совмещенной выборке $(T+m)$ за вычетом числа оцененных параметров $(k+m+1)$. В итоге значение F-статистики составит:

$$F(m, T-k-1) = \frac{(RSS_{T+m} - RSS^{D}_{T+m}) / m}{RSS^{D}_{T+m} / (T-k-1)}. \tag{10.80}$$

На самом деле для реализации теста даже не требуется оценивать уравнение регрессии с фиктивными переменными, поскольку значение RSS^{D}_{T+m} равно значению RSS_T — сумме квадратов отклонений для уравнения регрессии, оцененного на периоде выборки. Качество этой регрессии в точности такое же, как и у регрессии для первых T наблюдений в уравнении с фиктивными переменными, и отклонения здесь те же самые. Для последних m наблюдений в уравнении с фиктивными переменными нет отклонений, так как включение специальной фиктивной переменной для каждого наблюдения гарантирует точность уравнения для этих наблюдений. В итоге значение RSS^{D}_{T+m} в точности такое же, как и значение RSS_T, и F-статистика может быть переписана как

$$F(m, T-k-1) = \frac{(RSS_{T+m} - RSS_T) / m}{RSS_T / (T-k-1)}. \tag{10.81}$$

Этот тест известен как тест Чоу и был назван так по имени своего создателя

Г. Чоу (Chow, 1960), однако приводимая здесь интерпретация теста была предложена несколько позже Х. Песараном, Р. Смитом и С. Ео (Pesaran, Smith, Yeo, 1985).

Пример

Функция спроса на продукты питания сначала была оценена на данных за период 1959–1979 гг., и $RSS_T = 0{,}0052$, а затем — на данных за период 1959–1983 гг., $RSS_{T+m} = 0{,}0070$. Как следствие значение F-статистики равно:

$$F(4,18) = \frac{(0{,}0070 - 0{,}0052) / 4}{0{,}0052 / 18} = 1{,}56. \qquad (10.82)$$

Критическое значение F-статистики с 4 и 18 степенями свободы при 5-процентном уровне значимости равно 2,93, поэтому мы не отвергаем нулевую гипотезу о стабильности коэффициентов уравнения регрессии.

F-тест на стабильность коэффициентов

Если имеются приемлемые наблюдения за период предсказания, то можно провести F-тест на наличие структурного перелома, описанный в разделе 9.5, и оценить, значимо ли различаются коэффициенты периода выборки и периода предсказания. Для реализации этого теста сначала необходимо оценить раздельно уравнения регрессии для периода выборки и периода предсказания, а затем — совместно для этих двух периодов. После этого нужно проверить, значимо ли улучшается качество уравнения при разделении общего периода оценки регрессии на период выборки и период предсказания. Подтверждение этой гипотезы может служить свидетельством того, что коэффициенты регрессии нестабильны.

Пример

При оценивании функции спроса на продукты питания с использованием наблюдений за 1959–1979 гг. в качестве периода выборки, а за 1980–1983 гг. — в качестве периода предсказания, суммы квадратов отклонений для периода выборки, периода предсказания и совмещенного периода равнялись 0,0052; 0,0002 и 0,0070 соответственно. Оценка отдельных уравнений регрессии для двух подпериодов ведет к утрате трех степеней свободы, и число степеней свободы, остающееся после оценивания шести параметров (двух постоянных членов, двух коэффициентов при $\log x$, двух коэффициентов при $\log p$), равно 19. В итоге мы получаем следующую F-статистику, распределенную с 3 и 19 степенями свободы:

$$F(3,19) = \frac{(0{,}0070 - [0{,}0052 + 0{,}0002]) / 3}{(0{,}0052 + 0{,}0002) / 19} = 1{,}88. \qquad (10.83)$$

Критическое значение F-статистики с таким числом степеней свободы при 5-процентном уровне значимости равно 3,13, что позволяет нам сделать вывод об отсутствии явной нестабильности коэффициентов.

Оценка качества прогнозов

Задача становится гораздо более сложной, когда возникает необходимость оценить будущие значения независимых переменных до того, как спрогнозировать значения зависимой переменной, — это типичная ситуация при построении прогнозов. Как правило, мало что можно сказать о свойствах прогнозов, и это делает невозможной разработку формальных тестов на стабильность. Обычно прибегают к оценкам с позиций здравого смысла или к каким-то простым, очевидным критериям. Две наиболее популярные оценки — относительная ошибка прогноза и стандартная среднеквадратичная ошибка.

Относительная ошибка прогноза (*RFE*) определяется как $(\hat{y}_{T+p} - y_{T+p})/y_{T+p}$, т. е. как ошибка прогноза, деленная на действительное значение переменной. Ее можно применять как для оценки прогноза абсолютных значений, так и для оценки прогноза приростов. В первом случае используются сами значения $\hat{y}_{T+p}$ и y_{T+p}. Во втором случае вместо них используются предсказываемый и действительный прирост в прогнозный период $\Delta\hat{y}_{T+p}$ и Δy_{T+p}, где $\Delta\hat{y}_{T+p} = (\hat{y}_{T+p} - y_T)$ и $\Delta y_{T+p} = (y_{T+p} - y_T)$. Вывод для показателя абсолютных значений часто является слишком слабым. Если переменная y медленно изменяется во времени, то трудно сделать большую относительную ошибку при ее прогнозе. Соответствующий

Таблица 10.2

Прогнозируемая и действительная численность занятых по видам рабочей силы в Таиланде в 1976 г.

	1972 год (базовый)	*Уровни 1976 г.*			*Приросты 1976 г.*		
		$\hat{y}$	y	*RFE (%)*	$\Delta\hat{y}$	Δy	*RFE (%)*
Профессионально-технические специалисты	333	411	353	16	78	20	290
Управляющие	111	141	140	1	30	29	3
Служащие	224	293	241	22	69	17	306
Торговля и коммерция	1340	1790	1378	30	450	38	1084
Сельское хозяйство	11733	13293	13944	-5	1560	2211	-29
Транспорт и связь	319	408	360	13	89	41	117
Производство	1677	1890	1612	17	213	-65	оз*
Услуги	400	497	382	30	97	-18	оз
Всего	16137	18713	18410	2	2586	2273	14

** оз — ошибочный знак*

Источник: Pitayanon, 1987.

показатель приростов служит лучшим индикатором и часто является более адекватным в реальных задачах.

Оба варианта показателя проиллюстрированы в табл. 10.2 на примере построения прогноза численности занятых. Прогноз численности занятых — это попытка оценить, каков будет спрос на различные виды рабочей силы в заданный момент времени, с целью определить, сколько работников должны к этому времени получить соответствующую подготовку или образование. Такой прогноз регулярно делается плановиками в развивающихся странах, и он обычно приводит к большим ошибкам по причинам, связанным с особенностями рынка труда, которые здесь для нас несущественны.

Прогнозы были сделаны по видам рабочей силы, выделенным в соответствии с ее десятичной классификацией, на 1976 г. (с базовым 1972 г.). Как видно из табл. 10.2, прогнозы абсолютных значений выглядят вполне удовлетворительными — самая большая относительная ошибка прогноза равна 30%, что неудивительно, поскольку занятость не может сильно измениться всего за четыре года. Действительно, если бы в качестве прогноза на 1976 г. были взяты значения показателя в 1972 г., то самая большая относительная ошибка прогноза составляла бы всего 21%. С другой стороны, прогнозы приростов, которые гораздо более важны для политиков, мало чего стоят: два из них имеют неправильный знак, а большинство оставшихся имеют относительные ошибки, превышающие 100%.

Коэффициент Тейла (U)

Для многих видов оценивания критерий относительной ошибки прогноза является адекватным, но если вы захотите обобщить результаты оценивания с помощью одного числа, то средняя относительная ошибка прогнозов будет неудовлетворительным показателем, поскольку в ней большие положительные и отрицательные ошибки взаимно погашаются. Один возможный выход из этой ситуации — рассчитывать среднее для абсолютных значений относительных ошибок прогноза. Другой выход, предложенный Х. Тейлом (Theil, 1966), — рассчитывать среднеквадратичное значение ошибки прогноза приростов, обозначаемое через U:

$$U = \sqrt{\frac{\frac{1}{n}\sum(\Delta \hat{y}_{T+p}) - \Delta y_{T+p})^2}{\frac{1}{n}\sum(\Delta y_{T+p})^2}}. \qquad (10.84)$$

Преимущество данного показателя заключается в наличии двух естественных масштабирующих значений. Во-первых, значение показателя равно нулю, если сделан абсолютно точный прогноз, и, во-вторых, оно автоматически равно единице для «наивного» прогноза об отсутствии изменений. Если величина $\Delta\hat{y}_{T+p} = 0$ для каждого из прогнозов, числитель становится равным знаменателю, т. е. величине $(1/n)\Sigma(\Delta y_{T+p})^2$. Поскольку получаемый с помощью модели прогноз должен как минимум превосходить по качеству прогноз об отсутствии изменений, значение U должно лежать в границах между 0 и 1, и его близость к нулю будет свидетельствовать об относительном успехе сделанного прогноза.

Однако сравнение модельного прогноза с «наивным» прогнозом об отсутствии изменений — не очень сильный тест. Имеются и другие способы предсказания, которые менее тривиальны и как следствие дают лучшую базу для сравнения.

Один из них — прогнозировать значение прироста в этом году, равное действительному приросту в прошлом году, т. е. делать прогноз «о таком же приросте». Другой способ — приравнивать прирост среднему значению приростов в последние несколько (скажем, три или пять) лет. Затем можно рассчитать U-статистику как для прогнозной модели, так и для данного простого метода предсказания и сделать вывод об относительном успехе модели, если полученное значение статистики для нее окажется меньше.

В более сложной версии этого подхода с помощью регрессионного анализа может оцениваться лаговая структура простой базовой прогнозной модели, когда исследуется зависимость значения переменной от ее предыдущих значений с целью поиска наиболее подходящих весов. Поскольку предполагаемым преимуществом предлагаемой прогнозной модели является то, что она базируется на экономической теории, она должна давать лучшие прогнозы, чем стандартная авторегрессионная модель.

Однако нередко простые методы предсказания работают гораздо лучше, чем макроэкономические модели. Р. Купер сравнил функционирование шести главных макроэкономических моделей США и обнаружил, что получаемые с их помощью прогнозы уступают прогнозам, составляемым на основе авторегрессионной модели (Cooper, 1972). Ч. Нелсон использовал модель Бокса—Дженкинса (см. приложение 10.1) в качестве базы сравнения для оценки еще одной известной модели экономики США и получил такой же результат (Nelson, 1972). Оценка краткосрочных прогнозов для Великобритании, поступающих из различных организаций (Ash, Smyth, 1973), привела к тому же выводу. Стоит, однако, помнить, что макроэкономическое прогнозирование — одна из самых сложных задач прикладной экономической науки.

Упражнения

10.14. Оцените логарифмическую версию функции спроса на выбранный вами товар для периодов 1959—1979 и 1959—1983 гг. и проведите испытание с помощью теста Чоу на несостоятельность предсказания.

10.15. Оцените функцию спроса на данных за 1980—1983 гг. и, используя результаты упражнения 10.14, проведите испытание с помощью F-теста на стабильность коэффициентов.

Приложение 10.1

Метод Бокса—Дженкинса и анализ временных рядов

Целью эконометриста, работающего с временными рядами, является спецификация модели, которая была бы корректной с двух точек зрения: с точки

зрения включения в них соответствующих поведенческих уравнений и переменных и с точки зрения адекватной лаговой структуры внутри этих взаимосвязей. Иногда высказывается мнение, что лаговая структура модели должна определяться с помощью теоретической модели, а не просто как результат случайного поиска, но такое пожелание в большинстве случаев оказывается слишком оптимистичным. Теоретические модели в лучшем случае сами являются лишь приближенным описанием, и эмпирические эксперименты часто предшествуют теории.

Небольшая, но влиятельная группа исследователей впала в другую крайность, отстаивая мнение, что часто, особенно в макроэкономике, проблемы мультиколлинеарности, невольных подмен и т. д. совмещаются и делают почти невозможной правильную спецификацию поведенческих взаимосвязей и лаговой структуры модели. А модельер, сфокусировавший свое внимание на лаговой структуре, сможет сделать более качественные прогнозы.

Радикально сузив круг своих целей, они затратили громадные усилия, оценивая лаговую структуру модели с помощью небольшого числа параметров. Этот подход известен под названием «анализ временных рядов» или «метод Бокса—Дженкинса» (названный так по имени авторов первой работы Г. Бокса и Дж. Дженкинса [Box, Jenkins, 1972]; более простой текст, написанный экономистом, — работа Ч. Нелсона [Nelson, 1973]).

В стандартной модели с одной переменной искомое соотношение выводится только из текущего и лаговых значений зависимой переменной. Сначала для временного ряда рассчитывается первая разность или разность более высокого порядка для того, чтобы сделать его «стационарным», т. е. убрать тренд или другие свойства, которые делают распределение любого наблюдения зависящим от времени. После этого оценивается следующее уравнение регрессии:

$$\tilde{y}_t = \beta_0 + \beta_1 \tilde{y}_{t-1} + \ldots + \beta_p \tilde{y}_{t-p} + \gamma_0 \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \ldots + \gamma_q \varepsilon_{t-q}, \tag{10.85}$$

где $\tilde{y}_{t-s}$ — рассчитанные разности переменной y в период $(t-s)$; ε_t, ε_{t-1}, ... — независимо распределенные случайные члены с нулевым средним и постоянной дисперсией.

Как видно из уравнения (10.85), зависимость $\tilde{y}_t$ частично определена как авторегрессионный (AR) процесс (т. е. как зависимость переменной от своих прошлых значений), а с другой стороны — как скользящее среднее (MA) текущего и предыдущих значений случайного члена. Поэтому процесс в целом описывается как ARIMA (p, d, q) процесс, где p — порядок AR-части; q — порядок MA-части; d — порядок разностей, взятых из исходного ряда для достижения его стационарности.

Из уравнения (10.85) также видно, что оно совершенно не опирается на экономическую теорию. Это, однако, не означает, что оно *несовместимо* с экономической теорией. Например, если имеется модель из двух уравнений

$$y_t = \alpha + \beta_1 x_t + \beta_2 x_{t-1} + u_t; \tag{10.86}$$

$$x_t = \gamma + \delta y_{t-1} + v_t, \tag{10.87}$$

то в случае, когда не нужно брать разностей для достижения стационарности рядов, она может быть сведена к процессу ARIMA (2, 0, 1):

$$y_t = (\alpha + \beta_1\gamma + \beta_2\gamma) + \beta_1\delta y_{t-1} + \beta_2\delta y_{t-2} + u_t + \beta_1 v_t + \beta_2 v_{t-1}. \qquad (10.88)$$

Верно, тем не менее, и то, что этот подход не использует экономическую информацию, заключенную в модели. Единственное, чем может помочь экономическая информация, — это определить параметры p и q, выбор значений которых остается за исследователем.

Однако к досаде ортодоксальных экономистов прогнозы, сделанные с помощью моделей Бокса—Дженкинса в начале 1970-х гг., оказывались обычно не хуже, а часто и лучше прогнозов, полученных с помощью макроэкономических моделей (см.: Nelson, 1973). С тех пор была проделана огромная работа в обоих направлениях: с одной стороны, усложнялись сами эконометрические модели, а с другой — анализ временных рядов дополнялся многомерным анализом [когда правая часть уравнения (10.85) включает лаговые значения других переменных]. К сожалению, в последнее время не было проведено сравнения их прогнозной силы, которые позволили бы оценить продвижение в этих двух направлениях. В любом случае может оказаться совершенно необязательным строго придерживаться какой-либо одной точки зрения, поскольку наиболее вероятным исходом станет их сближение.

11

ОЦЕНИВАНИЕ СИСТЕМ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ

Если использовать МНК для оценивания параметров уравнения, которое является составной частью системы одновременных уравнений, то полученные оценки наверняка окажутся смещенными и несостоятельными, а статистические тесты — некорректными. Все это показано в первой части данной главы. Во второй части рассматриваются различные альтернативные процедуры оценивания, которые позволили бы преодолеть указанные затруднения.

11.1. Смещение при оценке одновременных уравнений

Ошибки измерения—не единственная возможная причина нарушения четвертого условия Гаусса—Маркова. Причиной может стать смещение, порождаемое системой одновременных уравнений, и этот случай лучше объяснить на примере.

Предположим, что вы хотите оценить параметры уравнения функции потребления в простой кейнсианской модели формирования доходов:

$$C_t = \alpha + \beta Y_t + u_t; \tag{11.1}$$

$$Y_t = C_t + I_t. \tag{11.2}$$

Модель описывает закрытую экономику без государственного вмешательства. В модели используются традиционные обозначения системы национальных счетов, где Y, C и I представляют совокупный выпуск, объем потребления и инвестиций соответственно. Здесь мы не принимаем во внимание концепцию Фридмена, поскольку рассматриваемая проблема является достаточно самостоятельной, и предполагаем, что уравнение (11.1) описывает поведенческую зависимость. В таком упрощенном виде это предположение не очень реалистично, но оно поможет нам в решении поставленной задачи.

После подстановки выражения (11.1) в (11.2) и преобразования мы сможем найти значение Y для любого момента времени:

$$Y_t = \frac{\alpha}{1-\beta} + \frac{I_t}{1-\beta} + \frac{u_t}{1-\beta}. \tag{11.3}$$

Первые два слагаемых в правой части уравнения знакомы каждому, кто имеет

даже поверхностное представление о кейнсианской теории формирования дохода. Эти слагаемые показывают, что совокупный уровень доходов зависит от постоянной составляющей объема потребления и от объема инвестиций. Если объем инвестиций возрастает на единицу, то совокупный доход увеличится на $1/(1-\beta)$ единиц. Это и есть знаменитый мультипликатор.

Здесь важно также заметить, что уровень совокупного дохода зависит и от величины u — случайного члена в уравнениях функции потребления. Как это происходит? Предположим, что в некоторый год в стране отдельные неэкономические причины вызвали увеличение объема потребления. Пусть какое-то важное событие вызвало рост общественных и личных расходов. Это будет отражено высоким положительным значением u в данном году, поскольку роль величины u и заключается в улавливании подобных воздействий. Поскольку объем потребления увеличился из-за таких необычно высоких расходов, объем выпуска также возрастет согласно базовому соотношению (11.2). Рост выпуска означает рост доходов, которые в свою очередь вызовут дополнительное увеличение объема потребления через переменную Y в функции потребления (11.1). Как следствие на такую же величину повысится и объем выпуска. Дополнительный прирост выпуска и, следовательно, доходов снова скажется на объеме потребления и т. д. Если u будет иметь отрицательные значения, то последствия окажутся аналогичными, только доходы и выпуск уменьшатся.

Описанный процесс представляет такой же эффект мультипликатора, как и в случае изменения объема инвестиций, и значение мультипликатора будет точно таким же: $1/(1-\beta)$. Отсюда — появление слагаемого $u/(1-\beta)$ в формуле (11.3). Если вы ставите перед собой единственную цель — увеличить выпуск и поднять уровень занятости, то для этого можно с одинаковым успехом расходовать деньги как на предметы роскоши, так и на инвестиции. Если вы не читали басню Б. Мандевиля «Ропчущий улей» (1705), перепечатанную позже как часть «Басен о пчелах», то советуем прочитать эту вещь.

Так или иначе, поскольку величина Y включает случайную составляющую $u/(1-\beta)$, она автоматически оказывается коррелированной со случайным членом в уравнении (11.1), и четвертое условие Гаусса—Маркова нарушается. Поэтому если попробовать оценить значения α и β с помощью МНК, то полученные оценки будут смещенными, а рассчитанные стандартные отклонения — некорректными.

О свойствах оценок на малых выборках мало что можно сказать. То, что происходит на больших выборках, зависит от поведения объясняющей переменной (переменных) модели. Далее в этой главе мы будем обычно предполагать, что дисперсии переменных и ковариации между ними на больших выборках стремятся к некоторым конечным пределам. Если это предположение выполняется, то оценки, полученные с помощью МНК, несостоятельны.

В рассматриваемой модели, если $\mathrm{Var}\,(I)$ на большой выборке стремится к пределу σ_I^2, то величина b будет стремиться к

$$\beta + \frac{(1-\beta)\sigma_u^2}{\sigma_I^2 + \sigma_u^2}, \tag{11.4}$$

и ошибка оценивания будет не нулевой (доказательство этого факта см. в приложении 11.1).

Содержательные экономические соображения позволяют нам предположить, что $0 \leq \beta \leq 1$, поэтому величина $(1 - \beta)$ будет положительной. Поскольку значения дисперсии всегда положительны, второе слагаемое в правой части формулы также будет положительным. Как следствие на больших выборках в данной модели величина b (оценка параметра β) окажется смещенной вверх.

Величина ошибки будет зависеть: 1) от отклонения β от единицы и 2) от отношения σ_u^2 к σ_I^2, т. е. от отношения дисперсии случайного члена к дисперсии объема инвестиций. Чем больше значения этих двух величин, тем серьезнее проблема. Предположим для примера, что отношение дисперсий равно $^1/_4$, а значение $\beta = 0{,}75$. В таком случае b будет стремиться к величине $0{,}75 + 0{,}25 \times (0{,}25/1{,}25)$, то есть к 0,80.

Проблема смещения, порождаемого системой одновременных уравнений, может быть разрешена путем замены МНК на другой метод оценивания. В следующих разделах мы обсудим три таких подхода. Все они — методы оценивания отдельного уравнения, в которых работа с каждым уравнением модели осуществляется самостоятельно. Системные методы, в которых все параметры уравнений оцениваются одновременно, в принципе более эффективны, однако мы оставляем их за рамками данной книги.

Что происходит, если значения дисперсии и ковариации объясняющих переменных не стремятся к некоторому конечному пределу? Предположим, например, что переменные имеют тренд и их дисперсии и ковариации между ними неограниченно возрастают. В таком случае МНК в итоге может оказаться состоятельным (работа Я. Кменты [Kmenta, 1984]). В разбираемой простой модели, если Var (I) неограниченно возрастает, то на больших выборках ошибка в оценке коэффициента регрессии исчезает и МНК обеспечивает состоятельную оценку. Однако даже в этом случае может оказаться более желательным использовать альтернативные методы оценивания, поскольку на малых выборках они имеют лучшие свойства.

Упражнение

11.1. В некоторой аграрной стране объем совокупного потребления обычно составляет 2000 единиц плюс случайная величина z, значение которой зависит от погоды. Среднее значение $z = 0$, стандартное отклонение — 100. Совокупный объем инвестиций изменяется согласно четырехлетнему циклу, начиная от 200, возрастая до 300 в следующем году, затем падая последовательно до 200 и 100, возвращаясь обратно к 200, и т. д. Совокупный доход Y равен сумме C и I. В таблице приведены данные о значениях C, I и Y за 20 лет (в качестве z использовалась нормально распределенная случайная величина с нулевым математическим ожиданием и единичным стандартным отклонением, умноженная на 100).

Традиционный экономист, используя приведенные данные, получит следующее уравнение регрессии для зависимости C от Y (в скобках приведены значения стандартных ошибок):

$$\hat{C} = 512 + 0{,}68\,Y; \qquad R^2 = 0{,}67;$$
$$(252)\ \ (0{,}11) \qquad F = 36{,}49.$$

t	C	I	Y	t	C	I	Y
1	1813	200	2013	11	1981	200	2181
2	1893	300	2193	12	2211	100	2311
3	2119	200	2319	13	2127	200	2327
4	1967	100	2067	14	1953	300	2253
5	1997	200	2197	15	2141	200	2341
6	2050	300	2350	16	1836	100	1936
7	2035	200	2235	17	2103	200	2303
8	2088	100	2188	18	2058	300	2358
9	2023	200	2223	19	2119	200	2319
10	2144	300	2444	20	2032	100	2132

Объясните, как получены данные результаты, несмотря на то что величина C совсем не зависит от Y. Объясните также, почему некорректны тесты для t-статистики и F-статистики.

11.2. Структурная и приведенная формы уравнений

Для удобства договоримся сначала о терминологии. В процессе оценивания параметров уравнений экономической модели важно различать эндогенные и экзогенные переменные. Приставки *эндо-* и *экзо-* обозначают относящееся соответственно к внутреннему и к внешнему. *Эндогенной* считается та переменная, значение которой определяется внутри модели. В модели формирования дохода, представленной уравнениями (11.1) и (11.2), C и Y являются эндогенными переменными, которые принимают свои значения в уравнении функции потребления и тождестве для совокупного дохода. *Экзогенной* является переменная, значение которой определяется вне модели и поэтому берется как заданное. В модели формирования дохода I — экзогенная переменная. Модель не объясняет, как получаются значения этой переменной, они просто используются как наперед заданные.

Приведенная классификация важна, поскольку она позволяет сказать, как *действительно* определяются значения эндогенных переменных. Модель формирования дохода, трактуемая буквально, говорит о том, что величина C зависит от Y и u, но это слишком упрощенное понимание. Учитывая тождество для совокупного дохода, так же верно (или так же неверно) было бы утверждать, что значение Y зависит от C. Для того чтобы выбраться из замкнутого круга, необходимо преобразовать уравнения и выразить как Y, так и C через их действительные детерминанты — I и u.

Это уже было сделано для Y в уравнении (11.3). Для того чтобы получить

аналогичное уравнение для C, подставим значение Y из (11.3) в (11.1) и после преобразования получим:

$$C_t = \frac{\alpha}{1-\beta} + \frac{\beta I_t}{1-\beta} + \frac{u_t}{1-\beta}. \qquad (11.5)$$

Уравнения, составляющие исходную модель, называются *структурными уравнениями* модели. Их можно разделить на две группы: *поведенческие уравнения*, описывающие эмпирические взаимосвязи между переменными, и уравнения-*тождества*. Уравнение (11.1) — пример поведенческого уравнения, а (11.2) — тождества. Единственное практическое различие между ними, с точки зрения эконометриста, заключается в том, что тождества не содержат каких-либо подлежащих оценке параметров, а также не включают случайного члена.

