

*Джеймс Сток,  
Марк Уотсон*

*Введение в эконометрику*



*James H. Stock and Mark W. Watson*

# **INTRODUCTION TO ECONOMETRICS**

*Third Edition*

*Addison-Wesley • 2011*



# РАНХиГС

РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА  
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ  
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

СЕРИЯ

«АКАДЕМИЧЕСКИЙ УЧЕБНИК»

Джеймс Сток, Марк Уотсон

## ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

Перевод с английского  
Под научной редакцией М.Ю. Турунцевой

Рекомендуется Российской академией народного хозяйства и государственной службы при Президенте Российской Федерации в качестве учебника для студентов ВПО, обучающихся по экономическим направлениям и специальностям, а также для студентов бакалавриата, углубленно изучающих микроэкономику, студентов магистратуры, аспирантов, преподавателей экономических факультетов вузов. (Основание – приказ Министерства образования и науки №130 от 22 февраля 2012 г.)



Большое преимущество

Москва • 2015

УДК 330.4

ББК 65.05

С81

Издание осуществлено при финансовой поддержке банка ВТБ 24 (ПАО)

*Перевод с английского:*

Василий Акимов (гл. 2)

Булат Гафаров (гл. 1, 5, 6, 7, 17)

Максим Леонов (гл. 18)

Юрий Пономарев (гл. 10–13)

Антон Скроботов (гл. 3, 4, 8, 9)

Марина Турунцева (гл. 14–16)

**Сток, Джеймс; Уотсон, Марк**

С81 Введение в эконометрику / Джеймс Сток, Марк Уотсон; пер. с англ.; под науч. ред. М.Ю. Турунцевой. — М.: Издательский дом «Дело» РАНХиГС, 2015. — 864 с. — (Академический учебник).

ISBN 978-5-7749-0865-3

«Введение в эконометрику» представляет собой учебник вводного уровня, включающий все основные разделы эконометрики, входящие в программы современных курсов обучения.

Книга хорошо продумана с методологической точки зрения. В основу издания положен принцип необходимости сочетания изучения теории с решением практических задач. Изложение материала представлено в четко заданной последовательности, любое новое понятие связывается с конкретным эмпирическим примером, который часто рассматривается на протяжении нескольких глав, если это согласуется с логикой изложения.

Дополнительные разделы учебника позволяют использовать его в курсах эконометрики более продвинутого уровня.

Для изучающих экономические науки.

ISBN 978-5-7749-0865-3

УДК 330.4

ББК 65.05

Authorized translation from the English language edition, entitled INTRODUCTION TO ECONOMETRICS, 3rd Edition; ISBN 0138009007; by STOCK, JAMES H.; and WATSON, MARK W.; published by Pearson Education, Inc.; publishing as Prentice Hall; Copyright © 2011 Pearson Education Limited

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc. RUSSIAN language edition published by DELO PUBLISHERS. Copyright © 2015 Лицензированный перевод английского издания под названием INTRODUCTION TO ECONOMETRICS, 3rd Edition; ISBN 0138009007; под авторством Джеймса Стока и Марка Уотсона, опубликованного Pearson Education, Inc. под маркой Prentice Hall; Copyright © 2011 Pearson Education Limited, publishing as Addison-Wesley.

Все права защищены. Ни одна часть настоящей книги не может быть распространена или передана ни в каком виде и никакими средствами, электронными или механическими, включая фотокопирование, запись или любые информационно-поисковые системы, без разрешения от Pearson Education, Inc. Издание на русском языке выпущено Издательским домом «Дело» © ФГБОУ ВПО «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации», 2015

---

# Оглавление

Предисловие научного редактора . . . . .	XI
Предисловие к русскому изданию . . . . .	XIII
Введение . . . . .	XV
<b>Часть I. Введение и обзор основных понятий теории вероятностей и математической статистики</b>	
<b>Глава 1. Типы данных и вопросы, которые интересуют экономистов . . . . .</b>	<b>3</b>
1.1. Какие вопросы рассматривают экономисты? . . . . .	3
1.2. Причинно-следственные связи и идеальные эксперименты . . . . .	8
1.3. Данные: источники и типы . . . . .	9
<b>Глава 2. Основные понятия теории вероятностей . . . . .</b>	<b>17</b>
2.1. Случайные величины и вероятностные распределения . . . . .	18
2.2. Математическое ожидание, среднее и дисперсия . . . . .	22
2.3. Двухмерные случайные величины . . . . .	28
2.4. Нормальное распределение, распределение хи-квадрат, распределения Стьюдента и Фишера . . . . .	38
2.5. Случайная выборка и распределение выборочного среднего . . . . .	45
2.6. Асимптотические распределения . . . . .	49
<b>Глава 3. Элементы математической статистики . . . . .</b>	<b>67</b>
3.1. Оценка среднего значения генеральной совокупности . . . . .	68
3.2. Тестирование гипотез о среднем значении генеральной совокупности . . . . .	73
3.3. Доверительные интервалы для среднего генеральной совокупности . . . . .	82
3.4. Сравнение средних значений различных генеральных совокупностей . . . . .	84
3.5. Оценка причинных эффектов при помощи разности средних значений, используя экспериментальные данные . . . . .	86
3.6. Использование $t$ -статистики при малом размере выборки . . . . .	89
3.7. Диаграммы рассеяния, выборочная ковариация и выборочная корреляция . . . . .	94
<b>Часть II. Основы регрессионного анализа</b>	
<b>Глава 4. Парная линейная регрессия . . . . .</b>	<b>111</b>
4.1. Модель парной линейной регрессии . . . . .	111
4.2. Оценка коэффициентов в модели парной линейной регрессии . . . . .	116
4.3. Критерии качества приближения данных моделью . . . . .	123

4.4. Предположения метода наименьших квадратов . . . . .	126
4.5. Выборочное распределение МНК-оценки . . . . .	131
4.6. Заключение . . . . .	134
<b>Глава 5. Парная регрессия: проверка гипотез и доверительные интервалы. . . . .</b>	<b>147</b>
5.1. Проверка гипотез о коэффициентах регрессии . . . . .	147
5.2. Доверительные интервалы для коэффициентов регрессии . . . . .	154
5.3. Регрессия с бинарной объясняющей переменной . . . . .	156
5.4. Гетероскедастичность и гомоскедастичность . . . . .	158
5.5. Теоретические основы обычного метода наименьших квадратов . . . . .	164
5.6. Использование <i>t</i> -статистики регрессии в малых выборках . . . . .	167
5.7. Заключение. . . . .	168
<b>Глава 6. Множественная линейная регрессия . . . . .</b>	<b>183</b>
6.1. Смещение из-за пропущенных переменных . . . . .	183
6.2. Модель множественной регрессии. . . . .	190
6.3. МНК-оценка множественной линейной регрессии. . . . .	194
6.4. Качество приближения данных моделью множественной линейной регрессии . . . . .	197
6.5. Предположения метода наименьших квадратов для множественной линейной регрессии . . . . .	200
6.6. Распределение МНК-оценок в модели множественной линейной регрессии . . . . .	202
6.7. Мультиколлинеарность . . . . .	203
6.8. Заключение . . . . .	207
<b>Глава 7. Множественная линейная регрессия: проверка гипотез и доверительные интервалы. . . . .</b>	<b>219</b>
7.1. Проверка гипотез и доверительные интервалы для одного коэффициента . . . . .	219
7.2. Проверка совместных гипотез . . . . .	224
7.3. Тестирование одного ограничения, включающего несколько коэффициентов модели. . . . .	232
7.4. Доверительные области для нескольких коэффициентов . . . . .	233
7.5. Выбор спецификации модели множественной регрессии. . . . .	235
7.6. Анализ данных по результатам тестов . . . . .	241
7.7. Заключение . . . . .	246
<b>Глава 8. Нелинейные регрессионные модели . . . . .</b>	<b>259</b>
8.1. Общая стратегия моделирования функции нелинейной регрессии. . . . .	261
8.2. Функции парных нелинейных регрессий . . . . .	269
8.3. Взаимодействия между независимыми переменными . . . . .	281
8.4. Нелинейные эффекты влияния изменения соотношения учеников и учителей на результаты тестов. . . . .	293
8.5. Заключение . . . . .	301
<b>Глава 9. Оценка исследований, основанных на множественной регрессии . . . . .</b>	<b>319</b>
9.1. Внутренняя и внешняя обоснованность. . . . .	320
9.2. Угрозы для внутренней обоснованности множественного регрессионного анализа . . . . .	323

9.3. Внутренняя и внешняя обоснованность при прогнозировании по модели регрессии. . . . .	337
9.4. Пример: результаты тестов и размеры классов. . . . .	339
9.5. Заключение . . . . .	349

### **Часть III. Регрессионный анализ: дополнительные главы**

<b>Глава 10. Регрессионный анализ панельных данных . . . . .</b>	359
10.1. Панельные данные. . . . .	360
10.2. Панельные данные с наличием двух периодов: сравнения «до и после» . . . . .	364
10.3. Регрессия с фиксированными эффектами. . . . .	367
10.4. Модель регрессии с фиксированными временными эффектами . . . . .	372
10.5. Предположения модели регрессии с фиксированными эффектами и стандартные ошибки модели регрессии с фиксированными эффектами . . . . .	375
10.6. Количество ДТП с летальным исходом и законы, направленные на сокращение случаев вождения в нетрезвом виде . . . . .	379
10.7. Заключение . . . . .	385
<b>Глава 11. Регрессии с бинарными зависимыми переменными . . . . .</b>	398
11.1. Бинарные зависимые переменные и линейная вероятностная модель . . . . .	399
11.2. Пробит- и логит-модели регрессии. . . . .	404
11.3. Оценка логит- и пробит-моделей и проверка статистических гипотез. . . . .	412
11.4. Применение к данным для Бостона . . . . .	416
11.5. Заключение . . . . .	424
<b>Глава 12. Регрессии с инструментальными переменными . . . . .</b>	439
12.1. ИП оценки с одним регрессором и одним инструментом . . . . .	440
12.2. Обобщенная модель регрессии с инструментальными переменными . . . . .	451
12.3. Проверка допустимости инструментов и проверка статистических гипотез. . . . .	458
12.4. Приложение метода к изучению спроса на сигареты . . . . .	465
12.5. Где найти допустимые инструменты? . . . . .	471
12.6. Заключение . . . . .	477
<b>Глава 13. Эксперименты и квазиэксперименты . . . . .</b>	493
13.1. Потенциальные исходы, причинные эффекты и идеализированные эксперименты . . . . .	494
13.2. Угрозы обоснованности экспериментов . . . . .	497
13.3. Эмпирические оценки эффектов уменьшения размера учебного класса . . . . .	503
13.4. Квазиэксперименты . . . . .	513
13.5. Потенциальные проблемы с квазиэкспериментами . . . . .	521
13.6. Экспериментальные и квазиэкспериментальные оценки в гетерогенных выборках . . . . .	524
13.7. Заключение . . . . .	530

## **Часть IV. Регрессионный анализ экономических временных рядов**

<b>Глава 14. Введение в модели временных рядов и прогнозирование . . . . .</b>	547
14.1. Использование регрессионных моделей для прогнозирования . . . . .	548
14.2. Введение во временные ряды и серийную корреляцию . . . . .	550
14.3. Авторегрессии . . . . .	557
14.4. Модели временных рядов с дополнительными переменными и авторегрессионные модели с распределенными лагами . . . . .	563
14.5. Информационные критерии и выбор глубины запаздывания . . . . .	573
14.6. Нестационарность I: тренды . . . . .	576
14.7. Нестационарность II: структурные сдвиги . . . . .	587
14.8. Заключение . . . . .	600
<b>Глава 15. Оценка динамического причинного влияния . . . . .</b>	615
15.1 Знакомство с данными по апельсиновому соку . . . . .	616
15.2. Динамическое причинное влияние . . . . .	620
15.3. Оценка динамического причинного влияния при помощи экзогенных регрессоров . . . . .	625
15.4. Стандартные ошибки, являющиеся состоятельными при наличии гетероскедастичности и автокорреляции . . . . .	629
15.5. Оценка динамического причинного влияния при помощи строго экзогенных регрессоров . . . . .	633
15.6. Цены на апельсиновый сок и заморозки . . . . .	642
15.7. Является ли экзогенность правдоподобным условием? Некоторые примеры . . . . .	650
15.8. Заключение . . . . .	653
<b>Глава 16. Модель регрессии временных рядов:</b>	
<b>дополнительные разделы . . . . .</b>	664
16.1. Векторные авторегрессии . . . . .	664
16.2. Многошаговые прогнозы . . . . .	669
16.3. Порядок интегрированности и DF-GLS-тест на единичные корни . . . . .	674
16.4. Коинтеграция . . . . .	681
16.5. Кластеризованная волатильность и авторегрессионные модели с условной гетероскедастичностью . . . . .	691
16.6. Заключение . . . . .	695
<b>Часть V. Эконометрическая теория регрессионного анализа</b>	
<b>Глава 17. Теория парной линейной регрессии . . . . .</b>	703
17.1. Расширенные предположения метода наименьших квадратов и оценка МНК . . . . .	704
17.2. Основные понятия асимптотической теории . . . . .	706
17.3. Асимптотические распределения МНК-оценки и <i>t</i> -статистики . . . . .	711
17.4. Точные выборочные распределения при нормально распределенных ошибках . . . . .	714
17.5. Взвешенный метод наименьших квадратов . . . . .	716

<b>Глава 18. Теория множественного регрессионного анализа . . . . .</b>	<b>730</b>
18.1. Линейная модель множественной регрессии и МНК-оценки в матричной форме . . . . .	731
18.2. Асимптотическое распределение МНК-оценок и <i>t</i> -статистик . . . . .	735
18.3. Проверка совместных гипотез . . . . .	738
18.4. Распределение статистик регрессии с нормальными ошибками . . . . .	740
18.5. Эффективность МНК-оценки при наличии гомоскедастичности ошибок . . . . .	743
18.6. Обобщенный метод наименьших квадратов . . . . .	745
18.7. Инструментальные переменные и обобщенный метод моментов. . . . .	751
<b>Приложения . . . . .</b>	<b>778</b>
<b>Список литературы . . . . .</b>	<b>787</b>
<b>Глоссарий. . . . .</b>	<b>794</b>
<b>Указатель. . . . .</b>	<b>811</b>



# Предисловие научного редактора

В русскоязычной литературе ощущается нехватка хороших как российских, так и переводных учебников по эконометрике вводного уровня. Книга Джеймса Стока<sup>1</sup> и Марка Уотсона<sup>2</sup> восполняет этот пробел. Оба автора являются признанными специалистами, внесшими большой вклад в развитие теоретической и прикладной эконометрики и имеющими немалый опыт в ее преподавании.

Учебник «Введение в эконометрику» отражает современные взгляды на эконометрику вводного уровня и методику ее преподавания. Он представляет собой наиболее полный и системный курс и включает все основные разделы эконометрики, входящие в программы современных курсов обучения.

Книга имеет ряд особенностей. Например, авторы нетрадиционным образом относятся к нарушению предпосылки о гомоскедастичности ошибок. В отличие от большинства учебников эконометрики, здесь не рассматриваются многочисленные процедуры тестирования для определения, гомоскедастичны ли остатки или нет. Авторы считают, что предпосылка о гомоскедастичности не является реалистичной для экономических данных. Поэтому вместо попытки выяснения типа гетероскедастичности для тестирования различных гипотез и построения доверительных интервалов они предлагают использовать стандартные ошибки коэффициентов регрессии, устойчивые к гетероскедастичности (и автокоррелированности для временных рядов).

Неоспоримым достоинством «Введения в эконометрику» является то, что книга содержит разделы, которые, как правило, не включаются в аналогичные издания. Упомянем две главы, посвященные оценке исследований, основанных на множественной регрессии (глава 9), и экспериментам и квазиэкспериментам (глава 13). Эти темы нечасто включают даже в учебники эконометрики более продвинутых уровней.

Еще одним достоинством «Введения в эконометрику» является то, что учебник хорошо продуман с методологической точки зрения. Изложение следует четко заданной последовательности, и любое новое понятие сразу же связывается с конкретным эмпирическим примером, который часто рассматривается на протяжении нескольких глав, если это согласуется с логикой изложения. Помимо этого в тексте выделяются основные понятия, которые необходимо запомнить, рассматриваются интересные примеры из реальной жизни. Также

---

<sup>1</sup> Джеймс Сток – профессор экономики Гарвардского университета, автор и научный редактор четырех книг, более чем 120 статей и глав в книгах.

<sup>2</sup> Марк Уотсон – профессор экономики Принстонского университета, автор и научный редактор четырех книг, более чем 100 статей и глав в книгах.

## Введение в эконометрику

---

к тексту учебника прилагаются вспомогательные материалы: базы данных для эмпирических упражнений, тексты возможных контрольных работ и т.д.

Несмотря на то что учебник содержит материал по всем разделам, изучаемым в стандартных вводных курсах эконометрики, его можно было бы расширить как за счет более подробного изучения ряда разделов, так и за счет включения отдельных не рассмотренных в нем тем. К числу таких разделов можно отнести некоторые специальные разделы эконометрики, например, такие как методы анализа панельных данных, методы анализа временных рядов, обобщенный метод моментов, методы анализа, используемые в финансовой эконометрике, системы одновременных уравнений или бутстреп.

В заключение хочу выразить признательность всем, кто принимал непосредственное участие в переводе книги: Василию Акимову, Булату Гафарову, Антону Скроботову и Максиму Леонову. Отдельная благодарность – Юрию Пономареву за добросовестный перевод глав 10–13.

Хочется поблагодарить всех, кто принимал участие в обсуждении перевода различных терминов: Сергея Дробышевского, Павла Трунина, Георгия Идрисова, Александра Кнобеля и Андрея Зубарева. Спасибо Юлии Груниной за оперативную техническую помощь.

Все недостатки и неточности перевода остаются на совести научного редактора перевода.

*Научный редактор перевода  
Марина Турунцева  
Январь 2014 г.*

## Предисловие к русскому изданию

Мы рады приветствовать новый перевод на русский язык в семье изданий учебника Стока и Уотсона «Введение в эконометрику». Как мы уже писали в предисловии к американскому изданию, «реальный мир экономики, бизнеса и правительства довольно сложен и беспорядочен, полон противоречивых идей и вопросов, которые требуют ответа». Эконометрика является набором инструментов, которые могут дать надежные ответы на количественные вопросы. Эти инструменты и методы эконометрики вытекают из основных научных принципов, которые делают акцент на воспроизводимости результатов и дают возможность оценить количественно неопределенность, которая обязательно возникает, когда мы делаем статистические выводы на основе выборок данных. Они также показывают, как лучше оценить причинно-следственные эффекты в имеющихся данных и когда скептически относиться к таким оценкам. Эти инструменты и принципы, стоящие за ними, являются устойчивыми и могут быть применены как к экономике, так и в смежных областях в разных странах – в России так же, как в Соединенных Штатах. Мы благодарим научного редактора перевода Марину Турунцеву и Российскую академию народного хозяйства и государственной службы при Президенте Российской Федерации за то, что они сделали этот перевод возможным.

*Джеймс Сток и Марк Уотсон*



# Введение

Эконометрика может стать развлечением и для преподавателя, и для студента. Реальный мир экономики, бизнеса и правительства довольно сложен и беспорядочен, полон противоречивых идей и вопросов, которые требуют ответа. Что является более эффективным: контролировать вождение в нетрезвом состоянии посредством принятия соответствующих законов или через увеличение налогов на алкоголь? Можете ли вы заработать деньги на фондовой бирже, покупая, когда цены на акции являются низкими относительно доходов, или вам следует выжидать, считая, что цены активов на фондовой бирже подчиняются закону случайного блуждания, как предполагает теория? Можем ли мы улучшить качество начального образования, уменьшив количество детей в классе, или наши дети должны просто слушать Моцарта по 10 минут в день? Эконометрика помогает нам отделить безумные идеи от здравых и найти количественные ответы на важные количественные вопросы. Эконометрика открывает окно в наш сложный мир и позволяет нам увидеть взаимосвязи, на основе которых люди, бизнес и правительства принимают свои решения.

Учебник «Введение в эконометрику» был написан для первого года бакалаврского курса эконометрики. Это отражение нашего опыта, что позволило нам адаптировать курс эконометрики для вводного уровня; интересные приложения должны обосновывать теорию, а теория должна согласовываться с приложениями. Этот простой принцип влечет за собой значимое отличие от книг по эконометрике старого поколения, в которых теоретические модели и предположения не связывались с приложениями. Неудивительно, что некоторые студенты задаются вопросом о важности эконометрики после того, как они потратили уйму времени на изучение предпосылок, нереалистичность которых они понимают впоследствии, для того чтобы затем изучать «решения» «проблем», возникающих в ситуации, когда приложения не связаны с предпосылками. Мы считаем, что гораздо лучше мотивировать необходимость использования какого-либо инструментария примерами, а потом показывать некоторые простые предположения, связанные с ними. Такой подход – немедленно связывающий теорию с эмпирическими примерами – может оживить эконометрику.

## Новое в этом издании

- обновлены трактовки стандартных ошибок в панельных регрессиях;
- обсуждается, когда и почему пропущенные данные могут представлять проблему для регрессионного анализа;

- модели разрывных регрессий используются как метод анализа квазиэкспериментов;
- обновлен раздел по слабым инструментам;
- раздел, в котором обсуждается использование и интерпретация контрольных переменных, включен в раздел об основных направлениях развития регрессионного анализа;
- введено понятие «потенциальные исходы» для экспериментальных данных;
- добавлены интересные вставки;
- добавлены новые задачи и компьютерные упражнения.

Третье издание книги основывается на той же философии и тех же принципах, что и первые два. Мы считаем, что эмпирические приложения должны вести за собой эконометрическую теорию, а не наоборот.

Одно из самых серьезных изменений в этом издании коснулось главы о панельных регрессиях (глава 10). В панельных данных данные для одного объекта обычно коррелированы во времени. Для того чтобы получаемые выводы были обоснованы, стандартные ошибки должны быть скорректированы методами, устойчивыми (робастными) к этой корреляции. Соответственно, теперь в главе по панельным данным рассматривается один из таких методов – кластеризованные стандартные ошибки. Метод кластеризованных стандартных ошибок является естественным для панельных данных расширением методов корректировки стандартных ошибок при наличии гетероскедастичности, которые вводятся во второй части книги, рассматривающей основы регрессионного анализа. Недавние исследования показывают, что кластеризованные стандартные ошибки обладают множеством желаемых свойств, которые обсуждаются в главе 10 и в приложении к этой главе.

Еще одним важным изменением является включение раздела об экспериментах и квазиэкспериментах (глава 13). Обсуждение регрессии «разности разностей» было упрощено и следует непосредственно из принципов множественной регрессии, рассмотренных в части II. В главе 13 мы вводим понятие потенциальных исходов и связываем эту все более распространяющуюся терминологию с понятиями, введенными в частях I и II.

Данное издание содержит множество других существенных изменений. Одним из них является включение в обсуждение аккуратного и доступного подхода с использованием контрольных переменных во множественной регрессии. В главе 7 теперь обсуждаются условия, необходимые для того, чтобы включение контрольных переменных приводило к желаемым результатам в том смысле, что коэффициент при интересующей нас переменной был бы несмещенным, в то время как коэффициенты при контрольных переменных, в общем случае, таковыми не являются. Другие изменения включают новое обсуждение проблемы пропусков в данных в главе 9, новое дополнительное, основанное на принципах математического анализа, приложение к главе 8, в котором выводятся формулы для угловых коэффициентов и эластичностей в функциях нелинейных

регрессий, и обновленное обсуждение проблемы наличия слабых инструментов в главе 12. Настоящее издание также включает в себя новые вставки, представляющие общий интерес, обновленные эмпирические примеры и дополнительные упражнения.

## Особенности книги

«Введение в эконометрику» отличается от других учебников по трем основным причинам. Во-первых, мы рассматриваем возникающие в реальной жизни вопросы и данные, чтобы обосновать развитие теории, и мы подробно обсуждаем основные выводы из результатов эмпирического анализа. Во-вторых, выбор обсуждаемых тем отражает современную теорию и практику. И, в-третьих, рассматриваемые нами теоретические вопросы и предположения соответствуют приложениям. Наша цель – научить студентов мудро использовать эконометрику и сделать это на математическом уровне, соответствующем вводному курсу.

### *Вопросы и данные, существующие в жизни*

Методологически мы рассматриваем каждую тему с точки зрения важного реально существующего вопроса, требующего получения конкретного численного ответа. Например, мы рассматриваем модели парной регрессии, множественной регрессии и анализ функциональной формы регрессии в контексте оценки влияния условий, в которых работают начальные школы, на результаты этой работы. (Имеют ли начальные школы с меньшим размером классов более высокие результаты тестов?) Мы учим методам анализа панельных данных в контексте анализа влияния законов о рождении в нетрезвом виде на количество ДТП со смертельным исходом. Мы используем возможное наличие расовой дискриминации на рынке ипотечных кредитов в качестве эмпирической иллюстрации при изучении моделей регрессии с бинарной зависимой переменной (логит и пробит). Мы изучаем метод инструментальных переменных в контексте оценки эластичности спроса на сигареты. Несмотря на то что все эти примеры связаны с экономикой, для их понимания достаточно вводного курса по экономике, а многие из них могут быть поняты вообще без изучения каких-либо курсов по экономике. Таким образом, преподаватель может сосредоточиться на обучении именно эконометрике, а не микроэкономике или макроэкономике.

Мы рассматриваем все наши эмпирические приложения серьезно, чтобы показать студентам, что они могут узнать из данных, одновременно оставаясь самокритичными и осознавая ограничения эмпирических методов анализа. В каждом эмпирическом приложении мы учим студентов исследовать альтернативные спецификации и тем самым оценивать, насколько устойчивы основные полученные ими результаты. Вопросы, заданные в эмпирических приложениях, важны, и мы даем серьезные и, как нам представляется, надежные ответы на них. Однако мы призываем студентов и преподавателей не просто

соглашаться с нами, а провести повторный анализ данных, которые можно найти на сайте, с данными учебника ([http://www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson)).

### ***Современный выбор тем***

Эконометрика прошла долгий путь с 1980 года. Темы, которые мы рассматриваем, отражают лучшее из современной прикладной эконометрики. Поскольку мы хотели написать вводный курс, мы ориентируемся лишь на процедуры и тесты, которые широко используются на практике. Например:

- **Регрессия с инструментальными переменными.** Мы представляем регрессии с инструментальными переменными как общий метод, используемый для учета корреляции между ошибкой и регрессором, которая может возникнуть по многим причинам, в том числе из-за пропущенных переменных и одновременной причинности. Два предположения о допустимости инструментов – экзогенность и релевантность – равнозначны. Мы проводим широкое обсуждение того, где брать инструменты, а также тестов на сверхидентифицирующие ограничения и методов диагностики слабых инструментов, а также объясняем, что делать, если эта диагностика вызывает проблемы.
- **Оценка программных документов.** Во все большем числе эконометрических исследований проводится анализ как случайных контролируемых экспериментов, так и квазиэкспериментов, также известных как естественные эксперименты. Мы рассматриваем эти темы, часто называемые методами оценки программных документов, в главе 13. Мы представляем стратегию такого исследования как альтернативный подход к решению проблемы смешения из-за пропущенных переменных, одновременной причинности и отбора данных и оцениваем как сильные, так и слабые стороны исследований с использованием экспериментальных или квазиэкспериментальных данных.
- **Прогнозирование.** В главе, посвященной прогнозированию (глава 14), рассматриваются одномерные (авторегрессионные) и многомерные прогнозы с использованием моделей временных рядов, а не больших систем одновременных уравнений. Мы обращаем внимание на простые и надежные инструменты, такие как авторегрессии и выбор моделей с помощью информационных критериев, которые хорошо работают на практике. В этом разделе также рассматриваются методы практического решения проблем, связанных с наличием стохастических трендов (единичных корней), тестирования наличия единичного корня и структурных сдвигов (в известные и неизвестные моменты времени), а также псевдовневыборочного прогнозирования. Все это обсуждается в контексте развития стабильных и надежных прогнозных моделей временных рядов.
- **Модели временных рядов.** Мы проводим четкое различие между двумя очень разными приложениями моделей временных рядов: прогнозированием и оценкой динамического причинного влияния. В главе

о причинном влиянии во временных рядах (глава 15) уделяется особое внимание вопросу о том, когда различные методы оценки, в том числе обобщенный метод наименьших квадратов, будут или не будут приводить к корректным выводам о причинности и когда можно посоветовать оценивать динамические регрессии с использованием МНК со стандартными ошибками, состоятельными при наличии гетероскедастичности и автокорреляции.

### ***Теория, которой соответствует практика***

Несмотря на то что изучение эконометрических методов лучше всего мотивируется эмпирическими приложениями, студенты должны достаточно хорошо знать эконометрическую теорию, чтобы понимать сильные и слабые стороны этих методов. Мы следуем современному подходу, в котором соответствие между теорией и практикой является максимально сильным, сохраняя при этом математические выкладки на уровне, требующем только умение делать алгебраические преобразования.

Современные эмпирические приложения имеют некоторые общие характеристики: как правило, большие наборы данных (сотни наблюдений, а часто и более); регрессоры, не являющиеся фиксированными в повторяющихся выборках, а собираемые при помощи случайного выбора (или какого-либо другого механизма, который делает их случайными); данные, не являющиеся нормально распределенными; отсутствие какой-либо априорной причины для предположения о наличии гомоскедастичности ошибок (напротив, часто есть основания полагать, что они гетероскедастичны). Все это приводит к важному отличию теоретического изложения в этом учебнике и других учебниках.