Уравнения, показывающие, как в действительности определяется значение эндогенных переменных, называются *уравнениями в приведенной форме*. Это уравнения, в которых эндогенные переменные выражены исключительно через экзогенные переменные и случайные составляющие. В разбираемом примере (11.3) и (11.5) являются уравнениями в приведенной форме.

Как быть, если одно (или более) структурное уравнение содержит в качестве объясняющих переменных лаговые значения эндогенных переменных? Предположим, что функция потребления имеет следующий вид:

$$C_t = \alpha + \beta_1 Y_t + \beta_2 C_{t-1} + u_t, \qquad (11.6)$$

как ее представляет, например, Т. Браун (Brown, 1952). Поскольку рассматриваемый период времени задан, значение C_{t-1} фиксировано и является предопределенным. В такой модели уравнения в приведенной форме отражают зависимость эндогенных переменных от *предопределенных переменных*, часть из которых являются экзогенными, значения других определяются в предшествующие периоды времени. Приведенная форма уравнения для Y_t в данном случае имеет вид:

$$Y_t = \frac{\alpha}{1-\beta_1} + \frac{I_t}{1-\beta_1} + \frac{\beta_2 C_{t-1}}{1-\beta_1} + \frac{u_t}{1-\beta_1}. \qquad (11.7)$$

Упражнения

11.2. Упрощенная модель закрытой экономики состоит из уравнений функции потребления, инвестиционной функции и тождества для национального дохода:

$$C_t = \alpha + \beta Y_t + u_t;$$

$$I_t = \delta + \varepsilon r_t + v_t;$$

$$Y_t = C_t + I_t + G_t,$$

где C_t — объем личных потребительских расходов в году t; I_t — объем инвестиций; G_t — совокупные государственные расходы; Y_t — валовой выпуск; r_t —

ставка процента в текущем году. Укажите, какие переменные в модели являются эндогенными, а какие — экзогенными.

11.3. В модель добавлены уравнение спроса на деньги и условие равновесия на денежном рынке:

$$M_t^d = \lambda + \mu Y_t + \theta r_t + w_t;$$
$$M_t^d = \overline{M}_t,$$

где M_t^d — спрос на деньги в году t; $\overline{M}_t$ — предложение денег, величина которого задана экзогенно. Установите, какие переменные являются эндогенными, а какие — экзогенными в такой расширенной модели.

11.3. Косвенный метод наименьших квадратов (КМНК)

Как мы убедились, попытка непосредственного оценивания параметров α и β уравнения функции потребления дает смещенные оценки, так как объясняющая переменная Y является эндогенной и частично зависит от u. Обратимся еще раз к (11.5) — приведенной форме уравнения для переменной C из исходной модели. Это уравнение может быть представлено в следующем виде:

$$C_t = \alpha' + \beta' I_t + u'_t, \tag{11.8}$$

где $\alpha' = \alpha/(1-\beta)$, $\beta' = \beta/(1-\beta)$ и $u'_t = u_t/(1-\beta)$.

Если использовать данные о величинах C и I для оценки параметров уравнения (11.8), то проблемы смещения, порождаемого одновременными уравнениями, не возникает. Объясняющей переменной является объем инвестиций, который предполагается экзогенным и, как следствие, не связанным с u'. В итоге случайный член u' удовлетворяет четвертому условию Гаусса—Маркова. Поэтому если вы оцените (11.8) с помощью МНК и получите:

$$\hat{C}_t = a' + b' I_t, \tag{11.9}$$

то a' будет несмещенной оценкой для α', а b' — для β'.

Возвращаясь к исходной форме уравнения, можно получить оценки a и b параметров α и β. Учитывая определения α' и β' в (11.8), можно выразить a' и b' через a и b следующим образом:

$$a' = \frac{a}{1-b} \quad \text{и} \quad b' = \frac{b}{1-b}. \tag{11.10}$$

Выражая a и b через a' и b', получим:

$$a = \frac{a'}{1+b'} \quad \text{и} \quad b = \frac{b'}{1+b'}. \tag{11.11}$$

Поскольку мы можем получить единственное выражение для a и b через оценки a' и b', уравнение называется *однозначно определенным (идентифицируемым)*. В следующем разделе мы рассмотрим случай, когда нельзя получить единственные исходные значения a и b, и такое уравнение называется *недоопределенным (неидентифицируемым)*, а также когда нельзя получить никакого решения, — в случае *переопределенного (сверхидентифицированного) уравнения*.

Пример

Проиллюстрируем наш подход с помощью эксперимента по методу Монте-Карло. Предположим, что реальная функция потребления представлена в виде:

$$C_t = 100 + 0{,}75Y_t + u_t, \tag{11.12}$$

а случайный член u равен умноженному на 50 случайному числу, взятому из выборки с нормальным распределением, нулевым математическим ожиданием и единичным стандартным отклонением. Значения C_t и Y_t получаются из уравнений в приведенной форме (11.3) и (11.5), которые в данном случае имеют вид:

$$C_t = 400 + 3I_t + 4u_t; \tag{11.13}$$

$$Y_t = 400 + 4I_t + 4u_t. \tag{11.14}$$

Взяв 20-летний период времени, предположим для простоты, что объем инвестиций равен 200 в первый год и возрастает на 10 в каждый последующий год, достигая 390 в 20-й год. Оценка уравнения регрессионной зависимости C от Y с помощью МНК дает:

$$\hat{C}_t = -84 + 0{,}87Y_t; \qquad R^2 = 0{,}99. \tag{11.15}$$
$$\text{(с. о.)}\ (38)\quad (0{,}02)$$

Как видим, получаемая оценка β слишком велика по сравнению с ожидаемым ее значением. Оценка α вообще имеет противоположный знак.

Проверим, насколько соответствуют полученные результаты выражению (11.4) для смещения на больших выборках. Значение Var $(I) = 3325$, а $\sigma_u^2 = 2500$, поэтому на больших выборках[1] имеем:

$$b \to 0{,}75 + \frac{(1 - 0{,}75) \times 2500}{3325 + 2500} = 0{,}75 + 0{,}11 = 0{,}86. \tag{11.16}$$

Как видим, в данном конкретном случае значение оценки на малой выборке весьма близко к значению, получаемому на большой выборке.

Используем теперь КМНК на тех же данных. Уравнение регрессионной зависимости C_t от I_t может быть оценено следующим образом:

$$\hat{C}_t = 167 + 3{,}84I_t; \qquad R^2 = 0{,}53. \tag{11.17}$$
$$\text{(с. о.)}(258)\ (0{,}86)$$

Используя (11.11), мы получим функцию потребления

$$\hat{C}_t = 34 + 0{,}79Y_t, \tag{11.18}$$

которая гораздо ближе к истинной модели. Тем не менее вы можете продолжать сомневаться в преимуществе КМНК, поскольку он дал меньшее значение коэффициента R^2 (0,53 вместо 0,99) и большие стандартные ошибки.

Однако высокое значение коэффициента R^2 в исходном уравнении является

[1] Мы предполагаем, что значение Var (I) остается постоянным на больших выборках; это верно, например, когда в последующие годы объем инвестиций принимает одно из своих 20 предыдущих значений с одинаковой вероятностью.

неизбежным. Даже если между C и Y *нет* никакой экономической связи, вы получите большое значение коэффициента R^2 при построении уравнения регрессионной зависимости C от Y. Причина заключается в том, что расходы на потребление составляют значительную часть совокупных доходов, и регрессия между C и Y не сильно отличается от регрессионной зависимости C от C.

Что касается стандартных ошибок, то в любом случае нарушения условий Гаусса—Маркова они рассчитываются неточно. Если вы определили 99-процентный доверительный интервал для b, используя оценку и стандартную ошибку, полученные на основе (11.15), но не принимая во внимание проблемы, вызванные смещением, то может оказаться, что этот интервал даже не включает истинное значение.

Конечно, слишком поспешно делать подобные обобщения на основе единственного эксперимента. В табл. 11.1 приведены результаты проведения экспериментов на десяти различных наборах случайных чисел для 20 наблюдений переменной u. В левой части табл. 11.1 показаны результаты построения уравнения регрессии между C и Y с помощью МНК. В середине таблицы даны оценки для приведенной формы уравнения регрессионной зависимости C от I. В правой части таблицы, производной от ее средней части, приведены соответствующие оценки функции потребления, полученные с помощью КМНК. Анализ таблицы показывает, что оценки, рассчитанные на основе КМНК, почти всегда лучше оценок на базе МНК и что оценки МНК параметра β близки к значению, полученному в уравнении (11.16).

Упражнения

11.4. Два исследователя пришли к выводу, что следующая простая модель формирования дохода применима для описания некоторой закрытой экономики:

$$C_t = \alpha + \beta Y_t + u_t;$$
$$Y_t = C_t + I_t,$$

где Y, C и I — совокупный доход, объем потребления и инвестиций соответственно; u — случайный член. Используя одинаковые временные ряды для Y, C и I, один исследователь построил уравнение регрессионной зависимости C от I, другой — регрессионной зависимости Y от I, и они получили следующие результаты:

$$\hat{C}_t = 4120 + 4{,}0I_t;$$

$$\hat{Y}_t = 4120 + 5{,}0I_t.$$

Покажите, что оба подхода дают одинаковые оценки α и β.

11.5. Обоснуйте математически, почему в предыдущем упражнении полученные результаты *должны* быть одинаковыми.

Таблица 11.1

	Метод наименьших квадратов ($\hat{C} = a + bY$)			Косвенный метод наименьших квадратов				
				Приведенная форма ($\hat{C} = a' + b'I$)			Функция потребления $\left(\hat{C} = \frac{a'}{1+b'} + \frac{b'}{1+b'}Y\right)$	
	a	b	R^2	a'	b'	R^2	$\frac{a'}{1+b'}$	$\frac{b'}{1+b'}$
1	–84	0,87	0,99	167	3,84	0,53	34	0,79
	(38)	(0,02)		(258)	(0,86)		(59)	(0,04)
2	–62	0,86	0,97	599	2,50	0,34	239	0,71
	(56)	(0,03)		(248)	(0,83)		(111)	(0,07)
3	–65	0,85	0,96	631	1,92	0,26	216	0,66
	(61)	(0,04)		(229)	(0,76)		(134)	(0,09)
4	–38	0,84	0,98	433	2,97	0,49	109	0,75
	(47)	(0,03)		(213)	(0,71)		(73)	(0,04)
5	–28	0,84	0,99	128	4,13	0,77	25	0,81
	(29)	(0,02)		(162)	(0,54)		(34)	(0,02)
6	–86	0,87	0,98	320	3,22	0,43	76	0,76
	(44)	(0,03)		(262)	(0,87)		(78)	(0,05)
7	–71	0,86	0,99	117	4,19	0,61	23	0,81
	(35)	(0,02)		(237)	(0,79)		(49)	(0,03)
8	–72	0,79	0,95	966	1,26	0,12	427	0,56
	(79)	(0,05)		(240)	(0,80)		(255)	(0,16)
9	–65	0,86	0,99	–34	4,51	0,72	–6	0,82
	(28)	(0,02)		(197)	(0,66)		(35)	(0,02)
10	–34	0,84	0,99	296	3,48	0,61	66	0,78
	(39)	(0,02)		(197)	(0,66)		(54)	(0,03)

11.4. Инструментальные переменные (ИП)

Как было показано в главе 8, проблема коррелированности объясняющей переменной со случайным членом может быть разрешена с помощью метода *инструментальных переменных* (ИП). Здесь, хотя и в силу других причин, мы сталкиваемся с такой же проблемой, и для ее решения в принципе можно использовать аналогичный подход.

Для применения данного метода необходимо сначала найти подходящую инструментальную переменную, которая обладала бы следующими свойствами: 1) она должна коррелировать, желательно тесно, с неудачной объясняющей переменной (в данном случае с Y); 2) она не должна коррелировать со случайным членом. Так получается, что модель сама предоставляет нам необходимую переменную. Величина I_t коррелирует с Y_t, поскольку Y_t зависит от

I_t, согласно приведенной форме уравнения (11.3), и I_t не коррелирует с u_t, поскольку является экзогенной переменной. Оценка β с помощью инструментальной переменной I_t определяется как

$$b_{\text{ИП}} = \frac{\text{Cov}(I, C)}{\text{Cov}(I, Y)}. \tag{11.19}$$

Можно показать, что полученная оценка $b_{\text{ип}}$ эквивалентна $b_{\text{кмнк}}$ — оценке β с помощью КМНК. Возвращаясь к (11.11) и учитывая, что b' рассчитывается как Cov (I, C)/Var (I), мы получим:

$$b_{\text{КМНК}} = \frac{b'}{1+b'} = \frac{\dfrac{\text{Cov}(I,C)}{\text{Var}(I)}}{1+\dfrac{\text{Cov}(I,C)}{\text{Var}(I)}} = \frac{\text{Cov}(I,C)}{\text{Var}(I)+\text{Cov}(I,C)} = \frac{\text{Cov}(I,C)}{\text{Var}(I,Y)} = b_{\text{ИП}}, \tag{11.20}$$

поскольку Cov (I, Y) может быть переписана как Cov (I,[I + C]) и далее преобразована в Var (I) + Cov (I, C).

Описанное правило применимо и в общем случае. Если уравнение в модели одновременных уравнений однозначно определено, метод ИП позволяет получить те же самые оценки коэффициентов, что и КМНК, если экзогенные переменные модели используются как инструментальные переменные. Поэтому КМНК можно рассматривать как частный случай метода ИП.

Пример

Если обратиться к модели в эксперименте по методу Монте-Карло, построенной для иллюстрации применения КМНК, и использовать вместо этого метод ИП, взяв I как инструментальную переменную для Y, то мы получим в точности те же оценки α и β, что и в последних двух столбцах табл. 11.1.

Свойства оценок, полученных методом ИП и КМНК

Анализируя КМНК, мы показали, как можно получить оценки структурных параметров из оценок коэффициентов уравнений в приведенной форме, но ничего не сказали о свойствах этих оценок: являются ли они несмещенными, состоятельными и т. д., не показали, как рассчитать их стандартные ошибки.

Теперь, когда показано, что в случае однозначной определенности КМНК эквивалентен методу ИП, мы можем ответить на все эти вопросы, обратившись к разделу 8.4. И хотя КМНК позволяет получить несмещенные оценки параметров уравнений в приведенной форме (что обеспечивается выполнением традиционных условий Гаусса—Маркова), нельзя делать какие-либо выводы об оценках коэффициентов структурных уравнений, рассчитанных с помощью КМНК и метода ИП на малых выборках. Приняв некоторые предпосылки, можно показать, что эти оценки состоятельны, и вывести выражения для их стандартных ошибок, применимые на больших выборках. В частности, в случае уравнения

с одной объясняющей переменной величина дисперсии распределения b с ростом числа наблюдений стремится к выражению (8.38). (Применительно к рассматриваемой модели в качестве x используется переменная Y, в качестве z — переменная I.)

Обычно, однако, на практике выборка оказывается не такой большой, чтобы можно было положиться на полученные результаты, и остается просто надеяться, что они приблизительно верны. Если для данной модели вам действительно необходимо исследовать свойства оценок, полученных с помощью КМНК и метода ИП, то можно выполнить соответствующие эксперименты по методу Монте-Карло.

Используя эксперименты по методу Монте-Карло, можно проверить состоятельность оценок на основе МНК и метода ИП, объединив 10 выборок в одну большую выборку из 200 наблюдений. Построение уравнения регрессии между C_t и Y_t методом ИП с использованием I_t как инструментальной переменной дает оценку предельной склонности к потреблению 0,76, что очень близко к действительной ее величине. Используя КМНК, мы оцениваем регрессионную зависимость C_t от I_t и получаем коэффициент при этой переменной, равный 3,20. С его помощью мы рассчитываем предельную склонность к потреблению, равную 0,76, т. е. имеем точно такое же значение, как и для оценки методом ИП.

11.5. Неидентифицируемость

Рассмотрим теперь несколько более сложную модель, состоящую из двух поведенческих уравнений. Допустим, что предложение товара на душу населения (y_d) и спрос на него (y_s) задаются следующими уравнениями:

$$y_d = \alpha + \beta p + \gamma x + u_d; \tag{11.21}$$

$$y_s = \delta + \varepsilon p + u_s, \tag{11.22}$$

где p — цена товара; x — доход на душу населения; u_d и u_s — случайные члены с дисперсиями $\sigma^2_{u_d}$ и $\sigma^2_{u_s}$ соответственно и выборочной ковариацией $\sigma^2_{u_d u_s}$. Переменная x предполагается экзогенной, p и y являются эндогенными, и их значения определяются в процессе установления рыночного равновесия. Когда рынок находится в равновесии, $y_d = y_s = y$. Выразив p и y через x, u_d и u_s, мы получим уравнения в приведенной форме:

$$p = \frac{\alpha - \delta}{\varepsilon - \beta} + \frac{\gamma}{\varepsilon - \beta} x + \frac{u_d - u_s}{\varepsilon - \beta}; \tag{11.23}$$

$$y = \frac{\alpha\varepsilon - \beta\delta}{\varepsilon - \beta} + \frac{\gamma\varepsilon}{\varepsilon - \beta} x + \frac{\varepsilon u_d - \beta u_s}{\varepsilon - \beta}. \tag{11.24}$$

Как видим, p зависит от u_d, поэтому использование МНК для уравнения (11.21) приведет к смещенным и несостоятельным оценкам. Переменная p зависит также от u_s, поэтому МНК даст смещенные и несостоятельные оценки для уравнения (11.22).

Перепишем для удобства уравнения в приведенной форме как

$$p = \alpha' + \beta' x + v_p; \tag{11.25}$$

$$y = \delta' + \varepsilon' x + v_y, \tag{11.26}$$

где

$$\alpha' = \frac{\alpha - \delta}{\varepsilon - \beta}; \quad \beta' = \frac{\gamma}{\varepsilon - \beta}; \quad \delta' = \frac{\alpha\varepsilon - \beta\delta}{\varepsilon - \beta}; \quad \varepsilon' = \frac{\gamma\varepsilon}{\varepsilon - \beta}, \tag{11.27}$$

и

$$v_p = \frac{u_d - u_s}{\varepsilon - \beta}; \quad v_y = \frac{\varepsilon u_d - \beta u_s}{\varepsilon - \beta}, \tag{11.28}$$

а v_p и v_y — составные случайные члены в приведенных уравнениях.

Рассмотрим теперь, можно ли использовать метод ИП или КМНК для получения состоятельных оценок коэффициентов. Начнем с первого из методов.

Метод инструментальных переменных

В нашей модели x — единственная экзогенная переменная, и в принципе ее можно использовать как инструментальную переменную вместо p, поскольку p зависит от x. Именно это мы и сделаем для уравнения предложения.

Однако в случае уравнения спроса решение оказывается невозможным. Переменная x уже присутствует в правой части уравнения, поэтому мы не можем использовать ее как инструментальную переменную вместо p. Если мы попробуем сделать это, то получим совершенную мультиколлинеарность. И что еще хуже, бесполезно искать подходящую инструментальную переменную за пределами модели. Как видно из (11.23), p является линейной функцией от x и составного случайного члена. Использование метода ИП на больших выборках ослабляет воздействие случайного члена. Поскольку правая часть уравнения спроса включает как x, так и линейную функцию от x, в пределе все равно проявляется совершенная мультиколлинеарность. В итоге мы можем получить оценки δ и ε, но не α, β или γ.

Косвенный метод наименьших квадратов

Тот же самый результат мы получим с помощью КМНК. Предположим, что мы применили МНК для оценивания параметров приведенной формы уравнений и имеем:

$$\hat{p} = a' + b'x; \tag{11.29}$$

$$\hat{y} = d' + e'x. \tag{11.30}$$

Предположим, что получены следующие оценки:

$$\hat{p} = 2{,}0 + 0{,}02x; \tag{11.31}$$

$$\hat{y} = 8{,}0 + 0{,}06x. \tag{11.32}$$

Используя (11.27), выведем следующие соотношения между оценками параметров уравнений в приведенной форме и оценками параметров уравнений в структурной форме:

$$a' = \frac{a-d}{e-b}; \quad b' = \frac{c}{e-b}; \quad d' = \frac{ae-bd}{e-b}; \quad e' = \frac{ce}{e-b}. \tag{11.33}$$

В численном примере для расчета структурных коэффициентов мы располагаем следующими уравнениями:

$$\frac{a-d}{e-b} = 2{,}0; \quad \frac{c}{e-b} = 0{,}02; \quad \frac{ae-bd}{e-b} = 8{,}0; \quad \frac{ce}{e-b} = 0{,}06. \tag{11.34}$$

Настораживает то, что имеется пять неизвестных, а уравнений — всего лишь четыре. Однако мы можем достичь некоторых результатов.

Во-первых, мы можем получить оценку e из второго и четвертого соотношений (11.34):

$$\frac{e'}{b'} = \frac{\dfrac{ce}{e-b}}{\dfrac{c}{e-b}} = e. \tag{11.35}$$

Следовательно, в нашем численном примере $e = 0{,}06/0{,}02 = 3$.

Во-вторых, хотя это и менее очевидно, первое и третье соотношения (11.33), а также оценка e дают возможность получить оценку d:

$$d' - ea' = \frac{ae-bd}{e-b} - \frac{ae-de}{e-b} = \frac{de-bd}{e-b} = d. \tag{11.36}$$

Следовательно, в нашем численном примере $d = 8{,}0 - (3 \times 2{,}0) = 2{,}0$. Это позволяет получить следующую оценку уравнения предложения:

$$\hat{y}_s = 2{,}0 + 3{,}0p. \tag{11.37}$$

Однако получить однозначные оценки a, b и c оказывается невозможным. У нас осталось два уравнения и три неизвестных. Можно, например, задать произвольное значение c, а затем найти значения a и b, но полученное решение будет, очевидно, непригодным. Проблема заключается в том, что связь между параметрами уравнений в структурной и приведенной формах слишком гибка.

Через оценки параметров уравнений в приведенной форме мы можем получить однозначные решения для d и e, но не для a, b или c. Это позволяет сделать вывод, что уравнение предложения определено, а уравнение спроса — недоопределено.

Что будет в случае, если спрос не зависит значимо от дохода? Это означает пренебрежимо малую величину параметра γ в уравнении (11.21) и как следствие параметров β' и ε' в (11.25) и (11.26) [см. определение b (11.27)]. Поэтому в уравнении (11.35) при расчете e мы будем делить оценку 0 на другую оценку 0, и полученный результат окажется бессмысленным. Следовательно, и оценка d, полученная из (11.36), будет бессмысленной. В итоге ни одно из уравнений не определено.

11.6. Предположим, что в долгосрочном периоде компании финансируют инвестиции I преимущественно из прибыли П, а объем получаемой прибыли зависит от инвестиций. На этой основе исследователь построил следующую модель корпоративного сектора экономики:

$$I_t = \alpha + \beta \Pi_t + u_t;$$

$$\Pi_t = \delta + \varepsilon I_t + \lambda I_{t-1} + v_t,$$

где индекс t обозначает текущий год, $(t-1)$ — предыдущий год, а u_t и v_t — случайные члены, не подверженные автокорреляции.

1) Определено ли какое-либо из уравнений? Объясните ваш ответ.

2) Что вы можете сказать о коэффициентах при переменных на основе приведенных ниже значений ковариации и дисперсии, рассчитанных на базе данных о промышленном секторе экономики за 25-летний период?

$$\text{Cov}\,(\Pi_t, I_t) = 57{,}0;\ \text{Var}\,(\Pi_t) = 113{,}0;$$

$$\text{Cov}\,(I_t, I_{t-1}) = 20{,}0;\ \text{Var}\,(I_t) = 30{,}0;$$

$$\text{Cov}\,(\Pi_t, I_{t-1}) = 45{,}0;\ \text{Var}\,(I_{t-1}) = 29{,}0.$$

(Величины Π_t, I_t и I_{t-1} измерены в миллиардах долларов в ценах 1985 г.)

11.7. Предположим, что в модели спроса и предложения товара как кривая спроса, так и кривая предложения сдвигаются со временем: первая — из-за изменения вкусов покупателей, вторая — из-за технического прогресса, делающего производство более дешевым. В этом случае структурные уравнения можно переписать в следующем виде:

$$y_{dt} = \alpha + \beta p_t + \gamma t + u_{dt};$$

$$y_{st} = \delta + \varepsilon p_t + \lambda t + u_{st};$$

$$y_{dt} = y_{st}.$$

Предположим, что уравнения в приведенной форме имеют вид:

$$\hat{p}_t = 1{,}2 + 0{,}04t;$$

$$\hat{y}_t = 7{,}6 - 0{,}38t.$$

Покажите, что уравнениям в приведенной форме соответствуют обе следующие модели:

(А) $y_{dt} = 10 - 2p_t - 0{,}3t;$

$y_{st} = 4 + 3p_t - 0{,}5t;$

и

(Б) $y_{dt} = 8{,}8 - p_t - 0{,}34t;$

$y_{st} = 5{,}2 + 2p_t - 0{,}46t.$

Комментарий: Только ли эти две модели в структурной форме соответствуют уравнениям в приведенной форме?

11.6. Сверхидентифицированность

Рассмотрим теперь модель, в которой спрос имеет временной тренд, скажем, потому что привычки медленно меняются со временем. Предположим, что спрос зависит также от дохода, и мы имеем:

$$y_{dt} = \alpha + \beta p_t + \gamma x_t + \rho t + u_{dt}; \tag{11.38}$$

$$y_{st} = \delta + \varepsilon p_t + u_{st}, \tag{11.39}$$

где t — переменная времени; ρ — коэффициент при ней. Исследуем, как и прежде, эффективность метода ИП и КМНК.

Метод инструментальных переменных

В модели две экзогенные переменные — x_t и t. Однако обе эти переменные присутствуют в уравнении спроса, и мы не можем использовать их как инструментальные для p_t. Как следствие уравнение спроса оказывается недоопределенным.

Однако в случае уравнения предложения у нас имеется широкое поле выбора. Модель, в которой экзогенных переменных, которые могут использоваться как инструментальные, больше, чем необходимо, называют переопределенной. Можно использовать как x_t, так и t в качестве инструментальных переменных для p_t. Полученные таким способом оценки δ и ε будут различаться, однако в обоих случаях они окажутся состоятельными. Какие из них использовать? Очевидно, те, которые более надежны, и поэтому на первом шаге можно рассчитать корреляцию переменных с p и выбрать ту из них, для которой корреляция выше. Однако можно сделать даже больше. Как мы покажем в следующем разделе, наилучшим решением в данном случае является применение так называемого *двухшагового метода наименьших квадратов* (ДМНК) и построение инструментальной переменной, которая является комбинацией x и t.

Косвенный метод наименьших квадратов

Приведенная форма уравнений имеет вид:

$$p_t = \frac{\alpha - \delta}{\varepsilon - \beta} + \frac{\gamma}{\varepsilon - \beta} x_t + \frac{\rho}{\varepsilon - \beta} t + \frac{u_{dt} - u_{st}}{\varepsilon - \beta}; \tag{11.40}$$

$$y_t = \frac{\alpha\varepsilon - \beta\delta}{\varepsilon - \beta} + \frac{\gamma\varepsilon}{\varepsilon - \beta} x_t + \frac{\rho\varepsilon}{\varepsilon - \beta} t + \frac{\varepsilon u_{dt} - \beta u_{st}}{\varepsilon - \beta}. \tag{11.41}$$

Переписав соответствующие уравнения регрессии как

$$\hat{p}_t = a' + b'x_t + c't; \tag{11.42}$$

$$\hat{y}_t = d' + e'x_t + f't, \tag{11.43}$$

мы увидим, что и e'/b', и f'/c' дают оценку ε. И хотя эти величины могут случай-

но совпасть, не существует выражения для e, которое одновременно удовлетворяло бы всем соотношениям между коэффициентами уравнений в приведенной и структурной формах. Так же нет удовлетворяющего всем соотношениям выражения для d. Уравнение предложения является переопределенным. В то же время уравнение спроса остается, как и прежде, недоопределенным.

Как следствие применение КМНК к рассматриваемой модели порождает сразу две проблемы: для уравнения предложения, поскольку связь между приведенной и структурной формами оказывается слишком тесной, и для уравнения спроса, поскольку эта связь становится слишком свободной. Однако, как и в случае применения метода ИП, эти проблемы асимметричны. В случае недоопределенности не хватает информации для фиксирования оценок параметров. Мы в принципе можем получить бесконечное число решений уравнений, не представляя, какое из них соответствует реальным значениям параметров. В случае переопределенности число решений больше одного, и, хотя эти решения различаются на малых выборках, все они состоятельны: разница между ними будет исчезать с ростом объема выборки. В разбираемом примере как e'/b', так и f'/c' являются состоятельными оценками ε. Проблема заключается в выборе между ними, если нужно делать такой выбор. Однако в случае применения ДМНК проблемы такого выбора не возникает.

11.7. Двухшаговый метод наименьших квадратов (ДМНК)

ДМНК как частный случай метода ИП

В предыдущем разделе уравнение предложения оказалось переопределенным, и сразу две переменные (x_t) и (t) могли использоваться как инструментальные для p_t. Однако вместо их раздельного применения можно предложить построить их комбинацию. Обозначим такую комбинацию как z_t, где

$$z_t = h_0 + h_1 x_t + h_2 t, \tag{11.44}$$

и требуется выбрать значения коэффициентов h_0, h_1 и h_2. В общем случае мы хотим, чтобы инструментальная переменная была как можно теснее коррелирована с заменяемой переменной, т. е. мы хотим выбрать такие h_1 и h_2, чтобы $r_{p,z}$ — коэффициент корреляции между p и z оказался максимальным.

На первый взгляд эта проблема может показаться сложной, но фактически она уже решена, поскольку можно использовать $\hat{p}_t$ из уравнения (11.42) вместо z_t. В процессе оценивания данного уравнения регрессии мы одновременно делали три вещи: 1) минимизировали сумму квадратов отклонений; 2) максимизировали значение коэффициента R^2; 3) максимизировали корреляцию между реальными и теоретическими значениями p (см. раздел 2.7). Именно это третье свойство мы и используем здесь.

В итоге мы имеем следующую двухшаговую процедуру:

1. Построить уравнения регрессии для уравнений приведенной формы и рассчитать теоретические значения эндогенных переменных.

2. Использовать теоретические значения как инструментальные переменные для действительных значений переменных.

Поскольку мы используем метод ИП, полученные таким образом оценки являются состоятельными, и можно вывести выражения для их стандартных отклонений на больших выборках. Однако, как всегда, мы мало что можем сказать об их свойствах на малых выборках.

ДМНК как метод «очищения» переменной

Вспомним, что причиной, по которой мы получили бы смещенные оценки, используя МНК для уравнения предложения, была корреляция между случайной составляющей переменной p_t и u_{st}. Отсюда следует, что если вам удастся очистить каждое наблюдение p_t от его случайной составляющей, то можно будет применить МНК.

К сожалению, невозможно точно выделить случайную составляющую в каждом наблюдении, однако мы можем получить ее оценку с помощью остатка для этого наблюдения, определяемого как $(p_t - \hat{p}_t)$. Если мы вычтем это выражение из исходных значений наблюдений вместо самих случайных составляющих, то получим $p_t - (p_t - \hat{p}_t)$, что равно $\hat{p}_t$. Следуя намеченному алгоритму, мы имеем альтернативную двухшаговую процедуру:

1. То же, что и раньше.
2. Использовать теоретические значения эндогенных объясняющих переменных вместо их действительных значений для оценки регрессии с помощью МНК.

Как это уже не раз случалось, можно показать, что полученные оценки в точности совпадают с оценками, рассчитанными на основе первой версии ДМНК (см. упражнение 11.11). Отсюда сразу же следует, что данная процедура эквивалентна предыдущей и дает состоятельные оценки несмотря на то, что мы вместо реальных значений случайных составляющих исключали остаточный член.

ДМНК в случае однозначной идентифицируемости

Как мы убедились, ДМНК может рассматриваться как способ конструирования наилучшей из возможных комбинаций инструментальных переменных в случае, когда в уравнении имеется избыток экзогенных переменных, которые могут использоваться как инструментальные. Поэтому совсем неудивительно будет обнаружить, что если такого избытка нет, то применение ДМНК не даст никаких преимуществ и приведет к точно таким же результатам, что и КМНК и метод ИП.

Покажем это на примере модели формирования дохода из раздела 11.1. Используя ДМНК для оценки уравнения функции потребления, рассчитаем регрессию приведенной формы уравнения для Y и получим $\hat{Y}$. Следуя первой версии ДМНК и используя $\hat{Y}$ как инструмент для Y, вычислим далее:

$$b_{\text{ДМНК}} = \frac{\text{Cov}(\hat{Y}, C)}{\text{Cov}(\hat{Y}, Y)}. \tag{11.45}$$

Если приведенную форму уравнения регрессии для Y записать как

$$\hat{Y} = g_0 + g_1 I, \tag{11.46}$$

то выражение (11.45) примет вид:

$$b_{\text{ДМНК}} = \frac{\text{Cov}(g_1 I, C)}{\text{Cov}(g_1 I, Y)} = \frac{\text{Cov}(I, C)}{\text{Cov}(I, Y)} = b_{\text{ИП}} \qquad (11.47)$$

[см. уравнение (11.19)], что также идентично $b_{\text{КМНК}}$ [см. уравнение (11.20)].

Это приводит к общему выводу о том, что в случае однозначной определенности уравнения все три метода являются эквивалентными. Заметим, в частности, что в случае однозначной определенности модели нет разницы между методом ИП и ДМНК. Предположим, что в данном уравнении три объясняющие переменные (x_1, x_2 и x_3) являются эндогенными и три экзогенные переменные (z_1, z_2 и z_3) могут выступать как инструментальные. Если используется метод ИП, то невозможно распределить роли инструментов между данными экзогенными переменными. Эти три инструментальные переменные используются совместно для трех эндогенных переменных, и оцененные регрессии окажутся теми же самыми, что и при применении ДМНК.

Упражнения

11.8. Как можно оценить параметры модели, описанной в упражнении 11.2? Как оценить ее параметры в расширенном виде, в каком она описана в упражнении 11.3?

11.9. Исследователь сформировал следующую простую макроэкономическую модель закрытой экономики:

$$C_t = \alpha + \beta Y_t + u_t;$$
$$Y_t = C_t + I_t + G_t,$$

где C — совокупное потребление; I — инвестиции; G — текущие расходы государственного сектора; Y — совокупный доход; u — случайный член. Переменные I и G могут рассматриваться как экзогенные. Исследователь располагает временными рядами годичных данных о значениях переменных за 25 лет. За время наблюдения значение C в среднем составляло примерно 70% от Y, I — примерно 20%, и G — примерно 10%. За время наблюдения дисперсия переменной I существенно превышала дисперсию G.

1) Объясните, почему применение МНК для оценки уравнения функции потребления дает несостоятельные оценки. В каком направлении, по-вашему, окажутся смещенными оценки α и β?

2) Исследователь оценивает уравнение функции потребления: (А) используя I как инструментальную переменную для Y; (Б) используя G как инструментальную переменную; (В) с помощью ДМНК. Соответствующие уравнения регрессии получились следующими (в скобках приведены стандартные ошибки):

$$\text{(А)} \quad \underset{(19,2)\ (0,13)}{\hat{C}_t = 1,7 + 0,69\, Y_t;} \qquad R^2 = 0,92;$$

$$\text{(Б)} \quad \underset{(25,5)\ (0,17)}{\hat{C}_t = -25,3 + 0,87\, Y_t;} \qquad R^2 = 0,85;$$

$$\text{(B)} \qquad \underset{(13,1)\ (0,09)}{\hat{C}_t = -4,0 + 0,72\ Y_t;} \qquad R^2 = 0,94.$$

В каждом из случаев автокорреляция не наблюдалась. Проанализируйте теоретические свойства каждого уравнения регрессии и установите, подтверждаются ли они полученными результатами.

11.10. Как бы вы предложили оценить в предыдущем упражнении приведенную форму уравнения для Y_t?[1]

11.11. В данном разделе были предложены две версии использования ДМНК для оценки уравнения предложения: 1) когда этот метод является частным случаем метода ИП, оценка равна $\text{Cov}(y, \hat{p})/\text{Cov}(p, \hat{p})$ и 2) когда метод используется как версия МНК для «очищения» переменных, оценка равна $\text{Cov}(y,\hat{p})\text{Var}(\hat{p})$. Докажите, что значение $\text{Cov}(p, \hat{p})$ равно значению $\text{Var}(\hat{p})$ и, следовательно, эти две версии эквивалентны.

11.8. Условие размерности для идентификации

Как мы уже убедились, в общем случае уравнение окажется идентифицируемым, если имеется достаточно экзогенных переменных, не включенных в само уравнение, которые можно использовать как инструментальные для всех эндогенных переменных уравнения. В полностью определенной модели будет столько уравнений, сколько имеется эндогенных переменных. Предположим, что число тех и других равно G. Максимальное число эндогенных переменных, которые могут появиться в правой части уравнения, равно $G-1$ (оставшаяся переменная — зависимая переменная этого уравнения). В таком случае нам необходимо по крайней мере $(G-1)$ экзогенных переменных, не включенных в это уравнение, которые использовались бы как инструментальные.

Предположим, однако, что в уравнение не включено j эндогенных переменных. Тогда нам понадобится лишь $(G-1-j)$ инструментальных переменных, то есть $(G-1-j)$ экзогенных переменных не должны быть включены в это уравнение. Однако общее число невключенных переменных остается прежним: j эндогенных переменных и $(G-1-j)$ экзогенных переменных составляют в сумме $G-1$.

Таким образом, мы приходим к общему выводу о том, что уравнение в модели с одновременными уравнениями наверняка окажется идентифицируемым, если в него не включено $(G-1)$ или более переменных. Если не включено точно $(G-1)$ переменных, оно, скорее всего, будет однозначно определенным, и в этом случае к одинаковым результатам приведет применение КМНК или метода ИП. Если не включено более $(G-1)$ переменных, уравнение будет переопределенным, и для его оценки используется ДМНК.

Это правило известно как условие размерности для идентификации. Здесь необходимо подчеркнуть, что данное условие является необходимым для идентификации, но вовсе не достаточным. Имеются случаи, которые мы не будем рассматривать здесь, когда уравнение является на самом деле недоопределенным, однако условие размерности для него выполняется.

[1] Используйте теоретические ограничения на коэффициенты.

Нулевые и не нулевые ограничения

Исключение $(G-1)$ переменных из уравнения может рассматриваться как утверждение, что коэффициенты при этих переменных в уравнении равны нулю. Представляя условие размерности формально, мы можем утверждать, что уравнение наверняка является идентифицируемым, если оно содержит $G-1$ (или больше) нулевых ограничений.

Однако это не единственный вид ограничений, который может приводить к идентифицируемости уравнения. Рассмотрим также три других вида ограничений.

Внешняя информация

При наличии внешней информации ее можно использовать для преодоления недоопределенности модели. Самый простой пример — это возможность получить независимую оценку одного из структурных параметров на другом множестве данных.

Рассмотрим снова вариант уравнений спроса и предложения из раздела 11.5 и предположим, что имеются приведенные формы уравнений регрессии (11.31) и (11.32). Из них мы получили четыре уравнения с пятью неизвестными (11.34). Уравнение предложения оказалось идентифицируемым, уравнение спроса — нет.

Однако допустим, что появилась возможность получить оценку коэффициента при показателе дохода из другого источника например, применяя регрессионный анализ к данным перекрестной выборки, как это описано в разделе 5.5. [До этого уравнения (11.31) и (11.32) были оценены на временных рядах.] Теперь имеются четыре уравнения с четырьмя неизвестными, и можно получить решение полностью, идентифицировав как уравнение предложения, так и уравнение спроса.

Предположим для примера, что на множестве структурных данных вы получили оценку c, равную 0,1. Кроме того, уже были оценены ранее $d=2$ и $e=3$ [см. уравнения (11.35) и (11.36)]. Теперь, зная величину c, можно использовать первые две части уравнения (11.34) для расчета a и b:

$$c/(e-b) = 0{,}1/(3-b) = 0{,}02; \tag{11.48}$$

$$(a-d)/(e-b) = (a-2)/(3-b) = 2{,}0. \tag{11.49}$$

Из первого уравнения можно получить $b=-2$, из второго $a=12$. В итоге оцененные структурные уравнения имеют вид:

$$\hat{y}_d = 12 - 2p + 0{,}1x; \tag{11.50}$$

$$\hat{y}_s = 2 + 3p. \tag{11.51}$$

Описанный подход скрывает две опасности, о которых всегда следует помнить. Во-первых, точность внешней оценки определяет точность получаемых с ее помощью оценок параметров. Во-вторых, имеется риск того, что значение коэффициента для внешней оценки отличается от его значения в модели. Обе эти проблемы рассматриваются в разделе 5.5.

Соотношение между коэффициентами

В некоторых случаях неидентифицируемая модель может быть идентифицирована заданием соотношения между структурными коэффициентами. Это можно объяснить на другом примере с использованием модели спроса и предложения. Предположим, что продавцы товара облагаются специальным налогом T, который они должны платить из выручки. Уравнение спроса остается неизменным, если переменная p обозначает рыночную цену на товар. Однако уравнение предложения изменяется под воздействием размера налога:

$$y_d = \alpha + \beta p + u_d; \tag{11.52}$$

$$y_s = \delta + \varepsilon p + \lambda T + u_s, \tag{11.53}$$

и λ, как ожидается, принимает отрицательное значение.

Прежде чем рассуждать далее, заметим, что уравнение спроса будет идентифицируемо, поскольку переменная T не включена в него и может выступать как инструментальная для p (мы предполагаем, что значение T изменялось во временном периоде, представленном выборкой данных), тем не менее уравнение предложения является неидентифицируемым. В то же время мы можем улучшить спецификацию модели. Вполне обоснованным является предположение о том, что продавцы товара реагируют на сумму, которую они получают после уплаты налога, т. е. на $(p - T)$, и уравнение (11.53) может быть переписано в виде:

$$y_s = \delta + \varepsilon (p - T) + u_s. \tag{11.54}$$

Другими словами, мы ввели ограничение $\lambda = -e$. Это сделало уравнение предложения идентифицируемым. Если использовать КМНК для оценивания исходной модели, то соотношения в приведенной форме, выражающие y и p через T, представляли бы четыре уравнения с пятью неизвестными. Введенное ограничение добавляет еще одно уравнение, и в итоге все структурные параметры могут быть однозначно оценены.

При использовании метода ИП можно рассматривать новую версию модели как систему из четырех уравнений:

$$y_d = \alpha + \beta p + u_d; \tag{11.55}$$

$$y_s = \delta + \varepsilon p^* + u_s; \tag{11.56}$$

$$p^* = p - T; \tag{11.57}$$

$$y_d = y_s, \tag{11.58}$$

где p^* — цена, получаемая продавцом товара, а уравнение (11.57) является тождеством. Переменная T не включена в уравнение спроса, поэтому она может использоваться как инструментальная для p. Точно так же эта переменная не включена в уравнение предложения, поэтому она может использоваться как инструментальная для p^*. В итоге оба уравнения оказываются определенными.

Не нулевое ограничение, как и нулевое ограничение, позволяет исключить одну объясняющую переменную из уравнения. Если эта переменная эндогенная, то уже не нужно искать для нее инструментальную переменную. Если эта переменная экзогенная, то она освобождается на роль инструментальной для одной из эндогенных переменных, оставшихся в уравнении.

Ограничения на распределение случайных членов

В модели спроса и предложения мы считали, что случайные члены u_d и u_s имеют дисперсию $\sigma^2_{u_d}$ и $\sigma^2_{u_s}$ соответственно, ковариация между ними равна $\sigma_{u_d u_s}$. Дисперсия случайных членов v_y и v_p в приведенной форме уравнений является линейной функцией этих величин. Если теоретические соображения позволяют наложить ограничения, включающие значения дисперсии и ковариации, в уравнения в структурной форме, это преобразуется в ограничения на их аналоги в приведенной форме, что может быть использовано для получения дополнительного соотношения между оценками структурных параметров.

Предположим, например, что имеется основание утверждать, что случайные члены в уравнениях спроса и предложения распределены независимо друг от друга, т. е. что значение $\sigma_{u_d u_s} = 0$. После некоторых преобразований можно показать, что это приводит к соотношению:

$$\sigma^2_{v_y} + (\beta\varepsilon)\sigma^2_{v_p} - (\beta + \varepsilon)\sigma_{v_y v_s} = 0. \qquad (11.59)$$

Если подставить в данное уравнение оценки дисперсий для v_y и v_s и ковариации между ними, полученные для уравнений регрессии приведенной формы, то получим простое соотношение, включающее оценки β и ε, которое поможет идентифицировать до этого неидентифицируемое уравнение (в качестве практического примера см. работу Я. Кменты [Kmenta, 1986, pp. 678–681]).

Как узнать, какие предположения необходимо сделать об экзогенных переменных?

Очевидно, возникает большое желание выявить экзогенные переменные, которые появляются в одних уравнениях и не встречаются в других в модели с одновременными уравнениями. Например, в модели из двух уравнений важно выявить одну экзогенную переменную для первого уравнения, которая не появляется во втором, и другую переменную для второго уравнения, которая не встречается в первом; в таком случае модель окажется однозначно идентифицируемой.

Возможно, что экономический смысл модели заставляет включать в нее такие переменные в первую очередь. И чем их будет больше — тем лучше. Однако если модель в своей исходной формулировке не содержит достаточное число экзогенных переменных, вполне естественно остановиться и подумать о выявлении дополнительных экзогенных переменных. Обычно это оказывается не так сложно. Приложив немного воображения, можно выявить достаточное число переменных для идентификации каждого уравнения, даже для его сверхидентификации. Затем использовать метод ИП или ДМНК для оценки параметров.

Однако насколько все это окажется удачным? Ответ на этот вопрос может быть дан на двух уровнях. Во-первых, включение новой переменной (пере-

менных) может базироваться лишь на вашем пожелании, и это выяснится, когда вы получите результаты оценивания регрессии, где стандартные ошибки окажутся значительными по сравнению с их коэффициентами, а t-статистики — незначимыми. Во-вторых, даже если использование новой переменной как инструментальной приводит к значимым результатам, это может быть следствием ошибочных установок. Рассмотрим следующую простую модель спроса и предложения:

$$y_d = \alpha + \beta p + u_d; \tag{11.60}$$

$$y_s = \delta + \varepsilon p + u_s. \tag{11.61}$$

В представленном виде оба уравнения являются недоопределенными. Предположим теперь, что исследователь установил причины включения временного тренда в уравнение предложения. При использовании времени как инструментальной переменной для p уравнение спроса будет определено, и в итоге могут быть получены оценки α и β. Чтобы подкрепить этот пример, предположим, что уравнение спроса оценено в виде:

$$\hat{y} = 9{,}0 - 2{,}0p. \tag{11.62}$$

Теперь допустим, что другой исследователь полагает, что воздействию временного тренда подвержен спрос. Если этот исследователь оказывается прав, то определено будет уравнение предложения, и время должно использоваться как инструментальная переменная для p в этом уравнении. Что получится, если для оценки берутся те же самые данные? Исследователь получит точно такой же результат:

$$\hat{y} = 9{,}0 - 2{,}0p. \tag{11.63}$$

Оба исследователя используют выражение $\mathrm{Cov}\,(y, t)/\mathrm{Cov}\,(p, t)$ для расчета коэффициента при переменной p, однако первый исследователь считает, что оценивает β, а второй — ε. Точно так же оба используют одинаковое выражение для расчета в уравнениях свободного члена. Поскольку получаемые результаты должны оказаться одинаковыми, нет никакого статистического основания для различения гипотез исследователей относительно спецификации модели. Можно привлечь лишь содержательные соображения. В данном случае они могут оказаться в пользу первого исследователя, который ожидает получить отрицательный коэффициент при p, но есть вероятность, что ни один из них не прав или что оба уравнения содержат временной тренд.

Упражнения

11.12. Спрос на товар в некоторой стране (q_d), его внутреннее предложение (q_s) и импорт этого товара (q_m) заданы следующими уравнениями:

$$q_d = \alpha_0 + \alpha_1 p + \alpha_2 Y + u_d;$$

$$q_s = \beta_0 + \beta_1 p + u_s;$$

$$q_m = \gamma_0 + \gamma_1 p + \gamma_2 w + u_m,$$

где p — цена товара на внутреннем рынке; w — цена товара на мировом рынке; Y — совокупный доход страны; u_d, u_s, u_m — случайные члены, распределен-

ные независимо друг от друга. Все переменные имеют индекс t, опущенный для удобства. В каждый момент времени рынок находится в равновесии:

$$q_d = q_s + q_m.$$

Имеются временные ряды значений каждой из переменных за 25 лет.

1) Объясните, почему попытка оценить эти три уравнения с помощью МНК приведет к получению несостоятельных оценок.

2) Как вы считаете, в каком направлении будет смещена оценка β_1, полученная с помощью МНК? Обоснуйте ваш ответ.

3) Объясните, возможно ли получение состоятельных оценок коэффициентов данных трех уравнений, и опишите ваши действия для достижения этого результата.

11.13. Пусть уравнения регрессии в приведенной форме в модели, задаваемой (11.52) и (11.53), имеют вид:

$$\hat{p} = 40 + 0{,}6T;$$

$$\hat{y} = 70 - 1{,}2T.$$

Выведите оценки α, β, δ и ε, используя ограничение $\lambda = -\varepsilon$.

11.9. Идентификация относительно стабильных зависимостей

Возможны случаи, когда вы можете оценить уравнение на основе некоторых предварительных предпосылок о его случайном члене. Это можно объяснить, возвращаясь к модели спроса и предложения, на этот раз в ее простейшем виде:

$$y_{d_t} = \alpha + \beta p_t + u_{d_t}; \tag{11.64}$$

$$y_{s_t} = \delta + \varepsilon p_t + u_{s_t}. \tag{11.65}$$

Здесь нет экзогенных переменных, и поэтому никакое из уравнений не идентифицируемо. Уравнения в приведенной форме имеют вид:

$$p_t = \frac{\alpha - \delta}{\varepsilon - \beta} + \frac{u_{d_t} - u_{s_t}}{\varepsilon - \beta}; \tag{11.66}$$

$$y_t = \frac{\alpha\varepsilon - \beta\delta}{\varepsilon - \beta} + \frac{\varepsilon u_{d_t} - \beta u_{s_t}}{\varepsilon - \beta}. \tag{11.67}$$

Другими словами, равновесные значения p и y для каждого наблюдения определяются константами $(\alpha - \delta)/(\varepsilon - \beta)$ и $(\alpha\varepsilon - \beta\delta)/(\varepsilon - \beta)$ плюс некоторые случайные компоненты.