- **Асимптотический подход.** Вследствие того что базы данных являются большими, с самого начала мы используем асимптотически нормальное приближение выборочных распределений для проверки гипотез и построения доверительных интервалов. Наш опыт показывает, что знакомство с основами асимптотической теории занимает меньше времени, чем изучение  $t$ -распределения Стьюдента и точных  $F$ -распределений, корректировку на степени свободы и так далее. Асимптотический подход также не дает студентам расстроиться, когда они обнаруживают, что из-за ненормальности ошибок теория точных распределений, которую они освоили, не работает на реальных данных. Преподнесенный в контексте выборочного среднего, асимптотический подход к проверке гипотез и построению доверительных интервалов вытекает непосредственно из множественного регрессионного анализа, логит- и пробит-моделей, метода инструментальных переменных и моделей временных рядов.
- **Случайный выбор.** Из-за того что регрессоры редко являются фиксированными в эконометрических приложениях, с самого начала мы относимся к данным по всем переменным (зависимым и независимым) как к результату случайного выбора. Это предположение используется

во введении в анализ межобъектных данных, оно легко обобщается на случай панельных данных и на временные ряды и из-за используемого асимптотического подхода не влечет за собой никаких дополнительных концептуальных или математических трудностей.

- **Гетероскедастичность.** В прикладной эконометрике обычно используются устойчивые к гетероскедастичности стандартные ошибки, для того чтобы не думать о ее наличии или отсутствии. В данной книге мы не рассматриваем гетероскедастичность как что-то исключительное или как «проблему», которую надо «решить»; вместо этого мы учитываем гетероскедастичность с самого начала и просто используем устойчивые к гетероскедастичности стандартные ошибки. Мы представляем гомоскедастичность как частный случай, который дает теоретическое обоснование МНК.

### ***Квалифицированные исследователи, мудрые потребители***

Мы надеемся, что студенты, использующие эту книгу, станут мудрыми специалистами в области эмпирического анализа. Чтобы достичь этого, они должны узнать не только то, как использовать методы регрессионного анализа, но и как оценивать достоверность эмпирических исследований, сделанных другими.

Наш подход к обучению тому, как оценить эмпирическое исследование, состоит из трех частей. Во-первых, сразу же после рассмотрения основных методов регрессионного анализа мы посвящаем главу 9 анализу угроз для внутренней и внешней обоснованности эмпирического исследования. В этой главе рассматриваются проблемы с данными и вопросы обобщения результатов на другие выборки, а также основные проблемы регрессионного анализа, в том числе проблемы пропущенных переменных, неправильной спецификации функциональной формы, ошибок в переменных, отбора наблюдений и одновременной причинности, а также способы выявления этих проблем на практике.

Во-вторых, мы применяем такие методы оценки эмпирических исследований к эмпирическому анализу рассматриваемых в книге примеров. Мы делаем это, рассматривая альтернативные спецификации, а также систематически анализируя различные угрозы обоснованности в эмпирических примерах, представленных в книге.

В-третьих, чтобы стать мудрыми потребителями, студенты должны получить свой собственный опыт как исследователи. Активное обучение лучше пассивного, и эконометрика является идеальным курсом для него. По этой причине на веб-сайте учебника представлены базы данных, программное обеспечение, и студентам предложено выполнить эмпирические упражнения в различных областях.

### ***Уровень математической строгости***

Наша цель состоит в том, чтобы развить у студентов глубокое понимание современных методов регрессионного анализа независимо от того, преподается ли курс на «высоком» или «низком» математическом уровне. Части I–IV учеб-

ника (которые охватывают основной материал) доступны студентам, изучившим математику до дифференциального исчисления. Части I–IV содержат меньше формул и больше примеров, чем многие вводные курсы эконометрики, и гораздо меньше формул, чем учебники по математике для бакалавриата. Но большее количество формул не подразумевает более сложные методы. По нашему опыту, более глубокий математический уровень обоснований не приводит к более глубокому пониманию методов большинством студентов.

Тем не менее студенты учатся по-разному, и для студентов с хорошей математической подготовкой сложность обучения может быть повышена за счет более глубокой математической базы. Поэтому часть V содержит вводную информацию об эконометрической теории и может быть использована студентами с сильной математической подготовкой. Мы считаем, что если математические главы части V использовать в сочетании с материалом из частей I–IV, то данная книга может быть применима при чтении продвинутых бакалаврских или магистерских курсов эконометрики.

## **Содержание и структура**

«Введение в эконометрику» состоит из пяти частей. Предполагается, что студенты прослушали курсы по теории вероятности и математической статистике, хотя мы рассматриваем этот материал в части I. Основы регрессионного анализа рассматриваются в части II. В частях III–V представлены дополнительные разделы, которые основываются на положениях, рассмотренных в части II.

### **Часть I**

В главе 1 мы знакомим читателя с предметом эконометрики и подчеркиваем важность получения количественных ответов на количественные вопросы. В ней рассматривается понятие причинности в статистических исследованиях, выделяются различные типы данных, встречающихся в эконометрике. Вопросы из области теории вероятности и математической статистики рассматриваются в главах 2 и 3 соответственно; будут ли эти главы изучаться в рамках конкретного курса или просто использоваться как вспомогательный материал, зависит от уровня подготовки студентов.

### **Часть II**

В главе 4 вводится модель парной линейной регрессии и понятие оценки метода наименьших квадратов (МНК), а в главе 5 обсуждается тестирование гипотез и построение доверительных интервалов в модели парной линейной регрессии. В главе 6 студенты узнают, как можно решить проблему смещения из-за пропущенной переменной, используя множественную регрессию и оценив тем самым влияние одной независимой переменной, в предположении постоянства других независимых переменных. В главе 7 описаны методы проверки гипотез, в том числе  $F$ -тесты и построение доверительных интервалов в модели

множественной регрессии. В главе 8 модель линейной регрессии расширяется до случая модели нелинейной регрессии, акцентируя внимание на функциях регрессии, линейных по параметрам (так, чтобы параметры могли быть оценены с помощью МНК). В главе 9 студенты сделают, в некотором смысле, шаг назад и узнают, как определить сильные и слабые стороны регрессионного анализа, применяя концепцию внутренней и внешней обоснованности исследования.

### **Часть III**

В части III рассматриваются некоторые расширения изученных базовых методов регрессионного анализа. В главе 10 студенты узнают, как использовать панельные данные для контроля ненаблюдаемых переменных, которые являются постоянными во времени. В главе 11 рассматриваются регрессии с бинарной зависимой переменной. В главе 12 показывается, как может быть использована регрессия с инструментальными переменными для решения проблемы корреляции между ошибкой регрессии и регрессорами, и обсуждается, как можно найти и оценить допустимые инструменты. В главе 13 студенты познакомятся с методами анализа данных из экспериментов и квази-, или естественных экспериментов, то есть с разделом, который часто называют «оценкой программных документов».

### **Часть IV**

В части IV рассматриваются методы регрессионного анализа временных рядов. Глава 14 посвящена прогнозированию, в ней вводятся различные современные методы, используемые для анализа временных рядов, такие как тесты на наличие единичного корня и тесты на стабильность. В главе 15 обсуждается использование временных рядов для оценки причинных зависимостей. В главе 16 рассматриваются некоторые более продвинутые методы анализа временных рядов, в том числе авторегрессионные модели с условной гетероскедастичностью.

### **Часть V**

Часть V представляет собой введение в эконометрическую теорию. Она является, скорее, приложением, которое заполняет исключенные из основного текста математические детали. Можно сказать, что это замкнутое изложение эконометрической теории и математических выводов в модели линейной регрессии. В главе 17 рассматривается теория регрессионного анализа модели парной линейной регрессии; здесь не используется матричная алгебра, но все равно требуется более высокий уровень математических знаний, чем в остальных частях книги. В главе 18 модели множественной регрессии, регрессии с инструментальными переменными и обобщенный метод моментов для оценки модели линейной регрессии изучаются с использованием матричной формы записи.

### **Предпосылки к книге**

Из-за того что преподаватели делают акценты на различных разделах эконометрики в своих курсах, мы написали эту книгу, имея в виду разнообразные

предпочтения преподавателей. Насколько это возможно, главы из частей III, IV и V являются «автономными» в том смысле, что для их изучения не требуется знакомства со всеми предыдущими частями. Разделы, необходимые для изучения каждой главы, приведены в таблице 1. Несмотря на то что мы поняли, что ряд вопросов, рассматриваемых в учебнике, может быть выделен в отдельные курсы, главы написаны таким образом, чтобы преподаватели могли переставлять их по своему усмотрению, если возникнет необходимость.

Таблица 1

## Разделы, необходимые для изучения глав из частей III, IV, и V

Главы	Необходимые части или главы								
	Часть I	Часть II		Часть III		Часть IV			Часть V
	1–3	4–7,9	8	10.1 10.2	12.1 12.2	14.1– 14.4	14.5– 14.8	15	17
10	X <sup>a</sup>	X <sup>a</sup>	X						
11	X <sup>a</sup>	X <sup>a</sup>	X						
12.1, 12.2	X <sup>a</sup>	X <sup>a</sup>	X						
12.3–12.6	X <sup>a</sup>	X <sup>a</sup>	X	X	X				
13	X <sup>a</sup>	X <sup>a</sup>	X	X	X				
14	X <sup>a</sup>	X <sup>a</sup>	b						
15	X <sup>a</sup>	X <sup>a</sup>	b			X			
16	X <sup>a</sup>	X <sup>a</sup>	b			X	X	X	
17	X	X	X						
18	X	X	X		X				X

В данной таблице приведены минимальные условия, необходимые для изучения материала в конкретной главе. Например, для изучения методов оценки динамического причинного влияния во временных рядами (глава 15) в первую очередь требуется изучить часть I (если это необходимо, в зависимости от уровня подготовки студентов и за исключением случаев, отмеченных в сноске a) и часть II (за исключением главы 8, см. сноска b), а также разделы 14.1–14.4.

<sup>a</sup> В главах 10–16 используются исключительно асимптотические приближения выборочных распределений, так что разделы 3.6 (*t*-распределение Стьюдента для тестирования гипотез о средних) и 5.6 (*t*-распределение Стьюдента для тестирования гипотез о коэффициентах регрессии) могут быть пропущены.

<sup>b</sup> Главы 14–16 (модели временных рядов) могут изучаться без предварительного знакомства с главой 8 (нелинейные функции регрессии), если преподаватель объяснит отдельно, как использовать логарифмические преобразования для приближения процентных изменений.

## Примеры формирования курсов

В данной книге совмещено несколько различных курсов.

### Стандартные вводные курсы эконометрики

Такой курс знакомит с предметом эконометрики (глава 1) и основными понятиями теории вероятности и математической статистики, если это необходимо

(главы 2 и 3). Затем изучаются модели парной линейной регрессии, множественной регрессии, основы анализа функциональной формы, а также методы оценки регрессионного исследования (т.е. вся часть II). В курсе могут быть рассмотрены панельные регрессии (глава 10), регрессии с ограниченной зависимой переменной (глава 11) и регрессии с инструментальными переменными (глава 12), если позволяет время. Завершить курс можно знакомством с экспериментами и квазиэкспериментами из главы 13, что дает возможность вернуться к вопросам оценки причинного влияния, поднятым в начале семестра, и повторить основные методы регрессионного анализа. *Необходимые предшествующие курсы: алгебра и введение в математическую статистику.*

### ***Введение в анализ временных рядов и прогнозирование***

Учебник также может быть использован при прочтении короткого вводного курса анализа временных рядов и прогнозирования, для изучения которого необходимым условием является курс эконометрики. Здесь нужно потратить некоторое время на повторение основ регрессионного анализа из части II в зависимости от уровня подготовки студентов. Затем можно перейти прямо к части IV и изучать методы прогнозирования (глава 14), оценки динамического причинного влияния (глава 15) и продвинутые разделы анализа временных рядов (глава 16), в том числе модели векторной авторегрессии и авторегрессионные модели с условной гетероскедастичностью. Важной составляющей такого курса является выполнение практических упражнений по прогнозированию, данные для которых доступны для преподавателей на веб-сайте книги. *Необходимые предшествующие курсы: алгебра и введение в эконометрику или любой эквивалентный курс.*

### ***Введение в эконометрическую теорию***

Книга также подходит для подготовки курсов лекций, читаемых на старших курсах бакалавриата, для студентов, имеющих хорошую математическую подготовку, или для курсов эконометрики магистерского уровня. Краткий обзор теории вероятности и математической статистики (часть I) изучается по мере необходимости. Курс знакомит с регрессионным анализом не на математическом, а на прикладном уровне на основе глав из части II. За этим следуют теоретические детали из глав 17 и 18 (до раздела 18.5). Затем изучаются регрессии с ограниченной зависимой переменной (глава 11) и оценка метода максимального правдоподобия (приложение 11.2). И, наконец, по выбору можно включить в курс регрессии с инструментальными переменными и обобщенный метод моментов (главы 12 и раздел 18.7), модели временных рядов (глава 14), а также методы оценки причинного влияния с использованием временных рядов и обобщенного метода наименьших квадратов (главы 15 и раздел 18.6). *Необходимые предшествующие курсы: дифференциальное исчисление и введение в математическую статистику. Глава 18 предполагает знакомство с матричной алгеброй.*

## Педагогические особенности

Данный учебник несет целый ряд педагогических функций, цель которых – помочь студентам понять, запомнить и научиться применять самые важные идеи. *Введения в главах* призваны объяснить необходимость изучения материала конкретной главы на примерах из реальной жизни и мотивировать студентов для работы. В них также кратко обсуждается последовательность изложения материала главы. *Основные термины* выделяются жирным цветом и определяются в каждой главе. *Вставки «Основные понятия»* напоминают главные идеи. Во *вставках с примерами* приводятся различные интересные примеры из смежных областей и рассматриваются реально проведенные исследования, в которых используются методы и концепции, обсуждаемые в учебнике. *Выводы*, заканчивающие каждую главу, являются полезным инструментом для обзора основных идей главы. *Вопросы для повторения и закрепления основных понятий* позволяют проверить, как студенты поняли основное содержание главы, а *Упражнения* дают возможность более глубоко проработать введенные в главе понятия и методы. *Эмпирические упражнения* позволяют студентам применить свои знания к решению реальных эмпирических задач. В конце учебника, в *Приложении*, даны статистические таблицы, в разделе *Список литературы* приведен список источников для дальнейшего чтения, в *Глоссарии* собраны основные понятия, используемые в книге.

## Дополнительные материалы

Дополнительные материалы, прилагаемые к третьему изданию «Введение в эконометрику», включают руководство по решению задач, файл с тестами, подготовленный Манфредом Кейлом (Manfred W. Keil, Clermont McKenna College) и слайды в PowerPoint® с рисунками, таблицами и основными понятиями. Руководство по решению задач включает в себя решения всех упражнений из конца каждой главы. Файл Test Item File, находящийся в программном обеспечении TestGen with QuizMaster, содержит большой выбор легко редактирующихся тестовых задач и вопросов различных типов. Все эти ресурсы доступны в Ресурсном центре преподавателя на сайте [http://www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson).

Кроме того, веб-сайт учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) предоставляет большое количество дополнительных ресурсов для студентов и преподавателей, включающих базы данных для эмпирических упражнений, базы данных для повторения эмпирических результатов, представленных в тексте, практические вопросы, ответы на вопросы для повторения и закрепления основных понятий из конца каждой главы и решения нечетных упражнений.

## Благодарности

Очень многие люди внесли свой вклад в первое издание этой книги. Мы выражаем самую большую благодарность нашим коллегам из Гарвардского

и Принстонского университетов, которые использовали первые версии этой книги на своих занятиях. Сьюзан Купер с факультета государственного управления им. Кеннеди Гарвардского университета подготовила неоценимые предложения и подробные комментарии для нескольких предварительных вариантов учебника. Как один из преподавателей, работающих с Джеймсом Стоком, она также помогла проверить большую часть материала из этой книги в то время, когда как она подготавливалась для обязательного магистерского курса на факультете им. Кеннеди. Мы также обязаны двум другим коллегам с факультета, Альберто Абади и Сью Динарски, за их терпеливое объяснение квазиэкспериментов и области оценки программных документов и за подробные комментарии ранних версий учебника. В Принстоне Эли Тэмер изучал ранние версии и также дал полезные комментарии к предпоследнему варианту книги.

Мы также обязаны очень многим нашим друзьям и коллегам-эконометристам, которые обсуждали с нами содержание этой книги и которые внесли очень много полезных предложений. Брюс Хансен (Университет Висконсин – Мэдисон) и Бо Оноре (Принстон) обеспечили полезную обратную связь на самом раннем этапе работы над основным материалом части II. Джошуа Энгрист (MIT) и Гвидо Имбенс (Калифорнийский университет, Беркли) внесли полезные предложения, касающиеся материалов об оценке программных документов. Нашему изложению материала о временных рядах способствовала дискуссия с Ясином Айт-Сахалиа (Принстон), Грэмом Эллиоттом (Университет Калифорнии, Сан-Диего), Эндрю Харви (Кембриджский университет) и Кристофером Симсом (Принстон). Наконец, очень многие наши коллеги внести полезные предложения в те части рукописи, которые близки к их области знаний: Дон Эндрюс (Йельский университет), Джон Баунд (Университет Мичигана), Грегори Чоу (Принстон), Томас Даунс (Университет Тафтса), Дэвид Друккер (Stata Corp.), Жан Болдуин Гроссман (Принстон), Эрик Ханусек (Институт Гувера), Джеймс Хекман (Университет Чикаго), Хань Хун (Принстон), Кэролайн Хоксби (Гарвард), Аллан Крюгер (Принстон), Стивен Левитт (Университет Чикаго), Ричард Лайт (Гарвард), Дэвид Наймарк (Мичиганский государственный университет), Джозеф Ньюхаус (Гарвард), Пьер Перрон (Бостонский университет), Кеннет Уорнер (Университет Мичигана) и Ричард Зекхаузер (Гарвард).

Многие люди проявили большую щедрость, предоставив нам данные. База данных по результатам тестов в Калифорнии были собрана при содействии Лекса Аксельрода из отдела стандартов и оценки результатов обучения Департамента образования штата Калифорния. Мы благодарны Чарли де Паскаль из службы по оценке знаний Департамента образования штата Массачусетс за помочь в работе с данными по результатам тестов в штате Массачусетс. Кристофер Рум (Университет Северной Каролины, Гринсборо) любезно предоставил нам свою базу данных по законам о рождении в нетрезвом виде и смертности в ДТП. Отдел исследований Федерального резервного банка Бостона заслуживает благодарности за данные о расовой дискриминации в сфере ипотечного кредитования; мы особенно благодарны Джейфри Тутеллу за предоставление

нам обновленной базы данных, которую мы используем в главе 9, и Линн Браун за объяснение ее особенностей. Мы благодарны Джонатану Груберу (MIT) за предоставление данных о продажах сигарет, которые мы анализируем в главе 10, и Алану Крюгеру (Принстон) за помощь с базой данных STAR по Теннесси, которая анализируется в главе 11.

Мы благодарим всех, кто тщательно проверял верстку на наличие ошибок. Керри Гриффин и Яир Листокин прочитали всю рукопись, а Эндрю Фрекер, Ори Хеффец, Амбер Хенри, Хонг Ли, Александро Тароцци и Мэтт Уотсон работали над отдельными главами.

В первом издании мы воспользовались помощью редактора исключительной квалификации, Джейн Тафтс, чей творческий подход, трудолюбие и внимание к деталям улучшили книгу во многих отношениях. Издательство Эддисон – Уэсли предоставило нам первоклассную поддержку, начиная от нашего отличного редактора, Сильвии Мэллори, и заканчивая всей командой, работавшей над подготовкой книги к публикации. Джейн и Сильвия терпеливо учили нас, как лучше писать, структурировать и представлять материал, и их усилия видны на каждой странице этой книги. Мы выражаем благодарность превосходной команде издательства Эддисон – Уэсли, которая работала с нами над вторым изданием: Эдриэну д'Амброзио (главному выпускающему редактору), Бриджит Пейдж (помощнику продюсера), Чарльзу Сполдингу (главному дизайнеру), Нэнси Фентон (ответственному редактору), в том числе за ее выбор Нэнси Фрейхофер и «Томпсон Стил Инкорпорейтед», которые взяли на себя весь процесс, связанный с выходом книги из печати, Хизер Макнелли (выпускающему редактору) и Дениз Клинтон (главному редактору). Наконец, мы обладали таким преимуществом, как квалифицированный редактор Кей Уэно, задействованный при подготовке второго издания. Мы также благодарны за отличное третье издание команде издательства Эддисон – Уэсли, Эдриэну д'Амброзио, Нэнси Фентон и Джилл Колонговски, а также Мэри Сэнджер, руководителю проекта с «Несбитт Графикс».

Мы также получили большую помощь и много предложений от преподавателей, студентов и исследователей, пока готовили третье издание. Изменения, внесенные в третье издание, включают или отражают предложения, исправления, комментарии, данные и помощь, которую мы получили от исследователей и преподавателей: Дональда Эндрюса (Йельский университет), Джеймса Кобба (Университет Флориды), Сьюзен Динарски (Университет Мичигана), Николь Эйхельбергер (Техасский технический университет), Бойда Фьюлстеда (Университет Юты), Мартина Грюноу, Дэниэла Хамермеша (Университет Техаса в Остине), Кейсуке Хирено (Университет Аризоны), Бо Оноре (Принстонский университет), Гвидо Имбенса (Гарвардский университет), Манфреда Кейла (колледж Клермонт Маккена), Дэйвида Лейбсона (Гарвардский университет), Дэйвида Ли (Принстонский университет), Бриджит Мэдриан (Гарвардский университет), Хорхе Маркесу (Университет Мэриленда), Карен Беннетт Матис (Управление по цитрусовым Флориды), Ульриха Мюллера (Принстонский университет), Серену Нг

## Введение в эконометрику

---

(Колумбийский университет), Гарри Патриноса (Всемирный банк), Питера Саммерса (Техасский технический университет), Андрея Васнова (Университет Сиднея) и Дугласа Янга (Университет штата Монтана). Нам также оказали помощь своим вкладом студенты Фернандо Осес де ла Гуардия и Кэрри Уилсон.

Вдумчивые отзывы на третье издание были подготовлены для издательства Эддисон – Уэсли Стивом де Лоахом (Элонский университет), Джейфри де Симоном (Университет Техаса в Арлингтоне), Гэри В. Энгельгардтом (Сиракузский университет), Лукой Флабби (Джорджтаунский университет), Штеффаном Хабермальцем (Северо-Западный университет), Кэролайн Дж. Хейнрих (Университет Висконсин – Мэдисон), Эммой М. Иглесиас-Васкес (Мичиганский государственный университет), Карлосом Ламаршем (Университет Оклахомы), Вики А. Мак-Кракен (Вашингтонский государственный университет), Клодини М. Перейрой (Тюлэнский университет) и Джоном Т. Уорнером (Клемсонский университет). Мы также получили очень полезные советы относительно предварительных вариантов глав 7–10 от Иоанны Берделл (Университет де Пола), Джанет Кохльхас (Университет Хьюстона), Апражит Махажан (Стэнфордский университет), Ся Мэн (Университет Брандейса) и Чан Шен (Джорджтаунский университет).

Прежде всего мы обязаны нашим семьям за их выносливость и терпение на протяжении всего этого проекта. Написание этой книги заняло много времени, и проект должен был казаться им бесконечным. Они больше, чем кто-либо, несли бремя наших обязательств, и мы глубоко благодарны им за их помощь и поддержку.

Часть I

ВВЕДЕНИЕ И ОБЗОР  
ОСНОВНЫХ  
ПОНЯТИЙ ТЕОРИИ  
ВЕРОЯТНОСТЕЙ  
И МАТЕМАТИЧЕСКОЙ  
СТАТИСТИКИ



# **Глава 1. Типы данных и вопросы, которые интересуют экономистов**

Если спросить полдюжины эконометристов, что такое эконометрика, то можно услышать полдюжины различных ответов. Кто-то скажет, что эконометрика – это наука о проверке экономических теорий. Другой, что эконометрика – это набор инструментов, используемых для прогнозирования будущих значений экономических показателей, таких как объем продаж фирмы, общий экономический рост или цены на акции. Третий эконометрист может сказать, что эконометрика – это процесс подбора параметров в математических моделях экономики для достижения их соответствия данным из реального мира. Ну а кто-то может сказать вам, что это наука и искусство обработки исторических данных для разработки численных, или количественных, рекомендаций по экономической политике для правительства и частного бизнеса.

На самом деле все эти ответы верны. В широком смысле эконометрика – это наука и искусство использования экономической теории и статистических методов для анализа экономических данных. Эконометрические методы используются в различных областях экономики, включая финансы, экономику труда, макроэкономику, микроэкономику, маркетинг и экономическую политику. Эконометрические методы также часто используются в других общественных науках, включая политологию и социологию.

Эта книга познакомит вас с основными методами, используемыми эконометристами. Мы применяем эти методы, чтобы ответить на множество конкретных количественных вопросов из мира бизнеса и экономической политики. Первая глава поднимает четыре таких вопроса и в общих чертах обсуждает эконометрические подходы к ответу на них. Глава заканчивается обзором основных видов данных, доступных эконометристу для ответа на эти и многие другие количественные экономические вопросы.

## **1.1. Какие вопросы рассматривают экономисты?**

Многие решения в экономике, бизнесе и государственном управлении основываются на понимании взаимосвязи между переменными в окружающем нас мире. Эти решения требуют количественных ответов на количественные вопросы.

Данная книга рассматривает несколько количественных вопросов, выбранных из числа современных экономических проблем. Мы рассматриваем четыре вопроса, которые касаются образовательной политики, расовой дискриминации

при выдаче ипотечных кредитов, потребления сигарет и макроэкономического прогнозирования.

**Вопрос № 1: Влияет ли уменьшение размера класса на качество начального образования?**

Предложения по реформе системы народного образования в США вызывают много споров. Большинство предложений касается самых маленьких – учеников начальной школы. Система начального образования преследует множество целей, касающихся развития социальных навыков. Однако для большинства родителей и учителей самым главным является получение базовых академических навыков: чтения, письма и основ математики. Одно из предложений для улучшения качества усвоения основных навыков заключается в сокращении размера класса в начальной школе. С меньшим числом ребят в классе, как утверждается, каждый ученик получает больше учительского внимания, во время занятий меньше поводов отвлекаться, можно изучить больше материала и, тем самым, поднять успеваемость.

Но влияет ли вообще размер класса в начальной школе на успеваемость детей? Уменьшение числа школьников в классе стоит денег: необходимо нанять больше учителей и, если школа загружена полностью, построить новые помещения для занятий. Чиновники, принимающие решения, должны сравнить все эти издержки с возможными выгодами от подобной реформы. Однако чтобы сравнить плюсы и минусы от сокращения размеров классов, нужно оценить количественно величину возможной выгоды. Насколько велико влияние данной меры на успеваемость по основным предметам? Возможно, меньший размер класса вообще никак не влияет на успеваемость?

Хотя на основании здравого смысла и повседневного опыта можно утверждать, что меньшее количество учеников в классе легче обучать, только лишь на основании здравого смысла невозможно дать количественный ответ о пользе сокращения размеров классов для повышения уровня успеваемости. Чтобы дать такой ответ, необходимо провести эмпирические исследования, то есть исследования, основанные на статистических данных и связывающие размер класса и успеваемость в начальной школе.

В данной книге мы рассмотрим взаимосвязь между размером класса и успеваемостью по основным предметам на основании данных из 420 школьных округов Калифорнии в 1999 году. По результатам анализа калифорнийских данных, оказывается, что ученики в округах с меньшими классами в среднем показывают лучшие результаты по стандартизованным тестам по сравнению с учениками из округов с большими классами. Несмотря на то что этот факт соответствует идеи о том, что малые классы обеспечивают более высокие результаты по стандартизованным тестам, этот результат может отражать просто-напросто прочие преимущества учеников в округах с малыми классами по сравнению с другими школьными округами. Например, школы в округах с малыми размерами классов обычно обучают учеников из более благополучных семей, чем школы в прочих округах, так что ученики из школ с малыми классами имеют больше возможностей для обучения вне школы. Может ока-

заться, что именно эти дополнительные возможности приводят к лучшим результатам тестов, а не малый размер классов. Во второй части мы используем множественный регрессионный анализ, чтобы изолировать влияние изменений в размере класса от изменений в других факторах, таких как социально-экономическая среда учеников.

**Вопрос № 2: Происходит ли дискриминация по расовым принципам на рынке ипотечных кредитов?**

Многие люди покупают свои дома, используя ипотеку, то есть большой по размеру кредит с залогом в виде приобретаемой недвижимости. По законам США, кредитные организации не имеют права принимать решение о выдаче кредита на основе расовых признаков: заявители, идентичные во всем, кроме расы, должны иметь равную возможность получения кредита. Таким образом, в теории не должно быть никакой расовой дискриминации при выдаче ипотеки.

Но исследователи из Федерального резервного банка Бостона, используя данные начала 1990-х годов, обнаружили в опровержение данного теоретического вывода, что 28 % чернокожих заявителей не смогли получить ипотечный кредит, в то время как среди белого населения только 9 % заявителей было отказано в займе. Говорят ли эти данные о том, что имеет место дискриминация на рынке ипотечных займов? Если так, то насколько она велика в действительности?