Ситуацию можно проиллюстрировать с помощью рис. 11.1. Пунктирные линии представляют фиксированные составляющие уравнений спроса и предложения, и их пересечение дает две только что полученные константы. На рисунке также показаны кривые спроса и предложения для четырех наблюдений. Для первого наблюдения величина u_s положительна, и S_1 находится правее линии

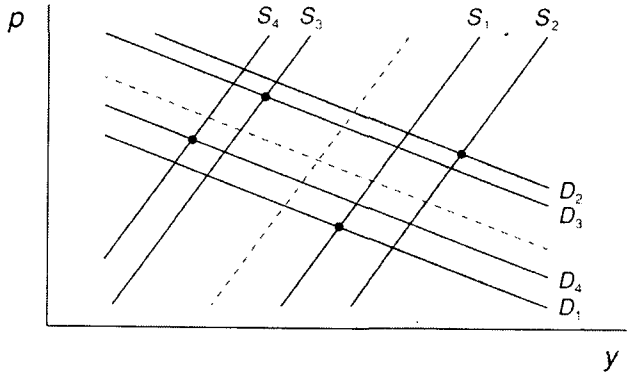


Рис. 11.1. Относительно стабильная функция спроса

с фиксированной составляющей, но величина u_d отрицательна, и D_1 находится левее линии с фиксированной составляющей. В результате величина p получается меньше своего среднего равновесного уровня, однако величина y изменяется несущественно. Аналогично можно рассуждать и для остальных наблюдений.

Очевидно, что точки равновесия случайно разбросаны вокруг фиксированной точки, и оцененное уравнение регрессионной зависимости y от p не будет соответствовать ни функции спроса, ни функции предложения.

Если вы окажетесь достаточно настойчивы в построении уравнения регрессии между y и p, то коэффициент при p, равный $\mathrm{Cov}(y, p)/\mathrm{Var}(p)$, на больших выборках будет стремиться к

$$\frac{\varepsilon\sigma_{u_d}^2 + \beta\sigma_{u_s}^2 - (\varepsilon+\beta)\sigma_{u_d u_s}}{\sigma_{u_d}^2 + \sigma_{u_s}^2 - 2\sigma_{u_d u_s}}. \tag{11.68}$$

Это выражение может быть переписано как

$$\beta + \frac{(\varepsilon-\beta)(\sigma_{u_d}^2 - \sigma_{u_d u_s})}{\sigma_{u_d}^2 + \sigma_{u_s}^2 - 2\sigma_{u_d u_s}} \quad \text{или} \quad \varepsilon - \frac{(\varepsilon-\beta)(\sigma_{u_s}^2 - \sigma_{u_d u_s})}{\sigma_{u_d}^2 + \sigma_{u_s}^2 - 2\sigma_{u_d u_s}}. \tag{11.69}$$

Как показывает первое выражение в (11.69), полученная оценка может рассматриваться как смещенная оценка β. Как показывает второе выражение в (11.69), она также может трактоваться как смещенная оценка ε. Можно делать выбор, но в общем случае оба варианта оказываются бесполезными.

Предположим, однако, что одна из зависимостей является относительно стабильной. Допустим, например, что рассматриваемый товар — мороженое, и спрос на него сильно изменяется от месяца к месяцу в зависимости от сезона, но кривая предложения остается относительно стабильной, поскольку это промышленный товар. Тогда значения $\sigma_{u_s}^2$ и $\sigma_{u_d u_s}$ окажутся относительно небольшими по сравнению с $\sigma_{u_d}^2$ и коэффициент при p будет ближе к ε, чем к β. Графически эта ситуация проиллюстрирована на рис. 11.2.

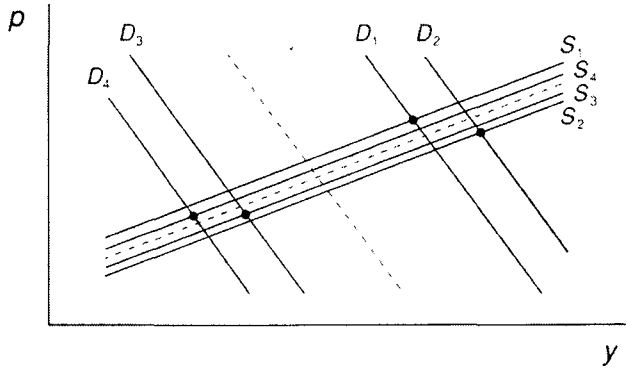


Рис. 11.2. Относительно стабильная функция предложения

Точно так же можно рассмотреть случай, когда относительно стабильной, по сравнению с функцией предложения, является функция спроса. Производство сельскохозяйственной продукции, предложение которой подвержено влиянию погоды, может служить примером такой ситуации.

Упражнение

11.14. Вы располагаете следующим набором данных:

Период времени (t)	1	2	3	4	5	6	7	8
Объем производства (y)	60	10	70	20	60	30	50	40
Цена за единицу (p)	10	70	15	70	20	60	50	30

1. Если рассматриваемый товар — обыкновенный промышленный продукт с малым лагом между принятием решения об объеме его выпуска и выполнением этого решения, сможете ли вы оценить уравнение спроса? Уравнение предложения?

2. Предположим, что этот же набор данных характеризует другой товар — сельскохозяйственный продукт, для которого объем выпуска запаздывает на один период времени относительно момента принятия решения. Сможете ли вы оценить уравнение спроса? Уравнение предложения? Нарисуйте соответствующие графики для каждого из случаев. (Здесь предполагается, что уравнения спроса и предложения не подвержены автокорреляции.)

Приложение 11.1

Величина смещения оценки для одновременных уравнений

Для смещения оценки в случае оценивания модели из одновременных уравнений с помощью МНК нет какой-либо однозначной формулы. Каждый раз она будет определяться структурой модели. Здесь мы исследуем смещение для случая, когда с помощью уравнения регрессионной зависимости C от Y оцениваются параметры функции потребления.

Как было показано в разделе 3.1, оценка β с помощью МНК может быть разбита на две составляющие: истинное значение коэффициента и ошибку:

$$b = \frac{\mathrm{Cov}(Y,C)}{\mathrm{Var}(Y)} = \beta + \frac{\mathrm{Cov}(Y,u)}{\mathrm{Var}(Y)}. \tag{11.70}$$

Мы хотели бы найти математическое ожидание b. К сожалению, нельзя сделать это непосредственно, поскольку $\mathrm{Cov}(Y, u)/\mathrm{Var}(Y)$ — отношение двух величин, каждая из которых частично зависит от одной и той же случайной переменной. Значение $\mathrm{Cov}(Y, u)$ непосредственно зависит от u. Величина $\mathrm{Var}(Y)$ также зависит от u, так как Y частично определяется u.

В то же время на больших выборках при принятии определенных предположений $\mathrm{Cov}(Y, u)$ и $\mathrm{Var}(Y)$ стремятся к своим аналогам в генеральной совокупности $\mathrm{pop.cov}(Y, u)$ и $\mathrm{pop.var}(Y)$, и отношение $\mathrm{Cov}(Y, u)/\mathrm{Var}(Y)$ будет стремиться к σ_{Yu}/σ^2_Y. Используя (11.3), можно представить σ_{Yu} как

$$\sigma_{Yu} = \mathrm{pop.cov}\left\{\left(\frac{\alpha}{1-\beta} + \frac{I}{1-\beta} + \frac{u}{1-\beta}\right), u\right\} =$$

$$= \mathrm{pop.cov}\left\{\frac{1}{1-\beta}, u\right\} + \mathrm{pop.cov}\left\{\frac{u}{1-\beta}, u\right\}. \tag{11.71}$$

Слагаемое $\alpha/(1-\beta)$ исчезает, поскольку это константа. Нет причин предполагать, что объем инвестиций коррелирует со случайной составляющей потребления, поэтому с достаточным основанием можно полагать, что $\mathrm{pop.cov}(I, u) = 0$. В итоге мы имеем:

$$\sigma_{Yu} = \mathrm{pop.cov}\left\{\frac{u}{1-\beta}, u\right\} = \frac{1}{1-\beta}\sigma^2_u. \tag{11.72}$$

Далее, если $\mathrm{Var}(I)$ на больших выборках стремится к своему пределу σ^2_I, то, снова убирая слагаемое $\alpha/(1-\beta)$ как константу и предполагая $\mathrm{pop.cov}(I, u) = 0$, мы получим:

$$\sigma^2_Y = \mathrm{pop.var}\left\{\frac{\alpha}{1-\beta} + \frac{I}{1-\beta} + \frac{u}{1-\beta}\right\} =$$

$$= \frac{1}{(1-\beta)^2}[\mathrm{pop.var}(I) + \mathrm{pop.var}(u) + 2\mathrm{pop.cov}(I,u)] =$$

$$= \frac{1}{(1-\beta)^2}(\sigma^2_I + \sigma^2_u). \tag{11.73}$$

В итоге на больших выборках

$$b \to \beta + \frac{\frac{\sigma_u^2}{1-\beta}}{\frac{1}{(1-\beta)^2}(\sigma_f^2 + \sigma_u^2)}, \tag{11.74}$$

что можно упростить до выражения (11.4).

12

ЧТО ДАЛЬШЕ?

В предыдущих главах было предложено краткое упрощенное рассмотрение базовых вопросов эконометрики. Однако при этом осталось несколько очевидных пробелов. Наиболее заметный из них — то, что различные проблемы, которые могли быть связаны, рассматривались раздельно. Например, мы анализировали автокорреляцию и оценку одновременных уравнений, но не рассматривали оценку одновременных уравнений с наличием автокорреляции и т. д.

Пришло, однако, время остановиться, и для этого есть три причины. Во-первых, существует предел объема материала, который может рассматриваться во вводном учебнике. Во-вторых, более глубокое изложение предъявляет свои технические требования, для него нужна более сложная математика и потребовалось бы переключиться на матричную алгебру. В-третьих, здесь необходимо было бы сделать два изменения в организации изложения. Мы должны были бы заменить метод наименьших квадратов как принцип, лежащий в основе получения оценок, на метод максимума правдоподобия и перейти от конструирования специальной модели для каждого конкретного случая к более систематическому подходу. В данной главе коротко рассматриваются оба этих направления анализа.

12.1. Метод максимального правдоподобия (ММП)

В классической модели линейной регрессии, где случайный член удовлетворяет условиям Гаусса—Маркова и отсутствуют другие сложности, базовым критерием для получения оценок коэффициентов является минимизация суммы квадратов отклонений. Этот выбор не был произвольным. Теорема Гаусса—Маркова гласит, что оценки, полученные методом наименьших квадратов, будут несмещенными и эффективными как на больших выборках, так и на малых, в случае если выполняются условия Гаусса—Маркова.

Тем не менее в последних четырех главах модель регрессии постепенно теряла связь с условиями Гаусса—Маркова и удалялась от своих первоначальных предпосылок. Мы рассмотрели использование лаговых значений зависимой переменной как регрессора, применили нелинейный регрессионный анализ (например, в оценке уравнений по методу Кокрана—Оркатта), кроме того, мы стали использовать инструментальные переменные. Оценка методом ИП не основана на методе наименьших квадратов. Оправдание применения метода ИП заключается в том, что он позволяет получать состоятельные оценки

в случаях, когда их не дает МНК, и, хотя минимизация дисперсии распределения оценки является приемлемым решением при наличии других альтернатив, поиск такого минимума — не главная задача.

К сожалению, для любителей простой жизни в тех случаях, когда нарушаются условия Гаусса—Маркова и мы вынуждены искать замену МНК, редко появляется единственный приемлемый вариант. Обычно различные исследователи могут предложить несколько конкурирующих оценок, каждая из которых является состоятельной. Например, в случае автокорреляции первого порядка состоятельные оценки дают как метод Кокрана—Оркатта, так и данный метод с поправкой Прайса—Уинстена. Если бы нам пришлось выбирать между двумя этими методами (хотя в действительности имеются и другие альтернативы, не упомянутые в главе 7), то как осуществить выбор? Привлекательным было бы выбрать асимптотически эффективную оценку. Причина, по которой метод Кокрана—Оркатта с поправкой Прайса—Уинстена оказался предпочтительнее метода Кокрана—Оркатта (и это подтверждено экспериментами по методу Монте-Карло), заключается в том, что на больших выборках его оценки параметров имеют меньшие стандартные ошибки.

Здесь наступает момент перехода к оценкам *методом максимального правдоподобия* (ММП). Они обычно не предъявляют требований к желательным свойствам малых выборок, но в случае корректной спецификации модели и при выполнении некоторых условий обеспечивают асимптотическую несмещенность, состоятельность и асимптотическую эффективность. Более того, они предоставляют возможность для проведения тестов, которые не могли использоваться в случае МНК.

Что такое оценивание на основе ММП? Оно может быть проиллюстрировано на простом примере. Предположим, что имеется непрерывная случайная переменная с неизвестным средним значением μ и (для простоты) известным стандартным отклонением, равным единице. Допустим также, что есть основания считать, что переменная имеет нормальное распределение. Предположим, наконец, что у вас имеются две альтернативные гипотезы: $\mu = \mu_0$ и $\mu = \mu_1$ и одно наблюдение x_1, как это показано на рис. 12.1. Какую из гипотез вы выберете? Принцип максимального правдоподобия утверждает, что следует выбрать ту из гипотез, которая обеспечивает наибольшую вероятность появления x_1. Поскольку x — непрерывная случайная величина, вероятность появления какого-либо отдельного ее значения является бесконечно малой. Вместо этого мы сравниваем величину плотности вероятности в точке x_1 для двух гипотез, представленную высотой графика функции плотности вероятности. Очевидно, что на рис. 12.1 функция с $\mu = \mu_0$ имеет более высокое значение плотности вероятности в точке x_1.

Следующий шаг — обобщение. Мы будем рассматривать все возможные значения μ и выберем то из них, которое дает максимальное значение плотности вероятности в точке x_1.

Функция плотности вероятности переменной x с заданным μ имеет вид:

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}. \qquad (12.1)$$

Зададимся теперь вопросом: какое значение μ максимизирует плотность вероятности для заданного наблюдения x_1? Графически очевидный ответ: $\mu = x_1$,

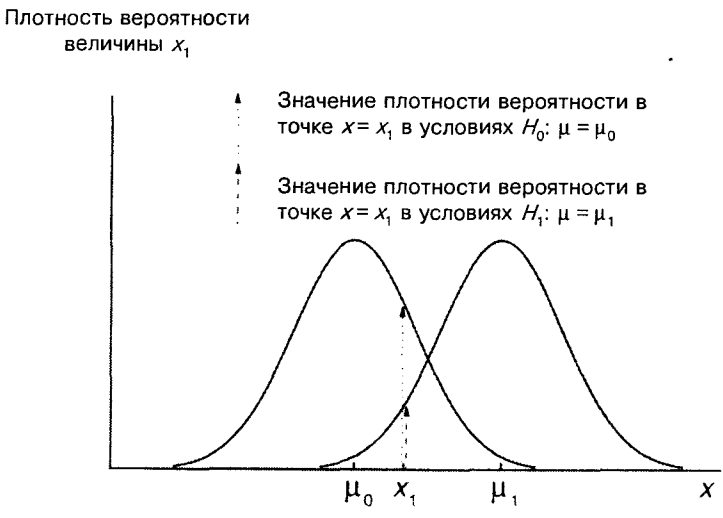


Рис. 12.1. Вероятность появления x_1 в условиях истинности H_0 и H_1

поскольку в этом случае распределение будет расположено вокруг x_1, и плотность вероятности окажется максимальной в центре распределения. Дадим также математическое доказательство этого факта.

Прежде всего заметим, что в решаемой задаче значение x_1 задано, а μ рассматривается как переменная величина. Как следствие мы можем рассматривать функцию плотности вероятности как функцию от μ при заданном x_1. В дальнейшем мы так и сделаем, а также дадим ей другое название — *функция правдоподобия* и, чтобы подчеркнуть происшедшие изменения, будем обозначать ее как L:

$$L(\mu|x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-\mu)^2}. \tag{12.2}$$

Отметим, во-вторых, что $\log L(\mu)$ будет иметь максимум при том же значении μ, которое максимизирует $L(\mu)$, поскольку логарифм от любой переменной возрастает или уменьшается с ростом или уменьшением значения переменной. Удобнее и математически проще найти максимум функции $\log L(\mu)$ (которая известна также под названием *логарифмическая функция правдоподобия*):

$$\log L(\mu) = -\log\sqrt{2\pi} - \frac{1}{2}(x_1 - \mu)^2. \tag{12.3}$$

Продифференцировав это выражение по μ, мы получим:

$$x_1 - \mu = 0, \tag{12.4}$$

поэтому оценка μ по ММП равна x_1. Вторая производная имеет отрицательное значение, и это подтверждает, что мы нашли максимум функции.

Все это, конечно, тривиально. Предположим теперь, что имеются два независимо распределенных наблюдения (x_1) и (x_2), и мы хотим оценить μ. Процедура оценивания по ММП заключается в нахождении значения μ, которое мак-

симизирует совместную функцию плотности вероятности, определяемую произведением частных функций плотности вероятности:

$$f(x_1, x_2|\mu) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-\mu)^2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2-\mu)^2}\right). \tag{12.5}$$

Это выражение может быть преобразовано в логарифмическую функцию правдоподобия для μ при заданных x_1 и x_2, и мы максимизируем его, максимизировав, как и прежде, логарифмическую функцию правдоподобия:

$$\log L(\mu) = -2\log\sqrt{2\pi} - \frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}(x_2-\mu)^2. \tag{12.6}$$

Из этого выражения можно получить условие первого порядка:

$$(x_1-\mu) + (x_2-\mu) = 0, \tag{12.7}$$

откуда следует, что оценка μ по ММП равна $(x_1 + x_2)/2$. Этот результат нетрудно обобщить на случай, когда имеется n наблюдений. Оценка μ по ММП равна среднему значению выборки и как таковая совпадает с оценкой по методу наименьших квадратов.

Если мы применим ММП в классической модели линейной регрессии, предполагая нормальное распределение случайного члена, то оценки всех коэффициентов (но не дисперсия случайного члена) будут равны оценкам, полученным с помощью МНК. Поэтому переход от метода наименьших квадратов к ММП в данном контексте является скорее эволюцией, чем радикальным преобразованием.

В настоящее время максимизируемая логарифмическая функция правдоподобия выдается почти всеми регрессионными пакетами диагностической статистики, сопровождающей результаты расчета регрессии. Ее можно использовать для проведения *теста на отношение правдоподобия*. Этот тест заключается в следующем. Пусть имеется две версии модели, и одна из них — версия другой с добавлением ограничений. Статистика 2(LLU − LLR), где LLU и LLR — логарифмические функции правдоподобия неограниченной и ограниченной версии модели, соответственно, на больших выборках подчиняется распределению χ^2 с s степенями свободы в случае принятия нулевой гипотезы о корректности ограниченной версии, где s — число наложенных ограничений. (Для дальнейшего обсуждения этого теста и двух других групп тестов, разработанных в рамках максимизации правдоподобия [тестов Вальда и тестов множителей Лагранжа] рекомендуется работа П. Кеннеди [Kennedy, 1985, pp. 58–59].)

Однако и помимо теоретических свойств принцип оценивания параметров, максимизирующих вероятность появления данной выборки, обладает большой привлекательностью для эконометристов. Принцип наименьших квадратов имеет, конечно, свои достоинства в случае выполнения условий Гаусса—Маркова, но все эти достоинства быстро теряются в противном случае.

Если ММП столь эффективен и привлекателен и если он столь приспособлен для проведения тестов, почему бы не использовать его все время? Суще-

ствуют четыре причины для осторожности. Во-первых, имеющиеся обычно выборки, особенно при анализе временных рядов, скорее являются малыми, чем большими. И вполне возможно, что методы, обладающие желаемыми свойствами на больших выборках, будут уступать другим методам на малых выборках. Единственный путь проверки в случае каждого конкретного исследования — проведение экспериментов по методу Монте-Карло. Во-вторых, наше решение поддерживается соображением о том, что столь популярные свойства состоятельности и асимптотической эффективности не являются безусловными. Они были установлены лишь для моделей определенного вида. Модели с трендовыми данными в этой связи могут порождать проблем не меньше, чем в случае применения анализа, базирующегося на методе наименьших квадратов. В-третьих, необходимо сделать предположение, что случайный член имеет определенное асимптотическое распределение; обычно предполагается нормальное распределение. Эта предпосылка не является необходимой при применении МНК в случае классической линейной регрессионной модели (она требуется лишь для проведения тестов). В-четвертых, оценка по ММП занимает слишком много времени — как времени исследователя, так и компьютера. Оценки часто выводятся в результате решения системы одновременных уравнений с использованием итеративной процедуры, поскольку они не могут быть выражены в виде явных математических формул.

Четвертый аргумент против применения ММП долгое время был, возможно, единственным важным препятствием на пути его широкого распространения, однако проблема быстро упрощается, и с удешевлением компьютерного времени оценки по ММП активно встраиваются в статистические пакеты. Например, оценка по ММП автокорреляции в авторегрессионной модели первого порядка, разработанная Ч. Бичем и Дж. Маккинноном (Beach, MacKinnon, 1978), теперь является стандартной принадлежностью статистических пакетов. Дж. Крамер (Cramer, 1986) полагает, что с расширением возможностей ММП до их пределов этот метод «в большой степени вытеснит линейный регрессионный анализ как главный инструмент прикладной эконометрики». Возможно, он прав. Тем не менее поскольку оценки по ММП и МНК совпадают (кроме оценки σ_u^2) при выполнении условий Гаусса—Маркова, вполне вероятно, что МНК останется исходным пунктом вводных курсов эконометрики.

12.2. Спецификация модели

В предыдущих главах мы использовали различные диагностические тесты, проверяющие адекватность спецификации регрессионной модели. Мы рассмотрели тесты спецификации объясняющих переменных: t-тест объясняющей способности (и следовательно, желательности включения в модель) отдельных объясняющих переменных и F-тест объясняющей способности группы объясняющих переменных. Были рассмотрены тесты адекватности спецификации остаточного члена: тест на автокорреляцию и тест на гетероскедастичность. Была предложена процедура, которая может использоваться как тест на функциональную форму зависимости: процедура Бокса—Кокса для определения наиболее подходящего вида преобразования переменных (внутри заданного класса пре-

образований). И это лишь небольшая часть всех диагностических тестов, применяемых на практике.

Общая стратегия исследования, которая неявно использовалась нами до настоящего времени, может быть обобщена следующим образом:

1. На основании экономической теории, опыта и интуиции сформировать предварительную модель.
2. Подобрать имеющиеся данные и оценить параметры модели.
3. Провести диагностические тесты.
4. Если хотя бы один из тестов указывает на неадекватность спецификации модели, пересмотреть ее с целью устранения этой неадекватности.
5. Когда спецификация оказывается удовлетворительной, поздравить себя с решением поставленной задачи и закончить работу.

Опасность этой стратегии (и «опасность» здесь — вполне подходящее слово) заключается в том, что причина, по которой окончательная версия модели окажется удовлетворительной, — искусная подгонка спецификации модели к имеющемуся набору данных, а вовсе не соответствие реальной модели. Эконометрическая литература заполнена двумя видами неявных свидетельств в пользу того, что это происходит все время, особенно с моделями, оцениваемыми на временных рядах, и в частности с моделями, отражающими макроэкономические взаимосвязи. Часто случается, что исследователи, анализирующие одно и то же явление, но имеющие доступ к различным источникам данных, строят внутренне согласованные, но абсолютно несравнимые модели; часто случается также, что модели, выдержавшие диагностические тесты на выборке, имеют крайне малую прогностическую способность. Особенно выделяется с обеих этих точек зрения литература, посвященная моделированию совокупных инвестиций. Еще одно свидетельство, если таковые все же необходимы, предоставляется экспериментами, показывающими, что совсем не сложно построить бессмысленные модели, выдерживающие все условные проверки (см. работу Дж. Пича и Дж. Уэбба [Peach, Webb 1983]). Как следствие все больше возрастает осознание того, что тесты позволяют отклонить лишь совершенно неверно специфицированные модели, и тот факт, что модель выдержала их, не может служить гарантией ее правильности.

«Но как насчет тестов на предсказательную способность модели, описанных в главе 10?» — спросите вы. Там модель подвергалась оценке своей способности соответствовать новым данным. С этими тестами возникает две проблемы. Во-первых, их эффективность довольно низка. Вполне возможно, что плохо специфицированная модель будет соответствовать наблюдениям периода предсказания, и нулевая гипотеза о стабильности модели не будет отвергнута, особенно если период предсказания невелик. «Хорошо», — скажете вы, удлиняя период предсказания и укорачивая период выборки. Это действительно может оказаться правильным направлением анализа, но здесь снова возникает проблема, особенно если выборка малая. Укорачивая период выборки, вы понизите соответствие модели выборочным данным, и определить, насколько существенно отличается ее поведение в период предсказания, будет еще труднее.

Другая проблема с тестами на стабильность предсказаний — вопрос, что делать

исследователю, если тест не удался. Понятно, что было бы неправильно закончить работу в этой точке, признав свое поражение. Естественное направление действий — продолжать переделывать модель, пока она не пройдет, и данный тест, однако после этого тест является не более «честным», чем тесты на периоде выборки. Такое неудовлетворительное положение дел порождает интерес к двум взаимосвязанным вопросам: возможности отклонения некоторых из конкурирующих моделей в результате сравнения их друг с другом и возможности принятия более систематической исследовательской стратегии, которая позволила бы сразу же исключить построение неадекватных моделей.

Сравнение альтернативных моделей

Техника сравнения альтернативных моделей может быть достаточно сложной, и здесь мы ограничимся очень кратким частичным рассмотрением некоторых вопросов. Начнем с введения различия между включенными и невключенными моделями. Одна модель называется *включенной* в другую, если первая модель может быть получена из второй путем наложения некоторых ограничений. Две модели называются *невключенными* (друг в друга), если ни одна из них не может быть представлена как ограниченная версия другой. Ограничения могут касаться любого аспекта спецификации модели, но в данном случае мы рассмотрим лишь ограничения, накладываемые на параметры объясняющих переменных в модели, состоящей из одного уравнения. Проиллюстрируем это на примере функции потребления домашних хозяйств, анализируемой в главах 7 и 10. В этих главах рассматривались три динамические модели: модель с лаговой зависимой переменной (обозначим ее как модель *B*); оценка Прайса—Уинстена статической модели (модель *C*); модель с лаговыми значениями всех переменных и без ограничений на ее параметры (модель *A*). Добавим для полноты картины простую статическую модель (модель *D*):

$$\text{(A)} \quad y_t = \lambda_0 + \lambda_1 y_{t-1} + \lambda_2 x_t + \lambda_3 x_{t-1} + \lambda_4 p_t + \lambda_5 p_{t-1} + u_{At}; \tag{12.8}$$

$$\text{(B)} \quad y_t = \lambda_0 + \lambda_1 y_{t-1} + \lambda_2 x_t + \lambda_4 p_t + u_{Bt}; \tag{12.9}$$

$$\text{(C)} \quad y_t = \lambda_0 + \lambda_1 y_{t-1} + \lambda_2 x_t - \lambda_1\lambda_2 x_{t-1} + \lambda_4 p_t - \lambda_1\lambda_4 p_{t-1} + u_{Ct}; \tag{12.10}$$

$$\text{(D)} \quad y_t = \lambda_0 + \lambda_2 x_t + \lambda_4 p_t + u_{Dt}. \tag{12.11}$$

Модели *B* и *C* включены в *A*, поскольку, как было отмечено в разделе 7.9, обе являются ее ограниченными версиями: модель *B* налагает нулевые ограничения на коэффициенты при x_{t-1} и p_{t-1} (это, конечно, просто другой способ отражения того, что эти переменные исключены из модели), модель *C* налагает ограничение на общий множитель, объясненное в разделе 7.9. Модель *D* может рассматриваться в качестве включенной как в модель *B*, так и в *C*. Она может быть получена из модели *B* приравниванием нулю коэффициента при y_{t-1}, а из модели *C* — наложением ограничения, согласно которому $\rho = 0$. Структура включения моделей представлена на рис. 12.2.

Лучше всего начать с самой общей модели — модели *A*. В случае расходов на жилье, если мы сравним *A* с *C*, то обнаружим, что ограничение на общий множитель, присутствующее в модели *C*, отвергается (см. раздел 7.9). Следова-

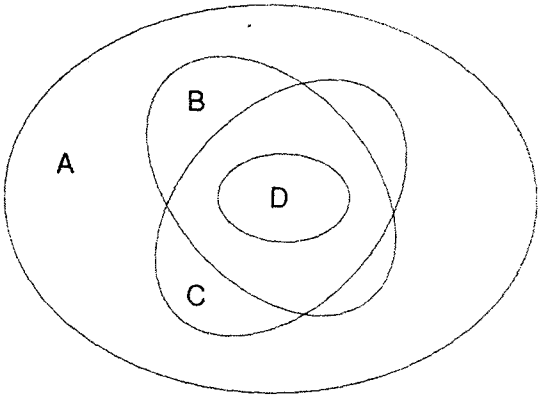


Рис. 12.2. Структура включения моделей *A, B, C* и *D*

тельно, модель *C* находится за пределами списка приемлемых моделей. Если мы сравним модель *A* с *B*, то обнаружим, что *B* является хорошей альтернативой *A*, поскольку оцененные коэффициенты при x_{t-1} и p_{t-1} незначимо отличаются от нуля, как это было показано в разделе 7.9. На самом деле вместо использования *t*-теста для отдельных коэффициентов следовало бы применить *F*-тест для их совместной объясняющей способности, и именно это будет сделано. Сумма квадратов отклонений для модели *A* равна 0,00076 и для модели *B* — 0,00080. Соответствующая *F*-статистика, имеющая 2 и 18 степеней свободы, равна 0,47. Она незначима даже при 5-процентном уровне значимости, поэтому модель *B* выдерживает данный тест. Наконец, модель *D* должна быть отвергнута, поскольку ограничение, согласно которому коэффициент при y_{t-1} равен нулю, отвергается с помощью простого *t*-теста. (Во всех этих рассуждениях мы предполагали, что процедуры тестов не зависят существенно от использования лаговых зависимых переменных в качестве объясняющих. Как вы знаете, это безусловно верно только для больших выборок.)

Приведенный пример иллюстрирует возможность как успеха, так и неудачи проведения тестов в случае включенной структуры моделей: успех — в отклонении двух из четырех моделей и неудача — поскольку некоторая неопределенность в итоге осталась. Есть ли основания предпочесть модель *A* модели *B* или наоборот? Многие авторы (например, Дэвид Хендри — уже упоминавшийся шотландский исследователь) сказали бы, что следует предпочесть модель *B* как более экономную, но не в смысле экономии денег, а в смысле экономии использования параметров: модель *A* требует оценки шести параметров, модель *B* — только четырех. Тем не менее преимущества экономности в данном контексте не до конца ясны. Представляется, что такая экономность связана в основном с принципом максимальной простоты предлагаемого объяснения (принцип KISS — keep it simple, stupid), известного также под названием «бритва Оккама». Поскольку все это с трудом может быть положено в основу четкого императива, постольку здесь остается проблема выбора между эффективностью и возможным смещением при включении и исключении переменных с незначимыми коэффициентами, рассмотренная в главе 6. Пожалуй, самый убедительный аргумент в пользу экономности заключается в том, что

обычно нетрудно придумать большое число потенциально подходящих объясняющих переменных, коэффициенты при которых оказываются незначимыми, и сохранение в итоге одной или двух из них в модели будет произвольным. Что делать в случае, когда конкурирующие модели оказываются невключенными моделями? Одна из возможных тактик — построить объединенную модель, включающую эти две модели как свои ограниченные версии, и проверить, имеется ли какой-либо прогресс при оценке каждой из моделей по сравнению с их объединением. Предположим, например, что модели F и G одинаковы во всем, кроме спецификации их независимых переменных. Обе модели включают переменные $x_1 \ldots x_e$ как регрессоры. Модель F содержит также переменные $x_{e+1} \ldots x_f$ (подмножество F). Модель G не включает их, но вместо этого содержит переменные $x_{f+1} \ldots x_g$ (подмножество G). Объединенная модель E будет вмещать все три подмножества независимых переменных. Теперь необходимо провести два теста. Приняв в качестве нулевой гипотезы предположение о том, что модель E имеет верную спецификацию, проверим совместную объясняющую способность переменных из подмножества F и, отдельно от них, из подмножества G. Здесь возможны четыре исхода: 1) оба подмножества F и G обладают значимой объясняющей способностью. Это довольно неожиданный исход, поскольку он влечет отклонение обеих моделей F и G в пользу не рассматривавшегося прежде их объединения; 2) подмножество F обладает значимой объясняющей способностью, а G — нет. Тогда следует отклонить модель G и принять модель F. Модель E также не может быть принята; 3) результат, противоположный (2), с соответствующими выводами; 4) ни подмножество F, ни G не обладают значимой объясняющей способностью. Обе модели F и G продолжают рассматриваться, как и модель E.

Для того чтобы сделать наши рассуждения более конкретными, рассмотрим следующий простой пример, в котором подмножества F и G содержат всего лишь по одной переменной. Модели F и G выглядят следующим образом:

$$\textit{Модель } F \qquad y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u_F; \tag{12.12}$$

$$\textit{Модель } G \qquad y = \alpha + \beta_1 x_1 + \beta_3 x_3 + u_G. \tag{12.13}$$

Подмножество F совпадает с x_2, подмножество G — с x_3. Объединенная модель E имеет следующий вид:

$$\textit{Модель } E \qquad y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_E. \tag{12.14}$$

В этом случае F-тест на объясняющую способность подмножеств F и G сводится к t-тесту для коэффициентов при x_2 и x_3 соответственно. После оценки модели E возможны четыре исхода: 1) оба коэффициента b_2 и b_3 значимо отличаются от нуля. Переменные x_2 и x_3 должны быть включены в модель. Тем самым, модели F и G отклоняются, модель E — принимается; 2) коэффициент b_2 значимо отличается от нуля, а b_3 — нет. В этом случае x_2 должна быть включена, а x_3 может быть включена в модель (тот факт, что коэффициент при переменной оказался незначимым, вовсе не означает, что переменная должна быть обязательно исключена из модели; это может быть действительно объясняющая переменная, но ее воздействие может быть слабым, а размер выборки — очень малым для того, чтобы коэффициент при переменной значимо отличался от нуля). Мы отклоняем модель G и принимаем модели F и E; 3) коэффициент b_3 значимо отличается от нуля, а b_2 — нет. Такой же вывод, как и в (2), только

F и G меняются местами; 4) ни b_2, ни b_3 не отличаются значимо от нуля. Все три модели принимаются, поскольку x_2 и x_3, несмотря на свои незначимые коэффициенты, могут оказаться действительно объясняющими переменными.

Очевидно, этот подход порождает множество возможных проблем. Во-первых, во всех тестах модель E используется как нулевая гипотеза, что может оказаться не совсем приемлемым в данном конкретном случае. Если модели F и G построены на основании различных принципов, их объединение может быть достаточно странным и неприемлемым с точки зрения экономической теории. В этом случае общих рамок для проведения тестов не существует. Во-вторых, четвертый исход, когда остается неопределенность, может появляться довольно часто. Если модели F и G были построены достаточно аккуратно, то вполне вероятно, что модель E будет обладать малой дополнительной объясняющей способностью и оба F-теста окажутся незначимыми. Для выхода из этой ситуации были предложены различные процедуры, но они выходят за рамки книги (более развернутый анализ этих проблем и дальнейшие ссылки см. в работе Я. Кменты [Kmenta, 1986, pp. 595–598]).

Применение подхода «от общего к частному» к спецификации модели

Как мы уже убедились, в случае когда мы начинаем с простой модели и в дальнейшем усложняем ее в соответствии с диагностическими тестами, имеется риск в конце концов получить ложную модель, которая удовлетворяет нас постольку, поскольку в процессе последовательной адаптации мы «подогнали» ее к выборке данных (именно подогнали, поскольку все вероятностные тесты оказываются некорректными из-за неправильно сформулированной нулевой гипотезы).

Некоторые авторы настаивают, что было бы лучше принять противоположный подход. Вместо попыток превратить частную начальную модель в более общую следует, считают они, начинать с общей модели и сводить ее к более частной путем последовательного наложения ограничений (после испытания их корректности).

Конечно, подход «от общего к частному» в принципе более предпочтителен. Проблема заключается в том, что в чистом виде он часто оказывается неприменим. Если размер выборки ограничен, а начальная спецификация включает большое число потенциальных объясняющих переменных, то мультиколлинеарность может вызвать незначимость коэффициентов у большинства из них или даже у всех переменных. Появление этой проблемы особенно реально для моделей на временных рядах. В крайнем случае число переменных может превышать число наблюдений, и оценить параметры модели окажется невозможным вообще. Когда оценка модели возможна, незначимость многих коэффициентов дает исследователю большую свободу в принятии решения о том, какие из переменных удалить. Конечная версия модели может оказаться очень чувствительной к такому произвольному начальному решению. Исключенная переменная с начальным незначимым коэффициентом могла бы иметь значимый коэффициент в сокращенной версии модели, если бы она была там оставлена. Добросовест-

ное систематичное применение принципа «от общего к частному» может потребовать исследования необозримого числа возможных путей упрощения модели. И даже если таких путей окажется немного, исследователь в итоге может остаться с большим числом моделей-конкурентов, ни одна из которых не предпочтительнее других. Этот исход напоминает «отпочковавшуюся» версию случая (4) для невключенных гипотез, который рассматривался в предыдущем разделе.

В итоге некоторый компромисс представляется нормальным явлением, и для него нет никаких правил, кроме принципов формирования исходной концепции модели. Более слабой, но действенной версией рассматриваемого подхода может быть осторожность при включении в исходную спецификацию ограничений, которые заранее имеют шанс быть отвергнутыми. Будет правильным, однако, заметить, что способность делать это свидетельствует об опыте исследователя, и в таком случае весь подход сводится к требованию накопления опыта. Скептики могут сравнить этот подход с развешиванием надписей «ПОДУМАЙ!», которые украшали офисы фирмы IBM в первые дни ее существования, и даже Д. Хендри — один из настойчивых защитников этого подхода — признает, что на практике здесь всегда неизбежна поисковая работа (Hendry, 1979, p. 228, ссылки на работы Дж. Дэвидсона и др. [Davidson, 1978]). Как неформальную поддержку данного подхода, полную занимательных едких замечаний о недостатках построения моделей по принципу «от простого к сложному» и иллюстративных примеров, рекомендуем работу Д. Хендри (Hendry, 1979) (с небольшой оговоркой, что χ^2 — тест на несостоятельность прогноза, используемый в данной работе, теперь широко признан недейственным и, как правило, больше не применяется).

Границы статистических выводов

Предыдущий раздел мог породить ощущение, что несколько технических проблем, иногда даже не до конца понятных, стоят между эконометристом и признанием теории, служащей информационной базой для модели. Если это так, то необходимо сделать поправку: даже если технические проблемы будут разрешены, то останутся другие, более глубокие содержательные проблемы, и эконометристу нужно их осознавать для того, чтобы не переоценивать, не недооценивать свой вклад.

В то время когда современная наука находилась еще в колыбели, была широко распространена вера в то, что природа подчиняется набору хорошо определенных законов, и задачей ученого является их обнаружение и доказательство их истинности с помощью искусных экспериментов. Научная деятельность представлялась чем-то вроде альпинизма с достижением абсолютной истины на горной вершине. Проблема этой простой идеи заключается в том, что никогда нельзя доказать абсолютную истинность любой теории. Во-первых, теория всегда строится для описания прошлых фактов, и из этого совсем не следует, что так всегда будет и в будущем. В любой момент это может быть нарушено каким-нибудь новым наблюдением. Во-вторых, достаточно часто случается, что один и тот же набор фактов (при некотором воображении) служит основанием для нескольких различных теорий.

Принимая все это во внимание, необходимо заново сформулировать определения прогресса науки и научного знания. Поскольку теории могут быть приняты лишь условно, то знание — собрание таких теорий — должно восприниматься как условная концепция. Научный прогресс определялся как замена старых условных теорий новыми, более эффективными, обычно появляющимися, когда новое наблюдение противоречит старой теории. Дальнейший прогресс произойдет, когда эта новая теория в свою очередь также будет заменена. Нужно признать, что «горная вершина» в науке не будет достигнута никогда. Лучшее, что можно делать, — это выбираться постепенно из долины невежества.

К сожалению, даже этот усложненный подход не до конца удовлетворителен. Поскольку нет строгих подходов к доказательству истинности теории, то нет и строгих подходов к доказательству ее ложности. Если теория вдруг стала противоречить результатам новых наблюдений, то защитник этой теории обычно может спасти ее, лишь слегка изменив. Он может сказать, что до этого теория была слишком упрощенной и происшедшие изменения эволюционны. Обладая воображением, всегда можно найти оправдание теории и примирить ее с возникшим противоречием (неважно, насколько неуклюже). Другими словами, нельзя даже определенно утверждать, что мы продвигаемся вверх.

Все это оставляет нас в весьма неуютном положении. Поскольку нет строгих критериев для определения как истинности, так и ложности теории, то как мы можем утверждать, что происходит прогресс науки? И какой смысл на самом деле имеют понятия «наука» и «ученый»?

Одна из школ науковедов утверждает, что ответы на эти вопросы являются в высшей мере субъективными и социально детерминированными и что наука состоит из всего, во что верят, — правильного или неправильного. Они отмечают, что, хотя и нельзя сразу отвергнуть теорию, некоторые теории становятся бесполезными с накоплением слишком большого количества неправдоподобных поправок, необходимых для спасения этих теорий от явно противоречащих им фактов. Согласно такому подходу, научное знание состоит из запаса принятых ранее правдоподобных теорий, а субъективность этого определения заключается в отсутствии абсолютного стандарта правдоподобности. Достаточно любопытно, что в своих рассуждениях о том, что такое знание, науковеды предпочитают брать примеры из естественных наук, а не из поведенческих, где эти проблемы являются гораздо более острыми.

Ученые обычно решают две задачи: наблюдение закономерностей и формулировка гипотез, позволяющих предсказать их. Ученые в естественных науках — может быть, за исключением астрономии — находятся в лучшем положении относительно наблюдений, чем эконометристы. Они, как правило, могут поставить и повторить специально построенный контролируемый эксперимент для обнаружения существования и установления природы закономерности. Эконометристы, напротив, обычно с трудом могут провести какой-либо эксперимент и никогда не могут осуществить контролируемый эксперимент. Вместо этого они должны удовлетвориться ограниченным набором несовершенных данных, и проблемы наблюдения обостряются двумя свойствами экономических моделей. Во-первых, экономические процессы являются, как правило, чрезвычайно сложными, и любая приемлемая теория может быть только их приближением, достаточно часто — весьма грубым.

Хорошая модель упростит наиболее важные анализируемые процессы, содержащиеся в наблюдениях, плохая же модель их просто проигнорирует. Во-вторых, поскольку экономика — поведенческая наука, часто случается, что наиболее важные предпосылки экономической теории берутся из области психологии, и поэтому их трудно или невозможно проверить. Примерами здесь могут служить предпосылки о максимизации прибыли или о максимизации полезности.

В попытке преодолеть все эти трудности эконометристы используют статистический анализ. В лучшем случае, когда эконометрист удачлив и опытен, он достигнет лишь начальной точки исследования с позиции представителя естественных наук. Более вероятно, что туман наблюдений в итоге не рассеется и науковедческие проблемы станут еще более сложными.

Отсюда следует, что если субъективность является принципиальной проблемой в естественных науках, то она еще более важна в экономике. Если вы скажете биохимику, который занят исследованием работы нервной системы, что вопрос о том, добавляет ли это исследование что-нибудь нового в научное знание, полностью субъективен, и выразите сомнение, можно ли вообще биохимика назвать ученым, то он, скорее всего, решит, что вам необходима помощь психиатра. Что касается биохимика, то он знает, что такое наука, в которой он работает, а если вы — нет, то это уже ваша проблема. Наоборот, в поведенческих науках имеются действительно яркие примеры субъективности. Например, в экономике наблюдается существенное разделение между монетаристами, неоклассиками и кейнсианцами, сторонниками экономической теории предложения, неорикардианцами, марксистами и пр., причем каждая из групп имеет свой набор теорий, не зависящих от подхода в эконометрическом анализе.

К счастью, наша повседневная жизнь мало зависит от обладания научным знанием, и часто даже от его существования. Фермер, вносящий удобрения, по опыту знает, что это повысит урожай при нормальном поливе и хорошей погоде. Выводы научного исследования о том, как химические соединения из удобрений распространяются в растениях и как они воздействуют на рост клеток, не очень заинтересуют фермера. Все, что ему необходимо, — это рабочее правило, сколько удобрения класть на акр земли для получения наилучших результатов.

В экономике также, к счастью, достаточно много примеров подобных рабочих правил, однако большинство из них касается микроэкономических связей. В макроэкономике до сих пор трудно найти рабочие правила, обладающие существенной прогнозной силой. Поэтому в предвидимом будущем останется огромный простор для субъективности и сосуществования различных школ и направлений.

Цель эконометриста — установить рабочие правила в своей области специализации. Очевидно, что есть предел тому, что может быть продемонстрировано с помощью эконометрики. Известно, что ничего нельзя *доказать* даже с помощью правильно примененной сложнейшей техники, поскольку другой исследователь всегда сможет по-иному интерпретировать те же самые результаты. Лучшее, что можно сделать, — это показать, что данная теория более правдоподобна, а другая — нет.

Неужели все это обесценивает эконометрический анализ? Конечно, нет. Политик так или иначе должен принимать решения, и если теоретическое совершенство недостижимо, то лучше, если эти решения будут основываться на правдоподобной теории, чем не основываться ни на какой теории вообще.

12.3. Послесловие к функциям спроса

Функции спроса дали основу для многих примеров и упражнений в этой книге, и было бы правильно закончить кратким рассмотрением перспектив их анализа. Насколько важно оценивание функций спроса на практике и какого уровня достигло искусство их построения?

Ответ на первый вопрос говорит, что функции спроса имеют долгую историю в развитии эконометрики. В самом деле, они дают один из старейших примеров оценки эмпирических взаимосвязей в экономике. Это оценка связи между изменением производства пшеницы в Британии и изменением равновесных цен в работе, опубликованной Чарльзом Давенантом примерно в 1700 г. (см. работу Ч. Уитворта [Whitworth, 1771]). Связь была выражена в виде эмпирической таблицы, а не в виде математической функции, но она позволяет определить эластичность спроса по цене на уровне 0,3 и объяснить, почему небольшие изменения урожая вызывают значительные изменения цен, которые немедленно отражаются на уровне доходов фермеров, стоимости жизни, торговом балансе, когда предприимчивые датчане в тяжелые времена приезжали с зерном, а уезжали со слитками золота. Неудивительно, что попытки смягчить эти потрясения с помощью законов о зерне были одной из главных политических проблем того времени.

В последние 50 лет анализ спроса был заслонен появлением других областей эмпирической экономической науки, особенно микроэкономики, но этот вопрос все равно остается важным: недавним примером может служить спрос на нефть (международные отношения после 1973 г. частично определялись ценовой эластичностью спроса на нефть и нефтепродукты). Правительства также обычно учитывают показатели ценовой эластичности при установлении налогов и тарифов, и самый наглядный пример здесь — использование эмпирически установленного факта низкой ценовой эластичности спроса на табак для повышения государственных доходов.

Учитывая высокую практическую важность функций спроса и их ключевую роль в микроэкономической теории, неудивительно, что они продолжают оставаться в центре внимания исследователей. Например, А. Дитон (Deaton, 1986) перечисляет более 200 работ в этой области, многие из которых совсем недавние.

Нынешний уровень искусства оценивания функций спроса нелегко привести к единому знаменателю. Некоторые исследователи стараются непосредственно связать свои модели с экономической теорией. Другие занимают более прагматичные позиции, ограничиваясь решением конкретной задачи.

Вставшие на первый путь пытаются связать анализ спроса с теорией полезности. Построение эконометрической модели на основе модели индивидуального поведения имеет три возможных преимущества. Во-первых, такая связь

ценна сама по себе; во-вторых, вполне вероятно, что она предотвратит спецификацию нереальных моделей; в-третьих, имеется вероятность того, что теория наложит ограничения на спецификацию и сделает ее более правдоподобной и эффективной. Для исследования этой связи потрачено огромное количество интеллектуальных усилий, и один из ведущих ученых в этой области А. Дитон (Deaton, 1986, p. 1768) в своем недавнем обзоре утверждает:

«Сила анализа потребительского спроса заключалась в его тесной связи с теорией и практикой, а важные теоретические предпосылки были столь точными, так как они обеспечивали тесную связь между теорией и интерпретацией опыта. Нельзя проводить прикладной анализ спроса, не принимая во внимание одновременно как статистику, так и экономическую теорию».

Такая точка зрения господствует в эконометрической литературе, посвященной анализу спроса, особенно в работах о системном оценивании функций спроса, которые в совокупности моделируют потребительские расходы. Все это не случайно: многие ограничения, налагаемые теорией полезности, касаются сразу нескольких уравнений, и поэтому они становятся важными лишь при системном оценивании. Самый ранний из подобных подходов, на который нередко ссылаются и сегодня, представлен линейной моделью Стоуна (Stone, 1954). Модель предполагает, что индивиды максимизируют функцию полезности, имеющую простую форму, с учетом своих бюджетных ограничений. Показано, что уравнения спроса в таком случае оказываются линейными как относительно своих переменных, так и относительно своих параметров. Одним из преимуществ модели является сравнительно небольшое число параметров: в системе из n уравнений спроса необходимо оценить лишь $(2n - 1)$ параметр — очевидное достоинство модели при ее оценивании на временных рядах. Наряду с другими приложениями модель была оценена для Великобритании (работы Р. Стоуна и Э. Дитона [Stone, 1954; Deaton, 1975]), а также для ряда развивающихся стран (работа К. Ллача, А. Пауэлла, Р. Уильямса [Lluch, Powell, Williams, 1977]), и исследователи во всех случаях заявляли, что они удовлетворены полученными результатами.

Тем не менее, как замечает после анализа методов оценивания один из сторонников «прямого» пути: «Совсем неудивительно, что авторы, которые при оценке системы уравнений спроса нагромоздили кучу ограничений, признали себя удовлетворенными полученными с таким трудом результатами. Влезшие на гору обычно не склонны критиковать вид, открывшийся им с вершины» (Deaton, 1986, p. 1787). Сторонники другого пути — «прагматики» — после восхищения виртуозностью оценки системы линейных уравнений спроса отмечают, что недостатков в ней столько же, сколько и достоинств. Во-первых, что наиболее очевидно, предположение о линейности не стыкуется с выводом многих исследований о предпочтительности использования других форм зависимости. Во-вторых, в исходной версии модели нет лагов, которые должны были бы присутствовать у функций спроса, оцениваемых на временных рядах. Примечательно и то, что два наиболее заметных эмпирических исследования функций спроса — С. Прайса и Х. Хаутеккера (Prais, Houthakker, 1954, 1971) и Х. Хаутеккера и Л. Тейлора (Houthakker, Taylor, 1970) — игнорируют теорию полезности несмотря на то, что Х. Хаутеккер внес вклад в ее развитие (см., например: Houthakker, 1960).