Обнаруженный Бостонским банком факт, что черным людям чаще отказывают в займах, чем белым, сам по себе не говорит о наличии дискриминации со стороны ипотечных агентств, поскольку черные и белые заявители могут отличаться многими показателями помимо цвета кожи. Прежде чем делать окончательные выводы о существовании дискриминации, нужно более детально рассмотреть имеющиеся данные, чтобы понять, есть ли разница в вероятности получить кредит людям с разным цветом кожи *при прочих равных*, и, если так оно и есть, понять, насколько эта разница велика. Для достижения этой цели в главе 11 мы познакомимся с эконометрическими методами, которые помогут нам оценить количественно влияние цвета кожи заявителя на шанс получения ипотеки при неизменных прочих характеристиках заявителя — в частности, его способности выплатить долги по займам.

**Вопрос № 3: Насколько сильно налоги на сигареты сокращают потребление табака?**

Курение сигарет является одним из основных источников проблем со здоровьем для людей во всем мире. Множество издержек курения, таких как медицинские расходы, ложащиеся на плечи самих курильщиков, а также сложнооценимые издержки некурящих людей, не желающих быть пассивными курильщиками, порождаются третьими лицами. Поскольку эти издержки создаются людьми, не являющимися конечными потребителями табака, правительство вмешательство могло бы помочь снизить потребление сигарет. Один из наиболее гибких методов для снижения потребления табака — повышение налогов на сигареты.

Элементарные экономические соображения говорят нам, что увеличение цены на сигареты приведет к снижению их потребления. Однако насколько

велико будет снижение? Если розничная цена вырастет на 1 %, то на сколько процентов сократится потребление сигарет? Процентное изменение в величине спроса, возникающее в результате 1 %-го роста цены, называется *эластичностью спроса по цене*. Если мы желаем сократить потребление сигарет на определенную величину, например, на 20 %, при помощи повышения налогов, то мы должны знать эластичность спроса по цене, чтобы рассчитать необходимый для достижения данной цели рост цен. Однако чему же равна эта эластичность для сигарет?

Несмотря на то что экономическая теория предлагает нам необходимые соображения, чтобы помочь ответить на этот вопрос, она не дает нам численного значения эластичности спроса по цене. Чтобы узнать эту эластичность, мы должны рассмотреть эмпирические свидетельства поведения курильщиков и потенциальных курильщиков. Другими словами, нам нужно проанализировать статистические данные по ценам и потреблению сигарет.

Мы рассмотрим данные по продажам сигарет, ценам, налогам и располагаемым доходам в США в 80-х и 90-х годах XX века. Из имеющихся данных следует, что штаты с низкими налогами и, соответственно, низкими ценами на сигареты имеют более высокий уровень потребления сигарет, в то время как штаты с высокими ценами имеют более низкий уровень. Однако анализ рассматриваемых данных осложняется тем, что причинно-следственная связь работает в обоих направлениях: низкие налоги ведут к высокому спросу, однако если число курильщиков в штате становится достаточно велико, местные политики могут пытаться удерживать налоги на низком уровне, чтобы удержать курящих избирателей. В главе 12 мы рассмотрим методы, при помощи которых можно учесть эту «двустороннюю причинно-следственную связь» или «взаимную причинность», и покажем, как можно использовать эти методы для оценки эластичности спроса на сигареты по цене.

#### **Вопрос № 4: Чему будет равен уровень инфляции в следующем году?**

По-видимому, люди всегда хотели заглянуть в будущее. Каков будет уровень продаж фирмы в результате инвестиций в новое оборудование? Вырастут ли цены на рынке акций в следующем месяце и, если это случится, насколько сильно? Покроют ли налоговые поступления в следующем году запланированные расходы на городские услуги? Будут ли в вашем следующем экзамене по микроэкономике преобладать вопросы по теории экстерналий или монополий? Можно ли будет в следующую субботу пойти на пляж?

Одним из показателей, в знании будущих значений которого макроэкономисты и финансисты особенно заинтересованы, является общий уровень инфляции. Финансовый специалист может порекомендовать взять кредит или отказаться от него при текущей процентной ставке в зависимости от прогноза инфляции в следующем году. Экономисты в центральных банках, таких как Федеральная резервная система США в Вашингтоне или Европейский центральный банк во Франкфурте (Германия), отвечают за поддержание инфляции на контролируемом уровне, так что их решения о процентной ставке зависят от прогноза инфляции в следующем году. Если они посчитают, что уровень инфляции вырас-

тет на один процентный пункт, то они могут увеличить процентную ставку более чем на один процентный пункт, чтобы замедлить экономику, которая, с их точки зрения, находится под угрозой «перегрева». Если их предположение окажется ошибочным, то они рискуют либо вызвать нежелательную рецессию, либо неожиданный скачок инфляции.

Профессиональные экономисты в своих суждениях основываются на численных прогнозах, полученных с использованием эконометрических моделей. Работа прогнозиста заключается в предсказании будущих значений показателей на основании предшествующей информации, и эконометристы делают это, основываясь на экономической теории и статистических методах, используемых для количественного анализа взаимосвязей в исторических данных.

Данные, которые мы будем использовать для прогноза инфляции,— это данные по уровням инфляции и безработицы в США. Важной эмпирической взаимосвязью в этом случае является «кривая Филлипса», которая связывает низкое текущее значение уровня безработицы с ростом инфляции в следующем году. Мы получим прогноз инфляции и оценим его правдоподобность на основе «кривой Филлипса» в главе 14.

## ***Количественные вопросы, количественные ответы***

Каждый из этих четырех вопросов требует численного ответа. Экономическая теория подсказывает, что ответ — «потребление сигарет должно пойти вниз, когда цена идет вверх», но фактическое значение этой величины должно быть получено эмпирически, то есть путем анализа данных. Так как мы используем данные для ответа на количественные вопросы, наши ответы всегда содержат некоторую неопределенность: разные наборы данных могут приводить к различным численным результатам. Таким образом, концептуально основа для анализа должна быть предоставлена одновременно и численным ответом на вопрос, и наличием меры того, насколько точен этот ответ.

Основной концепцией данной книги является основа эконометрики — модель множественной регрессии. Эта модель, представленная во второй части, дает математический способ оценить количественно то, как изменение в одной переменной влияет на другую переменную, считая другие переменные постоянными. Например, какой эффект оказывает изменение размера класса на результаты тестов, считая постоянными или контролируя характеристики учеников (например, такие характеристики, как доход семьи), которые администратор школьного округа не может контролировать? Какое влияние оказывает ваша раса на ваши шансы на получение одобрения на ипотечный кредит, считая постоянными такие факторы, как способность погашать кредит? Какое влияние оказывает увеличение на 1 % цен на сигареты на потребление сигарет, считая постоянным доход курильщиков и потенциальных курильщиков? Модель множественной регрессии и ее расширения дают возможность ответить на все эти вопросы, используя реальные данные, а также дать количественную оценку неопределенности, связанной с ответами на эти вопросы.

## 1.2. Причинно-следственные связи и идеальные эксперименты

Как и большинство вопросов, рассматриваемых в эконометрике, первые три вопроса в разделе 1.1 касаются причинно-следственных связей между переменными. Другими словами, действие считается причиной исхода, если исход является прямым результатом, то есть последствием этого действия. Прикоснение к горячей конфорке вызывает ожог, употребление воды утоляет жажду, накачивание воздуха в шину приводит к ее расширению, добавление удобрений к росткам помидоров приводит к росту урожая. Причинно-следственная связь означает, что некоторое действие (использование удобрений) ведет к определенному последствию (высокий урожай).

### Оценка причинно-следственных связей

Как нам лучше всего измерить причинно-следственное влияние определенного количества удобрений, например, 100 грамм на квадратный метр, на урожай помидоров (измеренных в килограммах)?

Один способ измерить такое влияние заключается в проведении эксперимента. В таком эксперименте исследователь-агроном сажает несколько растений томатов. Каждое растение находится в одинаковых условиях, за одним исключением: некоторые растения получают 100 грамм удобрений на квадратный метр, в то время как другие не получают удобрений. Более того, решение о распределении удобрений определяется случайным образом компьютером, тем самым гарантируется независимость решения о добавке удобрения от любых других характеристик между растениями. После того как растения дадут плоды, агроном взвешивает их для каждого растения. Разница между средним урожаем на квадратный метр для удобренной и не удобренной почвы показывает влияние удобрений на производство помидоров.

Рассмотренный пример является примером *случайного управляемого* (или *контролируемого*) эксперимента. Эксперимент является контролируемым в том смысле, что у нас есть две группы: *контрольная группа*, не получающая удобрений, и *исследуемая (экспериментальная)* группа, получающая удобрения в количестве равном 100 гр/м<sup>2</sup>. Эксперимент является случайным в том смысле, что воздействие (распределение удобрений) оказывается случайным образом. Это случайное распределение по группам устраниет любую возможность систематической связи между, например, количеством солнечного света, падающего на растение, и применением удобрения, так что единственной причиной систематического различия между контрольной и исследуемой группой остается применение удобрений. Если эксперимент качественно проведен в достаточно большом масштабе, он позволит оценить причинно-следственное влияние воздействия (применение 100 гр/м<sup>2</sup> удобрений) на интересующий нас результат (производство помидоров).

**Причинно-следственное влияние** определяется в данной книге как влияние определенного воздействия в идеальном случайном контролируемом экспери-

менте на исход этого эксперимента. В таком эксперименте единственной систематической причиной различия в исходах между исследуемой и контрольной группой может быть только само воздействие.

Можно представить идеальный случайный эксперимент, необходимый для ответа на каждый из первых трех вопросов, рассмотренных в разделе 1.1. Например, в примере с изучением влияния размера класса на успеваемость учеников можно представить себе случайное назначение «воздействия» различных размеров классов на разные группы учеников. Если эксперимент разработан и проведен так, что систематические различия между двумя группами учеников заключаются только в количестве учеников в классе, то в теории такой эксперимент позволил бы оценить влияние сокращения размера классов на результаты обучения при прочих равных условиях.

Концепция идеального случайного эксперимента полезна, поскольку она дает нам определение причинно-следственного влияния. На практике, тем не менее, невозможно провести идеальный эксперимент. Фактически эконометристы редко проводят эксперименты, поскольку они, как правило, неэтичны или их невозможно провести на удовлетворительном уровне, или же они чрезвычайно дорогие. Идея об идеальном эксперименте дает, однако, теоретический эталон для эконометрического анализа причинных связей с использованием данных.

### ***Прогнозирование и причинность***

Хотя первые три вопроса, рассмотренные в разделе 1.1, касаются причинно-следственных связей, четвертый вопрос – прогнозирование инфляции – с ним не связан. Вам не нужно знать о причинных связях, чтобы сделать хороший прогноз. Хороший способ «спрогнозировать», идет ли на улице дождь, – просто посмотреть, пользуются ли пешеходы на улице зонтами, хотя сам факт наличия зонта у пешеходов не является причиной дождя.

Несмотря на то что прогнозирование не требует наличия причинно-следственной связи, некоторые взаимосвязи между переменными, следующие из экономической теории, могут быть полезны для прогнозирования. Как мы увидим в главе 14, множественный регрессионный анализ позволяет нам получить количественные оценки исторических связей, предполагаемых экономической теорией, чтобы проверить, являются ли эти связи устойчивыми во времени, и дать количественные прогнозы относительно будущего, а также оценить точность этих прогнозов.

### **1.3. Данные: источники и типы**

Данные, с которыми мы сталкиваемся в эконометрике, появляются из двух источников: как результаты экспериментов или как неэкспериментальные наблюдения мира. В этой книге рассматриваются и экспериментальные данные, и наблюдения.

## **Сравнение экспериментальных данных с наблюдениями**

**Экспериментальные данные** появляются в результате экспериментов, проведенных с целью оценки влияния воздействия одного фактора на другой или последствий проведения той или иной политики, а также для исследования причинных взаимосвязей. Например, штат Теннеси профинансировал большой случайный управляемый эксперимент, посвященный исследованию влияния количества детей в классе на успеваемость школьников в 1980-х годах. В этом эксперименте (мы рассмотрим его в главе 13) тысячи учеников были случайным образом распределены в классы различной численности на несколько лет и писали ежегодные стандартные тесты.

Этот эксперимент стоил миллионы долларов и потребовал постоянного взаимодействия большого количества администраторов, родителей и учителей в течение нескольких лет. Поскольку ходом экспериментов с участием людей сложно управлять, они, в отличие от идеальных случайных экспериментов, имеют множество недостатков. Более того, при некоторых обстоятельствах проведение экспериментов не только очень дорогое и трудно управляемое мероприятие, но и зачастую оно противоречит этике. (Можно ли считать этичным предлагать случайно выбранным подросткам дешевые сигареты и смотреть, купят ли они их?) Из-за всех этих финансовых, практических и этических соображений эксперименты в экономике являются достаточно редким явлением. Вместо них мы используем данные, которые появляются при наблюдении за реальным поведением людей.

Данные, собранные посредством наблюдения за поведением реально существующих объектов, а не за поведением в эксперименте, называются *наблюдениями*. Такие данные собираются с использованием опросов, например, телефонных опросов потребителей или административных записей, таких как исторические данные кредитных организаций о заявках на получение ипотечного кредита.

Очевидно, что данные этого типа создают множество проблем, связанных с оценкой причинных влияний, и требуют эконометрических методов, позволяющих решить все эти проблемы. В реальном мире уровни «воздействия» (количество удобрений в примере с помидорами и пропорция между количеством учителей и учеников из примера про размер класса) не назначаются случайным образом, что затрудняет отделение причинного влияния рассматриваемых факторов от влияния других существенных факторов. Большая часть эконометрики, как и значительная часть данной книги, посвящена методам решения проблем, возникающих при работе с реальными данными для выявления причинных связей.

Вне зависимости от того, являются ли данные экспериментальными или нет, они разделяются на три типа: межобъектные, временные ряды и панельные данные. В данной книге мы встретимся со всеми тремя типами данных.

## **Межобъектные выборки**

Данные по различным объектам – работникам, потребителям, правительстенным учреждениям, странам и другим объектам – за один момент времени

называются *межобъектными данными*<sup>1</sup>. Например, данные о результатах тестов в калифорнийских школах являются межобъектными. Эти данные состоят из 420 наблюдений – школьных округов – за единственный период времени – 1999 год. В общем случае количество наблюдений в межобъектной выборке обозначается  $n$ . Так, например, для данных по Калифорнии  $n = 420$ .

Данные по результатам школьных тестов в Калифорнии содержат информацию о нескольких переменных для каждого округа. Данные выборочно приведены в таблице 1.1. В строках таблицы находятся данные для отдельного округа. Например, средний балл за тест для первого округа («округ № 1») равен 690,8. Это число равно среднему баллу оценок по математике и естествознанию для всех пятиклассников в данном округе в 1999 году по результатам стандартного теста (стэнфордский тест). Среднее отношение числа учеников и учителей в этом округе равно 17,89, то есть число учеников в округе № 1, разделенное на количество школьных учителей в округе № 1, равно 17,89. Средние расходы на ученика в округе № 1 равны 6 385 долл. США. Процент учеников в данном округе, все еще изучающих английский язык, то есть процент тех, для кого английский является вторым языком, и тех, кто пока не освоил английский, равен нулю.

Таблица 1.1

**Некоторые данные результатов тестов и других переменных  
для школьных округов Калифорнии в 1999 году**

Номер наблюдения (округ)	Среднее значение в округе	Количество учеников на одного учителя	Расходы на ученика (долл.)	Доля учеников, изучающих английский язык (%)
1	690,8	17,89	6 385	0,0
2	661,2	21,52	5 099	4,6
3	643,6	18,70	5 502	30,0
4	647,7	17,36	7 102	0,0
5	640,8	18,67	5 236	13,9
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
418	645,0	21,89	4 403	24,3
419	672,2	20,20	4 776	3,0
420	655,8	19,04	5 993	5,0

Примечание. Калифорнийские данные по результатам тестов приведены в Приложении 4.1.

<sup>1</sup> В русскоязычной литературе для перевода термина *cross-sectional data* используются такие термины, как «пространственные», «перекрестные» или «одномоментные выборки», а также прямая калька с английского варианта – «кросс-секционные данные». В данной книге мы будем использовать термин «межобъектные данные», который, на наш взгляд, наиболее точно передает смысл рассматриваемого понятия. – Примеч. науч. ред. перевода.

Остальные строки таблицы содержат данные по другим округам. Порядок округов выбран произвольно, как и номер округа, называемый *номером наблюдения*, назначен произвольным образом для упорядочивания данных. Как вы могли заметить, все указанные переменные существенно различаются от наблюдения к наблюдению.

Используя межобъектные данные, мы можем узнать о связях между переменными, изучая различия между людьми, фирмами или другими экономическими единицами за один период.

## **Временные ряды**

**Временные ряды** представляют собой наблюдения одного объекта (человека, фирмы, страны и проч.) в разные моменты (периоды) времени. Рассмотренные ранее данные по уровням инфляции и безработицы в США являются примерами временных рядов. Данные содержат наблюдения двух переменных (уровня инфляции и безработицы) для одного объекта (США) за 183 периода времени. Каждый временной период в этом наборе данных является кварталом (первый квартал – это январь, февраль и март; второй – апрель, май и июнь и т.д.). Наблюдения начинаются во втором квартале 1959 года, который обозначен как 1959: II, а заканчиваются в четвертом квартале 2004 года (обозначен как 2004: IV). Количество наблюдений (т.е. периодов времени) во временном ряду<sup>1</sup> обозначается  $T$ . Поскольку имеется 183 квартала с 1959: II по 2004: IV, рассматриваемый временной ряд содержит  $T=183$  наблюдения.

Таблица 1.2

### **Некоторые данные по уровню инфляции индекса потребительских цен и уровню безработицы в США: квартальные данные, 1959–2004 годы**

Номер наблюдения	Дата (год: квартал)	Инфляция по ИПЦ (%)	Уровень безработицы (%)
1	1959: II	0,7	5,1
2	1959: III	2,1	5,3
3	1959: IV	2,4	5,6
4	1960: I	0,4	5,1
5	1960: II	2,4	5,2
.	.	.	.
.	.	.	.
.	.	.	.

---

<sup>1</sup> Количество наблюдений временного ряда также называется *длиной* временного ряда. – Примеч. науч. ред. перевода.

## Глава 1. Типы данных и вопросы, которые интересуют экономистов

---

*Окончание таблицы 1.2*

Номер наблюдения	Дата (год: квартал)	Инфляция по ИПЦ (%)	Уровень безработицы (%)
181	2004: II	4,3	5,6
182	2004: III	1,6	5,4
183	2004: IV	3,5	5,4

*Примечание.* Данные по уровню инфляции и безработицы в США описаны в приложении 14.1.

Некоторые наблюдения рассматриваемых временных рядов приведены в таблице 1.2. Данные в каждой строке таблицы соответствуют различным временным периодам (году и кварталу). Во втором квартале 1959 года, к примеру, уровень инфляции был равен 0,7 %. Другими словами, если бы инфляция была в течение 12 месяцев на том же уровне, что и во втором квартале 1959 года, то общий уровень инфляции за год, измеренный при помощи индекса потребительских цен (ИПЦ), вырос бы на 0,7 %. Во втором квартале 1959 года уровень безработицы был равен 5,1 %, то есть 5,1 % от общего числа работоспособного населения заявили об отсутствии работы и о том, что они ищут новое рабочее место. В третьем квартале 1959 года уровень инфляции по ИПЦ был равен 2,1 %, а уровень безработицы – 5,3 %.

Временные ряды, наблюдения за отдельным объектом во времени можно использовать для изучения эволюции переменных во времени и для прогнозирования их будущих значений.

Таблица 1.3

### Некоторые данные по продажам сигарет, ценам, налогам по штатам США за различные годы, 1985–1995 годы

Номер наблюдения	Штат	Год	Продажи сигарет (пачек на человека, шт.)	Средние цены за пачку (с учетом налогов, долл.)	Общие налоги (акциз на сигареты и налог с продаж, долл.)
1	Алабама	1985	116,5	1,022	0,333
2	Арканзас	1985	128,5	1,015	0,370
3	Аризона	1985	104,5	1,086	0,362
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
47	Западная Вирджиния	1985	112,8	1,089	0,382
48	Вайоминг	1985	129,4	0,935	0,240
49	Алабама	1986	117,2	1,080	0,334
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
96	Вайоминг	1986	127,8	1,007	0,240

Окончание таблицы 1.3

Номер наблюдения	Штат	Год	Продажи сигарет (пачек на человека, шт.)	Средние цены за пачку (с учетом налогов, долл.)	Общие налоги (акциз на сигареты и налог с продаж, долл.)
97	Алабама	1987	115,8	1,135	0,335
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
528	Вайоминг	1995	112,2	1,585	0,360

Примечание. Данные по потреблению сигарет приведены в приложении 12.1.

## Панельные данные

Панельные данные, также называемые лонгитюдными данными, – это данные, содержащие информацию о различных объектах в различные (два и более) моменты времени. Рассмотренные нами данные по потреблению сигарет и ценам на них являются примером панельных данных. В таблице 1.3 приведены отдельные переменные и наблюдения из этого примера. Число объектов в панельных данных обозначается  $n$ , а число временных периодов –  $T$ . В данных по потреблению сигарет мы имеем  $n=48$  континентальных штатов США (объекты) за  $T=11$  лет (временных периодов) с 1985 по 1995 год. Таким образом, имеется  $n \times T = 48 \times 11 = 528$  наблюдений.

Некоторые данные по потреблению сигарет приведены в таблице 1.3. Первый блок из 48 наблюдений содержит данные по каждому штату в 1985 году. Данные упорядочены по алфавиту, начиная с Алабамы и заканчивая Вайомингом. В следующем блоке из 48 наблюдений содержатся данные за 1986 год, и так до 1995 года. Например, в 1985 году продажи сигарет в Арканзасе были равны 128,5 пачек на человека (т.е. общее число проданных пачек сигарет в Арканзасе в 1985 году, разделенное на общее население Арканзаса в 1985 году, равно 128,5). Средняя цена упаковки сигарет в Арканзасе в 1985 году с учетом налогов была равна 1,015 долл., из которых 37 центов составляют федеральные налоги, налоги штата и муниципальные налоги.

Панельные данные можно использовать для изучения экономических взаимосвязей на базе наблюдений за большим числом объектов из выборки и их эволюции с течением времени.

Определения межобъектных данных, временных рядов и панельных данных приведены во вставке «Основные понятия 1.1».

### Межобъектные данные, временные ряды и панельные данные

- Межобъектные данные состоят из множества объектов, наблюдаваемых в один момент времени.
- Временные ряды состоят из наблюдений одного объекта в разные моменты времени.
- Панельные данные (также известные как лонгитюдные данные) состоят из множества объектов, причем каждый объект наблюдается в течение двух и более периодов.

## **Выводы**

1. Многие решения в бизнесе и экономике требуют количественных оценок того, как изменение одной переменной влияет на другую переменную.
2. Концептуально правильным способом оценки причинных влияний является случайный управляемый эксперимент. Однако проведение подобных экспериментов в экономических приложениях обычно противоречит этике, не-практично и очень дорого.
3. Эконометрика дает инструменты для оценки причинных влияний с использованием либо неэкспериментальных данных, либо же данных, полученных в результате настоящих экспериментов со всеми их недостатками.
4. Межобъектные данные собираются посредством наблюдения за многими объектами в один период времени; временные ряды собираются путем наблюдения за одним объектом с течением времени; панельные данные получаются в результате наблюдения за большим числом объектов, причем каждый из объектов наблюдается два и более периодов.

## **Основные понятия**

Случайный управляемый эксперимент (с. 8).

Контрольная группа (с. 8).

Исследуемая (экспериментальная) группа (с. 8).

Причинно-следственное влияние, причинные связи (с. 8).

Экспериментальные данные (с. 10).

Наблюдения (неэкспериментальные данные) (с. 10).

Межобъектные данные (с. 11).

Временные ряды (с. 12).

Панельные данные (с. 14).

Лонгитюдные данные (с. 14).

## **Вопросы для повторения и закрепления основных понятий**

- 1.1. Разработайте гипотетический идеальный случайный эксперимент для изучения влияния времени, потраченного на учебу, на результаты экзамена по микроэкономике. Назовите препятствия, которые могут возникнуть при реализации этого эксперимента.
- 1.2. Предложите гипотетической идеальной случайный контролируемый эксперимент по исследованию влияния использования ремней безопасности на количество смертей в ДТП на шоссе. Назовите препятствия, которые могут возникнуть при реализации подобного эксперимента.
- 1.3. Вас попросили исследовать причинно-следственное влияние часов, затрачиваемых на программу повышения квалификации работников

## Часть I. Введение и обзор основных понятий теории вероятностей...

---

завода (измеренное в часах на одного работника в неделю), на производительность труда работников (производимая продукция на работника в час). Опишите:

- а) идеальный случайный управляемый эксперимент для измерения этого причинно-следственного влияния;
- б) неэкспериментальные межобъектные данные, при помощи которых вы могли бы оценить такое влияние;
- в) наблюдения в виде временных рядов для изучения этого влияния;
- г) панельные данные для изучения этого влияния.

## **Глава 2. Основные понятия теории вероятностей**

В данной главе описаны основные идеи, понятия и результаты теории вероятностей, которые необходимы для понимания главных концепций регрессионного анализа и эконометрики. Предполагается, что читатель уже имеет базовые знания по математической статистике и теории вероятности. Поэтому глава предназначена, в первую очередь, для тех, кто не очень уверен в своих знаниях в этой области либо чувствует необходимость освежить эти знания. Тем же, кто в совершенстве владеет материалом, представленным в главе, будет достаточно лишь бегло просмотреть ее содержание, основные определения и формулировки для предотвращения возможного непонимания, связанного с обозначениями.

Большинство явлений, окружающих нас, содержат в себе элемент случайности. Теория вероятностей предлагает математические инструменты для описания этих явлений. В разделе 2.1 дается описание вероятностного распределения случайной величины. Раздел 2.2 знакомит нас с понятиями математического ожидания, среднего значения и дисперсии одномерной случайной величины. Большинство задач экономики включает в себя более чем одну случайную величину, поэтому в разделе 2.3 дается описание базовых элементов теории вероятностей для двух случайных величин. Раздел 2.4 содержит в себе описание трех основных вероятностных распределений, играющих главную роль в статистике и эконометрике, – нормального распределения, распределения хи-квадрат и  $F$ -распределения.

Последние два раздела главы посвящены специальному источнику случайности, который занимает центральное место в эконометрике: случайности, возникающей при рассмотрении случайной выборки некоторой генеральной совокупности. Рассмотрим, например, выборку из десяти выпускников высших учебных заведений. Мы записываем величину их заработной платы и считаем средний заработка, основываясь на сделанных десяти наблюдениях. Так как выборка была проведена случайным образом, то каждый из десяти выпускников был выбран совершенно случайно из всего числа выпускников высших учебных заведений. Если мы выберем десять выпускников подобным образом еще раз, то обнаружим, что новая величина среднего заработка в сделанной нами выборке отличается от предыдущей. Таким образом, мы можем смело сказать, что средняя величина заработка также является случайной величиной. Следовательно, она также имеет вероятностное распределение, которое является распределением, соответствующим данной выборке, так как это распределение описывает различные средние значения по различным выборкам.

Раздел 2.5 описывает случайную выборку и распределение средних значений в выборках. Как правило, это распределение является весьма сложным.

Тем не менее при достаточно больших размерах выборки данное распределение среднего аппроксимируется нормальным распределением (даный результат доказывает центральная предельная теорема, описанная в разделе 2.6).

## 2.1. Случайные величины и вероятностные распределения

### **Вероятность, пространство исходов и случайные величины**

**Вероятности и исходы.** Пол человека, которого вы встретите первым, выйдя на улицу, оценка за экзамен, число сбоев вашего компьютера во время написания вами курсовой работы — все эти, и не только эти, события содержат в себе элемент случайности. Во всех вышеописанных ситуациях есть информация, неизвестная в данный момент, но которая становится известной с наступлением этих событий.

Взаимно исключаемые возможные результаты случайного процесса будем называть *исходами* или *элементарными событиями*. Например, тот же компьютер может никогда не выйти из строя, может дать сбой всего лишь только один раз, или два, или больше. Только один из этих исходов будет в действительности иметь место, то есть каждый из исходов исключает другой. Причем вероятности того, что будет иметь место тот или иной исход, неодинаковы.

**Вероятность исхода** представляет собой отношение числа наступлений данного исхода к общему числу событий. Если вероятность сбоя компьютера равна 80 %, то на нем можно написать лишь 80 % курсовых работ без сбоев компьютера.

**Пространство исходов и события.** Множество всех возможных исходов называется *пространством исходов* или *пространством элементарных событий*. Событие представляет собой некоторое подмножество пространства исходов. Таким образом, событие — это множество исходов. Событие «мой компьютер выйдет из строя не более одного раза» состоит из двух исходов: «один сбой» и «нет сбоя».

**Случайные величины.** Случайная величина — это численная интерпретация некоторого случайного исхода. Количество раз, когда компьютер выходит из строя во время написания курсовой работы, случайно и в то же время принимает численную величину, так что это случайная величина.

Существуют дискретные и непрерывные случайные величины. Как нетрудно догадаться из названия, *дискретные случайные величины* принимают только дискретные значения, такие как 0, 1, 2, 3, ..., в то время как *непрерывные случайные величины* принимают значения из континуума возможных значений.

### **Распределение вероятностей дискретной случайной величины**

**Распределение вероятностей.** *Распределение вероятностей* дискретной случайной величины представляет собой множество всех возможных значений этой случайной величины с вероятностями принятия этой величиной каждого из возможных исходов. При этом выполняется условие, что сумма этих вероятностей равна единице.

Например, пусть  $M$  – число сбоев компьютера, произошедших во время написания курсовой работы. Распределение вероятностей случайной величины  $M$  – это множество значений вероятностей каждого возможного исхода: вероятность  $M = 0$  обозначается как  $\Pr(M = 0)$  – это означает, что компьютер не давал сбоев;  $\Pr(M = 1)$  – это вероятность единственного сбоя компьютера и так далее. Пример распределения вероятностей числа сбоев  $M$  представлен во второй строке таблицы 2.1. Согласно этому распределению, вероятность отсутствия сбоев равна 80%, вероятность возникновения лишь одного сбоя – 10%, двух – 6%, трех – 3%, четырех – 1%. Сумма всех вероятностей в данном примере равна 100% (или единице). В соответствии с описанным распределением при возникновении четырех сбоев компьютера дальнейшее написание курсовой работы происходит от руки. Графическое представление данного распределения представлено на рисунке 2.1.

Таблица 2.1

Вероятность числа  $M$  сбоев компьютера

	Исход (количество сбоев)				
	0	1	2	3	4
Распределение вероятностей	0,80	0,10	0,06	0,03	0,01
Совместное распределение вероятностей	0,80	0,90	0,96	0,99	1,00

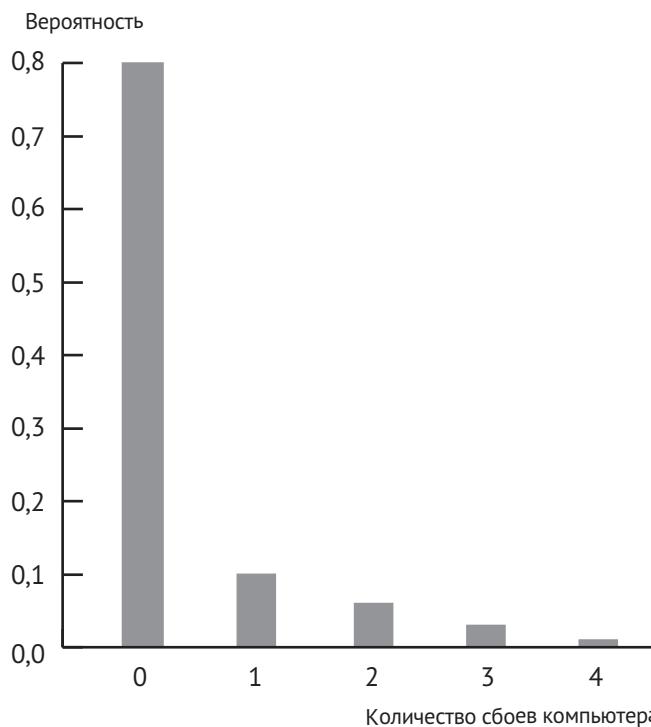


Рисунок 2.1. Распределение вероятностей числа сбоев компьютера

Высота каждого столбца – это вероятность указанного количества сбоев компьютера. Высота первого столбца равна 0,8, что означает, что вероятность отсутствия сбоев равна 80%. Высота второго столбца равна 0,1, что соответствует 10% вероятности одного сбоя и так далее.

**Вероятность событий.** Вероятность события определяется исходя из распределения вероятностей. Например, вероятность события, состоящего в том, что произошел один или два сбоя компьютера, равна сумме вероятностей каждого из этих исходов. Таким образом,  $\Pr(M=1 \text{ или } M=2) = \Pr(M=1) + \Pr(M=2) = 0,10 + 0,06 = 0,16$ , или 16 %.

**Интегральное распределение вероятностей.** Интегральное (кумулятивное) распределение вероятностей – это вероятность того, что случайная величина принимает значение меньшее или равное определенному числу. Последняя строка в таблице 2.1 представляет интегральное распределение вероятностей случайной величины  $M$ . Например, вероятность того, что может произойти хотя бы один сбой компьютера,  $\Pr(M \leq 1)$ , равна 90 %, что представляет собой сумму отсутствия сбоев (80 %) и единичный сбой (10 %). Интегральное распределение вероятностей часто называют интегральной (кумулятивной) функцией распределения, или интегральным (кумулятивным) распределением.

**Распределение Бернулли.** Важным частным случаем дискретной случайной величины является бинарная случайная величина. В данном случае наблюдаемая величина принимает значения либо 0, либо 1. Бинарную случайную величину называют случайной величиной Бернулли в честь швейцарского математика и ученого Яакоба Бернулли, жившего в XVII веке. Соответствующее распределение вероятностей называют распределением Бернулли.

Рассмотрим следующий пример. Пусть  $G$  – пол человека, которого вы встретите на улице первым. В таком случае  $G$  будет принимать значение, равное нулю, если первый встречный – мужчина, и значение, равное единице, если первый встреченный вами человек – женщина. В этом случае имеем:

$$G = \begin{cases} 1, & \text{с вероятностью } p; \\ 0, & \text{с вероятностью } 1-p, \end{cases} \quad (2.1)$$

где  $p$  – вероятность того, что тот, кого вы встретите первым, будет женщиной. Распределение вероятностей в уравнении (2.1) и есть распределение Бернулли.

### **Распределение вероятностей непрерывной случайной величины**

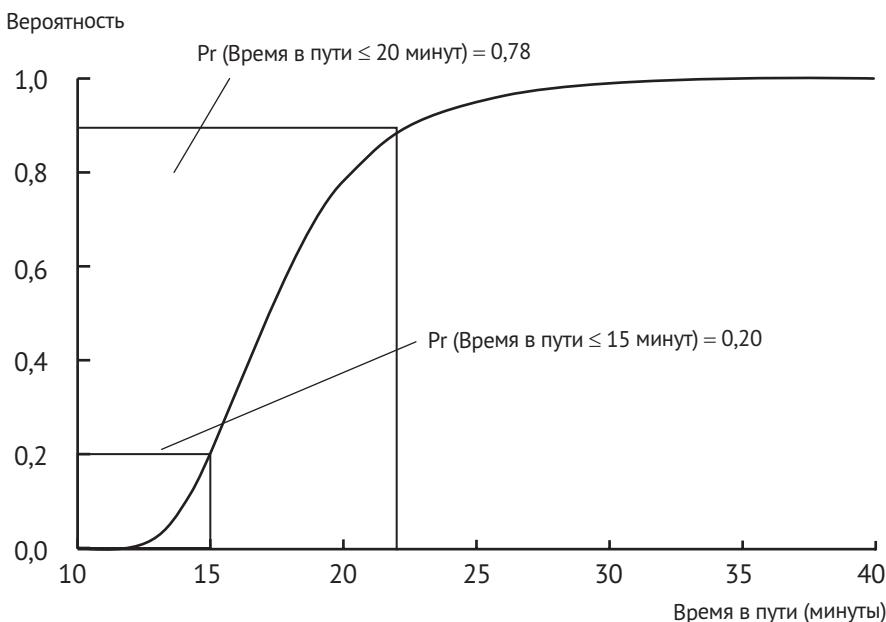
**Интегральное распределение вероятностей.** Интегральное распределение вероятностей для непрерывной случайной величины определяется аналогично дискретному случаю. Это означает, что интегральное распределение вероятностей также равно вероятности принятия случайной величиной значений ниже определенного числа либо равных ему.

В качестве примера представим студентку, едущую на автомобиле из дома в университет. Время его прибытия в учебное заведение может принять любое значение из континуума значений времени. Оно зависит от погодных условий, ситуации на дороге и прочих факторов. Таким образом, можно трактовать данное время как непрерывную случайную величину. На рисунке 2.2А

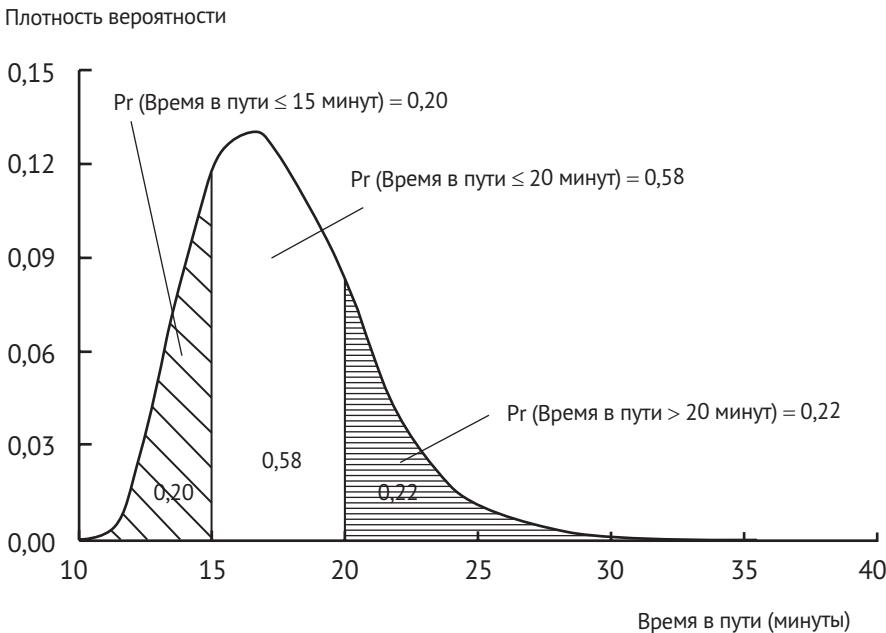
изображена гипотетическая совместная функция распределения времени, потраченного студенткой на дорогу. Например, вероятность того, что дорога займет менее 15 минут, равна 20 %, а вероятность того, что студентка будет в пути более 20 минут, – 78 %.

**Плотность вероятности.** В силу того что непрерывная случайная величина принимает значения из континуума возможных значений, распределение вероятностей, используемое для дискретного случая, не подходит для ее описания. Вместо этого в непрерывном случае используется функция плотности вероятности. Площадь под кривой функции плотности вероятности, заключенная между двумя точками, равна вероятности попадания значения случайной величины в интервал, границами которого являются данные точки. Функцию плотности вероятности также называют *функцией плотности* или просто *плотностью*.

На рисунке 2.2б изображена функция плотности времени, потраченного студенткой на дорогу до университета, соответствующего интегральной функции распределения на рисунке 2.2а. Вероятность того, что на дорогу уйдет от 15 до 20 минут, – это площадь под графиком функции плотности распределения между значениями 15 и 20 минут. Данное значение равно 0,58 или 58 %. Тот же результат можно получить, используя рисунок 2.2а: вероятность того, что студентка доберется до учебного заведения менее чем за 20 минут, равна 78 %, а того, что она доберется до него менее чем за 15 минут, – 20 %. Разность между этими значениями равна как раз 58 %. Таким образом, функции плотности распределения и интегрального распределения вероятностей представляют одну и ту же информацию, но в разных формах.



(a) Интегральная функция распределения времени в пути



(6) Функция плотности времени, затраченного на дорогу

**Рисунок 2.2. Интегральная функция распределения и функция плотности**

На рисунке 2.2а изображена интегральная функция распределения (c.d.f.) времени нахождения в пути студентки, из рассматриваемого выше примера. Вероятность того, что дорога займет больше 15 минут, равна 0,20 (20 %), вероятность того, что дорога займет меньше 20 минут, равна 0,78 (78 %).

Рисунок 2.2б представляет собой функцию плотности вероятности (p.d.f.) времени, необходимого студентке для того, чтобы добраться до учебного заведения. Вероятность того, что для этого ей потребуется от 15 до 20 минут, равна 0,58 (58 %), что определяется как площадь под графиком между двумя вертикальными прямыми, исходящими из точек 15 и 20 минут.

## 2.2. Математическое ожидание, среднее и дисперсия

### *Математическое ожидание случайной величины*

**Математическое ожидание.** Математическим ожиданием случайной величины  $Y$ , обозначаемым как  $E(Y)$ , называется среднее значение случайной величины, рассчитанное для большого числа повторяющихся испытаний в долгосрочном периоде. Математическое ожидание дискретной случайной величины рассчитывается как взвешенное среднее всех возможных значений случайной переменной с весами, равными вероятностям принятия случайной величиной соответствующего значения. Математическое ожидание переменной  $Y$  также называют *ожиданием*  $Y$  или его *средним значением* и обозначают  $\mu_Y$ .

Рассмотрим следующий пример. Предположим, вы одолжили другу 100 долл. под 10 % годовых. При благоприятном исходе событий вы получите обратно 110 долл. (100 долл.– основной долг и 10 долл. – процент). Но существует общий риск невозврата долга, равный 1 %. Следовательно, сумма, ожидаемая вами к возврату, равна 110 долл. с вероятностью 0,99 и 0 долл. с вероятностью 0,01.

Таким образом, при всех подобных займах в 99% случаев вы получите обратно 110 долл., но в одном проценте всех случаев вам ничего не вернут. Тогда получаем, что в среднем вы получите обратно  $110 \times 0,99 + 0 \times 0,01 = 108,90$  долл. Это и есть математическое ожидание суммы, которую вернет вам ваш друг.

Рассмотрим еще один пример. Предположим, компьютер ломается  $M$  раз, распределение вероятностей этого события представлено в таблице 2.1. Ожидаемое значение  $M$  равно среднему количеству сбоев компьютера в ходе написания нескольких (для чистоты эксперимента лучше взять бесконечно большое число) курсовых работ с весами, равными частоте возникновения определенного количества сбоев. Соответственно имеем:

$$E(M) = 0 \times 0,80 + 1 \times 0,10 + 2 \times 0,06 + 3 \times 0,03 + 4 \times 0,01 = 0,35. \quad (2.2)$$

Получается, что среднее количество сбоев компьютера в ходе написания курсовой работы равно 0,35. Конечно, данное значение должно быть целочисленным числом, так как бессмысленно говорить о том, что компьютер ломался 0,35 раз! Тем не менее полученная величина показывает, что в среднем при написании курсовых работ компьютер ломается в 35 случаях из 100.

Формула для математического ожидания дискретной случайной величины  $Y$ , принимающей  $k$  различных значений, приведена во вставке «Основные понятия 2.1» (вставка «Основные понятия 2.1» использует символ суммирования, описанного в упражнении 2.25).

### Математическое ожидание и среднее значения

Предположим, что случайная величина  $Y$  принимает  $k$  возможных значений  $y_1, \dots, y_k$ , где  $y_1$  – первое возможное значение,  $y_2$  – второе возможное значение и так далее. Вероятность того, что  $Y$  примет значение  $y_i$ , равна  $p_i$ . Аналогично определяются значения  $p_2, p_3$  и так далее. Математическое ожидание (ожидаемое значение) величины  $Y$  в таком случае определяется следующим образом:

$$E(Y) = y_1 p_1 + y_2 p_2 + \dots + y_k p_k = \sum_{i=1}^k y_i p_i, \quad (2.3)$$

где выражение  $\sum_{i=1}^k y_i p_i$  означает суммирование  $y_i p_i$  по индексу  $i$ , принимающему значения от 1 до  $k$ . Математическое ожидание  $Y$  также называется средним значением  $Y$  и обозначается  $\mu_Y$ .

## ОСНОВНЫЕ ПОНЯТИЯ

### 2.1

**Математическое ожидание бернуlliевской случайной величины.** Важным частным случаем понятий, рассмотренных выше, является математическое ожидание случайной величины, имеющей распределение Бернулли. Обозначим с помощью  $G$  бернуlliевскую случайную величину, имеющую распределение вероятностей, заданное уравнением (2.1). Математическим ожиданием  $G$  является

$$E(G) = 1 \times p + 0 \times (1 - p) = p. \quad (2.4)$$

Таким образом, ожидаемым значением случайной величины, имеющей распределение Бернулли, является  $p$ , вероятность того, что случайная величина примет значение, равное единице.

**Математическое ожидание непрерывной случайной величины.** Математическим ожиданием непрерывной случайной величины также является взвешенное по вероятностям среднее всех возможных ее значений. Так как случайная величина принимает значения из континуума возможных значений, формальное представление ожидаемого значения использует аппарат интегрального исчисления. Точное определение математического ожидания непрерывной случайной величины приводится в Приложении 17.1.

## Стандартное отклонение и дисперсия

Дисперсия и стандартное отклонение измеряют изменчивость (или «разброс») распределения вероятностей. Дисперсией случайной величины  $Y$ , обозначаемой как  $\text{var}(Y)$ , называется математическое ожидание квадрата отклонения  $Y$  от своего среднего значения:  $\text{var}(Y) = E[(Y - \mu_y)^2]$ .

В силу того что дисперсия  $Y$  включает в себя квадрат  $Y$ , ее единицами измерения являются значения  $Y$  в квадрате, что делает затруднительной интерпретацию дисперсии. По этой причине часто для измерения величины разброса распределения вероятности используют *стандартное отклонение*, равное квадратному корню из дисперсии случайной величины и обозначающееся как  $\sigma_y$ . Единицы измерения стандартного отклонения такие же, как и у  $Y$ . Описанные определения приведены во вставке «Основные понятия 2.2».

### ОСНОВНЫЕ ПОНЯТИЯ 2.2

#### Дисперсия и стандартное отклонение

Дисперсия дискретной случайной величины  $Y$ , которую принято обозначать  $\sigma_y^2$ , определяется следующим образом:

$$\sigma_y^2 = \text{var}(Y) = E[(Y - \mu_y)^2] = \sum_{i=1}^k (y_i - \mu_y)^2 p_i. \quad (2.5)$$

Стандартным отклонением случайной величины  $Y$  (обозначается как  $\sigma_y$ ) называется квадратный корень из дисперсии.

В рассмотренном выше примере дисперсия числа сбоев компьютера является взвешенным по вероятностям средним значением квадрата разности числа сбоев компьютера  $M$  и его среднего значения и равна 0,35:

$$\begin{aligned} \text{var}(M) &= (0 - 0,35)^2 \times 0,80 + (1 - 0,35)^2 \times 0,10 + (2 - 0,35)^2 \times \\ &\quad \times 0,06 + (3 - 0,35)^2 \times 0,03 + (4 - 0,35)^2 \times 0,01 = 0,6475. \end{aligned} \quad (2.6)$$

Стандартным отклонением  $M$  будет в данном случае величина  $\sigma_M = \sqrt{0,6475} \cong 0,80$ .

**Дисперсия случайной величины Бернулли.** Как мы уже выяснили, средним значением, или математическим ожиданием, случайной величины  $G$ , имеющей распределение вероятностей, заданное уравнением (2.1), является  $\mu_G = p$  [см. уравнение (2.4)]. В таком случае ее дисперсией будет:

$$\text{var}(G) = \sigma_G^2 = (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p). \quad (2.7)$$

Соответственно стандартным отклонением случайной величины, заданной законом распределения вероятностей Бернулли, будет  $\sigma_G = \sqrt{p(1 - p)}$ .

### **Математическое ожидание и дисперсия линейной функции от случайной величины**

В данном разделе рассматриваются случайные величины (назовем их  $X$  и  $Y$ ), связанные между собой линейно. Рассмотрим следующий пример. Пусть заработная плата работника облагается налогом, равным 20% от ее размера. Помимо заработной платы работник получает не облагаемый налогом трансферт, например грант в размере 2000 долларов. Тогда доходы работника после учета налога ( $Y$ ) связаны с доходами до уплаты налога ( $X$ ) следующим образом:

$$Y = 2000 + 0,8X. \quad (2.8)$$

Допустим, доходы до уплаты налога рассматриваемого нами индивида в следующем году представляют собой случайную величину с математическим ожиданием  $\mu_X$  и дисперсией  $\sigma_X^2$ . В таком случае доход работника после списания налога также является случайной величиной. Чему, в таком случае, будут равны математическое ожидание и дисперсия дохода индивида в следующем году после вычета налога? Мы знаем, что чистый доход индивида равен сумме 80% от заработной платы плюс 2000 долларов. Средним значением этой величины будет:

$$E(Y) = \mu_Y = 2000 + 0,8\mu_X. \quad (2.9)$$

Дисперсией чистого дохода индивида является математическое ожидание величины  $(Y - \mu_Y)^2$ . Так как  $Y = 2000 + 0,8X$ ,  $Y - \mu_Y = 2000 + 0,8X - (2000 + 0,8\mu_X) = 0,8(X - \mu_X)$ .

Получаем:  $E[(Y - \mu_Y)^2] = E\{[0,8(X - \mu_X)]^2\} = 0,64E[(X - \mu_X)^2]$ .

Следовательно,  $\text{var}(Y) = 0,64 \text{ var}(X)$ , а стандартным отклонением  $Y$  будет:

$$\sigma_Y = 0,8\sigma_X. \quad (2.10)$$

То есть стандартное отклонение распределения доходов индивида после списания налогов равно 80% стандартного отклонения распределения доходов до учета налогов.

В обобщенном варианте для данной ситуации зависимость  $Y$  от  $X$  задается следующим образом:

$$Y = a + bX. \quad (2.11)$$

Математическое ожидание и дисперсия в этом случае определяются как

$$\mu_Y = a + b\mu_X \text{ и} \quad (2.12)$$

$$\sigma_y^2 = b^2 \sigma_x^2, \quad (2.13)$$

и стандартным отклонением  $Y$  будет  $\sigma_y = b\sigma_x$ . Выражения (2.9) и (2.10) являются частными случаями более общих выражений (2.12) и (2.13) с  $a = 2000$  и  $b = 0,8$ .

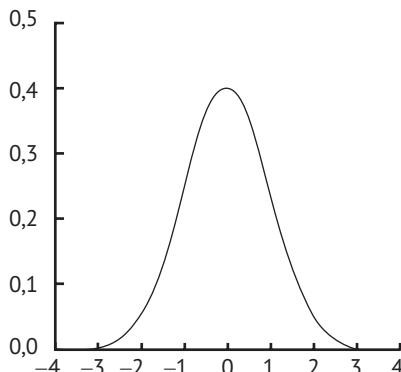
### **Другие характеристики формы функции плотности распределения вероятностей**

Среднее значение и стандартное отклонение – два наиболее важных параметра, которые характеризуют распределение вероятностей: математическое ожидание дает нам информацию о положении центра распределения, а стандартное отклонение – о его ширине. В данном разделе вводятся еще две характеристики функции плотности распределения – асимметрия, измеряющая асимметричность правого и левого хвостов распределения, и эксцесс (или куртозис) распределения, характеризующий «толщину» хвостов. Среднее значение, дисперсия, асимметрия и эксцесс рассчитываются на основе так называемых *моментов распределения*.

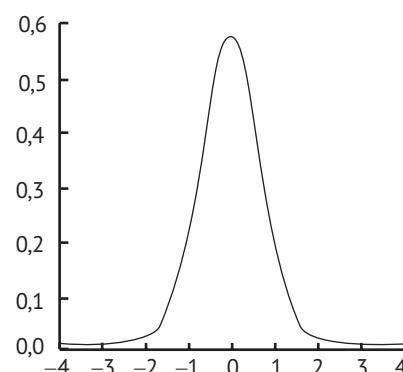
**Асимметрия.** На рисунке 2.3 изображено четыре распределения, два из которых симметричны относительно своего среднего (рис. 2.3а и 2.3б) и два несимметричны (рис. 2.3в и 2.3г). Несложно заметить, что распределение, изображенное на рисунке 2.3г, отклоняется от своего центра (т.е. асимметрично) больше, чем распределение, изображенное на рисунке 2.3в. Асимметрия распределения представляет собой численную величину, характеризующую степень симметричности распределения относительно своего математического ожидания. Если распределение симметрично относительно математического ожидания, то асимметрия равна нулю.

Асимметрия функции плотности распределения случайной величины  $Y$  вычисляется по формуле:

$$\text{Асимметрия} = \frac{E[(Y - \mu_Y)^3]}{\sigma_Y^3} \quad (2.14)$$



а) Асимметрия = 0,  
эксцесс = 3



б) Асимметрия = 0,  
эксцесс = 20

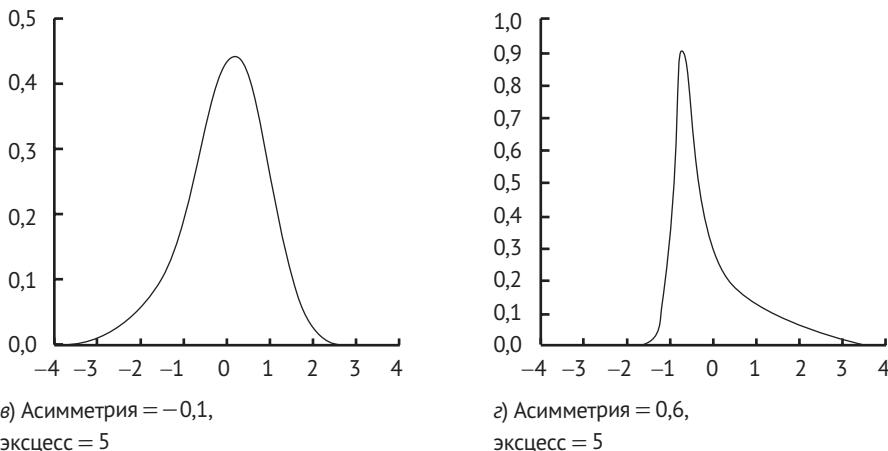


Рисунок 2.3. Четыре распределения с разными значениями асимметрии и эксцесса

Все изображенные распределения имеют математическое ожидание, равное нулю, и дисперсию, равную единице. Распределения, асимметрии которых равны нулю (случай *а* и *б*), симметричны относительно своего центра. Распределения, имеющие асимметрии, отличные от нуля, – несимметричны (случай *в* и *г*). Распределения с эксцессом, превышающим 3 (случаи *б*–*г*), имеют тяжелые хвосты и более острый пик по сравнению с нормальным распределением.

где  $\sigma_Y$  – стандартное отклонение случайной величины  $Y$ . В случае симметричных распределений вероятность того, что  $Y$  превысит свое среднее значение на какую-то фиксированную величину, будет равна вероятности того, что оно окажется меньше своего среднего на эту же величину. В таком случае положительные значения  $(Y - \mu_Y)^3$  будут в среднем (в терминах математического ожидания) компенсироваться отрицательными значениями. Поэтому, как уже отмечалось, для симметричных распределений имеем  $E[(Y - \mu_Y)^3] = 0$ , и, следовательно, значения асимметрии симметричных распределений равны нулю. Если распределение несимметрично относительно своего математического ожидания, то положительные значения величины  $(Y - \mu_Y)^3$  не всегда будут компенсировать отрицательные, в этом случае коэффициент асимметрии не будет равен нулю. Деление на  $\sigma_Y^3$  в формуле (2.14) позволяет избавиться от единиц измерения  $Y^3$ , делая, таким образом, все выражение для асимметрии безразмерной величиной. Другая интерпретация этого действия заключается в том, что изменение единиц измерения  $Y$  не окажет влияния на значение асимметрии.

На рисунке 2.3 приведены значения асимметрии для каждого из четырех распределений. Если у распределения длинный правый хвост (по сравнению с левым), то это означает, что положительные значения  $(Y - \mu_Y)^3$  не компенсируются полностью отрицательными значениями, и тогда значение асимметрии положительно. Аналогично, если у распределения длинный левый хвост, то асимметрия отрицательна.

**Эксцесс.** Эксцессом распределения называется величина, измеряющая тяжесть хвостов распределения (т.е. площадь под ними). Можно сказать, что значение эксцесса распределения говорит о том, насколько сильно дисперсия случайной

величины  $Y$  зависит от *выбросов*, или экстремальных значений. Чем больше эксцесс распределения, тем более вероятно наличие выбросов.

Эксцесс распределения случайной величины  $Y$  определяется следующим образом:

$$\text{Эксцесс} = \frac{E[(Y - \mu_Y)^4]}{\sigma_Y^4}. \quad (2.15)$$

Если распределение имеет тяжелые хвосты, то, вероятно, существуют значительные (экстремальные) отклонения  $Y$  от своего среднего, и именно эти значения приведут к большим значениям в среднем (в смысле математического ожидания) величины  $(Y - \mu_Y)^4$ . Следовательно, подобные распределения характеризуются большим значением эксцесса. В силу четности степени выражения  $(Y - \mu_Y)^3$ , значение эксцесса не может быть отрицательным.

Для случайной величины, имеющей нормальное распределение, эксцесс равен 3. Это значит, что случайная величина, эксцесс которой превышает 3, имеет более тяжелые, по сравнению с нормальным распределением, хвосты. Подобные распределения называются *островершинными*, или, что более часто используется, распределениями с тяжелыми хвостами. Подобно коэффициенту асимметрии эксцесс – это безразмерная величина.

На рисунке 2.3 приведены значения эксцесса каждого распределения. Распределения на рисунках 2.3 б–г являются острровершинными.

**Моменты распределения.** Среднее, или математическое ожидание  $E(Y)$  значений случайной величины  $Y$  называют также первым моментом  $Y$ , а математическое ожидание квадрата  $Y$ ,  $E(Y^2)$ , называют вторым моментом  $Y$ . Вообще говоря, математическое ожидание  $Y^r$  называют  $r$ -тым моментом случайной величины  $Y$ . Таким образом,  $r$ -тым моментом  $Y$  является  $E(Y^r)$ . Асимметрия – это функция от первого, второго и третьего моментов  $Y$ , в то время как эксцесс является функцией первых четырех моментов случайной величины  $Y$ .