На чисто практическом уровне есть ряд путей улучшения описанного здесь моделирования функций спроса. Во-первых, может быть предпринят переход от исследования временных рядов к анализу *перекрестных* (cross-section) данных, которые, как правило, более обширны и позволяют исследовать закономерности спроса с учетом характеристик домашних хозяйств (например, состава семьи). Анализ временных рядов обычно требуется для оценки ценовой эластичности спроса, и два очевидных направления улучшения построенных здесь моделей — это учет роста численности населения (например, путем перехода к спросу на душу населения и доходу на душу населения) и учет смещения оценок, имеющего место в системах одновременных уравнений (которое, между прочим, в большой степени игнорировалось теми, кто такие системы оценивал на практике, и было отмечено лишь позже; работа К. Бронсара и Л. Сальва-Бронсар [Bronsard, Salvas-Bronsard, 1984]).

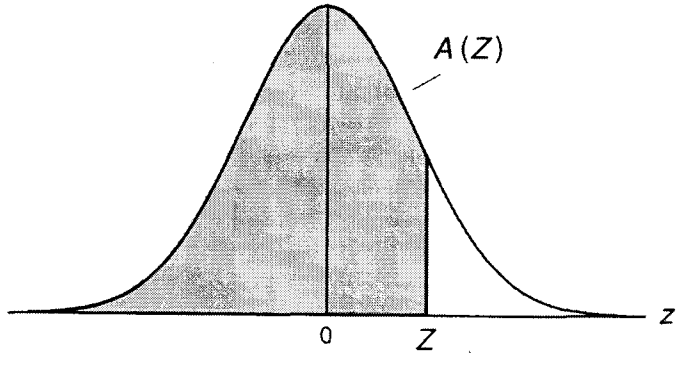
Для тех, кто захочет изучать анализ спроса дальше, рекомендуем доступно написанную работу Дж. Томаса (Thomas, 1987). В ней автор настолько обходит технические аспекты, насколько это возможно при серьезном изложении вопроса, и работа может служить введением как в теорию полезности, так и в эмпирические исследования. Отметим еще одну работу Х. Хаутеккера и Л. Тейлора (Houthakker, Taylor, 1970), где исследуются такие же временные ряды для США (только за более ранний период), какие используются в нашей книге для построения функции спроса, что позволит провести их прямое сравнение. Образцовый анализ связей потоков и запасов в этой работе может служить хорошей иллюстрацией того, что для достижения успеха при построении эконометрических моделей жизненно важны как математические методы, так и здравый смысл.

Приложение А

СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Таблица А.1

Кумулятивное стандартизированное нормальное распределение



$A(Z)$ — это интеграл функции плотности вероятности стандартизированного нормального распределения от $-\infty$ до Z (другими словами, площадь под кривой слева от Z). $A(Z)$ дает вероятность того, что величина нормально распределенной случайной переменной не превысит среднее значение больше, чем на Z стандартных отклонений.

Значения Z, играющие важную роль в книге:

Z	$A(Z)$	
1,645	0,950	нижняя граница правой 5-процентной области
1,960	0,975	нижняя граница правой 2,5-процентной области
2,326	0,990	нижняя граница правой однопроцентной области
2,576	0,995	нижняя граница правой 0,5-процентной области

	A(Z)									
Z	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9700	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987									

Источник: Pearson E.S., Harley H.O. (editors), Biometrika Tables for Statisticians, Cambridge, Cambridge University Press, 1970 (перепечатано с любезного разрешения Biometrika Trustees).

Таблица А.2

t-распределение: критические значения *t*

Число степеней свободы	*Тесты*	*Уровень значимости*					
	Двусторонний	*10%*	*5%*	*2%*	*1%*	*0,2%*	*0,1%*
	Односторонний	*5%*	*2,5%*	*1%*	*0,5%*	*0,1%*	*0,05%*
1		6,314	12,706	31,821	63,657	318,31	636,62
2		2,920	4,303	6,965	9,925	22,327	31,598
3		2,353	3,182	4,541	5,841	10,214	12,924
4		2,132	2,776	3,747	4,604	7,173	8,610
5		2,015	2,571	3,365	4,032	5,893	6,869
6		1,943	2,447	3,143	3,707	5,208	5,959
7		1,895	2,365	2,998	3,499	4,785	5,408
8		1,860	2,306	2,896	3,355	4,501	5,041
9		1,833	2,262	2,821	3,250	4,297	4,781
10		1,812	2,228	2,764	3,169	4,144	4,587
11		1,796	2,201	2,718	3,106	4,025	4,437
12		1,782	2,179	2,681	3,055	3,930	4,318
13		1,771	2,160	2,650	3,012	3,852	4,221
14		1,761	2,145	2,624	2,977	3,787	4,140
15		1,753	2,131	2,602	2,947	3,733	4,073
16		1,746	2,120	2,583	2,921	3,686	4,015
17		1,740	2,110	2,567	2,898	3,646	3,965
18		1,734	2,101	2,552	2,878	3,610	3,922
19		1,729	2,093	2,539	2,861	3,579	3,883
20		1,725	2,086	2,528	2,845	3,552	3,850
21		1,721	2,080	2,518	2,831	3,527	3,819
22		1,717	2,074	2,508	2,819	3,505	3,792
23		1,714	2,069	2,500	2,807	3,485	3,767
24		1,711	2,064	2,492	2,797	3,467	3,745
25		1,708	2,060	2,485	2,787	3,450	3,725
26		1,706	2,056	2,479	2,779	3,435	3,707
27		1,703	2,052	2,473	2,771	3,421	3,690
28		1,701	2,048	2,467	2,763	3,408	3,674
29		1,699	2,045	2,462	2,756	3,396	3,659
30		1,697	2,042	2,457	2,750	3,385	3,646
40		1,684	2,021	2,423	2,704	3,307	3,551
60		1,671	2,000	2,390	2,660	3,232	3,460
120		1,658	1,980	2,358	2,617	3,160	3,373
∞		1,645	1,960	2,326	2,576	3,090	3,291

Источник: Pearson E.S., Harley H.O. (editors), Biometrika Tables for Statisticians, Cambridge, Cambridge University Press, 1970 (перепечатано с любезного разрешения Biometrika Trustees).

Пример. Для распределения с 25 степенями свободы вероятность того, что t будет больше 2,060, равна 0,025 и вероятность того, что t будет меньше 2,060, составит 0,025. Если гипотеза отвергается в обеих крайних областях, т. е. в двустороннем тесте, то уровень значимости равен 0,05 (5%). Если гипотеза отвергается в одной крайней области, т. е. при одностороннем тесте, то уровень значимости составит 0,025 (2,5%). Более подробные разъяснения см. в главе 3.

Таблица А.3

F-распределение: критические значения F с v_1 и v_2 степенями свободы, уровень значимости в 5%

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,86	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица А.3 (продолжение)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999,5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Источник: Pearson E.S., Harley H.O. (editors), Biometrika Tables for Statisticians, Cambridge, Cambridge University Press, 1970 (перепечатано с любезного разрешения Biometrika Trustees).

Таблица А.4

Распределение χ^2: критические значения χ^2 для уровней значимости в 5, 1 и 0,1%

Число степеней свободы	*5%*	*1%*	*0,1%*
1	3,8415	6,6349	10,828
2	5,9915	9,2103	13,816
3	7,8147	11,3449	16,266
4	9,4877	13,2767	18,467
5	11,0705	15,0863	20,515
6	12,5916	16,8119	22,458
7	14,0671	18,4753	24,322
8	15,5073	20,0902	26,125
9	16,9190	21,6660	27,877
10	18,3070	23,2093	29,588
11	19,6751	24,7250	31,264
12	21,0261	26,2170	32,909
13	22,3620	27,6882	34,528
14	23,6848	29,1412	36,123
15	24,9958	30,5779	37,697
16	26,2962	31,9999	39,252
17	27,5871	33,4087	40,790
18	28,8693	34,8053	42,312
19	30,1435	36,1909	43,820
20	31,4104	37,5662	45,315
21	32,6706	38,9322	46,797
22	33,9244	40,2894	48,268
23	35,1725	41,6384	49,728
24	36,4150	42,9798	51,179
25	37,6525	44,3141	52,618
26	38,8851	45,6417	54,052
27	40,1133	46,9629	55,476
28	41,3371	48,2782	56,892
29	42,5570	49,5879	58,301
30	43,7730	50,8922	59,703
40	55,7585	63,6907	73,402
50	67,5048	76,1539	86,661
60	79,0819	88,3794	99,607
70	90,5312	100,425	112,317
80	101,879	112,329	124,839
90	113,145	124,116	137,208
100	124,342	135,807	149,449

Источник: Pearson E.S., Harley H.O. (editors), Biometrika Tables for Statisticians, Cambridge, Cambridge University Press, 1970 (перепечатано с любезного разрешения Biometrika Trustees).

Таблица А.5

d-статистика Дарбина–Уотсона: d_L и d_U, уровень значимости в 5%

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Таблица А.5 (продолжение)

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

Примечание: n — число наблюдений, k — число объясняющих переменных (без учета постоянного члена).
Источник: Durbin, Watson (1951) (перепечатано с любезного разрешения Biometrika Trustees).

Приложение Б

НАБОР ДАННЫХ

В этом приложении представлены три набора данных для упражнений, приведенных в книге.

Таблица Б.1

В таблице приведены данные о величине личного дохода (PI), личного располагаемого дохода (DPI), совокупных личных расходах (TPE) и о 20 видах потребительских расходов населения США за период 1959–1983 гг. Все данные приводятся в миллиардах долларов в ценах 1972 г. Данные о величине DPI и о разных видах потребительских расходов впервые используются в упражнении 2.4. Данные о переменной времени (TIME) объясняются и применяются в первый раз в упражнении 2.5. Данные о величине PI используются в упражнении 6.17. Данные о величине TPE не применяются в упражнениях и являются дополнением.

Каждой переменной, которая используется в регрессионной модели, необходимо дать название. В большинстве статистических пакетов вы можете выбрать для переменных любое название при некоторых ограничениях. Обычно это требования использовать в названии стандартные буквы или числа и ограничение на длину этого названия. В третьем столбце таблицы приведены названия, используемые для обозначения соответствующих переменных.

Данные последнего столбца таблицы дают возможность обнаружить допущенные ошибки. Если статистический пакет позволяет без затруднений рассчитывать средние значения, сделайте это для соответствующего ряда данных и сравните полученную величину с числом в последнем столбце. Они должны совпадать. Если это не так, то вы сделали по крайней мере одну ошибку. В случае если статистический пакет не позволяет без проблем получить среднее значение, рассчитайте уравнение регрессии — любое уравнение с соответствующим рядом в качестве зависимой переменной. Среднее значение зависимой переменной всегда выводится как одна из контрольных статистик.

Таблица Б.1

Личные потребительские расходы населения США (млрд. долл., в ценах 1972 г.)*

Категория	*Статья*	*Название переменной*	*1959*	*1960*	*1961*	*1962*	*1963*	*1964*	*1965*
	Дата	DATE	1959,0	1960,0	1961,0	1962,0	1963,0	1964,0	1965,0
	Время	TIME	1,0	2,0	3,0	4,0	5,0	6,0	7,0
ДОХОД	Личный доход	PI	544,9	559,7	575,4	602,0	622,9	658,0	700,4
	Личный располагаемый доход	DPI	479,7	489,7	503,8	524,9	542,3	580,8	616,3
РАСХОДЫ	Совокупные личные расходы	TPE	440,4	452,0	461,4	482,0	500,5	528,0	557,5
Текущие расходы	Питание	FOOD	99,7	100,9	102,5	103,5	104,6	108,8	113,7
	Одежда	CLOT	36,3	36,6	37,3	38,9	39,6	42,6	44,2
	Бензин	GASO	13,7	14,2	14,3	14,9	15,3	16,0	16,8
	Моторное масло	FUEL	5,2	5,0	4,7	4,7	4,9	5,2	5,5
	Табак	TOB	10,7	10,9	11,2	11,2	11,4	11,3	11,6
	Косметика	COSM	3,1	3,5	3,9	4,2	4,5	4,8	5,3
	Лекарства	PHAR	3,5	3,9	4,3	4,7	4,9	5,1	5,3
Услуги	Плата за жилье	HOUS	60,9	64,0	67,0	70,7	74,0	77,4	81,6
	Газ	GAS	3,9	4,1	4,3	4,7	4,9	5,1	5,3
	Вода	WAT	2,0	2,2	2,3	2,5	2,7	2,8	2,9
	Телефон	TELE	4,7	5,0	5,4	5,7	6,1	6,6	7,3
	Местный транспорт	LOCT	3,9	3,9	3,6	3,6	3,5	3,4	3,3
	Воздушный транспорт	AIR	0,9	0,9	1,0	1,1	1,2	1,4	1,6
	Медицинские услуги	DOC	8,8	9,0	9,1	9,8	10,2	11,9	12,1
	Услуги стоматологов	DENT	3,2	3,2	3,3	3,5	3,4	3,9	4,0
	Отдых	REC	9,6	10,0	10,4	10,9	11,3	11,6	11,9
	Частное образование	PRIV	5,6	6,0	6,3	6,6	7,0	7,4	8,1
Товары длительного пользования	Кухонное оборудование	KIT	4,2	4,2	4,2	4,4	4,6	5,1	5,2
	Посуда	TAB	2,6	2,5	2,5	2,6	2,5	2,8	3,1
	Ювелирные изделия	JEWL	2,2	2,2	2,2	2,3	2,5	2,6	2,9

* Таблица является неполной: в ней приведены данные лишь о некоторых статьях потребительских расходов каждого вида (текущие расходы, услуги, товары длительного пользования).

Источники: National Income and Product Accounts of the U.S., 1929–76, Table 2.5; Survey of Current Business, July 1982, 1983, 1984, Table 2.5.

Таблица Б.1 (продолжение)

Категория	*Статья*	*Название переменной*	*1966*	*1967*	*1968*	*1969*	*1970*	*1971*	*1972*	*1973*	*1974*	*1975*
	Дата	DATE	1966,0	1967,0	1968,0	1969,0	1970,0	1971,0	1972,0	1973,0	1974,0	1975,0
	Время	TIME	8,0	9,0	10,0	11,0	12,0	13,0	14,0	15,0	16,0	17,0
ДОХОД	Личный доход	PI	740,6	774,4	816,2	853,5	876,8	900,0	951,4	1007,9	1004,8	1010,8
	Личный располагаемый доход	DPI	646,8	673,5	701,3	722,5	751,6	779,2	810,3	865,3	858,4	875,8
РАСХОДЫ	Совокупные личные расходы	TPE	585,7	602,7	634,4	657,9	672,1	696,8	737,1	768,5	763,6	780,2
Текущие расходы	Питание	FOOD	116,6	118,6	123,4	125,9	129,4	130,0	132,4	129,4	128,1	132,3
	Одежда	CLOT	46,9	46,9	49,0	50,0	49,4	51,8	55,4	59,3	58,7	60,9
	Бензин	GASO	17,8	18,4	19,9	21,4	22,9	24,2	25,4	26,2	24,8	25,6
	Моторное масло	FUEL	5,6	5,6	5,3	5,0	4,7	4,6	5,0	5,4	4,2	4,2
	Табак	TOB	11,7	11,8	11,7	11,4	11,7	11,8	12,2	12,8	13,0	12,9
	Косметика	COSM	5,9	6,3	6,6	6,8	7,0	7,1	7,4	7,9	7,8	7,4
	Лекарства	PHAR	5,5	5,8	6,4	7,0	7,7	8,0	8,7	9,3	9,8	9,7
Услуги	Плата за жилье	HOUS	85,3	89,1	93,5	98,4	102,0	106,4	112,5	118,2	124,2	128,3
	Газ	GAS	5,4	5,7	5,9	6,2	6,3	6,4	6,6	6,4	6,5	6,6
	Вода	WAT	3,0	3,0	3,1	3,3	3,5	3,6	3,9	4,1	4,3	4,4
	Телефон	TELE	8,1	8,7	9,5	10,4	11,2	11,7	12,4	13,7	14,4	15,9
	Местный транспорт	LOCT	3,3	3,2	3,3	3,5	3,4	3,4	3,4	3,4	3,5	3,5
	Воздушный транспорт	AIR	1,7	2,1	2,4	2,8	2,7	2,8	3,1	3,4	3,7	3,6
	Медицинские услуги	DOC	12,1	12,5	12,8	13,6	14,4	14,8	15,7	16,9	17,2	17,8
	Услуги стоматологов	DENT	4,1	4,3	4,4	4,8	5,1	5,1	5,3	6,1	6,2	6,4
	Отдых	REC	12,4	12,7	13,4	14,1	14,6	15,1	15,8	16,9	17,6	17,9
	Частное образование	PRIV	8,8	9,3	10,0	10,6	10,9	11,2	11,7	11,9	11,7	12,1
Товары длительного пользования	Кухонное оборудование	KIT	5,8	6,0	6,6	7,0	7,3	7,9	8,9	9,9	9,9	9,3
	Посуда	TAB	3,5	3,7	3,8	3,8	3,7	3,8	4,0	4,2	4,1	3,7
	Ювелирные изделия	JEWL	3,6	3,9	4,1	4,1	4,1	4,3	4,6	5,2	5,4	5,5

Таблица Б.1 (продолжение)

Категория	*Статья*	*Название переменной*	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*	*1982*	*1983*	*Среднее*
	Дата	DATE	1976,0	1977,0	1978,0	1979,0	1980,0	1981,0	1982,0	1983,0	1971,000
	Время	TIME	18,0	19,0	20,0	21,0	22,0	23,0	24,0	25,0	13,000
ДОХОД	Личный доход	PI	1056,2	1105,4	1162,3	1200,7	1209,5	1248,6	1254,4	1284,6	908,856
	Личный располагаемый доход	DPI	906,8	942,9	988,8	1015,5	1021,6	1049,3	1058,3	1095,4	780,032
РАСХОДЫ	Совокупные личные расходы	TPE	823,1	864,3	903,2	927,6	931,8	950,9	963,3	1009,2	707,184
Текущие расходы	Питание	FOOD	139,7	145,2	146,1	149,3	153,2	153,0	154,6	161,2	128,084
	Одежда	CLOT	63,8	67,5	73,6	76,7	77,9	82,6	84,2	88,5	56,744
	Бензин	GASO	26,8	27,7	28,3	27,4	25,1	25,1	25,3	26,1	21,744
	Моторное масло	FUEL	4,6	4,4	4,7	4,7	3,9	3,6	3,6	4,0	4,732
	Табак	TOB	13,7	13,1	13,5	13,7	13,6	14,0	13,7	13,0	12,304
	Косметика	COSM	7,5	7,8	8,1	8,4	8,3	8,3	8,1	8,1	6,564
	Лекарства	PHAR	10,0	10,2	10,4	10,8	10,7	10,6	10,3	10,2	7,712
Услуги	Плата за жилье	HOUS	134,9	141,3	148,5	154,8	159,8	164,8	167,5	171,3	111,856
	Газ	GAS	6,7	6,5	6,7	6,6	6,6	6,3	6,4	6,0	5,844
	Вода	WAT	4,3	4,4	4,5	4,8	5,1	5,1	5,1	5,1	3,680
	Телефон	TELE	17,1	18,3	20,0	21,6	22,7	23,3	24,1	24,2	13,124
	Местный транспорт	LOCT	3,6	3,6	3,7	3,8	3,5	3,2	3,2	3,1	3,472
	Воздушный транспорт	AIR	4,0	4,3	4,7	5,1	4,6	4,1	3,7	3,8	2,828
	Медицинские услуги	DOC	18,0	19,2	18,6	20,1	21,5	22,0	22,4	23,3	15,352
	Услуги стоматологов	DENT	6,9	7,2	8,1	7,9	8,1	8,5	8,6	8,5	5,668
	Отдых	REC	19,1	20,4	21,8	22,2	23,4	26,1	27,7	29,8	16,668
	Частное образование	PRIV	12,2	12,2	12,7	13,1	13,3	13,7	13,6	13,7	10,388
Товары длительного пользования	Кухонное оборудование	KIT	9,7	10,5	11,1	11,9	12,1	12,4	11,9	12,7	8,112
	Посуда	TAB	3,9	4,1	4,3	4,5	4,4	4,4	4,3	4,7	3,660
	Ювелирные изделия	JEWL	6,1	6,3	6,8	6,7	6,3	6,6	6,7	7,0	4,568

Таблица Б.2

В таблице представлены дефляторы цен для TPE и для 20 видов потребительских расходов из табл. Б.1. Эти ряды в форме индекса (1972 = 100%) показывают, как росли цены во времени. В каждом конкретном случае часть происшедших изменений объяснялась общим ростом цен, а часть — изменениями рынка товара или технологии его производства. Эти данные впервые используются в упражнении 5.1, где вам предлагается построить относительный индекс цены товара, т. е. индекс, показывающий, происходило ли изменение цены на товар более или менее быстрыми темпами, чем общий рост цен. Именно такая относительная цена товара, а не его номинальная цена, определяет спрос на этот товар. Названия переменных в третьем столбце таблицы те же, что и в табл. Б.1, только с добавлением «P» в начале. Средние значения для рядов данных представлены в последней колонке для обнаружения ошибок, как и в табл. Б.1.

Таблица Б.3

Данные этой таблицы используются только в упражнении 9.5. Средние значения для рядов данных представлены в последней колонке для обнаружения ошибок, как и в табл. Б.1 и Б.2.

Таблица Б.2

Дефляторы цен для личных потребительских расходов (1972=100%)

Категория	*Статья*	*Название переменной*	*1959*	*1960*	*1961*	*1962*	*1963*	*1964*	*1965*	*1966*
ЦЕНЫ	Совокупные личные расходы	PTPE	70,6	71,9	72,6	73,7	74,8	75,9	77,2	79,4
Текущие расходы	Питание	PFOOD	69,0	69,8	70,6	71,4	72,4	73,7	75,5	79,4
	Одежда	PCLOT	72,0	72,9	73,4	73,7	74,5	75,0	75,8	78,0
	Бензин	PGASO	82,2	84,5	83,9	84,5	84,5	84,4	87,5	89,5
	Моторное масло	PFUEL	77,7	76,2	79,3	79,3	81,0	79,1	80,8	83,0
	Табак	PTOB	61,0	63,2	63,8	64,4	65,8	67,1	69,8	72,9
	Косметика	PCOSM	84,6	84,5	84,3	85,1	85,4	85,5	85,1	84,1
	Лекарства	PPHAR	98,8	98,9	97,8	96,2	95,4	95,2	94,8	95,1
Услуги	Плата за жилье	PHOUS	73,8	75,1	76,3	77,4	78,4	79,3	80,3	81,5
	Газ	PGAS	74,9	79,8	80,9	80,8	80,8	81,1	81,4	81,9
	Вода	PWAT	58,4	60,2	61,8	63,4	65,4	66,5	68,0	70,5
	Телефон	PTELE	88,3	89,6	89,9	89,9	90,0	90,1	88,9	87,0
	Местный транспорт	PLOCT	51,1	52,6	54,7	57,1	58,5	60,3	61,7	64,5
	Воздушный транспорт	PAIR	68,9	73,6	78,4	82,4	76,4	76,1	76,5	76,6
	Медицинские услуги	PDOC	56,1	57,6	59,1	60,8	62,2	63,7	65,8	69,9
	Услуги стоматологов	PDENT	60,9	62,1	62,4	64,0	65,9	67,5	69,6	72,0
	Отдых	PREC	61,1	63,7	65,8	67,5	69,4	71,8	73,9	76,6
	Частное образование	PPRIV	61,5	62,8	63,8	65,3	66,9	68,5	70,5	73,6
Товары длительного пользования	Кухонное оборудование	PKIT	103,6	102,2	100,1	97,8	95,9	94,9	92,6	91,1
	Посуда	PTAB	68,2	70,3	71,1	72,9	75,3	76,6	76,5	78,1
	Ювелирные изделия	PJEWL	86,5	86,4	86,5	86,5	86,9	92,2	89,6	85,4

Таблица Б.2 (продолжение)

Категория	*Статья*	*Название переменной*	*1967*	*1968*	*1969*	*1970*	*1971*	*1972*	*1973*	*1974*	*1975*
ЦЕНЫ	Совокупные личные расходы	PTPE	81,4	84,6	88,4	92,5	96,5	100,0	105,7	116,3	125,2
Текущие расходы	Питание	PFOOD	80,0	83,1	87,5	92,5	94,9	100,0	114,3	130,8	140,1
	Одежда	PCLOT	81,4	86,0	91,0	94,8	97,8	100,0	103,6	110,5	114,2
	Бензин	PCASO	92,4	93,8	97,0	97,9	98,7	100,0	109,4	147,7	157,7
	Моторное масло	PFUEL	85,6	88,3	90,2	93,6	99,5	100,0	114,8	182,4	197,4
	Табак	PTOB	75,5	80,3	85,6	92,0	95,7	100,0	102,8	107,8	115,5
	Косметика	PCOSM	85,5	88,1	92,1	94,4	97,4	100,0	102,7	113,9	128,3
	Лекарства	PPHAR	94,7	94,9	95,9	98,0	99,7	100,0	100,3	103,8	112,5
Услуги	Плата за жилье	PHOUS	83,2	85,4	88,4	92,1	96,5	100,0	104,8	110,6	116,8
	Газ	PGAS	81,7	82,5	84,0	88,6	95,0	100,0	104,5	117,7	140,9
	Вода	PWAT	72,2	75,6	80,7	87,0	96,3	100,0	105,2	111,5	122,3
	Телефон	PTELE	88,0	88,1	89,2	90,3	94,7	100,0	102,6	107,0	110,4
	Местный транспорт	PLOCT	69,0	73,3	78,3	88,6	95,2	100,0	101,2	104,6	112,7
	Воздушный транспорт	PAIR	76,8	78,4	84,3	90,9	97,5	100,0	103,9	112,7	122,7
	Медицинские услуги	PDOC	74,8	78,9	84,4	90,7	97,0	100,0	103,1	112,5	126,4
	Услуги стоматологов	PDENT	75,6	79,7	85,4	90,2	96,0	100,0	103,0	110,9	122,3
	Отдых	PREC	79,6	84,4	88,6	93,1	97,5	100,0	103,3	109,8	117,2
	Частное образование	PPRIV	76,8	80,3	85,1	90,6	95,3	100,0	107,3	120,4	131,2
Товары длительного пользования	Кухонное оборудование	PKIT	91,2	93,1	95,1	97,5	99,5	100,0	100,1	105,2	116,9
	Посуда	PTAB	80,3	86,8	90,4	93,2	95,8	100,0	105,7	118,6	139,5
	Ювелирные изделия	PJEWL	86,5	89,2	94,2	95,8	97,7	100,0	103,6	109,5	117,4

Таблица Б.2 (продолжение)

Категория	*Статья*	*Название переменной*	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*	*1982*	*1983*	*Среднее*
ЦЕНЫ	Совокупные личные расходы	PTPE	131,7	139,3	149,1	162,5	179,0	194,5	206,0	213,6	114,496
Текущие расходы	Питание	PFOOD	143,4	149,6	164,9	182,4	196,6	213,3	222,1	226,5	120,152
	Одежда	PCLOT	117,9	122,5	125,5	129,2	134,3	138,4	141,0	143,6	101,080
	Бензин	PGASO	164,3	173,7	181,3	243,2	337,9	376,4	356,6	344,9	154,156
	Моторное масло	PFUEL	212,0	239,9	252,7	340,2	470,8	571,7	565,3	531,2	194,880
	Табак	PTOB	120,4	126,2	133,0	141,0	152,0	164,2	182,7	218,4	104,844
	Косметика	PCOSM	135,6	143,3	151,1	161,5	176,2	194,4	210,5	222,9	119,060
	Лекарства	PPHAR	119,3	127,0	135,9	145,7	159,1	176,6	194,7	211,4	117,668
Услуги	Плата за жилье	PHOUS	123,4	131,6	141,2	152,5	166,5	183,2	199,3	212,1	112,388
	Газ	PGAS	164,8	195,6	214,9	249,2	297,0	336,8	404,2	473,4	154,896
	Вода	PWAT	135,9	150,5	167,4	175,1	186,1	208,4	233,2	252,6	114,968
	Телефон	PTELE	114,3	115,6	117,0	116,6	118,7	130,1	143,5	152,6	103,696
	Местный транспорт	PLOCT	122,7	129,3	134,9	143,6	166,3	196,1	215,0	220,3	104,464
	Воздушный транспорт	PAIR	132,9	141,0	147,5	159,2	217,0	273,6	302,0	319,9	126,768
	Медицинские услуги	PDOC	140,6	153,6	166,2	181,6	200,5	222,5	243,5	262,4	117,356
	Услуги стоматологов	PDENT	130,2	139,9	149,7	162,3	181,6	198,9	214,3	228,7	111,724
	Отдых	PREC	122,4	127,6	134,2	142,6	153,1	163,4	171,8	178,8	104,688
	Частное образование	PPRIV	140,2	149,1	161,0	177,4	198,2	218,4	231,7	242,4	117,532
Товары длительного пользования	Кухонное оборудование	PKIT	123,4	127,8	134,4	141,7	148,8	157,4	167,0	171,9	113,968
	Посуда	PTAB	148,3	154,5	163,8	177,3	195,7	215,3	224,4	228,6	120,288
	Ювелирные изделия	PJEWL	121,2	123,0	129,5	141,5	176,7	183,7	179,5	182,4	113,256

Источники: National Income and Product Accounts of the U.S., 1929–76, Table 7.12; Survey of Current Business, July 1982, 1983, 1984, Table 7.12.

Таблица Б.3

Личные потребительские расходы населения США, 1977—1982 гг. (млрд. долл. 1972 г., без учета сезонных колебаний)*

Категория	*Название переменной*	*77Q1*	*77Q2*	*77Q3*	*77Q4*	*78Q1*	*78Q2*	*78Q3*	*78Q4*	*79Q1*	*79Q2*	*79Q3*	*79Q4*
Совокупные личные расходы	TOTEX	201,8	214,5	216,6	230,9	211,6	224,8	226,3	240,0	221,0	229,0	231,2	245,7
Текущие расходы	NONDUR	74,8	82,2	83,0	93,1	77,9	84,2	85,9	95,9	80,5	86,2	87,6	98,2
Питание	FOOD	37,4	42,7	44,5	45,8	40,7	42,9	43,6	44,5	40,6	43,8	44,9	46,6
Одежда	CLOTHING	14,9	19,6	19,4	23,9	20,2	21,6	20,9	20,5	21,9	21,6	20,0	20,8
Бензин	GASOLINE	6,4	7,1	7,3	6,8	6,4	7,2	7,4	7,2	6,9	6,8	7,0	6,6
Моторное масло	FUELOIL	1,7	0,8	0,7	1,3	1,7	0,9	0,8	1,3	1,8	1,0	0,8	1,3
Услуги	SERVICES	97,8	97,2	99,1	99,0	102,6	102,4	103,9	103,1	107,2	106,0	107,0	107,2
Плата за жилье	HOUSING	34,8	35,1	35,5	35,7	36,3	36,9	37,5	37,7	38,1	38,5	39,0	39,1
Газ и электричество	GASELEC	7,3	4,7	5,1	5,5	7,7	4,9	5,2	5,6	8,0	5,0	5,0	5,6
Транспорт	TRANSPT	7,7	8,3	8,5	8,3	8,0	8,8	8,9	8,4	8,3	9,0	9,1	8,7
Товары длительного пользования	DURABLES	29,2	35,2	34,5	38,9	30,8	38,3	36,5	41,0	33,2	36,9	36,6	40,3
Автомобили	VEHICLES	13,8	17,3	16,5	15,9	14,6	19,0	17,1	16,3	15,5	16,5	15,9	14,7
Мебель	FURNITUR	11,1	12,6	12,9	16,4	11,5	13,5	13,8	17,6	12,6	14,3	15,0	18,5
Фиктивная переменная для I квартала	D1	1	0	0	0	1	0	0	0	1	0	0	0
Фиктивная переменная для II квартала	D2	0	1	0	0	0	1	0	0	0	1	0	0
Фиктивная переменная для III квартала	D3	0	0	1	0	0	0	1	0	0	0	1	0
Фиктивная переменная для IV квартала	D4	0	0	0	1	0	0	0	1	0	0	0	1
Временной тренд	TREND	1	2	3	4	5	6	7	8	9	10	11	12

* Таблица является неполной: в ней приведены данные лишь о некоторых статьях потребительских расходов каждого вида (текущие расходы, услуги товары длительного пользования).

Таблица Б.3 (продолжение)

Категория	*Название переменной*	*80Q1*	*80Q2*	*80Q3*	*80Q4*	*81Q1*	*81Q2*	*81Q3*	*81Q4*	*82Q1*	*82Q2*	*82Q3*	*82Q4*	*Среднее*
Совокупные личные расходы	TOTEXP	224,2	230,2	231,7	244,1	229,4	239,3	240,0	247,5	232,8	243,4	242,1	251,5	231,2
Текущие расходы	NONDUR	81,4	88,8	89,1	95,7	83,0	90,9	91,3	97,1	84,4	91,8	91,9	96,1	88,0
Питание	FOOD	42,8	45,5	45,0	47,3	40,2	45,0	45,5	50,9	40,9	46,0	47,2	49,8	44,3
Одежда	CLOTHING	14,5	15,9	15,2	21,6	13,7	16,9	17,8	25,0	15,1	17,9	18,3	25,2	19,3
Бензин	GASOLINE	5,8	6,3	6,5	6,4	5,9	6,4	6,6	6,3	6,1	6,7	6,6	6,3	6,6
Моторное масло	FUELOIL	1,0	1,0	1,0	1,0	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,8	1,1
Услуги	SERVICES	110,3	108,3	109,9	110,3	113,8	112,3	113,6	113,5	117,0	115,7	116,7	116,8	107,9
Плата за жилье	HOUSING	39,5	39,7	40,2	40,4	40,9	41,5	42,1	42,2	42,5	42,8	43,1	43,0	39,3
Газ и электричество	GASELEC	7,7	5,1	5,7	5,9	8,0	5,3	5,5	6,0	8,3	5,5	5,4	5,8	6,0
Транспорт	TRANSPT	8,2	8,4	8,6	8,0	7,8	8,2	8,4	7,9	7,6	8,2	8,3	7,7	8,3
Товары длительного пользования	DURABLES	32,5	33,1	33,5	38,2	32,5	36,3	35,0	37,1	31,1	35,9	33,5	39,2	35,4
Автомобили	VEHICLES	14,5	13,1	13,8	13,1	14,3	14,9	14,9	11,9	13,8	15,3	14,2	14,1	15,0
Мебель	FURNITUR	13,2	14,2	14,7	18,0	13,3	14,9	15,1	18,3	12,7	14,5	14,6	17,9	14,6
Фиктивная переменная для I квартала	D1	1	0	0	0	1	0	0	0	1	0	0	0	0,25
Фиктивная переменная для II квартала	D2	0	1	0	0	0	1	0	0	0	1	0	0	0,25
Фиктивная переменная для III квартала	D3	0	0	1	0	0	0	1	0	0	0	1	0	0,25
Фиктивная переменная для IV квартала	D4	0	0	0	1	0	0	0	1	0	0	0	1	0,25
Временной тренд	TREND	13	14	15	16	17	18	19	20	21	22	23	24	12,50

Источник: Рассчитано по данным Survey of Current Business, July 1982, 1983, Tables 9.2 and 7.2.

Almon Shirley. The distributed lag between capital appropriations and expenditures (1965). — Econometrica 33(1): 178–196.

Ash J.C.K., Smyth David J. Forecasting the United Kingdom Economy. — Farnborough, Hampshire: Saxon House (1973).

Beach Charles M., MacKinnon James G. A maximum likelihood procedure for regression with autocorrelated errors (1978). — Econometrica 46(1): 51–58.

Betancourt Roger, Kelejian Harry. Lagged endogenous variables and the Cochrane-Orcutt procedure (1981). — Econometrica 49(4): 1073–1078.

Box G.E.P., Cox D.R. An analysis of transformations (1964). — Journal of the Royal Statistical Society Series B 26(2): 211–243.

Box G.E.P., Jenkins G.M. Time Series Analysis. — San Francisco: Holden Day (1970).

Breusch T.S., Godfrey L. A review of recent work on testing for auto-correlation in dynamic simultaneous models. — In: Currie D., Nobay R., Peel D. (eds.), Macroeconomic Analysis. — London: Croom Helm (1981).

Bronsard Camille, Salvas-Bronsard Lise. On price exogeneity in complete demand systems (1984). — Journal of Econometrics 24(3): 235–247.

Brown T.M. Habit persistence and lags in consumer behaviour (1952). — Econometrica 20(3): 355–371.

Cagan P.D. The monetary dynamics of hyperinflation. — In: Friedman Milton (ed.), Studies in the Quantity Theory of Money. — Chicago: University of Chicago Press (1956).

Chow Gregory C. Tests of equality between sets of coefficients in two linear regressions (1960). — Econometrica 28(3): 591–605.

Cobb Charles W., Douglas Paul H. A theory of production (1928). — American Economic Review 18(1, suppl.): 139–165.

Cochrane D., Orcutt G.H. Application of least squares regression to relationshi ps containing autocorrelated error terms (1949). — Journal of the American Statistical Association 44(245): 32–61.

Cooper R.L. The predictive performance of quarterly econometric models of the United States. — In: Hickman B.G. (ed.), Econometric Models of Cyclic Behavior. — New York: Columbia University Press (1972).

Cramer J.S. Econometric Applications of Maximum Likelihood Methods. — Cambridge: Cambridge University Press (1986).

Davidson James E.H., Hendry David F., Srba Frank, Yeo Stephen. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom (1978). — Economic Journal 88(352): 661–692.

Deaton Angus S. Models and Projections of Demand in Post-War Britain. — London: Chapman and Hall (1975).

Deaton Angus S. Demand analysis. — In: Griliches Zvi , Intriligator Michael D. (eds.), Handbook of Econometrics, Vol. 3. — Amsterdam: North-Holland (1986).

Dougherty Christopher, Jones A. David. The determinants of birth weight (1982). — American Journal of Obstetrics and Gynecology 144(2): 190–200.

Dufour Jean-Marie. Dummy variables and predictive tests for structural change (1980). — Economic Letters 6(3): 241–247.

Durbin J. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables (1970). — Econometrica 38(3): 410–421.

Durbin J., Watson G.S. Testing for serial correlation in least squares regression I (1950). — Biometrica 37(3–4): 409–428.

Eisner Robert. A permanent income theory for investment: some empirical explorations (1967). — American Economic Review 57(3): 363–390.

Ezekiel Mordecai The cobweb theorem (1938). — Quarterly Journal of Economics 52(2): 255–280.

Friedman Milton. A Theory of the Consumption Function. — Princeton: Princeton University Press (1957).

Friedman Milton. The demand for money: some theoretical and empirical results (1959). — Journal of Political Economy 67(4): 327–351.

Geary R.C. A note on residual heterovariance and estimation efficiency in regression (1966). — The American Statistician 20(4): 30–31.

Glejser H. A new test for heteroskedasticity (1969). — Journal of the American Statistical Association 64(325): 316–323.

Goldfeld Stephen M., Quandt Richard E. Some tests for homoscedasticity (1965). — Journal of the American Statistical Association 60(310): 539–547.

Harvey Andrew C. The Econometric Analysis of Time Series.— Deddington, Oxford: Philip Allan (1981).

Hendry David F. Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. — In: Ormerod Paul (ed.), Modelling the Economy. — London, Heinemann (1979).

Hendry David F., Mizon Grayham E. Serial correlation as a convenient simplification, not a nuisance (1978). — Economic Journal 88(351): 549–563.

Houthakker H.S. Additive preferences (1960). — Econometrica 28(2): 244–257.

Houthakker H.S., Taylor Lester D. Consumer Demand in the United States: Analysis and Projections (2 ed.). — Cambridge, Mass.: Harvard University Press (1970).

Johnston John. Econometric Methods (3 ed.). — London: McGraw-Hill (1984).

Kaldor Nicholas. Causes of the Slow Rate of Economic Growth of the United Kingdom. — Cambridge: Cambridge University Press (1966).

Kendall M.G. Hiawatha designs an experiment (1959). — The American Statistician 13(5): 23–24.

Kendrick John W., Grossman Elliot S. Productivity in the United States: Trends and Cycles. — Baltimore: John Hopkins (1980).

Kennedy Peter E. A Guide to Econometrics. — Oxford: Blackwell (1985).

Keynes John Maynard. The General Theory of Employment, Interest and Money. — London: Macmillan (1936).

Kish Leslie. Survey Sampling. — New York: John Wiley (1965).

Kmenta Jan. Elements of Econometrics (2 ed.). — New York: Macmillan (1986).

Koyck L.M. Distributed Lags and Investment Analysis. — Amsterdam: North Holland (1954).

Kuh Edwin, Meyer John R. How extraneous are extraneous estimates? (1957). — Review of Economics and Statistics 39(4): 380–393.

Lintner John. Distribution of incomes of corporations among dividends, retained earnings and taxes (1956). — American Economic Review 46(2): 97–113.

Liviatan Nissan. Tests of the Permanent-Income Hypothesis based on a reinterview savings survey (1963). — In: Christ Carl (ed.), Measurement in Economics. — Stanford: Stanford University Press.

Lluch Constantino, Powell Alan A., Williams Ross A. Patterns in Household Demand and Saving. — New York: Oxford University Press (1977).

Mandeville Bernard. The Grumbling Hive (1705), published anonymously as a pamphlet, reprinted in: The Fable of the Bees, or Private Vices, Public Benefits (modern edition with preface by F.B. Kaye). — Oxford: Oxford University Press (1924).

Mood A.M., Graybill F.P. Introduction to the Theory of Statistics. – New York: McGraw-Hill (1963).

Moser Claus, Kalton Graham. Survey Methods in Social Investigation (2 ed.). — London: Heinemann (1970).

Muth John F. Rational expectations and the theory of price movements (1961). — Econometrica 29(3): 315–335.

Nelson Charles R. The prediction performance of the FRB-MIT-PENN model of the U.S. economy (1972). — American Economic Review 62(5): 902–917.

Nelson Charles R. Applied Time Series Analysis. — San Francisco: Holden Day (1973).

Nerlove Marc. Returns to scale in electricity supply. — In: Christ Carl (ed.), Measurement in Economics. — Stanford: Stanford University Press (1963).

Oxley Leslie T., Roberts Colin J. Pitfalls in the application of the Cochrane-Orcutt technique (1982). — Oxford Bulletin of Economics and Statistics 44(3): 227–240.

Park Rolla Edward, Mitchell Bridger M. Estimating the autocorrelated error model with trended data (1980). — Journal of Econometrics 13(2): 185–201.

Peach James T., Webb James L. Randomly specified macroeconomic models: some implications for model selection (1983). — Journal of Economic Issues 17(3): 697–720.

Pesaran M. Hashem, Smith R.P., Yeo J. Stephen. Testing for structural stability and predictive failure: a review (1985). — Manchester School 53(3): 280–295.

Pitayanon Sumalee. Thailand's experience in manpower planning and labour market analysis. — In: Amjad Rashid (ed.), Human Resource Planning: The Asian Experience. New Delhi: International Labour Office Asian Employment Programme (1988).

Prais S.J., Houthakker H.S. The Analysis of Family Budgets. — Cambridge: Cambridge University Press (1954).

Prais S.J., Winsten C.B. Trend Estimators and Serial Correlation. — Cowles Commission Discussion Paper no. 383, Chicago (1954).

Salkever David S. The use of dummy variables to compute predictions, prediction errors and confidence intervals (1976). — Journal of Econometrics 4(4): 393–97.

Sheffrin S.M. Rational Expectations. — Cambridge: Cambridge University Press (1983).

Spitzer John J. A primer on Box-Cox estimation (1982). — Review of Economics and Statistics 64(2): 307–313.

Stone J. Richard N. The Measurement of Consumer Expenditure and Behaviour in the United Kingdom 1920–38, — Cambridge: Cambridge University Press (1954).

Theil H. Applied Economic Forecasting. — Amsterdam: North Holland (1966).

Thomas J. James. An Introduction to Statistical Analysis for Economists (2 ed.). — London: Weidenfeld and Nicholson (1983).

Thomas J. James, Wallis Kenneth F. Seasonal variation in regression analysis (1971). — Journal of the Royal Statistical Association, Series A 134(1): 57–72.

Thomas R. Leighton. Applied Demand Analysis. — Harlow, Essex: Longman Economics Series (1987).

Whitworth Charles. The Political and Commercial Works of Charles D'Avenant. — London: Horsfield, Becket, de Hondt and Cadell (1771).

Wonnacott Thomas H., Wonnacott Ronald J. Introductory Statistics for Business and Economics (4 ed.). — New York: John Wiley (1990).

Zarembka Paul. Functional form in the demand for money (1968). — Journal of the American Statistical Association 63(322): 502–511.

АЛМОН (Almon) Ширли 303
БЕТАНКУР (Betancourt) Роже 228
БИЧ (Beach) Чарльз М. 354
БОКС (Box) Дж. Э. П. 130, 132, 319
БРАУН (Brown) Т. М. 294, 326
БРОЙШ (Breusch) Т. С. 238
БРОНСАР (Bronsard) Камиль 365
ГИРИ (Geary) Р. С. 215
ГОДФРИ (Godfrey) Л. 238
ГОЛДФЕЛД (Goldfeld) Стивен М. 210
ГРОССМЕН (Grossman) Эллиот С. 182
ГРЭЙБИЛЛ (Graybill) Ф. П. 109
ДАВЕНАНТ (Davenant) Чарльз 363
ДАРБИН (Durbin) Джеймс 227
ДЖЕНКИНС (Jenkins) Дж. М. 320
ДЖОНС (Jones) А. Дэвид 264
ДЖОНСТОН (Johnston) Джон 312
ДИТОН (Deaton) Ангус С. 363–364
ДОУГЕРТИ (Dougherty) Кристофер 264
ДУГЛАС (Douglas) Поль Х. 142, 144, 146, 192
ДЭВИДСОН (Davidson) Джеймс Э. Х. 360
ДЮФОР (Dufour) Жан-Мари 314
ЕЗЕКИЕЛ (Ezekiel) Мордекай 307
ЕО (Yeo) Дж. Стивен 316
ЗАРЕМБКА (Zarembka) Пол 130
КАЛДОР (Kaldor) Николас 50
КАЛТОН (Kalton) Грэхем 156
КВАНДТ (Quandt) Ричард Э. 210
КЕЙГАН (Cagan) Филип Д. 297–298, 300
КЕЙНС (Keynes) Джон Мейнард 295
КЕЛЕЙАН (Kelejian) Харри 228
КЕНДАЛЛ (Kendall) М. Дж. 22, 24
КЕНДРИК (Kendrick) Джон У. 182
КЕННЕДИ (Kennedy) Питер Э. 353
КИШ (Kish) Лэсли 156
КЛЕЙН (Klein) Лоуренс 294
КМЕНТА (Kmenta) Ян 324, 343, 359
КОББ (Cobb) Чарльз У. 142, 144, 146, 192
КОЙК (Koyck) Л. М. 289, 294
КОКС (Cox) Д. Р. 130, 132
КРАМЕР (Cramer) Дж. С. 354
КУ (Kuh) Эдвин 158
КУПЕР (Cooper) Р. Л. 319
ЛИВИАТАН (Liviatan) Ниссан 260
ЛИНТНЕР (Lintner) Джон 292
ЛЛАЧ (Lluch) Константино 364
МАЙЗОН (Mizon) Грейхем Э. 230
МАККИННОН (MacKinnon) Джеймс Дж. 354
МАНДЕВИЛЬ (Mandeville) Бернард 323
МЕЙЕР (Meyer) Джон Р. 158
МИТЧЕЛЛ (Mitchell) Бриджер М. 234, 236, 237

МОЗЕР (Moser) Клаус 156
МУД (Mood) А. М. 109
МУС (Muth) Джон Ф. 309
НЕЛСОН (Nelson) Чарльз Р. 319–321
НЕРЛОВ (Nerlove) Марк 192
ОКСЛИ (Oxley) Лэсли 228
ПАРК (Park) Ролла Эдвард 234, 236, 237
ПАУЭЛЛ (Powell) Ален А. 364
ПЕСАРАН (Pesaran) М. Хашем 316
ПИТАЯНОН (Pitayanon) Сумали 317
ПИЧ (Peach) Джеймс Т. 355
ПРАЙС (Prais) С. Дж. 223, 364
РОБЕРТС (Roberts) Колин Дж. 228
САЛКЕВЕР (Salkever) Дэвид С. 314
САЛЬВА-БРОНСАР (Salvas-Bronsard) Лиз 365
СМИТ (Smith) Р. П. 316
СМИТ (Smyth) Дэвид Дж. 319
СПИЦЕР (Spitzer) Джон Дж. 133
СТОУН (Stone) Дж. Ричард Н. 364
ТЕЙЛОР (Taylor) Лестер Д. 364–365
ТОМАС (Thomas) Дж. Джеймс 84, 237, 312
ТОМАС (Thomas) Р. Лейтон 365
УИКСТИД (Wicksteed) Филип 142
УИЛЬЯМС (Williams) Росс А. 364
УИНСТЕН (Winsten) С. Б. 223
УИТВОРТ (Whitworth) Чарльз 363
УОЛЛИС (Wallis) Кеннет Ф. 237
УОННАКОТ (Wonnacott) Рональд Дж. 104
УОННАКОТ (Wonnacott) Томас Х. 104
УОТСОН (Watson) Дж. С. 227
УЭББ (Webb) Джеймс Л. 355
ФРИДМЕН (Friedman) Милтон 253–255, 257–261, 298–302
ХАРВИ (Harvey) Эндрю С. 310
ХАУТЕККЕР (Houthakker) Х. С. 364–365
ХЕНДРИ (Hendry) Дэвид Ф. 230, 357, 360
ЧОУ (Chow) Грегори С. 315
ШЕФФРИН (Sheffrin) С. М. 309
ЭЙСНЕР (Eisner) Роберт 258
ЭШ (Ash) Дж. С. К. 319

А

Автокоррелированные случайные члены (autocorrelated disturbance terms), см. *Автокорреляция*
Автокорреляция (autocorrelation) 217–240
авторегрессионная схема первого порядка (first-order autoregressive) 219–220
авторегрессионная схема более высокого порядка (higher order autoregressive) 237–240
интервал наблюдения и А. 217–218
как источник смещения при оценивании модели 229–233
как лаговая структура с ограничением 229–231
определение 217
отрицательная 218–219
оценки, снижающие А. 222–225
положительная 217–218
порожденная преобразованием Койка 290–291, 302
последствия 217, 224–225
причины 217–221
с лаговой зависимой переменной 227–228, 290–291
тесты на наличие А. 219–221, 227–228, 237–238
эксперименты по методу Монте-Карло 228–229
Адаптивные ожидания (adaptive expectations) 292–296 (см. также *Гипотеза о постоянном доходе*)
модель гиперинфляции Кейгана 297–298
Алмон лаги (Almon lags), см. *Полиномиально распределенные лаги*
Альтернативные гипотезы (alternative hypothesis), см. *Гипотеза, Проверка гипотез*

Б

Бокса—Дженкинса анализ временных рядов (Box-Jenkins time-series analysis) 319–321
Бокса—Кокса тест на вид функции (Box-Cox test of functional form) 129–133, 354

В

Вероятности распределение (probability distribution) 4, 8–9
Вероятностный предел (предел по вероятности) (plim, probability limit) 26–28
для отношения двух величин 27–28
определение 26
Вероятность (probability) 3–4, 8–13
Взвешенная регрессия (weighted regression) 211