## 2.3. Двухмерные случайные величины

Большинство интересующих нас задач в экономике включает в себя две и более случайные величины. Например, действительно ли, что выпускники университетов имеют больше шансов получить работу, чем те, кто таковыми не являются? Как соотносятся между собой распределения доходов женщин и доходов мужчин? Оба этих вопроса используют распределение двух случайных величин, которые рассматриваются одновременно (образование и статус трудовой занятости в первом случае, доход и половой признак во втором). Поиск ответов на подобные вопросы требует введения таких понятий, как совместное, безусловное и условное распределение вероятностей.

### Совместное и безусловное распределение вероятностей

**Совместное распределение.** Совместным распределением вероятностей двух дискретных случайных переменных  $X$  и  $Y$  называется вероятность того, что

обе величины одновременно принимают определенные значения, скажем,  $x$  и  $y$ . Сумма вероятностей принятия двумерной случайной величиной  $(X, Y)$  всех возможных значений  $(x, y)$  равна 1. Совместное распределение вероятностей обозначается как  $\Pr(X = x, Y = y)$ .

Для примера рассмотрим влияние погодных условий на время, которое тратит студентка на путь из дома до университета (см. раздел 2.1). Условие «идет дождь или нет» оказывает влияние на продолжительность времени, прошедшего студенкой в пути. Обозначим с помощью  $Y$  бинарную случайную переменную, которая может принимать только два значения – 0 или 1.  $Y = 1$  соответствует случаю, когда дорога занимает менее 20 минут;  $Y = 0$  означает, соответственно, противное. Другая бинарная величина,  $X$ , принимает значение, равное единице в случае, если идет дождь, и равное нулю – в случае ясной погоды. Тогда имеем четыре различные комбинации этих случайных величин:  $(X = 0, Y = 0)$  – идет дождь, дорога занимает много времени;  $(X = 0, Y = 1)$  – идет дождь, но студент находится в пути менее 20 минут;  $(X = 1, Y = 0)$  – несмотря на то что дождя нет, студентка все же проводит много времени в пути;  $(X = 1, Y = 1)$  – дождя нет, студентка быстро добирается до места учебы. Совместное распределение вероятностей – это частота, с которой каждый из указанных исходов имеет место при многократном повторении подобной ситуации.

Пример совместного распределения из рассмотренной задачи представлен в таблице 2.2. Согласно этой таблице, 15 % дней из рассматриваемого периода идет дождь, и дорога занимает много времени ( $X = 0, Y = 0$ ), то есть вероятность того, что будет пасмурно и быстро добраться до учебы не получится, равна 15 %. Также имеем:  $\Pr(X = 0, Y = 1) = 0,15$ ,  $\Pr(X = 1, Y = 0) = 0,07$  и  $\Pr(X = 1, Y = 1) = 0,63$ . Эти четыре возможных исхода взаимно исключаемы и образуют пространство элементарных событий для рассматриваемой ситуации. Сумма их вероятностей равна 1.

Таблица 2.2

**Совместное распределение погодных условий и времени, затрачиваемого на дорогу**

	Пасмурно ( $X = 0$ )	Ясно ( $X = 1$ )	Сумма вероятностей
Много времени в пути ( $Y = 0$ )	0,15	0,07	0,22
Мало времени в пути ( $Y = 1$ )	0,15	0,63	0,78
Сумма вероятностей	0,30	0,70	1,00

**Безусловное распределение вероятностей.** Безусловное распределение вероятностей случайной величины  $Y$  – это просто другое название распределения вероятностей этой величины. Оно используется для установления различия распределения самой переменной  $Y$  (ее безусловного распределения) от совместного распределения  $Y$  и другой случайной величины.

Безусловное распределение  $Y$  может быть определено через совместное распределение  $X$  и  $Y$  посредством суммирования вероятностей всех возмож-

ных исходов, при которых  $Y$  принимает определенное значение. Если  $X$  может принимать  $l$  различных значений  $x_1, \dots, x_l$ , то безусловная вероятность того, что  $Y$  примет значение  $y$ , определяется следующим образом:

$$\Pr(Y = y) = \sum_{i=1}^l \Pr(X = x_i, Y = y). \quad (2.16)$$

Например, как следует из таблицы 2.2, вероятность продолжительного нахождения в пути во время дождя равна 15 %, а вероятность того, что на дорогу будет потрачено много времени, но при этом будет ясная погода, равна 7 %. Получаем, что вероятность того, что дорога займет много времени (при наличии или отсутствии дождя), равна 22 %. Безусловные распределения для каждой продолжительности пути приведены в последнем столбце таблицы 2.2. Аналогично, безусловная вероятность того, что будет пасмурная погода, равна 30 %, что показано в последней строке таблицы 2.2.

## Условные распределения вероятностей

**Условное распределение вероятностей.** Распределение случайной величины  $Y$ , обусловленной другой случайной величиной  $X$ , принимающей определенные значения, называется условным распределением случайной величины  $Y$  при условии  $X$ . Условная вероятность того, что  $Y$  примет значение  $y$ , при  $X$ , равном  $x$ , обозначается как  $\Pr(Y = y | X = x)$ .

Например, чему равна вероятность того, что студентка потратит много времени на дорогу, ( $Y = 0$ ), если известно, что в это время будет идти дождь, ( $X = 0$ )? Используя данные таблицы 2.2, получим, что совместная вероятность долгого пути во время дождя равна 15 %, а вероятность добраться до университета быстро во время дождя также равна 15 %. Отсюда во время дождя дорога может занять много или мало времени с одинаковой вероятностью. Получаем, что вероятность провести в дороге много времени, ( $Y = 0$ ), при условии, что будет идти дождь, ( $X = 0$ ), равна 50 %:  $\Pr(Y = 0 | X = 0) = 0,50$ . Аналогично, так как безусловная вероятность того, что будет идти дождь, равна 30 %, то дождь будет идти в 30 случаях из 100 независимо от затраченного на дорогу времени. Из этих 30 % вероятность того, что студент проведет в дороге много времени, равна 50 % (0,15/0,30).

В общем случае условное распределение величины  $Y$  при условии  $X = x$  есть

$$\Pr(Y = y | X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}. \quad (2.17)$$

Таблица 2.3

### Совместное распределение сбоев компьютера ( $M$ ) и возраста компьютера ( $A$ )

A. Совместное распределение						
	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Всего
Старый компьютер ( $A = 0$ )	0,35	0,065	0,05	0,025	0,01	0,50
Новый компьютер ( $A = 1$ )	0,45	0,35	0,01	0,005	0,00	0,50
Всего	0,80	0,10	0,06	0,03	0,01	1,00

## Окончание таблицы 2.3

Б. Условное распределение $M$ при условии $A$						
	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Всего
$\Pr(M A=0)$	0,70	0,13	0,10	0,05	0,02	1,00
$\Pr(M A=1)$	0,90	0,07	0,02	0,01	0,00	1,00

В рассмотренном примере вероятность того, что дорога займет много времени при условии, что будет идти дождь, равна

$$\Pr(Y=0 | X=0) = \Pr(X=0, Y=0) / \Pr(X=0) = 0,15 / 0,30 = 0,50.$$

В качестве второго примера рассмотрим модификацию задачи про сбои компьютера. Допустим, что написание курсовой работы происходит на компьютере университетской библиотеки. Библиотекарь случайным образом выбирает компьютер и предоставляет его нам для соответствующей цели. Половина компьютеров библиотеки – это старые модели, другая половина – новые. В силу случайности выбора компьютера возраст компьютера  $A$  также является случайной величиной. Пусть  $A=1$  в случае, если компьютер новый, и  $A=0$  в противном случае. Совместное распределение количества сбоев компьютера  $M$  и его возраста  $A$  приведено в части А таблицы 2.3. Условное распределение данных величин представлено в части Б той же таблицы. Например, совместное распределение такого события, при котором  $M=0$  и  $A=0$ , равно 0,35. В силу того что половина компьютеров являются старыми, условная вероятность отсутствия сбоев при работе на старом компьютере равна:  $\Pr(M=0 | A=0) = \Pr(M=0, A=0) / \Pr(A=0) = 0,35 / 0,50 = 0,70$ , то есть 70 %. В противоположность этому результату отсутствие сбоев при работе на новом компьютере происходит в 90 случаях из 100, то есть вероятность такого события равна 90 %. Согласно условным распределениям вероятностей, представленным в части Б таблицы 2.3, новые компьютеры менее подвержены сбоям, чем старые: например, вероятность того, что за время написания курсовой работы произойдет три сбоя, равна 5 % при работе на старом компьютере и 1 % при работе на новом.

**Условное математическое ожидание.** Условным математическим ожиданием (или условным средним) случайной величины  $Y$  при условии  $X$  называется математическое ожидание условного распределения  $Y$  относительно  $X$ . Таким образом, условное математическое ожидание – это ожидаемое значение величины  $Y$ , вычисленное на основе условного распределения вероятностей  $Y$  относительно  $X$ . Если  $Y$  принимает  $k$  значений  $y_1, \dots, y_k$ , то условное среднее значение величины  $Y$  при  $X=x$  определяется следующим образом:

$$E(Y | X=x) = \sum_{i=1}^k y_i \Pr(Y=y_i | X=x). \quad (2.18)$$

Например, основываясь на условных распределениях вероятностей из таблицы 2.3, получаем, что ожидаемое число сбоев компьютера при условии, что этот компьютер является старым, равно:  $E(M | A=0) = 0 \times 0,70 + 1 \times 0,13 +$

$+2 \times 0,10 + 3 \times 0,05 + 4 \times 0,02 = 0,56$ . В случае если компьютер является новым, соответствующая величина примет значение  $E(M | A = 0) = 0,14$ , которое, как и следовало ожидать, меньше, чем для старого компьютера.

Условное математическое ожидание  $Y$  при  $X = x$  есть среднее значение  $Y$  при условии, что  $X$  принимает значение  $x$ . В случае, представленном в таблице 2.3, среднее количество сбоев равно 0,56 для старых компьютеров, так что условное математическое ожидание  $Y$  при условии, что компьютер является старым, равно 0,56. Аналогично, для новых компьютеров среднее количество сбоев равно 0,14, так что условное математическое ожидание  $Y$  при условии, что компьютер является новым, равно 0,14.

**Закон повторного математического ожидания.** Математическим ожиданием случайной величины  $Y$  является взвешенное по распределению вероятностей  $X$  среднее значение  $Y$  при условии  $X$ . Например, средним ростом совершеннолетнего человека является сумма взвешенных значений среднего роста совершеннолетних мужчин и среднего роста совершеннолетних женщин. Коэффициентами (или, как часто говорят, весами) в данном случае будут являться доли совершеннолетних мужчин и женщин от их общего количества. Математически это можно записать следующим образом. Пусть случайная величина  $X$  принимает  $l$  значений  $x_1, \dots, x_l$ , тогда

$$E(Y) = \sum_{i=1}^l E(Y | X = x_i) \Pr(X = x_i). \quad (2.19)$$

Данное уравнение выводится из уравнений (2.18) и (2.17) (подробнее см. упражнение 2.19).

Иными словами, математическим ожиданием случайной величины  $Y$  называется математическое ожидание ее условного математического ожидания при условии  $X$ , то есть получаем:

$$E(Y) = E[E(Y | X)], \quad (2.20)$$

где внутреннее математическое ожидание в правой части выражения (2.20) посчитано с использованием условного распределения вероятностей  $Y$  при условии  $X$ , а внешнее определяется посредством безусловного распределения  $X$ . Уравнение (2.20) называется *законом повторного математического ожидания*.

Например, среднее число сбоев компьютера  $M$  определяется как взвешенное среднее условных математических ожиданий  $M$  при условии, что компьютер является новым в одном случае и старым – в другом. Таким образом,  $E(M) = E(M | A = 0) \times \Pr(A = 0) + E(M | A = 1) \times \Pr(A = 1) = 0,56 \times 0,50 + 0,14 \times 0,50 = 0,35$ . Это значение, очевидно, совпадает с безусловным средним значением  $M$ , вычисленным выше в уравнении (2.2).

Закон повторного математического ожидания предполагает, что если условное среднее  $Y$  при условии  $X$  равно нулю, то среднее  $Y$  равно нулю. Данное утверждение является прямым следствием уравнения (2.20): если  $E(Y | X) = 0$ , то  $E(Y) = E[E(Y | X)] = E[0] = 0$ . То есть если среднее  $Y$  при условии  $X$  равно нулю, то и средневзвешенное по вероятностям значение также равно нулю, то есть среднее  $Y$  равно нулю.

Закон повторного математического ожидания также распространяется и на несколько случайных величин. Рассмотрим три случайные совместно распределенные величины  $X$ ,  $Y$  и  $Z$ . В этом случае закон повторного математического ожидания утверждает, что  $E(Y) = E[E(Y | X, Z)]$ , где  $E(Y | X, Z)$  – это условное математическое ожидание  $Y$  при одновременном выполнении условий  $X$  и  $Z$ . Пусть в примере с поломкой компьютеров, описанном в таблице 2.3,  $P$  означает количество программ, установленных на компьютере. Тогда выражение  $E(M | A, P)$  представляет собой математическое ожидание количества сбоев компьютера возраста  $A$ , на котором установлено  $P$  программ. Ожидаемое общее количество сбоев компьютера,  $E(M)$ , равно взвешенному значению среднего ожидаемого количества сбоев компьютера, имеющего возраст  $A$ , с различным количеством установленных на нем программ  $P$ , взвешенному по долевому количеству компьютеров с разными комбинациями  $A$  и  $P$ .

В упражнении 2.20 представлены некоторые дополнительные свойства условного математического ожидания для нескольких случайных величин.

**Условная дисперсия.** Дисперсия  $Y$  при условии  $X$  представляет собой условную дисперсию случайной величины  $Y$  при заданном  $X$ . Таким образом, *условная дисперсия* случайной величины  $Y$  при заданном  $X$  может быть записана как

$$\text{var}(Y | X = x) = \sum_{i=1}^k [y_i - E(Y | X = x)]^2 \Pr(Y = y_i | X = x). \quad (2.21)$$

Например, условная дисперсия числа поломок компьютера при условии, что компьютер является старым, равна:  $\text{var}(M | A = 0) = (0 - 0,56)^2 \times 0,70 + (1 - 0,56)^2 \times 0,13 + (2 - 0,56)^2 \times 0,10 + (3 - 0,56)^2 \times 0,05 + (4 - 0,56)^2 \times 0,02 \approx 0,99$ . Соответственно стандартное отклонение  $M$  при условии  $A = 0$  равно  $\sqrt{0,99} = 0,99$ . Условная дисперсия  $M$  при  $A = 1$  равна дисперсии распределения вероятностей во второй строке части Б таблицы 2.3. Она равна 0,22, так что стандартное отклонение распределения вероятностей параметра  $M$  для новых компьютеров равно  $\sqrt{0,22} = 0,47$ . Ожидаемое количество сбоев компьютеров, соответствующее условным распределениям таблицы 2.3, равно 0,14 для новых компьютеров и 0,56 для старых, а соответствующие условные стандартные отклонения равны 0,47 и 0,99.

### **Независимость случайных переменных**

Две случайные величины  $Y$  и  $X$  являются *независимо распределенными* (или просто *независимыми*), если наличие информации о значении одной из них не дает никакой информации о величине значения другой. Другими словами,  $Y$  и  $X$  являются независимыми в том случае, если условное распределение  $Y$  при условии  $X$  совпадает с безусловным распределением  $Y$ . Таким образом, случайные величины  $Y$  и  $X$  независимо распределены, если для всех значений  $y$  и  $x$  выполняется:

$$\Pr(Y = y | X = x) = \Pr(Y = y) \quad (\text{независимость } X \text{ и } Y). \quad (2.22)$$

При подстановке уравнения (2.22) в уравнение (2.17) можно получить формулу независимых случайных величин в терминах их совместного распределения. То есть получаем, что если  $Y$  и  $X$  независимы, то выполняется следующее равенство:

$$\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y). \quad (2.23)$$

Таким образом, совместное распределение вероятностей двух независимых случайных величин равно произведению их безусловных распределений.

## **Ковариация и корреляция двух случайных величин**

**Ковариация.** Одной из мер степени зависимости совместного изменения двух случайных величин является их ковариация. Ковариацией случайных величин  $Y$  и  $X$  называется математическое ожидание  $E[(X - \mu_X)(Y - \mu_Y)]$ , где  $\mu_X$  – среднее значение  $X$ , а  $\mu_Y$  – среднее значение  $Y$ . Ковариацию обозначают как  $\text{cov}(Y, X)$ , или как  $\sigma_{XY}$ . Пусть  $X$  принимает  $l$  значений, а  $Y$  принимает  $k$  значений. Тогда в соответствии с определением их ковариация вычисляется по формуле:

$$\begin{aligned} \text{cov}(X, Y) = \sigma_{XY} &= E[(X - \mu_X)(Y - \mu_Y)] = \\ &= \sum_{i=1}^k \sum_{j=1}^l (x_j - \mu_X)(y_i - \mu_Y) \Pr(X = x_j, Y = y_i). \end{aligned} \quad (2.24)$$

Для интерпретации данной формулы предположим, что если рассматриваемое значение  $X$  больше своего среднего (таким образом, разность  $X - \mu_X$  положительна), то и величина  $Y$  больше своего среднего (т.е.  $Y - \mu_Y > 0$ ), а когда  $X$  меньше своего среднего (и, соответственно,  $X - \mu_X < 0$ ), то и  $Y$  меньше своего среднего ( $Y - \mu_Y < 0$ ). В обоих случаях произведение  $(X - \mu_X) \times (Y - \mu_Y)$  принимает положительные значения, что означает, что ковариация больше нуля. В противоположность сделанным предположениям, в случае если  $X$  и  $Y$  изменяются в разных направлениях ( $X$  принимает большие значения при малом  $Y$  и наоборот) ковариация будет отрицательной. В случае независимости  $X$  и  $Y$  ковариация равна нулю (упражнение 2.19).

**Корреляция.** Так как ковариация представляет собой математическое ожидание произведения отклонений  $X$  и  $Y$  от их средних значений, ее единицей измерения является произведение единиц измерения случайных величин  $X$  и  $Y$ . Поэтому в некоторых случаях качественная интерпретация значений ковариации является затруднительной.

Корреляция – альтернативная мера измерения зависимости  $X$  и  $Y$ , для которой данная проблема решена. Корреляцией случайных переменных  $X$  и  $Y$  называют величину, равную отношению их ковариаций к их стандартным отклонениям:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \quad (2.25)$$

Поскольку единицы измерения числителя в уравнении (2.25) такие же, как и единицы измерения его знаменателя, то корреляция является безразмерной величиной. Случайные величины  $X$  и  $Y$  являются некоррелированными, если  $\text{corr}(X, Y) = 0$ .

Корреляция всегда принимает значения только из отрезка  $[-1, 1]$  (приложение 2.1):

$$-1 \leq \text{corr}(X, Y) \leq 1. \quad (2.26)$$

**Корреляция и условное математическое ожидание.** Если условное среднее  $Y$  не зависит от  $X$ , то  $Y$  не коррелирована с  $X$ . Таким образом, если

$$E(Y|X) = \mu_Y, \text{ то } \text{cov}(Y, X) = 0 \text{ и } \text{corr}(Y, X) = 0. \quad (2.27)$$

Докажем последний результат. Предположим, что  $Y$  и  $X$  имеют нулевое среднее, так что  $\text{cov}(Y, X) = E[(Y - \mu_Y)(X - \mu_X)] = E(YX)$ . В соответствии с законом повторного математического ожидания [уравнение (2.20)]  $E(YX) = E[E(YX|X)] = E[E(Y|X)X] = 0$ , так как  $E(Y|X) = 0$ . Таким образом,  $\text{cov}(Y, X) = 0$ . Результат, представленный в уравнении (2.27), получается подстановкой  $\text{cov}(Y, X) = 0$  в определение корреляции из уравнения (2.25). Если средние значения  $Y$  и  $X$  не равны нулю, то данный результат получается аналогичным образом.

Однако обратное утверждение, состоящее в том, что условное среднее  $Y$  при условии  $X$  не зависит от  $X$  для некоррелированных величин  $Y$  и  $X$ , не всегда является верным. Другими словами, возможна такая ситуация, когда условное среднее  $Y$  является функцией  $X$ , но, тем не менее,  $Y$  и  $X$  не коррелированы. Подобный пример рассматривается в упражнении 2.33.

### **Математическое ожидание и дисперсия суммы случайных величин**

Математическим ожиданием суммы двух случайных величин  $X$  и  $Y$  является сумма их математических ожиданий:

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y. \quad (2.28)$$

◊ ◊ ◊

### **Распределение заработных плат в США в 2008 году**

Некоторые родители убеждают своих детей в том, что те будут работать на более хорошей и высокооплачиваемой работе, если получат высшее образование. Но действительно ли это так? Имеются ли различия в распределении заработной платы между работниками, закончившими высшее учебное заведение, и работниками, не имеющими высшего образования? Существуют ли различия среди работников с одинаковым уровнем образования или в доходах мужчин и женщин? Например, действительно ли женщина, имеющая высшее образование, зарабатывает столько же, сколько зарабатывает мужчина, имеющий такое же образование?

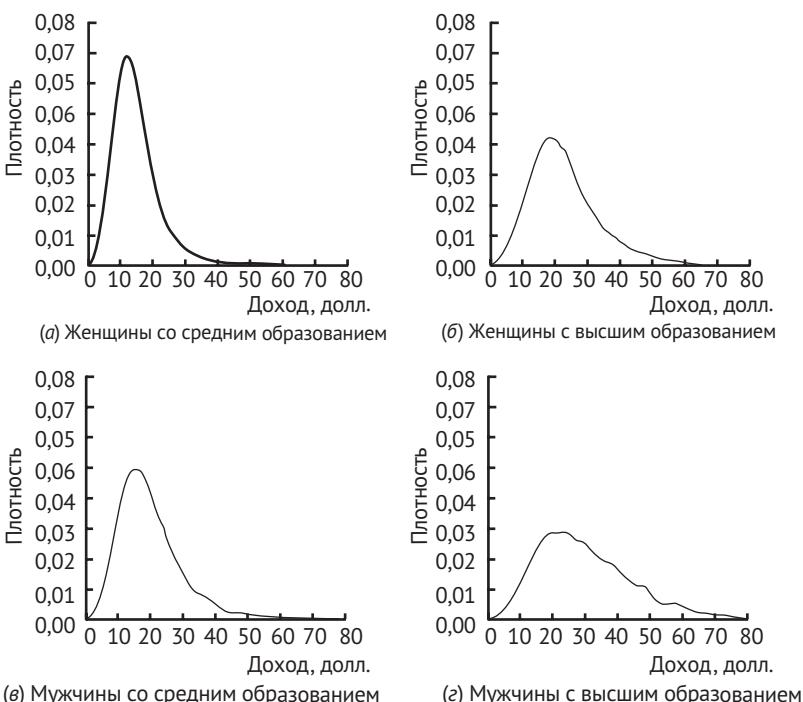
Как получить ответ на данный вопрос? Одним из способов является исследование распределения доходов занятой части населения, представители которой различаются по уровню образования (среднее или высшее) и по половому признаку. Четыре соответствующих условных распределения представлены на рисунке 2.4, а их средние значения, стандартные отклонения и некоторые процентили условных распределений даны в таблице 2.4<sup>1</sup>. Например, условным математическим ожиданием доходов среди женщин, имеющих среднее образование (т.е. величина  $E(\text{Доход}=\text{Образование}=Среднее, Пол=Женский))$  составляет 14,73 долларов в час.

Распределение почасового дохода женщин, имеющих высшее образование (рис. 2.4б), смещено вправо относительно доходов женщин, имеющих только среднее образование (рис. 2.4а). Аналогичное смещение доходов наблюдается у мужчин (рис. 2.4г и 2.4в, соответственно). Как для мужчин, так и для женщин средние доходы выше в случае, если работник имеет высшее образование (табл. 2.4, первый столбец).

## Часть I. Введение и обзор

Интересен тот факт, что разброс распределения доходов, оцененный посредством стандартного отклонения, больше в случае наличия у индивидов высшего образования. В дополнение к этому, как для мужчин так и для женщин, 90-й процентиль доходов гораздо выше у работников, имеющих высшее образование, чем у работников, окончивших только среднее учебное заведение. Эти результаты соответствуют родительским наставлениям о том, что высшее образование открывает в карьерном плане для человека больше дверей, чем среднее.

Другим свойством представленных распределений является смещение распределения доходов мужчин вправо относительно распределения доходов у женщин. Данное явление — расхождение в зарплатах между мужчинами и женщинами (gender gap) — является очень важным и неприятным аспектом распределения доходов. Мы вернемся к этой теме в следующих главах.



**Рисунок 2.4. Условное распределение среднего почасового дохода занятого населения в США в 2008 году с учетом уровня образования и пола**

На графике представлены распределения доходов мужчин и женщин, имеющих среднее (распределения а и в) и высшее образование (б и г).

**Таблица 2.4**

### Основные характеристики условного распределения средних почасовых доходов занятого населения в США в 2008 году с учетом уровня образования и пола

	Среднее	Стандартное отклонение	Процентили			
			25 %	50 % (медиана)	75 %	90 %
(а) Женщины со средним образованием	14,73 долл.	7,77 долл.	9,62 долл.	13,19 долл.	17,50 долл.	23,96 долл.

Окончание таблицы 2.4

	Среднее	Стандартное отклонение	Процентили			
			25 %	50 % (медиана)	75 %	90 %
(б) Женщины с высшим образованием	23,93	12,59	15,38	21,63	28,85	39,42
(в) Мужчины со средним образованием	19,64	10,21	12,50	17,48	24,04	32,69
(г) Мужчины с высшим образованием	30,97	16,08	19,23	28,45	39,34	52,88

*Примечание.* Средний почасовой доход представляет собой сумму заработной платы (служащих или рабочих) до учета налогов, бонусов, премий. Распределения были получены на основе данных текущего обследования населения (Current Population Survey) за март 2009 года, описанного в приложении 3.1.



### Среднее, дисперсия и ковариация суммы случайных величин

Пусть  $X$ ,  $Y$  и  $V$  – случайные величины и пусть  $\mu_X$  и  $\sigma_X^2$  – среднее и дисперсия  $X$ , а  $\sigma_{XY}$  – ковариация случайных величин  $X$  и  $Y$  (для остальных случайных величин приняты аналогичные обозначения). Пусть  $a$ ,  $b$  и  $c$  – константы. Приведенные ниже соотношения (2.29) – (2.35) следуют из определений среднего, дисперсии и ковариации:

$$E(a+bX+cY) = a + b\mu_X + c\mu_Y, \quad (2.29)$$

$$\text{var}(a+bY) = b^2\sigma_Y^2, \quad (2.30)$$

$$\text{var}(aX+bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \quad (2.31)$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \quad (2.32)$$

$$\text{cov}(a+bX+cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \quad (2.33)$$

$$E(XY) = \sigma_{XY} + \mu_X \mu_Y, \quad (2.34)$$

$$|\text{corr}(X, Y)| \leq 1 \text{ и } |\sigma_{XY}| \leq \sqrt{\sigma_X^2 \sigma_Y^2} \\ (\text{корреляционное неравенство}). \quad (2.35)$$

**ОСНОВНЫЕ ПОНЯТИЯ**  
**2.3**

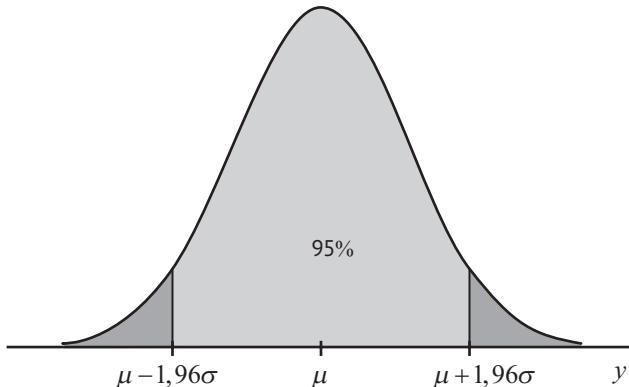
Дисперсия суммы  $X$  и  $Y$  есть сумма их дисперсий и удвоенной ковариации:

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \quad (2.36)$$

Если  $X$  и  $Y$  являются независимыми случайными величинами, то их ковариация равна нулю, а дисперсия суммы независимых случайных величин равна сумме их дисперсий:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) = \sigma_X^2 + \sigma_Y^2 \quad (X \text{ и } Y \text{ независимы}). \quad (2.37)$$

Во вставке «Основные понятия 2.3» приведены наиболее часто используемые выражения для среднего, дисперсии и ковариации суммы случайных величин. Вывод из данных выражений приведен в приложении 2.1.



**Рисунок 2.5. Плотность нормального распределения**

Рисунок функции плотности нормального распределения с математическим ожиданием  $\mu$  и дисперсией  $\sigma^2$  представляет собой колоколообразную кривую с центром в точке  $\mu$ . Площадь под данной кривой, ограниченная интервалом  $\mu - 1,96\sigma$  и  $\mu + 1,96\sigma$ , равна 0,95. Нормальное распределение обозначается как  $N(\mu, \sigma^2)$ .

## 2.4. Нормальное распределение, распределение хи-квадрат, распределения Стьюдента и Фишера

В экономике наиболее часто встречаются распределениями вероятностей случайных величин являются нормальное распределение, распределение хи-квадрат ( $\chi_m^2$ ),  $t$ -распределение Стьюдента и  $F$ -распределение Фишера.

### Нормальное распределение

График функции непрерывной *нормально распределенной* случайной величины имеет колоколообразную форму (рис. 2.5). Функция, определяющая нормальное распределение вероятностей, приведена в приложении 17.1. Как видно на рисунке 2.5, график функции плотности нормального распределения с математическим ожиданием  $\mu$  и дисперсией  $\sigma^2$  симметричен относительно своего среднего, а 95 % его площади заключено в интервале между  $\mu - 1,96\sigma$  и  $\mu + 1,96\sigma$ .