Включенные модели (nested models), см. *Модели спецификация*
Внешняя информация (extraneous information)
используемая для идентификации 341
используемая для снижения мультиколлинеарности 157–158
Временной тренд (time trend)
как замещающая переменная для показателя технического прогресса 157–158, 182–183
оценивание 123–124
Выборочная дисперсия (sample variance) 44–45
ожидаемое значение 44
определение 44
Выборочная ковариация (sample covariance) 34–42, 49–51
альтернативное выражение 42–43
определение 35
почему не является хорошей мерой связи 49–51
Выборочное среднее (sample mean)
дисперсия 16, 25–26, 47
несмещенность 17–18
распределение 16–17
Var, см. *Выборочная дисперсия*

Г

Гайавата (Hiawatha) 22–25
Гаусса—Маркова теорема (Gauss-Markov theorem) 87, 147, 244
Гаусса—Маркова условия (Gauss-Markov conditions) 80–83, 146, 201–201, 217, 227, 244, 256, 312–313, 322–323, 327, 350–351
Генеральная совокупность (population) 4
Геометрически распределенный лаг (geometric distributed lag), см. *Койка распределение*
Гетероскедастичность (heteroscedasticity) 201–215
определение 201
оценки, снижающие Г. 210–214
последствия 202–204
причины 204
тесты на наличие Г. 204–209
эксперименты по методу Монте-Карло 215–216
Гиперинфляция (hyperinflation)
Кейгана модель Г. 297–298
Гипотеза (hypothesis)
альтернативная 90
нулевая 90, 104–108
Гипотеза о постоянном доходе (permanent income hypothesis)
динамические свойства 300–302
критика Фридменом стандартной функции потребления 253–258
оценивание функции потребления методом инструментальных переменных 260–261
оценивание функции потребления методом решетчатого поиска 299–300
переменный компонент потребления и доход (transitory consumption and income), их определения 253
эксперимент по методу Монте-Карло 255–257

постоянный компонент потребления и доход, их определения 253
применения к экономической политике 257
Фридмена модель 253, 298–302
Глейзера тест на гетероскедастичность (Glejser test for heteroscedasticity) 208–209, 213
Голдфелда—Квандта тест на гетероскедастичность (Goldfeld-Quandt test for heteroscedasticity) 207–208, 210
Гомоскедастичность (homoscedasticity)
как нулевая гипотеза 206, 208, 212–213
определение 201–202

Д

Дарбина h-статистика (Durbin h statistic) 227–228
Дарбина—Уотсона d-статистика (Durbin-Watson d statistic) 219-221, 227, 240–241
как индикатор неправильной спецификации модели 230, 233 таблицы d_L и d_U 372–373
Двухшаговый метод наименьших квадратов (ДМНК) (two-stage least squares, TSLS) 337–339
как метод «очищения» переменной 338
как частный случай метода инструментальных переменных 337–338
Дискретные случайные переменные (discrete random variables), см. *Случайные переменные*
Дисперсии правила (variance rules) 45–46
Дисперсия (variance), см. *Выборочная дисперсия, Теоретическая дисперсия случайной переменной*
ДМНК, см. *Двухшаговый метод наименьших квадратов*
Доверительный интервал (confidence interval) 102–104
для предсказания 311
определение 103

З

Зависимая переменная (dependent variable) 53
Замещающие переменные (proxy variables) 182–186
идеальные (ideal) 184–185
непреднамеренные (unintentional) 186
несовершенные (imperfect) 185–186, 249–250
эффект использования 182
Зарембки шкалирование (Zarembka scaling), см. *Бокса—Кокса тест на вид функции*
Значимости уровень (significance level) 93–95
в сравнении с показателем мощности критерия 107

И

Идентификация (identification)
неполная идентифицированность 327, 332–334

нулевые и ненулевые ограничения 341
относительно стабильных зависимостей 345–347
размерности условие 340–344
сверхидентифицированность 327, 336–337
точная идентифицированность 327
Измерения ошибки (measurement errors) 55, 243, 247–252 (см. также *Гипотеза о постоянном доходе*)
зависимой переменной 251–252
объясняющих переменных 248–249
Инструментальные переменные (ИП) (instrumental variables) 243, 258–261, 330–332
идентификация и ИП 330, 333, 338–339
определение 259
применение для оценивания одновременных уравнений 330–332, 336–339
применение для оценивания функции потребления Фридмена 260–261
состоятельность 259
теоретическая дисперсия оценки 260
Интерпретация уравнения регрессии (interpretation of regression equation)
линейной регрессии 66–69
логарифмической регрессии 120–124
множественной регрессии 139–141, 146–148
ИП, см. *Инструментальные переменные*

K

Качественные зависимые переменные (qualitative dependent variables) 285–287
Качественные объясняющие переменные (qualitative explanatory variables), см. *Фиктивные переменные*
Качество оценки (goodness of fit) 69–72
F-тест на К.о. 109–111, 113–114
КМНК, см. *Косвенный метод наименьших квадратов*
Кобба—Дугласа производственная функция (Cobb-Douglas production function) 144–146, 157–158, 182–183, 188–190
проверка ограничения на постоянство эффекта от масштаба производства 188–190
свойства 144–146, 182–183
Ковариации правила (covariance rules) 38–42
Ковариация (covariance), см. *Выборочная ковариация, Теоретическая ковариация*
Койка преобразование (Koyck transformation) 289–291, 300
Койка распределение (Koyck distribution) 289–291, 303
динамические свойства 290, 300
оценивание с помощью преобразования Койка 290
оценивание с помощью решетчатого поиска 289–290
Кокрана—Оркатта итеративная процедура (Cochrane-Orcutt iterative procedure) 224–225, 228–237
как метод оценивания нелинейной регрессии 230
обобщенная для применения при автокорреляционной схеме более высокого порядка 239–240
Корректировки дивидендов модель (dividend adjustment model) 292–293
Косвенный метод наименьших квадратов (КМНК) (indirect least squares) 327–329, 331–332

определенность и КМНК 333–334, 336–337
эквивалентность методу ИП при однозначной определенности 331
Коэффициент детерминации R^2 (coefficient of determination) 70–72, 113, 163–164
определение 70
скорректированный, поправленный (adjusted, corrected) 113, 163–164
эквивалентность другим измерителям качества оценивания 70–71, 113–114
Коэффициент корреляции (correlation coefficient)
выборочный (sample) 47–49, 52
теоретический (population) 47
t-тест для К.к. 112–114
частный (partial) 52

Л

Лаговая зависимая переменная (lagged dependent variable)
как причина несостоятельности в случае автокорреляции 227–228
как причина смещения в случае малой выборки 247
Лаговая переменная (lagged variable)
определение 196–198
Лаговая структура (lag structure), см. также *Койка распределение, Полиномиально распределенные лаги*
определение 197
Лагранжа множитель, см. *Тест с множителем Лагранжа на автокорреляцию*
Линейная вероятностная модель (linear probability model) 286
Линейной регрессии модель (linear regression model) 53
линейность по параметрам 116, 141–142
линейность по переменным 116, 141–142
Логарифмические преобразования (logarithmic transformations) 119–124
правила Л.п. 120
Логарифмическое правдоподобие (log-likelihood) 352
Логит-анализ (logit analysis) 286–287

М

Максимального правдоподобия оценка (МП-оценка) (maximum likelihood estimation, ML estimation) 286, 350–354
логарифмическая функция правдоподобия 352
правдоподобия функция 352–353
Математическое ожидание (expected value) 5–8
дискретной случайной переменной 5–6
непрерывной случайной переменной 14, 32
правила 7–8, 14
функции дискретной случайной переменной 6–7
функции непрерывной случайной переменной 31–32
Метод наименьших квадратов (МНК) (ordinary least squares, OLS) 57
(см. также *Наименьших квадратов принцип, Регрессии коэффициенты*)
численные примеры 58–60
МНК, см. *Метод наименьших квадратов*

Модели неправильная спецификация (вид функции) (model misspecification, functional form) 55
выявление путем анализа остатков 195
как возможная причина появления автокорреляции 232–233
Модели неправильная спецификация (ненужные переменные) (model misspecification, irrelevant variables) 55
проявления 177–179
эксперимент по методу Монте-Карло 179–180
Модели неправильная спецификация (пропущенные переменные) (model misspecification, omitted variables) 54–55, 166–174, 262–263
аналитический вывод формулы смещения 167
воздействие на коэффициент R^2 173–174
выявление путем анализа остатков 193–196
как возможная причина появления автокорреляции 229–231
направление смещения 171–173
неприменимость статистических тестов 168
эксперимент по методу Монте-Карло 168–171
Модели спецификация (model specification) 165–180, 354–360 (см. также *Модели неправильная спецификация, Тесты на устойчивость*)
включенные и невключенные модели 356–357
«от общего к частному» подход 359–360
сравнение альтернативных моделей 356–359
Монте-Карло метод, см. *Эксперименты по методу Монте-Карло*
Мощность теста (power of a test) 107–108
МП, см. *Максимального правдоподобия оценка*
Мультиколлинеарность (multicollinearity) 155–158, 183–184
определение 155
оценки, снижающие М. 156–158
порожденная текущими и лаговыми значениями объясняющих переменных в регрессионной модели 289–290
Мультипликатор (multiplier) 257, 323

Н

Наборы данных для упражнений 374–383
Наименьших квадратов принцип (least squares principle) 57–59, 62–64, 137–139
обоснование 58
определение 57
сравнение с принципом максимального правдоподобия 350–351, 353–354
Научное знание (scientific knowledge) 360–362
Независимая переменная (independent variable), см. *Объясняющая переменная, Стохастические объясняющие переменные*
Независимость двух случайных переменных (independence of two random variables) 8
Неидентифицируемость, см. *Идентификация*
Нелинейной регрессии модель (nonlinear regression model)
оцениваемая методом решетчатого поиска 132–133
оцениваемая с помощью итеративной процедуры 126–129
сводимая к линейной с помощью логарифмического преобразования 119–124, 142–144
сводимая к линейной с помощью переопределения переменных 116–119

Ненужные переменные (irrelevant variables), см. *Модели неправильная спецификация* (ненужные переменные)
Неправильная спецификация вида функции (functional misspecification), см. *Модели неправильная спецификация* (вид функции)
Непрерывные случайные переменные (continuous random variables), см. *Случайные переменные*
Несмещенность (unbiasedness) 17–18, 20–22
выборочного среднего 17–18
коэффициентов регрессии 82–83
определение 17
связь с дисперсией 21–22
Несостоятельность (inconsistency) 27–28
Нормального распределения таблица (normal distribution table) 367 (см. также *Случайный член*)
Нулевая гипотеза (null hypothesis), см. *Гипотеза, Проверка гипотез*

О

Область принятия гипотезы (acceptance region) 94
Обобщенный метод наименьших квадратов (ОМНК) (generalized least squares, GLS) 234–236
Общая инфляция/инфляция, вызванная ростом заработной платы (пример) (price inflation/wage inflation example) 90, 92–94, 98–100, 310
Объясняющая переменная (explanatory variable) 53–54 (см. также *Стохастические объясняющие переменные*)
Ограничение (restriction)
используемое для идентификации 341
используемое для снижения эффекта мультиколлинеарности 157–158
определение 157
тест на общий фактор 229–231
F-тест для линейного ограничения 188–189
эффект от использования ограничения 166
Одновременных уравнений оценивание (simultaneous equations estimation), см. также *Идентификация, Одновременных уравнений смещение*
двухшаговый метод наименьших квадратов 337–339
инструментальные переменные 330–332
косвенный метод наименьших квадратов 327–329, 331–332, 333–334
приведенная форма уравнений 325–326, 333, 336–337
структурные уравнения 326
экзогенные переменные 325, 343–344
эксперимент по методу Монте-Карло 328–330
эндогенные переменные 325
Одновременных уравнений смещение (simultaneous equations bias) 322–324, 328–329, 348–349
Односторонний t-тест (one-tailed *t* test), см. *t-тест*
Ожидание (expectation), см. *Математическое ожидание*
ОМНК, см. *Обобщенный метод наименьших квадратов*
Отклонение (остаток) (residual)
графическое представление 61–62
использование в тесте Спирмена на ранговую корреляцию 206

использование для улучшения спецификации модели 193–196
определение 56
Оценка (формула, метод оценивания) (estimator) 15–22 (см. также *Двухшаговый метод наименьших квадратов, Инструментальные переменные, Косвенный метод наименьших квадратов, Максимального правдоподобия оценка, Метод наименьших квадратов*)
выборочного среднего 15
выборочной дисперсии 15
доказательство несмещенности оценки 32–33
коэффициентов регрессии, см. *Регрессии коэффициенты*
несмещенная 17–18
определение 15
разница между оцениваемым значением и О. 15
состоятельная 26–28
эффективная 18–20
Ошибки в переменных (errors in variables), см. *Измерения ошибки*
Ошибки I и II рода (Type I, Type II errors) 94–95, 97, 105–107, 112, 236–237 (см. также *Проверка гипотез*)

П

Парка—Митчелла анализ методом Монте-Карло (Park-Mitchell Monte Carlo study), см. *Эксперименты по методу Монте-Карло*
Паутинообразная модель (cobweb cycle) 307
Первых разностей уравнение регрессии (first differences regression) 224
Переменной неправильная спецификация (variable misspecification), см. *Модели неправильная спецификация* (ненужные переменные), *Модели неправильная спецификация* (пропущенные переменные)
Плотности вероятности функция (probability density function) 10–14
Плотности вероятности функция для выборочного среднего (probability density function of sample mean) 16–17
связь с размером выборки 25–27
Плотности вероятности функция для коэффициента регрессии (probability density function of regression coefficient) 91–92
Полиномиально распределенные лаги (лаги Алмон) (polynomial distributed lags, Almon lags) 303–306
Потерь функция (loss function) 21–22
Потребления функция (consumption function), см. также *Одновременных уравнений оценивание*
Брауна модель 294, 302
смещение при оценивании одновременных уравнений 322–324, 328, 348–349
Фридмена гипотеза о постоянном доходе, см. *Гипотеза о постоянном доходе*
Правдоподобия функция (likelihood function) 352
Прайса—Уинстена поправка (Prais-Winsten correction) 223, 225, 235, 240
Предопределенные переменные (predetermined variables) 326
Предсказание (prediction) 309–319
доверительные интервалы 313–314
несмещенность 312
ошибка 310
прогнозы 310, 312, 317–319

стандартные ошибки 313–314
теоретическая дисперсия 312
тест на устойчивость 315–317, 355
Приведенная форма уравнений (reduced form equation) 326
Проверка гипотез (hypothesis tests) 89–100, 104–108 (см. также *Регрессии коэффициенты, Ошибки I и II рода*)
цели проведения 90–91, 97
Прогнозы (forecasts) 310, 312, 317–319 (см. также *Предсказание*)
относительная ошибка прогноза (relative forecasting error, RFE) 317–318
оценивание 317–319
Пропущенные переменные (omitted variables), см. *Модели неправильная спецификация* (пропущенные переменные)

P

Распределенные лаги (distributed lags), см. *Койка распределение, Полиномиально распределенные лаги*
Размерности условие для идентификации (order condition for identification), см. *Идентификация*
Рациональные ожидания (rational expectations) 306–309
Регрессии коэффициенты (ИП), см. *Инструментальные переменные*
Регрессии коэффициенты (МНК, две объясняющие переменные)
аналитическое представление 146–147
вывод выражений для Р.к. 137–138
выражения для Р.к. 137–138
несмещенность 147
стандартные ошибки 153–154
теоретическая дисперсия 148
t-тесты и доверительные интервалы 154
эксперимент по методу Монте-Карло 150–153
Регрессии коэффициенты (МНК, одна объясняющая переменная)
аналитическое представление 73–74, 243–244
вывод выражений для Р.к. 62–64
выражения для Р.к. 62
доверительные интервалы 102–104
несмещенность 82–83, 244–246
несостоятельность в модели одновременных уравнений 322–324
несостоятельность, вызванная ошибками измерения объясняющей переменной 248–251
плотности вероятности функция 91–92
проверка гипотез 89–100
смещение в малой выборке, вызванное лаговой зависимой переменной 246
стандартные ошибки 84
теоретическая дисперсия 84, 260
t-статистика 97
t-тесты двусторонние 96–100
t-тесты односторонние 104–108
эксперимент по методу Монте-Карло 74–79, 84–85
Регрессор (regressor), см. *Объясняющая переменная, Стохастические объясняющие переменные*

Решетчатый поиск (grid search)
использование в процедуре Хилдрета—Лу для автокорреляции 224, 228
использование в тесте Бокса—Кокса 132—133
использование при оценивании модели гиперинфляции Кейгана 297—298
использование при оценивании распределения Койка 290
использование при оценивании функции потребления Фридмена 299—301
R^2, см. *Коэффициент детерминации* R^2

C

Сверхидентифицированность (overidentification), см. *Идентификация*
Сезонные фиктивные переменные (seasonal dummy variables) 273—276
СКО, см. *Сумма квадратов отклонений*
Скорректированный коэффициент детерминации
(adjusted, corrected R^2) 163—164
Случайные переменные (random variables)
дискретные (discrete) 3—10
независимость (independence of) 8
непрерывные (continuous) 3, 10—13, 31—32
Случайный член (disturbance term) 14, 53—55 (см. также *Гаусса—Маркова условия*)
в нелинейных моделях 125—126, 214
ограничения, налагаемые при идентификации 343
предположение о нормальности (распределения) 81—82, 91
происхождение 54—55
Смещение (bias), см. также *Измерения ошибки, Одновременных уравнений смещение*
определение 17
Состоятельность (consistency) 26—28
Спецификации ошибка (specification error), см. *Модели неправильная спецификация*
Спецификация вида функции, см. *Бокса—Кокса тест, Нелинейной регрессии модель*
Спирмена тест ранговой корреляции на гетероскедастичность
(Spearman rank correlation test for heteroscedasticity) 206—207
Спроса функция (demand function) 363—365
вопросы экономической политики и С. ф. 363
интерпретация 64—65
набор данных для упражнений 374—383
теория полезности и С.ф. 363—365
Спроса функция на продукты питания (пример)
Бокса—Кокса тест на вид функции 131
доверительный интервал для коэффициентов регрессии 103
интерпретация линейного уравнения 64—66, 134—137
интерпретация логарифмического уравнения 142
оценивание с помощью итеративной процедуры Кокрана—Оркатта с поправкой Прайса—Уинстена 224—225
оценивание с помощью обобщенной процедуры Кокрана—Оркатта для автокорреляции более высокого порядка 240
предсказания, их стандартные ошибки и доверительные интервалы 311, 314
тест с множителем Лагранжа на наличие автокорреляции более высокого порядка 238
t-тест для коэффициентов уравнения регрессии 96—97
F-тест и *t*-тест для линейного ограничения 190—191

F-тест на объясняющую способность 111
F-тест на стабильность коэффициентов 316
Чоу тест на неудачу предсказания 315–316
Среднеквадратичная ошибка (mean square error) 21
Стандартная ошибка коэффициента регрессии (standard error of regression coefficient), см. *Регрессии коэффициенты, Метод наименьших квадратов*
Стандартная ошибка предсказания (standard error of prediction) 313–314
Стандартная ошибка уравнения регрессии (standard error of regression equation, s_u) 84–85
Стандартное отклонение (standard deviation) 9, 32
Статистические выводы (statistical inference) 360–362
Степени свободы (degrees of freedom), см. *Хи-квадрат тест, F-тест, t-тест*
Стохастические объясняющие переменные (stochastic regressors) 243–247
Структурные уравнения (structural equations) 326
Сумма квадратов отклонений (необъясненная, остаточная) (residual sum of squares, RSS) 109–110
использование в тестах Чоу 282–285, 315–316
в тесте Голдфелда—Квандта на гетероскедастичность 207–208
в тесте на общий фактор 230
в *F*-тесте на качество оценки 109, 160–162
в *F*-тесте на линейное ограничение 188–189, 191
в *F*-тесте на объясняющую способность группы переменных 161–162
Σ знак суммы (объяснение) 29–31
доказательство несмещенности оценки 32–33

T

Тейла коэффициент U для оценивания качества прогнозов (Theil's U coefficient for evaluating forecasts) 318–319
Теоретическая дисперсия выборочного среднего (population variance of sample mean) 16–17, 47
Теоретическая дисперсия случайной переменной (для генеральной совокупности) (population variance of random variable) 8–10, 14, 32–33, 47–48
дискретной случайной переменной 8–10, 32–33
непрерывной случайной переменной 32–33
несмещенные оценки 17–18
определение 8–9
Теоретическая ковариация (pop. cov, population covariance) 43–44
определение 43
Теоретическая характеристика (population characteristic) 15, 17
Теоретическое среднее значение (математическое ожидание) (population mean) 6
несмещенные оценки 17–18
Тест на общий фактор (common factor test) 229–232
Тест с множителем Лагранжа на автокорреляцию (Lagrange multiplier test for autocorrelation) 228, 238
Тесты на значимость (tests of significance), см. *Автокорреляция, Бокса—Кокса тест, Гетероскедастичность, Проверка гипотез, Ограничение, Тест на общий фактор, Тесты на отношение правдоподобия, t-тест, F-тест, Чоу тест*
Тесты на отношение правдоподобия (likelihood ratio tests) 353

Тесты на устойчивость (stability tests) 315–319, 355
Тейла коэффициент U 318–319
тест Чоу на неудачу предсказания 312–13
F-тест на стабильность коэффициентов 315–316
Технический прогресс (technical progress) 157–158, 182–183
t-распределения таблица (критические значения) 368
t-статистика (t statistic) 97–98
t-тест (t test)
вывод результатов 98–100
двусторонний 96–100
для коэффициента корреляции 112–114
интерпретация как предельного случая для F-теста 162–163
односторонний 90, 104–108
степени свободы 96–97

У

Уровень значимости, см. *Значимости уровень*

Ф

Фиктивные переменные (dummy variables) 262–287
взаимодействия (interactive) 280–282
выбор эталонной категории (choice of reference category) 272–273
для коэффициента наклона линейной зависимости 280–282
интерпретация уравнения регрессии с Ф.п. 266, 270–271, 274–275, 278–279, 280–281
использование при расчете ошибок предсказания и стандартных ошибок 314
ловушка при применении Ф.п. (dummy variable trap) 273
множественные совокупности (multiple sets of) 277–279
набор 270–276
определение 265
польза от применения 262
проверка гипотез 266–267, 272
сезонные 273–276
F-тест на совместную объясняющую способность набора Ф.п. 272, 276
Чоу тест 282–285
F-распределения критические значения 369–370
F-статистика (F statistic) 160–163
F-тест (F test)
на неудачу предсказания 315–316
на обоснованность наложения линейного ограничения 188–189
на обоснованность объединения двух выборок при оценивании регрессионной модели (Чоу тест) 282–285, 315
объясняющей способности уравнения регрессии 160–161
совместной объясняющей способности группы переменных 161–163, 272, 276, 357–359

Х

Хи-квадрат распределение (chi-square distribution)
критические значения 371
Хи-квадрат тест (применение) (chi-square test)
тест Бокса—Кокса 130
тест на общий фактор 230—231
тест на отношение правдоподобия 353
Хилдрета—Лу метод для моделей с автокорреляцией (Hildreth-Lu procedure for autocorrelation) 224

Ц

Центральная предельная теорема (Central Limit Theorem) 82

Ч

Частичная корректировка (partial adjustment) 291—293
распределение Койка и Ч.к. 292
Чоу тест на неудачу предсказания (Chow test of predictive failure) 315—316
Чоу тест на обоснованность объединения двух выборок при оценивании регрессионной модели (Chow test of validity of combining two subsamples to fit regression model) 282—285
как тест на стабильность коэффициентов 315—316

Ш

Шум (noise) 55

Э

Экзогенные переменные (exogenous variables) 325—326, 343
определение 325
Эксперименты по методу Монте-Карло (Monte Carlo experiments)
включения ненужных переменных 179—180
генерация случайного члена 75
гетероскедастичность (пример) 215—216
иллюстрация воздействия автокорреляции 227—228
иллюстрация критики Фридменом стандартной функции потребления 255—257
иллюстрация свойств коэффициентов, полученных с помощью МНК 75—79, 84—85
иллюстрация смещения, наблюдаемого в системах одновременных уравнений, и использования КМНК и ИП для оценивания

одновременных уравнений 328–329
иллюстрация факторов, воздействующих на точность оценивания коэффициентов множественной регрессии 150–153
невключения объясняющих переменных 168–171
определение 75
Парка—Митчелла анализ процедур оценивания в моделях с автокорреляцией 234–237
Экспоненциальные временные тренды (exponential time trends)
определение 123
оценивание 123–124
Эластичность (elasticity)
интерпретация Э., оцененной на перекрестных выборках и на временных рядах 157–158
определение 121
оценивание 121–122
Энгеля кривая (Engel curve) 123–124, 131
Эндогенные переменные (endogenous variables) 325–326
определение 325
Эффективность (efficiency)
оценок регрессии 18–20

Кристофер Доугерти

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

Пер. с англ.

Учебное издание

Редактор *С.М. Рыловский*
Корректор *Е.А. Морозова*
Компьютерная верстка *А.Р. Комлев*
Художественное оформление *«Ин-Арт»*

Подписано в печать 10.10.97. Формат 70x100/16
Усл. печ. л. 33,54. Печать офсетная. Гарнитура "Таймс"
Доп. тираж 5000 экз. Цена договорная. Заказ 289

ЛР № 070824 от 21.01.93 г.

ISBN 5-86225-458-7

9 785862 254587

ООО "Издательский Дом ИНФРА-М"
127214, Москва, Дмитровское ш., 107
Тел.: (095) 485-7063, 485-7177
Факс (095) 485-5318. Робофакс (095) 485-5444
E-mail: books@infra-m.ru

Отпечатано в ГУП ИПК "Ульяновский Дом печати"
432601, г. Ульяновск, ул. Гончарова, 14