Для описания некоторых типов нормального распределения введены специальные термины и обозначения. Так, например, нормальное распределение со средним  $\mu$  и дисперсией  $\sigma^2$  обозначается как  $N(\mu, \sigma^2)$ . *Стандартным нормальным распределением* называют нормальное распределение, у которого  $\mu = 0$  и  $\sigma^2 = 1$ . Соответственно оно обозначается как  $N(0, 1)$ . Часто случайную

величину, имеющую распределение вероятностей  $N(0, 1)$ , обозначают через  $Z$ , а стандартную нормальную кумулятивную функцию распределения обычно обозначают через  $\Phi$ . В введенных обозначениях получаем:  $\Pr(Z \leq c) = \Phi(c)$ , где  $c$  — константа. Значения, которые принимает функция распределения стандартной нормальной случайной величины, приведены в таблице 1 приложения.

Для определения значений вероятностей нормально распределенной случайной величины приведем ее к *стандартизированному виду*. Для этого вычтем из нее ее среднее значение и полученную разность разделим на ее стандартное отклонение. Например, предположим, что случайная величина  $Y$  распределена в соответствии с законом  $N(1, 4)$ . Какова вероятность того, что  $Y \leq 2$ ? Данный случай изображен на рисунке 2.6а. Стандартизированное значение случайной величины  $Y$  определяется как  $(Y - 1) / \sqrt{4} = \frac{1}{2}(Y - 1)$ . Соответственно, в отличие от  $Y$ , случайная величина  $\frac{1}{2}(Y - 1)$  распределена по стандартному нормальному закону распределения и имеет нулевое среднее и единичную дисперсию (упражнение 2.8). График функции плотности случайной величины  $\frac{1}{2}(Y - 1)$  изображен на рисунке 2.6б. Теперь выражение  $Y \leq 2$  эквивалентно неравенству  $\frac{1}{2}(Y - 1) \leq \frac{1}{2}(2 - 1)$ , так что  $\frac{1}{2}(Y - 1) \leq \frac{1}{2}$ . Таким образом,

$$\Pr(Y \leq 2) = \Pr\left[\frac{1}{2}(Y - 1) \leq \frac{1}{2}\right] = \Pr\left(Z \leq \frac{1}{2}\right) = \Phi(0,5) = 0,691, \quad (2.41)$$

где значение 0,691 взято из таблицы 1 приложения.

### Вычисление вероятностей принятия определенных значений нормально распределенной случайной величиной

Предположим, что случайная величина  $Y$  распределена нормально со средним значением  $\mu$  и дисперсией  $\sigma^2$ , то есть  $Y$  имеет распределение  $N(\mu, \sigma^2)$ . Тогда стандартизированным значением случайной величины  $Y$  будет случайная величина  $Z = (Y - \mu) / \sigma$ .

Пусть через  $c_1$  и  $c_2$  — константы, для которых верно, что  $c_1 < c_2$ . И пусть  $d_1 = (c_1 - \mu) / \sigma$  и  $d_2 = (c_2 - \mu) / \sigma$ . Тогда

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2), \quad (2.38)$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1), \quad (2.39)$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \quad (2.40)$$

Значения функции  $\Phi$  приведены в таблице 1 приложения.

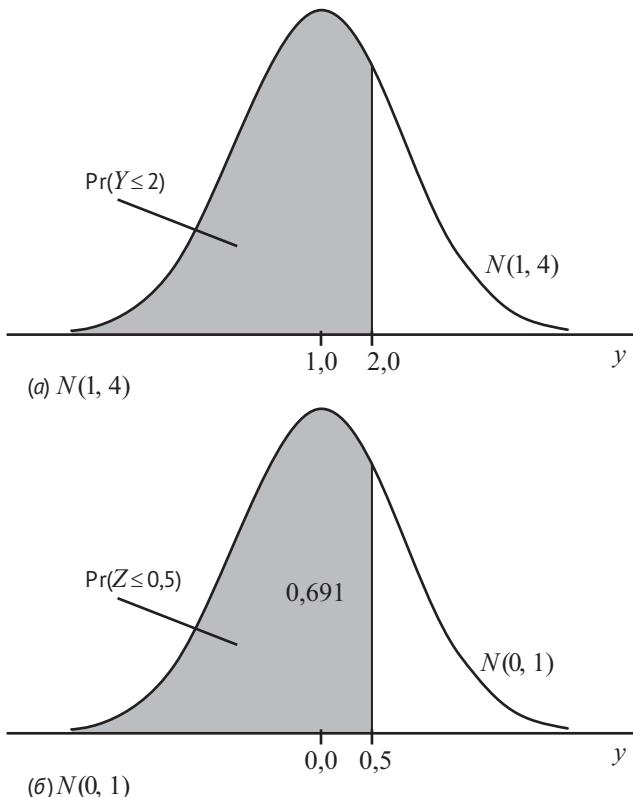
## ОСНОВНЫЕ ПОНЯТИЯ

### 2.4

Аналогичным образом можно вычислить вероятности превышения нормально распределенной случайной величиной некоторого определенного значения или вероятность того, что ее значение попадает в заданный интервал. Во вставке «Основные понятия 2.4» описано, каким образом это можно сделать.

В дополнительной вставке «Черный день Уолл-Стрит» показан необычный способ использования нормального распределения.

Нормальное распределение является симметричным, так что его асимметрия равна нулю. Эксцесс нормального распределения равен 3.



**Рисунок 2.6. Вычисление вероятности события  $Y \leq 2$  для случайной величины  $Y$ , распределенной как  $N(1, 4)$**

Чтобы вычислить вероятность того, что  $Y \leq 2$ , стандартизируем  $Y$ , а затем воспользуемся таблицей значений стандартного нормального распределения.  $Y$  приводится к стандартному виду вычитанием его среднего значения ( $\mu = 1$ ) и делением полученной разности на его стандартное отклонение ( $\sigma = 2$ ). Графически вероятность того, что  $Y \leq 2$ , изображена на рисунке 2.6а, а соответствующая вероятность для стандартизированного значения — на рисунке 2.6б. В силу того что стандартизированная случайная величина  $(Y-1)/2$  является случайной величиной, распределенной по стандартному нормальному закону распределения ( $Z$ ), получаем, что  $\Pr(Y \leq 2) = \Pr\left(\frac{Y-1}{2} \leq \frac{2-1}{2}\right) = \Pr(Z \leq 0,5)$ . Используя таблицу 1 приложения, имеем  $\Pr(Z \leq 0,5) = \Phi(0,5) = 0,691$ .

**Многомерное нормальное распределение.** Нормальное распределение может быть расширено на случай многомерной случайной величины. Такое распределение называется *многомерным нормальным распределением*, а в случае если рассматриваются только две случайные переменные — *двумерным нормальным распределением*. Формула функции плотности двумерного нормального распределения приведена в приложении 17.1, а для случая многомерного нормального распределения — в приложении 18.1.

Многомерное нормальное распределение имеет четыре важных свойства. Пусть случайные величины  $X$  и  $Y$  распределены по двумерному нормальному закону распределения, их ковариация равна  $\sigma_{XY}$ , а  $a$  и  $b$  — две константы. Тогда сумма

$aX + bY$  распределена по (одномерному) нормальному закону распределения. То есть  $aX + bY$  имеет распределение:

$$N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}), \quad (2.42)$$

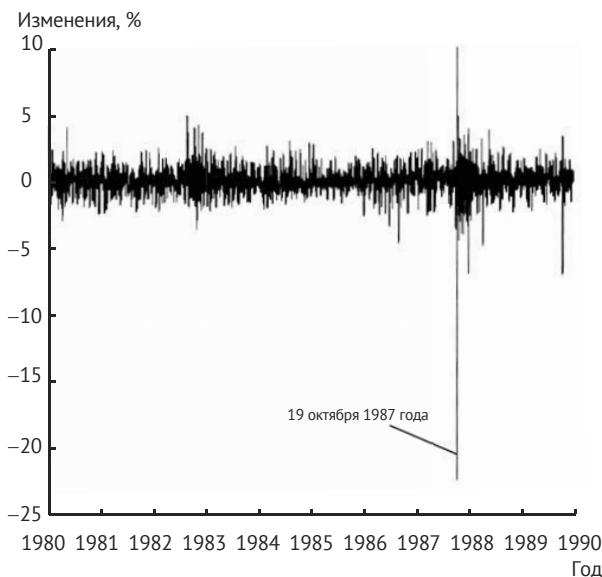
где  $X$  и  $Y$  – двумерная нормальная случайная величина.

◇ ◇ ◇

### **Черный день Уолл-Стрит**

В обычный торговый день стоимость акций, торгуемых на американской фондовой бирже, может как вырасти, так и упасть примерно на 1% или даже больше. Это много, но ничто в сравнении с тем, что произошло на бирже в «Черный понедельник», 19 октября 1987 года. В этот день индекс Доу–Джонса (индекс средней стоимости акций 30 крупнейших американских промышленных предприятий) упал на 22,6%! Для справки: с 1 января 1980 по 31 декабря 2009 года стандартное отклонение дневного процентного изменения индекса составило 1,13%, падение же на 22,6% составляет  $20 (=22,6/1,13)$  стандартных отклонений. Масштаб такого падения изображен на рисунке 2.7, на котором приведена динамика индекса в 1980-х годах.

Если считать, что процентные изменения значений индекса имеют нормальное распределение, то вероятность 20-кратного падения значения индекса равнялась бы  $\Pr(|Z| \geq 20) = 2 \times \Phi(-20)$ . Вы не найдете данного значения функции плотности в таблице 1 приложения, но его можно вычислить, используя компьютер. Попробуйте! Вы получите, что вероятность такого события равна  $5,5 \times 10^{-89}$ , что в десятичной форме равно 0,000...00055, где в общей сложности 88 нулей!



**Рисунок 2.7. Динамика процентных изменений индекса Доу–Джонса в 1980-х годах, ежедневные данные**

В течение 1980-х годов среднее дневное отклонение индекса Доу–Джонса составляло 0,05%, а его стандартное отклонение равнялось 1,16%. 19 октября 1987 года, в «Черный понедельник», индекс обвалился на 22,6%, что составляет 20 стандартных отклонений.

Насколько мало число  $5,5 \times 10^{-89}$ ? Рассмотрим несколько примеров:

- Население Земли, в грубом приближении, составляет 7 млрд человек, так что вероятность выиграть в лотерее, в которой примут участие сразу все люди, равна  $1,4 \times 10^{-10}$  (одна семимиллионная).
- Согласно исследованиям, наша Вселенная существует на протяжении 14 миллиардов лет, что равно  $5 \times 10^{17}$  секундам, так что вероятность наступления каждой секунды равна  $2 \times 10^{-18}$ .
- В атмосфере содержится примерно  $10^{43}$  молекул воздуха в пределах 1 км от земли. Соответственно вероятность случайного выбора одной из них составляет  $10^{-43}$ .

Несмотря на то что тот день был действительно «черным днем» для Уолл-Стрит, сам факт его наступления говорит о том, что вероятность его возникновения была все же больше, чем  $5,5 \times 10^{-89}$ . В действительности было много плохих и хороших для фондового рынка дней, в которые изменение индекса было достаточно большим по сравнению со стандартным отклонением. В таблице 2.5 приведено 10 самых значительных изменений значений индекса в течение 7571 торгового дня за период с 1 января 1980 по 31 декабря 2009 года со стандартизованными значениями их изменений. Все рассмотренные изменения индекса превышали 6,4 стандартных отклонения, что является очень маловероятным событием даже в случае нормальности распределения цен на акции, входивших в индекс.

Говоря научным языком, распределение процентных изменений значения индекса имеет более тяжелые хвосты по сравнению с нормальным. По этой причине профессиональные финансисты используют другие методы моделирования цен на акции. Один из них определяет изменение стоимости акций как нормально распределенные величины с изменяющейся во времени дисперсией. В данном случае на периоды, подобные октябрю 1987 года, и финансовые кризисы, подобные осени 2008 года, приходятся более высокие значения дисперсии (модели с непостоянными дисперсиями рассматриваются в главе 16). Другие методы отказываются от использования нормального распределения в пользу распределений с более тяжелыми хвостами, как, например, в книге Насима Талеба «Черный лебедь» (2007) [Nassim Taleb, 2007]. Подобные модели более состоятельны в описании очень плохих и очень хороших дней на Уолл-Стрит.

Таблица 2.5

**Десять крупнейших дневных изменений значения индекса Доу–Джонса  
в 1980–2009 годах и их вероятность в соответствии с нормальным  
распределением**

Дата	Процентное изменение ( $x$ )	Стандартизированная величина отклонения $z = (x - \mu) / \sigma$	Вероятность подобного изменения $\Pr( Z  \geq z) = 2\Phi(-z)$
19 октября 1987 г.	-22,6	-20,1	$5,5 \times 10^{-89}$
13 октября 2008 г.	11,1	9,8	$1,1 \times 10^{-22}$
28 октября 2008 г.	10,9	9,6	$6,5 \times 10^{-22}$
21 октября 1987 г.	10,1	9,0	$2,8 \times 10^{-19}$
26 октября 1987 г.	-8,0	-7,2	$7,4 \times 10^{-13}$

Окончание таблицы 2.5

Дата	Процентное изменение ( $x$ )	Стандартизированная величина отклонения $z = (x - \mu) / \sigma$	Вероятность подобного изменения $\Pr( Z  \geq z) = 2\Phi(-z)$
15 октября 2008 г.	-7,9	-7,0	$2,1 \times 10^{-12}$
1 декабря 2008 г.	-7,7	-6,9	$6,3 \times 10^{-12}$
9 октября 2008 г.	-7,3	-6,5	$5,9 \times 10^{-11}$
27 октября 1997 г.	-7,2	-6,4	$1,4 \times 10^{-10}$
17 сентября 2001 г.	-7,1	-6,4	$2,0 \times 10^{-10}$

◊ ◊ ◊

В целом, если совместное распределение  $n$  случайных величин является многомерным нормальным, то их любая линейная комбинация (например их сумма) также является нормальным распределением.

Во-вторых, если несколько случайных величин распределены по многомерному нормальному закону, то безусловное распределение каждой из этих случайных величин является нормальным [данное утверждение следует из уравнения (2.42) при подстановке в него значений  $a = 1$  и  $b = 0$ ].

В-третьих, если ковариация нескольких случайных переменных, распределенных по многомерномуциальному нормальному закону, равна нулю, то эти переменные являются независимыми. Так, например, если  $X$  и  $Y$  имеют двумерное нормальное распределение и  $\sigma_{XY} = 0$ , то случайные величины  $X$  и  $Y$  являются независимыми. В разделе 2.3 было определено, что если случайные величины  $X$  и  $Y$  независимы, то, несмотря на вид их совместного распределения,  $\sigma_{XY} = 0$ . Если  $X$  и  $Y$  имеют совместное нормальное распределение, то обратное также верно. Данный результат – равенство нулю ковариации независимых переменных – является особенным свойством многомерного нормального распределения, которое не всегда выполняется в общем случае.

И наконец, если случайные величины  $X$  и  $Y$  распределены по двумерному нормальному закону, то условное математическое ожидание  $Y$  при заданном  $X$  линейно по  $X$ . Математически данное свойство выражается следующим образом:  $E(Y | X = x) = a + bx$ , где  $a$  и  $b$  – константы (упражнение 17.11). Совместная нормальность распределения предполагает линейность условных математических ожиданий совместно распределенных случайных величин, но обратное свойство неверно.

### Распределение хи-квадрат

Для проверки разного рода гипотез в статистике и эконометрике часто используется распределение хи-квадрат ( $\chi^2$ -квадрат).

Распределение хи-квадрат представляет собой распределение вероятностей суммы  $m$  квадратов случайных величин, имеющих стандартное нормальное распределение. Соответственно, рассматриваемое распределение зависит от  $m$  – степени свободы распределения хи-квадрат. Например, пусть  $Z_1$ ,  $Z_2$  и  $Z_3$  – независимые случайные величины, распределенные по стандартному

нормальному закону. Тогда случайная величина  $Z_1^2 + Z_2^2 + Z_3^2$  распределена по хи-квадрат распределению с тремя степенями свободы. Название этого распределения происходит от греческой буквы хи ( $\chi$ ), которая обычно используется для его обозначения. Стандартно используемое обозначение распределения хи-квадрат с  $m$  степенями свободы —  $\chi_m^2$ .

Некоторые процентили для распределения  $\chi_m^2$  приведены в таблице 3 приложения. Например, 95-й процентиль для  $\chi_3^2$  равен 7,81, так что  $\Pr(Z_1^2 + Z_2^2 + Z_3^2 \leq 7,81) = 0,95$ .

### *t-распределение Стьюдента*

*t-распределением Стьюдента с  $m$  степенями свободы* называется распределение случайной величины, которая является отношением случайной величины, распределенной по стандартному нормальному закону распределения, к квадратному корню независимой случайной величины, распределенной по закону хи-квадрат с  $m$  степенями свободы, деленной на  $m$ . Пусть  $Z$  — стандартная нормальная случайная величина,  $W$  — случайная переменная, распределенная по закону хи-квадрат с  $m$  степенями свободы, и пусть  $Z$  и  $W$  независимы. В таком случае случайная величина, определяемая как  $Z / \sqrt{W/m}$ , имеет *t-распределение Стьюдента* (часто его называют просто *t-распределение*) с  $m$  степенями свободы. Распределение Стьюдента обычно обозначают  $t_m$ . В таблице 2 приложения приведены некоторые процентили для распределения Стьюдента.

*t-распределение Стьюдента* зависит от степеней свободы  $m$ . График данного распределения имеет колоколообразную форму, однако, в отличие от нормального распределения, при малом количестве степеней свободы  $m$  его график имеет более тяжелые хвосты. При  $m=30$  график *t-распределения Стьюдента* хорошо аппроксимируется стандартным нормальным распределением, а распределение  $t_\infty$  совпадает со стандартным нормальным распределением.

### *F-распределение*

*F-распределением Фишера* со степенями свободы  $m$  и  $n$ , которое принято обозначать  $F_{m,n}$ , называется распределение, вычисляемое как отношение значения случайной величины, распределенной в соответствии с законом хи-квадрат с  $m$  степенями свободы, деленного на  $m$ , к значению независимой случайной величины, имеющей распределение хи-квадрат с  $n$  степенями свободы, деленного на  $n$ . Для математической интерпретации сказанного обозначим через  $W$  случайную величину, распределенную по закону хи-квадрат, с  $m$  степенями свободы, а через  $V$  — случайную величину, имеющую распределение хи-квадрат с  $n$  степенями свободы. Тогда случайная величина  $\frac{W/m}{V/n}$  имеет

*F-распределение  $F_{m,n}$* .

В статистике и эконометрике наиболее важными являются случаи с большими значениями степеней свободы *F-распределения* в знаменателе. Аппрокси-

мацией здесь выступает распределение  $F_{m, \infty}$ . В данном случае  $V$  является математическим ожиданием бесконечно большого числа случайных величин, имеющих распределение хи-квадрат, и это математическое ожидание равно 1 в силу равенства единице среднего квадрата случайной величины, распределенной по стандартному нормальному закону распределения (упражнение 2.24). Таким образом, распределение  $F_{m, \infty}$  является распределением хи-квадрат случайной величины с  $m$  степенями свободы, деленной на  $m$ , т.е.  $W/m$  имеет распределение  $F_{m, \infty}$ . Например, согласно таблице 4 приложения, 95-й процентиль распределения  $F_{3, \infty}$  равен 2,60, что в точности равно значению 95-го процентиля распределения  $\chi^2_3$ , 7,81 (таблица 2 приложения), деленного на количество 3 – степеней свободы ( $7,81/3=2,60$ ).

90, 95 и 99-й процентили распределения  $F_{m, n}$  приведены в таблице 5 приложения для некоторых значений  $m$  и  $n$ . Например, 95-й процентиль распределения  $F_{3,30}$  равен 2,92, а 95-й процентиль  $F_{3,90}$  равен 2,71. При возрастании числа степеней свободы  $n$  в знаменателе 95-й процентиль распределения  $F_{3, n}$  стремится к предельному значению 2,60 распределения  $F_{3, \infty}$ .

## 2.5. Случайная выборка и распределение выборочного среднего

Почти все статистические и эконометрические процедуры, используемые в книге, опираются на значение выборочного среднего или взвешенного выборочного среднего. В силу этого характеристика распределения выборочного среднего (среднего по выборке) играет важную роль в понимании процесса реализации эконометрических процедур.

В данном разделе рассмотрены базовые понятия о выборках и распределениях выборочных средних. Мы начнем обсуждение этой темы с определения случайной выборки. Случайная выборка представляет собой совокупность значений, выбранных случайным образом из некоторого распределения (генеральной совокупности). Таким образом, выборочное среднее является случайной величиной. В конце раздела рассмотрены свойства случайной выборки и выборочного среднего.

### Случайная выборка

**Простая случайная выборка.** Предположим, что наша студентка (см. пример из раздела 2.1) мечтает стать статистиком и решила проанализировать свои поездки из дома в университет. В некоторые дни она записывает время, потраченное на дорогу, причем эти дни выбираются случайным образом, а записываемое ею время имеет кумулятивную функцию распределения, изображенную на рисунке 2.2а. В силу случайности выбора дней запись, сделанная в конкретный день, не дает нам никакой информации о значении времени, потраченного студенткой на дорогу, в какой-либо другой день. То есть время, потраченное на дорогу, является независимо распределенной случайной величиной.

Ситуация, описанная выше, является примером простейшей схемы выбора, используемой в статистике, – *простой случайной выборки*. В ней  $n$  объектов выбраны случайным образом из общей генеральной совокупности (в данном примере генеральная совокупность – это множество дней, в течение которых студентка посещала университет), причем каждый элемент генеральной совокупности (каждый день в рассматриваемом случае) может быть включен в выборку с такой же вероятностью, что и любой другой элемент (день) генеральной совокупности.

$n$  наблюдений в выборке обозначаются  $Y_1, \dots, Y_n$ , где  $Y_1$  – первое наблюдение,  $Y_2$  – второе и так далее. В нашем примере  $Y_1$  – время, потраченное студентом на дорогу в первый из  $n$  дней выборки, а  $Y_i$ , соответственно, в  $i$ -й день.

Так как элементы выборки отобраны из генеральной совокупности случайным образом, значения, которые принимают  $Y_1, \dots, Y_n$ , также являются случайными. Если в выборку включить другие элементы генеральной совокупности, то и их значения будут другими. Таким образом, элементы случайной выборки  $Y_1, \dots, Y_n$  можно считать случайными величинами. До момента формирования выборки  $Y_1, \dots, Y_n$  могут принимать множество других возможных значений из генеральной совокупности. Но после того как выборка сформирована, каждому ее элементу ставится в соответствие определенное значение.

**ОСНОВНЫЕ  
ПОНЯТИЯ**

2.5

**Простая случайная выборка и независимые  
одинаково распределенные случайные величины (i.i.d.)**

В простой случайной выборке  $n$  объектов отбираются случайным образом из генеральной совокупности, и каждый элемент может быть выбран с одинаковой вероятностью. Значение  $i$ -го объекта выборки обозначается как  $Y_i$ . Так как каждый элемент генеральной совокупности с одинаковой вероятностью может стать элементом случайной выборки, а распределение  $Y_i$  одинаково для всех  $i$ , то случайные величины  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными (i.i.d.). Это означает, что распределение  $Y_i$  одинаково для всех  $i = 1, \dots, n$  и  $Y_1$  не зависит от  $Y_2, \dots, Y_n$  и так далее.

**Независимые одинаково распределенные (i.i.d.) случайные величины.** Так как  $Y_1, \dots, Y_n$  выбираются случайным образом из одной и той же генеральной совокупности, безусловное распределение вероятностей случайной величины  $Y_i$  одинаково для каждого  $i = 1, \dots, n$ . Это безусловное распределение есть распределение  $Y$  в генеральной совокупности, из которой формируется выборка. Когда  $Y_i$  имеют одинаковые безусловные распределения для всех  $i = 1, \dots, n$ , случайные величины  $Y_1, \dots, Y_n$  называются *одинаково распределенными случайными величинами*.

В случае простой случайной выборки знание значения  $Y_1$  не дает никакой информации о значении  $Y_2$ . Другими словами, в простой случайной выборке  $Y_1$  не зависит от  $Y_2, \dots, Y_n$ .

Если  $Y_1, \dots, Y_n$  выбраны из одного распределения и являются независимыми, то их называют *независимыми одинаково распределенными случайными величинами*.

Понятия простой случайной выборки и независимых одинаково распределенных случайных величин приведены во вставке «Основные понятия 2.5».

### **Выборочное распределение выборочного среднего**

*Выборочным средним значением (sample average) или средним по выборке (sample mean),* которое принято обозначать  $\bar{Y}$ ,  $n$  наблюдений  $Y_1, \dots, Y_n$  называется величина, определяемая как

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \dots + Y_n) = \frac{1}{n} \sum_{i=1}^n Y_i. \quad (2.43)$$

Важным моментом в статистике и эконометрике является тот факт, что при формировании случайной выборки из генеральной совокупности выборочное среднее  $\bar{Y}$  также является случайной величиной. В силу случайности выбора элементов выборки каждый из элементов  $Y_i$  является случайной величиной. Получаем, что все  $Y_1, \dots, Y_n$  являются случайными величинами. Соответственно, их среднее – тоже случайная величина. В случае выбора других значений из генеральной совокупности среднее значение новой выборки, вообще говоря, будет другим. Таким образом, выборочное среднее значение  $\bar{Y}$  меняется от одной случайной выборки к другой.

Рассмотрим пример. Допустим, рассматриваемая нами ранее студентка фиксирует время, которое она тратила на дорогу до университета, в течение пяти случайно выбранных дней, а после этого считает среднее по полученной выборке. Значение этого среднего зависит от того, в какие именно дни делала свои записи студентка.

Так как величина выборочного среднего  $\bar{Y}$  является случайной, то она также имеет свое вероятностное распределение. Распределение  $\bar{Y}$  называется *выборочным распределением  $\bar{Y}$* , так как оно представляет собой распределение вероятностей возможного значения случайной величины  $\bar{Y}$ , которое может быть вычислено для различных случайных выборок  $Y_1, \dots, Y_n$ .

В статистике и эконометрике выборочные распределения среднего и взвешенного среднего являются одними из самых важных понятий. Мы начнем наше обсуждение свойств выборочного среднего  $\bar{Y}$  с вычисления его среднего и дисперсии в терминах соответствующих величин распределения генеральной совокупности  $Y$ .

**Математическое ожидание и дисперсия выборочного среднего  $\bar{Y}$ .** Допустим, наблюдаемые значения  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами. Пусть  $\mu_Y$  и  $\sigma_Y^2$  – соответственно среднее

и дисперсия  $Y_i$  (в силу того что наблюдения независимо одинаково распределены, их средние значения и дисперсии одинаковы для всех  $i = 1, \dots, n$ ). При  $n=2$  среднее значение суммы  $Y_1 + Y_2$  находится из выражения (2.28):  $E(Y_1 + Y_2) = \mu_Y + \mu_Y = 2\mu_Y$ . Таким образом, математическое ожидание выборочного среднего равно  $E[\frac{1}{2}(Y_1 + Y_2)] = \frac{1}{2} \times 2\mu_Y = \mu_Y$ . В общем случае имеем:

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \mu_Y. \quad (2.44)$$

Дисперсия  $\bar{Y}$  находится из выражения (2.37). Например, для  $n=2$   $\text{var}(Y_1 + Y_2) = 2\sigma_Y^2$ . Следовательно, используя выражение (2.31) при  $a=b=\frac{1}{2}$  и тот факт, что для нашей выборки  $\text{cov}(Y_1, Y_2) = 0$ , получаем, что  $\text{var}(\bar{Y}) = \frac{1}{2}\sigma_Y^2$ . В общем случае для  $n$  наблюдений, имеющих i.i.d. распределение,  $Y_i$  и  $Y_j$  независимы для  $i \neq j$ , так что  $\text{cov}(Y_i, Y_j) = 0$ . Следовательно,

$$\begin{aligned} \text{var}(\bar{Y}) &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(Y_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{cov}(Y_i, Y_j) = \frac{\sigma_Y^2}{n}. \end{aligned} \quad (2.45)$$

Стандартным отклонением случайной величины  $\bar{Y}$  является квадратный корень из ее дисперсии, то есть  $\sigma_Y / \sqrt{n}$ .

◇ ◇ ◇

### ***Финансовая диверсификация и портфельные инвестиции***

Принцип диверсификации гласит, что можно снизить инвестиционный риск путем создания портфеля из множества небольших инвестиций в разные активы. Риск у полученного портфеля будет значительно ниже, чем риск одной крупной инвестиции. Этот принцип означает, что не нужно «класть все яйца в одну корзину».

Математически принцип диверсификации следует из уравнения (2.45). Предположим, мы инвестируем 1 доллар в  $n$  активов равными частями. Пусть  $Y_i$  – доход за один год от инвестиции в  $i$ -й актив. Поскольку инвестиции в один актив составляют  $1/n$  долларов, доходность всего портфеля составит  $(Y_1 + Y_2 + \dots + Y_n) / n = \bar{Y}$ . Для простоты допустим, что каждый актив имеет одинаковую ожидаемую доходность,  $\mu_Y$ , одинаковую дисперсию  $\sigma^2$  и одинаковую положительную корреляцию с другими активами  $\rho$ , так что  $\text{cov}(Y_i, Y_j) = \rho\sigma^2$ . Ожидаемая доходность портфеля в этом случае равна  $E(\bar{Y}) = \mu_Y$ , и для больших значений  $n$  дисперсия портфеля равна  $\text{var}(\bar{Y}) = \rho\sigma^2$  (упражнение 2.26). Инвестирование всех денег в один актив или их равномерная диверсификация среди  $n$  активов не влияет на величину ожидаемой доходности портфеля, но в случае диверсификации риск от инвестиционной деятельности снижается с  $\sigma^2$  до  $\rho\sigma^2$ .

Математические принципы финансовой диверсификации лежат в основе существования таких финансовых продуктов, как паевые фонды. Вкладывая свои средства в паевые фонды, инвестор владеет долей фонда, в управлении которого находится большое количество активов. Таким образом, инвестор владеет небольшой долей многих активов, находящихся под управлением фонда. Но диверсификация имеет свои пределы: доходность одного актива часто коррелирует с доходностью других активов, так что  $\text{var}(\bar{Y})$  остается положительной даже для очень большого количества активов  $n$ . В случае с акциями риск инвестиционной деятельности можно снизить путем создания портфеля акций, но данный портфель все равно подвержен непредсказуемым флюктуациям всего рынка акций.



В заключение выпишем формулы для математического ожидания, дисперсии и стандартного отклонения выборочного среднего случайной величины  $\bar{Y}$  еще раз:

$$E(\bar{Y}) = \mu_Y, \quad (2.46)$$

$$\text{var}(\bar{Y}) = \sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}, \quad (2.47)$$

$$\text{и std.dev}(\bar{Y}) = \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}. \quad (2.48)$$

Эти результаты верны для любых распределений  $Y_i$ . То есть распределение случайной величины  $Y_i$  не должно иметь какой-то специальный вид, например, подчиняться нормальному закону распределения, для того чтобы выполнялись выражения (2.46) – (2.48).

Выражение  $\sigma_{\bar{Y}}^2$  обозначает дисперсию выборочного среднего  $\bar{Y}$  всей рассматриваемой выборки. В свою очередь  $\sigma_Y^2$  представляет собой дисперсию каждого отдельного значения  $Y_i$  в выборке, то есть это дисперсия генеральной совокупности, из которой выбирается наблюдение. Аналогично,  $\sigma_{\bar{Y}}$  обозначает стандартное отклонение выборочного среднего  $\bar{Y}$ .

*Выборочное распределение  $\bar{Y}$  для нормально распределенной случайной величины.* Пусть  $Y_1, \dots, Y_n$  – независимые одинаково распределенные случайные величины из нормального распределения  $N(\mu_Y, \sigma_Y^2)$ . Как следует из уравнения (2.42), сумма  $n$  нормально распределенных случайных величин также нормально распределена. Так как математическое ожидание  $\bar{Y}$  равно  $\mu_Y$ , а дисперсия  $\bar{Y}$  равна  $\sigma_Y^2 / n$ , то в случае когда  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами из нормального распределения  $N(\mu_Y, \sigma_Y^2)$ ,  $\bar{Y}$  имеет распределение  $N(\mu_Y, \sigma_Y^2 / n)$ .

## 2.6. Асимптотические распределения

Выборочные распределения играют ключевую роль в разработке статистических и эконометрических процедур. Поэтому очень важно знать, какой вид имеет выборочное распределение  $\bar{Y}$ . Существует два подхода к определению выборочных распределений: «точный» и «приблизительный».

Результатом точного подхода является вывод формулы плотности рассматриваемого выборочного распределения, верной для любого  $n$ . Распределение, которое точно описывает распределение случайной величины  $\bar{Y}$ , называется *точным* или *распределением в конечных выборках*. Например, если  $Y$  имеет нормальное распределение, а  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами, то (как было показано в разделе 2.5) точным распределением  $\bar{Y}$  является нормальное распределение со средним  $\mu_Y$  и дисперсией  $\sigma_Y^2/n$ . К сожалению, если распределение  $Y$  не является нормальным, то точное распределение  $\bar{Y}$  является очень сложным и зависит от распределения  $Y$ .

При использовании приблизительного метода рассматриваются распределения, аппроксимирующие большую по объему выборку. Аппроксимацию распределения случайной выборки при помощи больших выборок часто называют *асимптотическим распределением*, поскольку такие распределения становятся точными при  $n \rightarrow \infty$ . Как мы увидим в данном разделе, подобные аппроксимации могут очень точно описывать истинное распределение даже для выборки, состоящей из  $n=30$  наблюдений. Так как в эконометрике объемы выборок обычно исчисляются сотнями или тысячами, то асимптотические распределения оказываются очень близкими к точным распределениям.

В данном разделе рассматриваются два основных инструмента, которые необходимы для аппроксимации точных распределений: закон больших чисел Чебышева и центральная предельная теорема. По закону больших чисел для случая больших выборок значение  $\bar{Y}$  с большой вероятностью очень близко к  $\mu_Y$ . Центральная предельная теорема утверждает, что для больших выборок распределение стандартизированного выборочного среднего  $(\bar{Y} - \mu_Y)/\sigma_{\bar{Y}}$  является «приблизительно» нормальным.

В отличие от точных распределений, сложных и зависящих от распределения  $Y$ , асимптотические распределения являются более простыми. Более того, примечательно, что асимптотическое нормальное распределение случайной величины  $(\bar{Y} - \mu_Y)/\sigma_{\bar{Y}}$  не зависит от вида распределения  $Y$ . Такое приближение к нормальному распределению позволяет сделать множество упрощений и лежит в основе регрессионного анализа.

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**2.6**

**Сходимость по вероятности, состоятельность и закон  
больших чисел Чебышева**

Выборочное среднее  $\bar{Y}$  сходится по вероятности к  $\mu_Y$  (или, что эквивалентно, выборочное среднее  $\bar{Y}$  является состоятельной оценкой  $\mu_Y$ ), если вероятность того, что  $\bar{Y}$  принадлежит отрезку  $[\mu_Y - c, \mu_Y + c]$ , приближается к 1 по мере возрастания  $n$  для любого значения  $c > 0$ . Сходимость  $\bar{Y}$  к  $\mu_Y$  по вероятности обозначается как  $\bar{Y} \xrightarrow{P} \mu_Y$ .

Закон больших чисел утверждает, что если  $Y_i, i = 1, \dots, n$  являются независимыми одинаково распределенными случайными величинами с  $E(Y_i) = \mu_Y$  и если наличие больших выбросов маловероятно (т. е.  $\text{var}(Y_i) = \sigma_Y^2 < \infty$ ), то  $\bar{Y} \xrightarrow{P} \mu_Y$ .

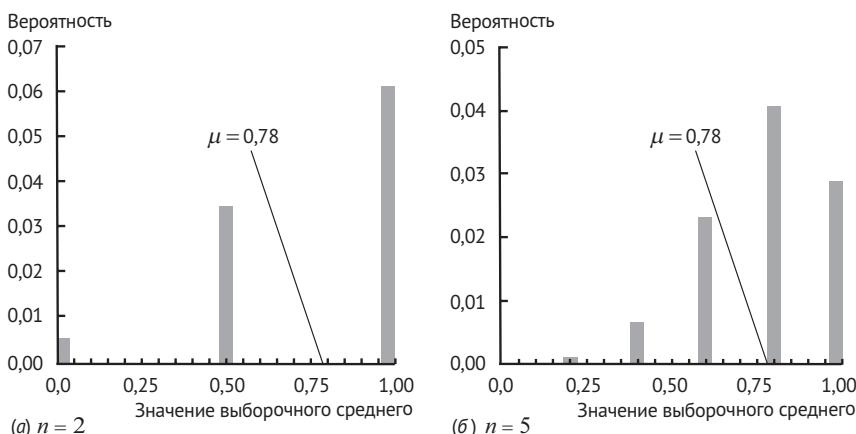
## Закон больших чисел Чебышева и состоятельность

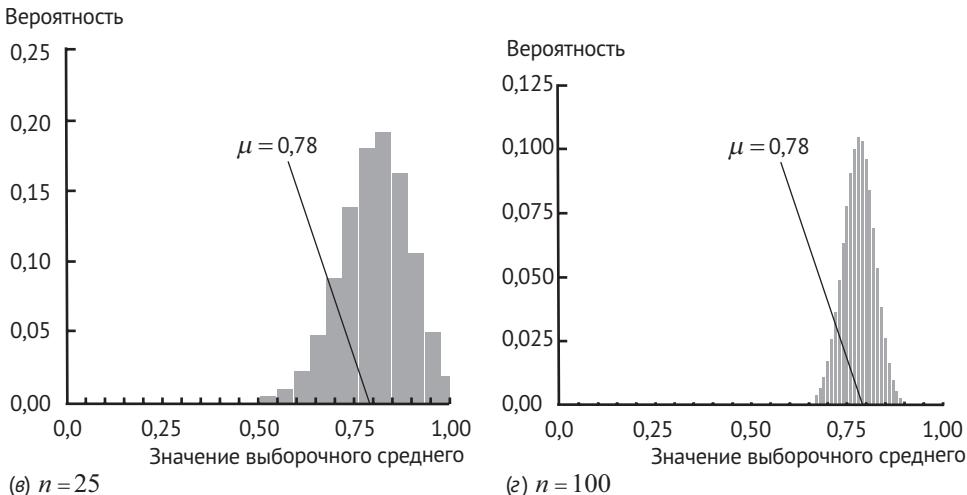
Закон больших чисел Чебышева утверждает, что при стандартных предположениях выборочное среднее  $\bar{Y}$  при больших  $n$  с достаточно высокой вероятностью будет принимать значения, примерно равные  $\mu_Y$ . Иногда данное утверждение называют «законом о средних». При усреднении большого числа случайных величин с одинаковым математическим ожиданием большие значения уравновешиваются маленькими и, таким образом, выборочное среднее равно их среднему.

В качестве примера рассмотрим упрощенную постановку задачи про студентку, которая записывает время, потраченное на дорогу до университета, фиксируя лишь тот факт, что путь был коротким (на дорогу у него ушло менее 20 минут) или длинным. Пусть  $Y_i$  принимает значения, равные 1, если в случайно выбранный день  $i$  на дорогу было потрачено менее 20 минут, и 0 – в противном случае. Так как выборка наблюдений из всего множества записей случайна, то  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами. Таким образом,  $Y_1, \dots, Y_n$  являются исходами экспериментов Бернулли, для которых, согласно таблице 2.2,  $Y_i$  принимает значение, равное 1, с вероятностью 0,78. Поскольку математическое ожидание случайной величины Бернулли равно вероятности положительного исхода эксперимента, то мы имеем  $E(Y_i) = \mu_i = 0,78$ . Выборочное среднее  $\bar{Y}$  – это доля дней во всей выборке, в которых дорога домой заняла мало времени.

На рисунке 2.8 изображено выборочное распределение  $\bar{Y}$  для выборок разных объемов  $n$ . При  $n=2$  (рис. 2.8а)  $\bar{Y}$  принимает только лишь одно из трех значений: 0,  $\frac{1}{2}$  или 1 (соответственно в обоих наблюдениях на дорогу было потрачено много времени; только лишь в одном случае дорога заняла мало времени; в обоих случаях на дорогу ушло мало времени). Каждое из этих значений не является близким к 0,78. При увеличении  $n$  (рис. 2.8 б – г) число возможных значений выборочного среднего  $\bar{Y}$  возрастает и приближается к  $\mu_Y$ .

Свойство сходимости  $\bar{Y}$  к среднему  $\mu_Y$  при увеличении  $n$  называется *сходимостью по вероятности* или, более коротко, *состоятельностью* (см. вставку «Основные понятия 2.6»). Закон больших чисел гласит, что при определенных предположениях  $\bar{Y}$  сходится по вероятности к  $\mu_Y$  или, что эквивалентно, выборочное среднее  $\bar{Y}$  есть состоятельная оценка  $\mu_Y$ .





**Рисунок 2.8. Выборочное распределение выборочного среднего  $n$  бернульиевых случайных величин**

Распределения являются выборочными распределениями величины  $\bar{Y}$   $n$  независимых случайных величин Бернулли с  $p = \Pr(Y_i = 1) = 0,78$  (вероятность того, что дорога займет мало времени, равна 78%). Дисперсия выборочного распределения  $\bar{Y}$  снижается с увеличением  $n$ . Таким образом, выборочное распределение концентрируется вокруг своего среднего  $\mu = 0,78$  при увеличении размера выборки  $n$ .

Условиями для выполнения закона больших чисел, который мы будем использовать в данной книге, являются предположения о том, что случайные величины  $Y_i$  ( $i = 1, \dots, n$ ) независимо и одинаково распределены (i.i.d.) и дисперсия  $Y_i$ ,  $\sigma_{Y_i}^2$ , конечна. Почему необходимо выполнение именно этих условий, объясняется в разделе 17.2, в котором приводится доказательство закона больших чисел Чебышева. Если данные выбраны случайным образом, то условие о независимом одинаковом распределении выполняется автоматически. Предположение о конечности дисперсии означает, что очень большие значения  $Y_i$ , то есть выбросы, маловероятны и наблюдаются нечасто. Другими словами, подобные большие значения могли бы сильно изменить  $\bar{Y}$  и выборочное среднееисказилось бы. Это предположение выглядит правдоподобным для всех рассматриваемых в книге примеров. Например, в рассмотренном выше примере существует верхний предел времени, которое тратит студентка на дорогу в университет (она может оставить машину на парковке и пойти пешком, если пробка будет чудовищной), следовательно, дисперсия распределения затрачиваемого на дорогу времени конечна.

### Центральная предельная теорема

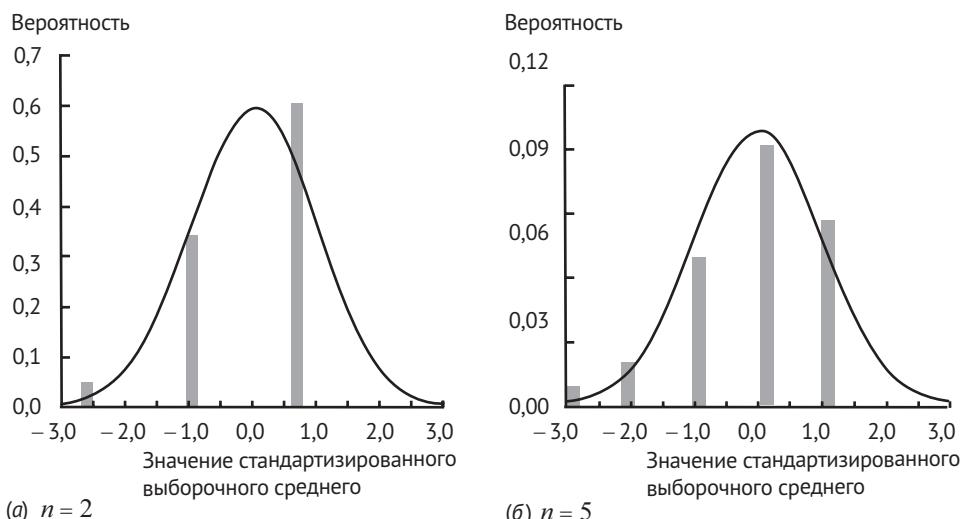
**Центральная предельная теорема** утверждает, что при выполнении некоторых общих предположений распределение  $\bar{Y}$  хорошо приближается нормальному распределением при достаточно больших значениях  $n$ . Напомним, что средним значением  $\bar{Y}$  является  $\mu_Y$ , а  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$  – ее дисперсией. Тогда, согласно центральной предельной теореме, распределение  $\bar{Y}$  аппроксимируется распределением  $N(\mu_Y, \sigma_{\bar{Y}}^2)$ . Как обсуждалось ранее в разделе 2.5, распределение

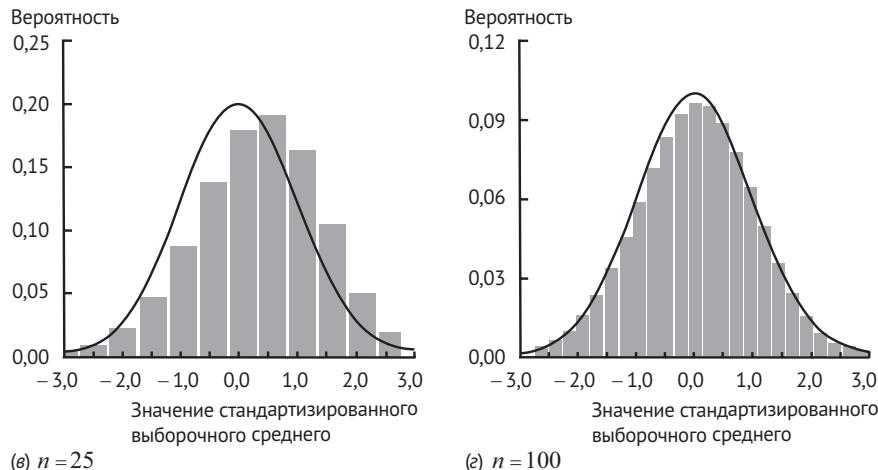
$N(\mu_Y, \sigma_{\bar{Y}}^2)$  является точным распределением  $\bar{Y}$  в случае, когда  $Y_1, \dots, Y_n$  является выборкой из генеральной совокупности с функцией распределения  $N(\mu_Y, \sigma_Y^2)$ . Центральная предельная теорема показывает, что то же самое верно при больших  $n$ , даже если каждая случайная величина  $Y_1, \dots, Y_n$  не распределена по нормальному закону.

Сходимость функции плотности распределения  $\bar{Y}$  к колоколообразному виду или ее аппроксимацию нормальным распределением можно отчасти увидеть на рисунке 2.8. Тем не менее из-за того что при больших  $n$  распределение в некотором смысле сжимается, для того чтобы более четко увидеть это свойство, нужно предпринять некоторые действия. Один из самых простых способов – посмотреть на распределение  $\bar{Y}$  через увеличительное стекло. Или использовать любой другой способ, чтобы растянуть график по оси абсцисс.

Например, можно стандартизировать  $\bar{Y}$ , вычитая его среднее и разделив на стандартное отклонение. Тогда стандартизированное выборочное среднее будет принимать значения из интервала от 0 до 1. В этом случае следует рассмотреть функцию распределения стандартизированной версии  $\bar{Y}$ , то есть  $(\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$ . Согласно центральной предельной теореме, это распределение должно достаточно хорошо аппроксимироваться стандартным нормальным распределением  $N(0, 1)$  при больших значениях  $n$ .

График функции плотности распределения стандартизированного выборочного среднего  $\bar{Y}, (\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$ , изображен на рисунке 2.9 для тех же случаев, что и на рисунке 2.8. Графики функций плотности распределений, изображенные на рисунках 2.8 и 2.9, в точности совпадают, с той лишь разницей, что на рисунке 2.9 изменены значения горизонтальной оси таким образом, что стандартизированная случайная величина имеет среднее 0 и дисперсию 1. После подобных изменений легко увидеть, что при достаточно больших  $n$  функция плотности распределения  $\bar{Y}$  хорошо аппроксимируется функцией плотности нормального распределения.





**Рисунок 2.9. Функции плотности распределения стандартизированного выборочного среднего  $n$  случайных величин Бернулли с  $p=0,78$**

На графиках изображены стандартизованные функции плотности распределений  $\bar{Y}$  с рисунка 2.8. На них соответствующие (см. рис. 2.8) функции плотности распределений отцентрированы и увеличены (в  $\sqrt{n}$  раз) по горизонтальной оси. В случае больших выборок выборочные функции плотности хорошо аппроксимируются функциями плотности нормального распределения (сплошная линия), что соответствует утверждению центральной предельной теоремы.

## ОСНОВНЫЕ ПОНЯТИЯ 2.7

### Центральная предельная теорема

Предположим, что случайные величины  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами с  $E(Y_i) = \mu_Y$  и  $\text{var}(Y_i) = \sigma_Y^2$ , где  $0 < \sigma_Y^2 < \infty$ . При  $n \rightarrow \infty$  распределение случайной величины  $(\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$  (где  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$ ) стремится к стандартному нормальному.

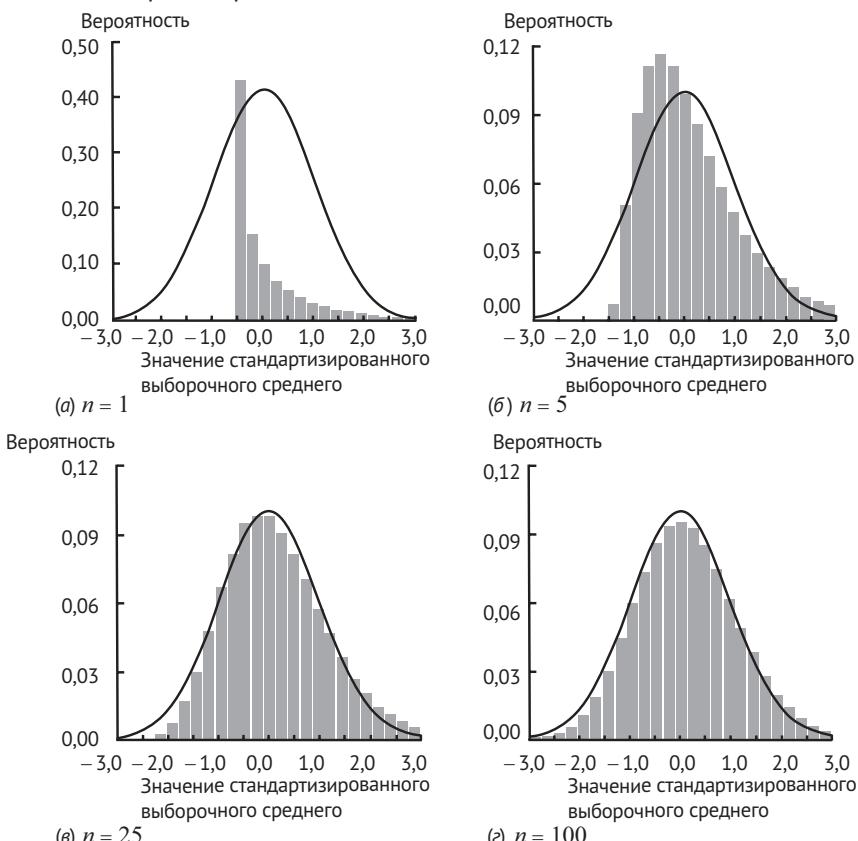
Здесь возникает вопрос о том, какую выборку считать «достаточно большой». То есть насколько большим должно быть значение  $n$  для того, чтобы распределение выборочного среднего  $\bar{Y}$  приближалось к нормальному? Ответ: «Это зависит от...». Качество аппроксимации нормальным распределением зависит от того, какому закону распределения подчиняются  $Y_i$ . Если сами  $Y_i$  распределены по нормальному закону, то выборочное среднее  $\bar{Y}$  также будет распределено по нормальному закону при любых  $n$ . В противном случае, если  $Y_i$  распределены в соответствии с распределением, сильно отличающимся от нормального, то для того чтобы выборочное среднее хорошо аппроксимировалось нормальным распределением, необходимо, чтобы  $n$  было достаточно большим, равным 30 или даже больше.

Данный факт представлен на рисунке 2.10 для генеральной совокупности с распределением, значительно отличающимся от распределения Бернулли. Рассматриваемое распределение имеет длинный правый хвост (оно скошено вправо). Выборочное распределение  $\bar{Y}$  после центрирования и корректировки масштаба изображено на рисунках 2.10 б – г для случаев с  $n = 5, 25$  и  $100$ , соответственно. Несмотря на то что полученное распределение принимает колоколообразную форму уже при  $n = 25$ , оно все равно значительно отличается

от нормального. При  $n=100$ , тем не менее, аппроксимация функцией нормального распределения дает гораздо лучшие результаты. Фактически при  $n \geq 100$  аппроксимация  $\bar{Y}$  функцией нормального распределения дает хорошие результаты для большинства видов распределений.

Центральная предельная теорема является замечательным результатом. В то время как при «небольших значениях  $n$ » функции распределения  $\bar{Y}$ , соответствующие случаям (б) и (в) на рисунках 2.9 и 2.10, значительно отличаются друг от друга, при «больших  $n$ » распределения на рисунках 2.9г и 2.10г очень похожи, они имеют одинаковую форму. В силу того что распределение выборочного среднего  $\bar{Y}$  аппроксимируется функцией нормального распределения при больших значениях  $n$ , говорят, что  $\bar{Y}$  имеет *асимптотически нормальное распределение*.

Удобство использования аппроксимации нормальным распределением с его широкой применимостью в силу центральной предельной теоремы делает его основой современной прикладной эконометрики. Формулировка центральной предельной теоремы приведена во вставке «Основные понятия 2.7».



**Рисунок 2.10. Распределение стандартизированного выборочного среднего  $n$  случайных величин из асимметричного распределения**

На графиках изображены выборочные распределения стандартизированного выборочного среднего  $n$  для случайных величин, распределенных по некоторому асимметричному закону, который изображен на рисунке 2.10а. При небольших значениях  $n$  (например  $n=5$ ) выборочное распределение, подобно распределению всей генеральной совокупности, также оказывается асимметричным. Но при больших значениях  $n$  ( $n=100$ ) выборочное распределение в соответствии с центральной предельной теоремой аппроксимируется стандартным нормальным распределением (сплошная линия).

## Выходы

1. Вероятность, с которой случайная величина принимает различные значения, задается ее кумулятивной функцией распределения, функцией распределения вероятностей (для дискретных случайных величин) и функцией плотности вероятности (для непрерывных случайных величин).
2. Математическим ожиданием случайной величины  $Y$  (или ее средним значением  $\mu_Y$ ), обозначаемым  $E(Y)$ , называется ее среднее значение, взвешенное по вероятности. Дисперсией  $Y$  называется величина  $\sigma_Y^2 = E[(Y - \mu_Y)^2]$ . Стандартное отклонение  $Y$  равно квадратному корню из ее дисперсии.
3. Совместная вероятность двух случайных величин  $X$  и  $Y$  описывается совместным распределением их вероятностей. Условным распределением вероятности  $Y$  при условии  $X = x$  называется распределение вероятностей  $Y$  в случае, когда  $X$  принимает только значение  $x$ .
4. График функции плотности вероятности нормально распределенной случайной величины имеет колоколообразную форму (см. рис. 2.5). Для вычисления вероятности того, что нормально распределенная случайная величина принимает значение, меньшее либо равное некоторой величине, необходимо стандартизировать эту случайную величину, а затем использовать соответствующие значения из таблицы для стандартного нормального кумулятивного распределения (см. табл. 1 приложения).
5.  $n$  случайных величин  $Y_1, \dots, Y_n$ , выбранные случайным образом из некоторой генеральной совокупности, образуют выборку из  $n$  независимых одинаково распределенных (i.i.d.) случайных величин.
6. Выборочное среднее  $\bar{Y}$  изменяется при изменении выборки и, таким образом, также является случайной величиной. Если  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами, то:
  - a) математическим ожиданием выборочного среднего  $\bar{Y}$  является  $\mu_Y$ , а его дисперсией  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$ ;
  - b) по закону больших чисел  $\bar{Y}$  сходится по вероятности к  $\mu_Y$ ;
  - c) по центральной предельной теореме стандартизированная величина выборочного среднего  $\bar{Y} - (\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$ , является случайной величиной, распределенной по стандартному нормальному закону распределения  $N(0, 1)$  при больших значениях  $n$ .

## Основные понятия

Исход, элементарное событие (с. 18).

Вероятность исхода (с. 18).

Пространство исходов (элементарных событий) (с. 18).

Событие (с. 18).

Дискретная случайная величина (с. 18).

Непрерывная случайная величина (с. 18).

Распределение вероятностей (с. 18).

- Интегральное распределение вероятностей (с. 20).  
Интегральная функция распределения (с. 20).  
Случайная величина Бернулли (с. 20).  
Распределение Бернулли (с. 20).  
Функция плотности вероятности (с. 21).  
Функция плотности (с. 21).  
Плотность (с. 21).  
Математическое ожидание (с. 22).  
Ожидание (с. 22).  
Среднее (с. 22).  
Дисперсия (с. 24).  
Стандартное отклонение (с. 24).  
Моменты распределения (с. 26).  
Асимметрия (с. 26).  
Эксцесс (с. 27).  
Выброс (с. 28).  
Островершинное распределение (с. 28).  
 $r$ -тый момент (с. 28).  
Совместное распределение вероятностей (с. 28).  
Безусловное распределение вероятностей (с. 29).  
Условное распределение (с. 30).  
Условное математическое ожидание (с. 31).  
Условное среднее (с. 31).  
Закон повторного математического ожидания (с. 32).  
Условная дисперсия (с. 33).  
Независимо распределенные случайные величины (с. 33).  
Независимые случайные величины (с. 33).  
Ковариация (с. 34).  
Корреляция (с. 34).  
Некоррелированные случайные величины (с. 34).  
Нормальное распределение (с. 38).  
Стандартное нормальное распределение (с. 38).  
Стандартизовать случайную величину (с. 39).  
Многомерное нормальное распределение (с. 40).  
Двухмерное нормальное распределение (с. 40).  
Распределение хи-квадрат (с. 43).  
 $t$ -распределение Стьюдента (с. 44).  
 $t$ -распределение (с. 44).  
 $F$ -распределение (с. 44).  
Простая случайная выборка (с. 46).  
Генеральная совокупность (с. 46).  
Однаково распределенные случайные величины (с. 46).  
Независимые одинаково распределенные случайные величины (с. 47).  
Выборочное среднее значение (с. 47).  
Среднее по выборке (с. 47).

- Выборочное распределение (с. 47).
- Точное (конечное) распределение (с. 50).
- Асимптотическое распределение (с. 50).
- Закон больших чисел (с. 51).
- Сходимость по вероятности (с. 51).
- Состоятельность (с. 51).
- Центральная предельная теорема (с. 52).
- Асимптотически нормальное распределение (с. 55).

### **Вопросы для повторения и закрепления основных понятий**

- 2.1. Примерами случайных величин, рассматриваемых в данной главе, являются:
  - a) пол человека, которого вы встретите первым;
  - b) число поломок компьютера;
  - c) время, необходимое студентке для того, чтобы добраться до учебного заведения;
  - d) является ли новым или старым компьютер, которым вы будете пользоваться в библиотеке;
  - e) будет идти дождь или нет.

Объясните, почему каждый из приведенных примеров может считаться случайным событием?

- 2.2. Предположим, что случайные переменные  $X$  и  $Y$  являются независимыми и известны их распределения. Объясните, почему значение величины  $X$  ничего не говорит относительно значения величины  $Y$ ?
- 2.3. Предположим, что  $X$  – количество дождей в вашем городе в течение месяца, а  $Y$  – количество детей, рожденных в Лос-Анджелесе за тот же период. Являются ли  $X$  и  $Y$  независимыми? Объясните ваш ответ.
- 2.4. Группа по эконометрике состоит из 80 студентов. Средний вес студентов в этой группе равен 66 кг (145 lb). Пусть случайным образом было выбрано четыре студента и рассчитан их средний вес. Будет ли он равен 66 кг? Объясните ваш ответ. Используйте данный пример для объяснения, почему выборочное среднее  $\bar{Y}$  является случайной величиной.
- 2.5. Предположим,  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами с функцией распределения  $N(1, 4)$ . Определите плотность распределения  $\bar{Y}$  при  $n = 2$ . Проделайте то же самое упражнение для случаев  $n = 10$  и  $n = 100$ . Другими словами, покажите, в чем различие между функциями плотности распределения для указанных случаев? Определите связь полученного результата с законом больших чисел.
- 2.6. Пусть  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами, чья функция плотности вероятности изображена на рисунке 2.10a. Вам необходимо рассчитать  $\Pr(\bar{Y} \leq 0,1)$ . Будет ли рациональным использовать для этого аппроксимацию с  $n = 5$ ? Как изменится

нится ваш ответ для случаев  $n = 25$  и  $n = 100$ ? Объясните полученные результаты.

- 2.7.  $Y$  – случайная переменная с  $\mu_Y = 0$ ,  $\sigma_Y = 1$ , коэффициентом асимметрии, равным нулю, и коэффициентом эксцесса, равным 100. Определите гипотетическую функцию распределения вероятностей  $Y$ . Объясните, почему среди  $n$  выбранных случайным образом величин могут быть большие выбросы?

### **Упражнения**

- 2.1. Пусть  $Y$  означает количество «орлов» в эксперименте с подбрасыванием монет.
- Выведите функцию плотности случайной величины  $Y$ .
  - Выведите функцию распределения случайной величины  $Y$ .
  - Вычислите среднее и дисперсию случайной величины  $Y$ .
- 2.2. Используя информацию о распределении вероятностей из таблицы 2.2, вычислите:
- $E(Y)$  и  $E(X)$ .
  - $\sigma_X^2$  и  $\sigma_Y^2$ .
  - $\sigma_{XY}$  и  $\text{corr}(X, Y)$ .
- 2.3. Используя информацию для  $X$  и  $Y$  из таблицы 2.2, рассмотрите следующие случайные переменные:  $W = 3 + 6X$  и  $V = 10 - 7Y$ . Вычислите:
- $E(W)$  и  $E(V)$ .
  - $\sigma_W^2$  и  $\sigma_V^2$ .
  - $\sigma_{WV}$  и  $\text{corr}(W, V)$ .
- 2.4. Пусть  $X$  – бернуlliевская случайная величина с  $P(X = 1) = p$ .
- Покажите, что  $E(X^3) = p$ .
  - Покажите, что  $E(X^k) = p$  при  $k > 0$ .
  - Пусть  $p = 0,3$ . Вычислите среднее, дисперсию, асимметрию и эксцесс случайной величины  $X$  (Подсказка: используйте формулы из упражнения 2.21).
- 2.5. В сентябре в Сиэтле средняя дневная температура равна  $70^\circ$  по Фаренгейту, а стандартное отклонение –  $7^\circ F$ . Определите среднее, стандартное отклонение и дисперсию данных показателей в градусах Цельсия.
- 2.6. В приведенной ниже таблице представлена информация о совместном распределении вероятностей между статусом занятости (занятый или безработный) и наличием высшего образования среди населения трудоспособного возраста США в 2008 году.

**Совместное распределение статуса занятости  
и наличия высшего образования среди населения США старше 25 лет в 2008 году**

	Безработный ( $Y = 0$ )	Занятый ( $Y = 1$ )	Итого
Без высшего образования ( $X = 0$ )	0,037	0,622	0,659
С высшим образованием ( $X = 1$ )	0,009	0,332	0,341
Итого	0,046	0,954	1,000

- a) Вычислите  $E(Y)$ .
- б) Уровень безработицы определяется как доля рабочей силы, которая в рассматриваемый момент является безработной. Покажите, что уровень безработицы определяется как  $1 - E(Y)$ .
- в) Вычислите  $E(Y | X = 1)$  и  $E(Y | X = 0)$ .
- г) Определите уровень безработицы для:
- i) выпускников высших учебных заведений;
  - ii) людей, не имеющих высшего образования.
- д) Человек, выбранный случайным образом из всей генеральной совокупности, оказывался безработным. Какова вероятность того, что он имеет высшее образование? Какова вероятность того, что он не имеет высшего образования?
- е) Являются ли уровень образования и рабочий статус независимыми? Объясните ваш ответ.
- 2.7. В генеральной совокупности, состоящей из пар одновременно работающих женщин и мужчин, средняя зарплата мужчин составляет 40тыс. долларов в год со стандартным отклонением 12тыс. долларов. Средняя зарплата женщин равна 45тыс. долларов в год при стандартном отклонении 18тыс. долларов. Корреляция заработных плат мужчин и женщин внутри пар равна 0,80. Пусть  $C$  обозначает суммарный доход для случайно выбранной пары.
- а) Чему равно среднее значение  $C$ ?
- б) Чему равна ковариация между заработками мужчины и заработками женщины?
- в) Чему равно стандартное отклонение  $C$ ?
- г) Ответьте на вопросы (а) – (в), переведя доллары в евро.
- 2.8. Случайная переменная  $Y$  имеет среднее значение, равное 1, и дисперсию, равную 4. Пусть  $Z = \frac{1}{2}(Y - 1)$ . Покажите, что  $\mu_Z = 0$  и  $\sigma_Z^2 = 1$ .
- 2.9.  $X$  и  $Y$  – дискретные случайные переменные, чье распределение вероятностей описывается в приведенной ниже таблице.

		Значение $Y$				
		14	22	30	40	65
Значение $X$	1	0,02	0,05	0,10	0,03	0,01
	5	0,17	0,15	0,05	0,02	0,01
	8	0,02	0,03	0,15	0,10	0,09

Таким образом,  $\Pr(X = 1, Y = 14) = 0,02$  и так далее.

- а) Вычислите распределение вероятностей, среднее и дисперсию  $Y$ .
- б) Вычислите распределение вероятностей, среднее и дисперсию  $Y$  при условии  $X = 8$ .
- в) Вычислите ковариацию и корреляцию между  $X$  и  $Y$ .
- 2.10. Вычислите:
- а)  $Y$  имеет распределение  $N(1, 4)$ , определите  $\Pr(Y \leq 3)$ .
- б)  $Y$  имеет распределение  $N(3, 9)$ , определите  $\Pr(Y > 0)$ .

- в)  $Y$  имеет распределение  $N(50, 25)$ , определите  $\Pr(10 \leq Y \leq 52)$ .  
 г)  $Y$  имеет распределение  $N(5, 2)$ , определите  $\Pr(6 \leq Y \leq 8)$ .
- 2.11. Вычислите:
- а) Пусть  $Y$  имеет распределение  $\chi_4^2$ . Чему равно  $\Pr(Y \leq 7,78)$ ?
  - б) Пусть  $Y$  имеет распределение  $\chi_{10}^2$ . Чему равно  $\Pr(Y > 18,31)$ ?
  - в) Пусть  $Y$  имеет распределение  $F_{10, \infty}$ . Чему равно  $\Pr(Y > 1,83)$ ?
  - г) Объясните, почему задания в пунктах (б) и (в) имеют одинаковые ответы?
  - д) Если случайная величина  $Y$  распределена в соответствии с  $\chi_1^2$ , определите  $\Pr(Y \leq 1,0)$ . (*Подсказка:* используйте определение распределения  $\chi_1^2$ .)
- 2.12. Вычислите:
- а) Пусть  $Y$  имеет распределение  $t_{15}$ . Вычислите  $\Pr(Y > 1,75)$ .
  - б) Пусть  $Y$  имеет распределение  $t_{90}$ . Вычислите  $\Pr(-1,99 \leq Y \leq 1,99)$ .
  - в) Пусть  $Y$  имеет распределение  $N(0,1)$ . Вычислите  $\Pr(-1,99 \leq Y \leq 1,99)$ .
  - г) Почему, по вашему мнению, ответы в пунктах (б) и (в) примерно одинаковые?
  - д) Пусть  $Y$  имеет распределение  $F_{7,4}$ . Вычислите  $\Pr(Y > 4,12)$ .
  - е) Пусть  $Y$  имеет распределение  $F_{7,120}$ . Вычислите  $\Pr(Y > 2,79)$ .
- 2.13. Пусть  $X$  – случайная величина Бернулли с  $\Pr(X = 1) = 0,99$ ,  $Y \sim N(0,1)$ ,  $W \sim N(0,100)$ , и  $X$ ,  $Y$  и  $W$  – независимы. Пусть  $S = XY + (1-X)W$ . (То есть  $S = Y$ , если  $X = 1$ , и  $S = W$ , если  $X = 0$ .)
- а) Покажите, что  $E(Y^2) = 1$  и  $E(W^2) = 100$ .
  - б) Покажите, что  $E(Y^3) = 0$  и  $E(W^3) = 0$ . (*Подсказка:* чему равна асимметрия для симметричного распределения?)
  - в) Покажите, что  $E(Y^4) = 3$  и  $E(W^4) = 3 \times 100^2$ . (*Подсказка:* используйте тот факт, что коэффициент эксцесса нормального распределения равен 3.)
  - г) Вычислите  $E(S)$ ,  $E(S^2)$ ,  $E(S^3)$  и  $E(S^4)$ . (*Подсказка:* используйте закон повторного математического ожидания при условии  $X = 0$  и  $X = 1$ .)
  - д) Вычислите асимметрию и эксцесс для  $S$ .
- 2.14. Пусть у генеральной совокупности  $\mu_Y = 100$  и  $\sigma_Y^2 = 43$ . Используя центральную предельную теорему, ответьте на следующие вопросы:
- а) Чему равна вероятность  $\Pr(\bar{Y} \leq 101)$  в случайной выборке размера  $n = 100$ ?
  - б) Чему равна вероятность  $\Pr(\bar{Y} > 98)$  в случайной выборке размера  $n = 165$ ?
  - в) Чему равна вероятность  $\Pr(101 \leq \bar{Y} \leq 103)$  в случайной выборке размера  $n = 64$ ?
- 2.15. Допустим,  $Y_i$ ,  $i = 1, 2, \dots, n$  – независимые одинаково распределенные случайные величины, каждая из которых имеет распределение  $N(10, 4)$ .
- а) Вычислите  $\Pr(9,6 \leq \bar{Y} \leq 10,4)$  для случаев:
    - (i)  $n = 20$ ;
    - (ii)  $n = 100$ ;
    - (iii)  $n = 1000$ .

- б) Предположим, что  $c$  – некоторое положительное число. Покажите, что  $\Pr(10 - c \leq \bar{Y} \leq 10 + c)$  стремится к 1 с ростом  $n$ .
- в) Используя результат, полученный в пункте (б), проверьте сходимость по вероятности  $\bar{Y}$  к 10.
- 2.16. Пусть  $Y \sim N(5, 100)$  и вам нужно вычислить  $\Pr(Y < 3,6)$ . Но, к сожалению, вы не имеете возможности воспользоваться таблицей нормального распределения, приведенной в таблице 1 приложения. Тем не менее у вас есть возможность сгенерировать независимые одинаково распределенные случайные величины из распределения  $N(5, 100)$  на компьютере. Объясните, как, используя компьютер, можно достаточно точно определить, чему равно  $\Pr(Y < 3,6)$ .
- 2.17. Пусть  $Y_i, i = 1, \dots, n$  – независимые одинаково распределенные случайные величины Бернулли с  $p = 0,4$ . Обозначим через  $\bar{Y}$  соответствующее выборочное среднее:
- Используя центральную предельную теорему, вычислите:
    - $\Pr(\bar{Y} \geq 0,43)$  при  $n=100$ ;
    - $\Pr(\bar{Y} \leq 0,37)$  при  $n=400$ .
  - Каким должно быть значение  $n$ , для того чтобы выполнялось  $\Pr(0,39 \leq \bar{Y} \leq 0,41) \geq 0,95$ ?
- 2.18. В течение года ураган может нанести повреждения дому. От года в год повреждения происходят случайным образом. Пусть  $Y$  – размер повреждений в долларах США в данном году. Предположим, что в 95 % случаев  $Y = 0$  долл., но в 5 % случаев  $Y = 20$  тыс. долл.
- Найдите математическое ожидание и стандартное отклонение размера повреждений в течение года.
  - Рассмотрите случайную выборку из 100 домов (наблюдения выборки можно считать независимыми одинаково распределенными случайными величинами). Пусть  $\bar{Y}$  – выборочное среднее размера повреждений этих ста домов за год.
    - Каким является ожидаемое значение  $\bar{Y}$ ?
    - Какова вероятность того, что  $\bar{Y}$  превысит 2000 долл.?
- 2.19. Рассмотрим две случайные величины  $X$  и  $Y$ . Предположим, что  $Y$  принимает  $k$  значений  $y_1, \dots, y_k$ , в то время как  $X$  принимает  $l$  значений  $x_1, \dots, x_l$ .
- Покажите, что  $\Pr(Y = y_i) = \sum_{j=1}^l \Pr(Y = y_j | X = x_i) \Pr(X = x_i)$ . (Подсказка: используйте определение  $\Pr(Y = y_j | X = x_i)$ .)
  - Используйте ответ на вопрос пункта (a), чтобы проверить уравнение (2.19).
  - Предположим, что  $X$  и  $Y$  независимы. Покажите, что  $\sigma_{XY} = 0$  и  $\text{corr}(X, Y) = 0$ .
- 2.20. Рассмотрим три случайные величины  $X, Y$  и  $Z$ . Предположим, что  $Y$  принимает  $k$  значений  $y_1, \dots, y_k$ ,  $X$  принимает  $l$  значений  $x_1, \dots, x_l$ , а  $Z$  принимает  $m$  значений  $z_1, \dots, z_m$ . Пусть  $\Pr(X = x, Y = y, Z = z)$  – совместное распределение вероятностей  $X, Y$  и  $Z$ . Условное распределение

ние вероятностей для  $Y$  задается выражением  $\Pr(Y = y | X = x, Z = z) = \frac{\Pr(Y = y, X = x, Z = z)}{\Pr(X = x, Z = z)}$ .

- a) Объясните, как безусловная вероятность того, что  $Y = y$ , может быть получена из совместного распределения вероятностей? (Подсказка: данное утверждение является обобщением уравнения (2.16).)
- б) Покажите, что  $E(Y) = E[E(Y | X, Z)]$ . (Подсказка: данное утверждение является обобщением уравнений (2.19) и (2.20).)
- 2.21. Пусть  $X$  – случайная величина с моментами  $E(X)$ ,  $E(X^2)$ ,  $E(X^3)$  и так далее.
- Покажите, что  $E(X - \mu)^3 = E(X^3) - 3[E(X^2)][E(X)] + 2[E(X)]^3$ .
  - Покажите, что  $E(X - \mu)^4 = E(X^4) - 4[E(X)][E(X^3)] + 6[E(X)]^2[E(X^2)] - 3[E(X)]^4$ .
- 2.22. Предположим, что вы инвестируете деньги в размере, например, 1 долл. Вы планируете инвестировать долю  $w$  ваших денег в паевой фонд акций, а остальную часть,  $1 - w$ , в паевой фонд облигаций. Допустим, 1 долл., инвестированный в акции, приносит доходность  $R_s$  за 1 год, в то время как  $R_b$  – доходность по облигациям. Предположим, что  $R_s$  – случайная величина со средним 0,08 (8 %) и стандартным отклонением 0,07, а  $R_b$  – случайная величина со средним 0,05 (5 %) и стандартным отклонением 0,04. Корреляция между  $R_s$  и  $R_b$  равна 0,25. Согласно вашим планам инвестирования, общая доходность инвестиции составит  $R = wR_s + (1 - w)R_b$ .
- Допустим,  $w = 0,5$ . Вычислите среднее и стандартное отклонение  $R$ .
  - Допустим,  $w = 0,75$ . Вычислите среднее и стандартное отклонение  $R$ .
  - Каким должно быть значение  $w$  для того, чтобы математическое ожидание  $R$  приняло наибольшее значение?
  - (Задание повышенной сложности!) Какое значение  $w$  минимизирует стандартное отклонение  $R$ ? (Приведите графическую иллюстрацию.)
- 2.23. В данном упражнении рассматривается пара случайных величин  $X$  и  $Y$ , для которых условное среднее  $Y$  при условии  $X$  зависит от  $X$ , но  $\text{corr}(X, Y) = 0$ . Пусть  $X$  и  $Z$  – две независимые случайные величины, имеющие стандартное нормальное распределение, и пусть  $Y = X^2 + Z$ .
- Покажите, что  $E(Y | X) = X^2$ .
  - Покажите, что  $\mu_Y = 1$ .
  - Покажите, что  $E(XY) = 0$ . (Подсказка: используйте тот факт, что нечетные моменты стандартного нормального распределения равны нулю.)
  - Покажите, что  $\text{cov}(X, Y) = 0$  и, следовательно,  $\text{corr}(X, Y) = 0$ .
- 2.24. Предположим, что  $Y_i \sim i.i.d. N(0, \sigma^2)$ ,  $i = 1, 2, \dots, n$ .
- Покажите, что  $E(Y_i^2 / \sigma^2) = 1$ .
  - Покажите, что  $W = (1 / \sigma^2) \sum_{i=1}^n Y_i^2$  имеет распределение  $\chi_n^2$ .
  - Покажите, что  $E(W) = n$ . (Подсказка: используйте ваш ответ, полученный в пункте (a).)

г) Покажите, что  $V = Y_1 \sqrt{\frac{\sum_{i=2}^n Y_i^2}{n-1}}$  имеет распределение  $t_{n-1}$ .

- 2.25. Пусть  $x_1, \dots, x_n$  и  $y_1, \dots, y_n$  – числовые последовательности, и пусть  $a, b$  и  $c$  – константы. Покажите, что

а)  $\sum_{i=1}^n ax_i = a \sum_{i=1}^n x_i$ ;

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

в)  $\sum_{i=1}^n a = na$ ;

г)  $\sum_{i=1}^n (a + bx_i + cy_i)^2 = na^2 + b^2 \sum_{i=1}^n x_i^2 + c^2 \sum_{i=1}^n y_i^2 + 2ab \sum_{i=1}^n x_i + 2ac \sum_{i=1}^n y_i + 2bc \sum_{i=1}^n x_i y_i$ .

- 2.26. Предположим, что  $Y_1, Y_2, \dots, Y_n$  – случайные величины с математическим ожиданием  $X$ , дисперсией  $\sigma_Y^2$  и корреляцией  $\rho$  (для всех пар  $Y_i$  и  $Y_j$ , таких, что  $i \neq j$ ).

а) Покажите, что  $\text{cov}(Y_i, Y_j) = \rho \sigma_Y^2$  для  $i \neq j$ .

б) Допустим, что  $n = 2$ . Покажите, что  $E(\bar{Y}) = \mu_Y$  и  $\text{var}(\bar{Y}) = \frac{1}{2} \sigma_Y^2 + \frac{1}{2} \rho \sigma_Y^2$ .

в) Для случая  $n \geq 2$  покажите, что  $E(\bar{Y}) = \mu_Y$  и  $\text{var}(\bar{Y}) = \sigma_Y^2 / n + [(n-1)/n] \rho \sigma_Y^2$ .

г) При очень больших значениях  $n$  покажите, что  $\text{var}(\bar{Y}) \approx \rho \sigma_Y^2$ .

- 2.27. Пусть  $X$  и  $Z$  – две совместно распределенные случайные величины. Предположим, что значение  $Z$  известно, но неизвестно значение  $X$ . Пусть  $\tilde{X} = E(X | Z)$  – результат попытки угадать значение  $X$  на основе информации о  $Z$ , и пусть  $W = X - \tilde{X}$  – соответствующая ошибка измерения.

а) Покажите, что  $E(W) = 0$ . (Подсказка: используйте закон повторного математического ожидания.)

б) Покажите, что  $E(WZ) = 0$ .

в) Обозначим через  $\hat{X} = g(Z)$  другой способ угадать  $X$  на основе информации о  $Z$ , и ошибка этого измерения равна  $V = X - \hat{X}$ . Покажите, что  $E(V^2) \geq E(W^2)$ . (Подсказка: пусть  $h(Z) = g(Z) - E(X | Z)$ , так что  $V = [X - E(X | Z)] - h(Z)$ . Выполните  $E(V^2)$ .)

## Приложения

### Приложение 2.1. Вывод результатов вставки «Основные понятия 2.3»

В данном приложении приводятся доказательства уравнений из вставки «Основные понятия 2.3».

Уравнение (2.29) следует напрямую из определения математического ожидания.

Для вывода уравнения (2.30) используем определение дисперсии и запишем:

$$\text{var}(a+bY) = E\{(a+bY - E(a+bY))^2\} = E\{[b(Y - \mu_Y)]^2\} = b^2 E[(Y - \mu_Y)^2] = b^2 \mu_Y^2.$$

Чтобы вывести уравнение (2.31), используем определение дисперсии и запишем выражение:

$$\begin{aligned} \text{var}(aX + bY) &= E\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} = \\ &= E\left\{\left[a(X - \mu_X) + b(Y - \mu_Y)\right]^2\right\} = \\ &= E\{[a^2(X - \mu_X)^2] + 2E[ab(X - \mu_X)(Y - \mu_Y)] + E[b^2(Y - \mu_Y)^2]\} = \\ &= a^2 \text{var}(X) + 2ab \text{cov}(X, Y) + b^2 \text{var}(Y) = \\ &= a^2 \sigma_X^2 + 2ab \sigma_{XY} + b^2 \sigma_Y^2, \end{aligned} \tag{2.49}$$

в котором второе равенство получается путем группировки членов, третье равенство получается в результате раскрытия квадратных скобок и четвертое равенство следует из определения дисперсии и ковариации.

Для вывода уравнения (2.32), в силу того что  $E(Y - \mu_Y) = 0$ , можно записать  $E(Y^2) = E\{[(Y - \mu_Y) + \mu_Y]^2\} = E[(Y - \mu_Y)^2] + 2\mu_Y E(Y - \mu_Y) + \mu_Y^2 = \sigma_Y^2 + \mu_Y^2$ .

Чтобы получить уравнение (2.33), используем определение ковариации. Тогда имеем:

$$\begin{aligned} \text{cov}(a+bX+cV, Y) &= E\{[a+bX+cV - E(a+bX+cV)][Y - \mu_Y]\} = \\ &= E\{[b(X - \mu_X) + c(V - \mu_V)][Y - \mu_Y]\} = \\ &= E\{[b(X - \mu_X)][Y - \mu_Y]\} + E\{[c(V - \mu_V)][Y - \mu_Y]\} = \\ &= b\sigma_{XY} + c\sigma_{VY}, \end{aligned} \tag{2.50}$$

что является уравнением (2.33).

Для вывода уравнения (2.34) запишем:

$$\begin{aligned} E(XY) &= E\{[(X - \mu_X) + \mu_X][(Y - \mu_Y) + \mu_Y]\} = \\ &= E[(X - \mu_X)(Y - \mu_Y)] + \mu_X E(Y - \mu_Y) + \mu_Y E(X - \mu_X) + \mu_X \mu_Y = \sigma_{XY} + \mu_X \mu_Y. \end{aligned}$$

Теперь докажем корреляционное неравенство из уравнения (2.35), то есть соотношение  $|\text{corr}(X, Y)| \leq 1$ . Пусть  $a = -\sigma_{XY} / \sigma_X^2$  и  $b = 1$ . Воспользовавшись уравнением (2.31), имеем:

$$\begin{aligned} \text{var}(aX + Y) &= a^2 \sigma_X^2 + \sigma_Y^2 + 2a\sigma_{XY} = \\ &= (-\sigma_{XY} / \sigma_X^2)^2 \sigma_X^2 + \sigma_Y^2 + 2(-\sigma_{XY} / \sigma_X^2) \sigma_{XY} = \\ &= \sigma_Y^2 - \sigma_{XY}^2 / \sigma_X^2. \end{aligned} \tag{2.51}$$

## Часть I. Введение и обзор

---

В силу того что дисперсия не может быть отрицательной, выражение  $\text{var}(aX + Y)$  положительно. Тогда из уравнения (2.51) имеем  $\sigma_Y^2 - \sigma_{XY}^2 / \sigma_X^2 \geq 0$ . Перепишем данное выражение и получим ковариационное неравенство:

$$\sigma_{XY}^2 \leq \sigma_X^2 \sigma_Y^2. \quad (2.52)$$

Ковариационное неравенство подразумевает, что  $\sigma_{XY}^2 / (\sigma_X^2 \sigma_Y^2) \leq 1$  или, что эквивалентно,  $|\sigma_{XY} / (\sigma_X \sigma_Y)| \leq 1$ , из чего по определению корреляции следует корреляционное неравенство  $|\text{corr}(X, Y)| \leq 1$ .

## **Глава 3. Элементы математической статистики**

Статистика – это наука о том, как изучать мир вокруг нас, используя имеющиеся у нас данные. Статистические методы помогают нам ответить на вопрос об интересующих нас неизвестных характеристиках распределения генеральной совокупности. Например, каково среднее значение доходов среди молодых людей, недавно закончивших колледж? Различаются ли средние доходы между мужчинами и женщинами, и если различаются, то насколько?

Все эти вопросы связаны с распределением доходов в генеральной совокупности работников. Один из способов ответить на них – это опросить всех работников, измеряя доход каждого из них, и найти, таким образом, распределение доходов в генеральной совокупности. На практике, однако, такой всеобъемлющий опрос является слишком дорогим. Перепись населения в США проводится, например, раз в десять лет. При этом затраты на проведение переписи населения в 2000 году составили 10 млрд долл., а перепись 2010 года, по оценкам, обошлась в 15 млрд долл. или даже больше. Процесс разработки опросных листов для переписи, управление и проведение обследования, а также сбор и анализ данных занимают десять лет. Несмотря на всю эту подготовку и прочее, многие домохозяйства не попадают под обследование по тем или иным причинам. Следовательно, необходим другой, более практичный подход.

Ключевым моментом в математической статистике является то обстоятельство, что мы распространяем на генеральную совокупность свойства случайной выборки из нее. Вместо того чтобы обследовать все население США, мы могли бы обследовать, скажем, 1000 домохозяйств, выбранных случайно методом простого случайного выбора. Используя статистические методы, мы можем использовать полученную выборку для того, чтобы сделать предварительные выводы о характеристиках всей генеральной совокупности, то есть провести статистическую проверку.

В эконометрике используются три типа статистических методов: оценка, проверка гипотез и построение доверительных интервалов. Оценка предполагает вычисление «лучшего предположения» о численном значении для неизвестной характеристики распределения генеральной совокупности по выборке данных, например ее среднего. Тестирование гипотез предполагает формулирование конкретных гипотез о генеральной совокупности и принятие решения о том, верна она или нет, с использованием информации о выборке. На основе информации о выборке мы строим доверительные интервалы, то есть получаем интервал или диапазон для неизвестных характеристик генеральной совокупности.

В разделах 3.1, 3.2 и 3.3 вы найдете обзоры свойств оценок и основных понятий, используемых при тестировании гипотез и построении доверительных интервалов в контексте статистической проверки гипотез о неизвестных средних генеральной совокупности.

Большинство интересных вопросов в экономике связаны с отношениями между двумя или более переменными или сравнениями между различными генеральными совокупностями. Например, существует ли разрыв между доходами мужчин и женщин среди недавно закончивших колледж выпускников. В разделе 3.4 методы, используемые для изучения среднего в генеральной совокупности из разделов 3.1–3.3, обобщаются для сравнения средних в двух различных генеральных совокупностях. В разделе 3.5 обсуждается, как методы для сравнения средних в двух генеральных совокупностях могут быть использованы для оценки причинных эффектов в экспериментах. В разделах 3.2–3.5 основной акцент сделан на использовании нормального распределения для проверки гипотез и для построения доверительных интервалов в случае, когда размер выборки большой. В некоторых особых случаях тестирование гипотез и построение доверительных интервалов могут быть основаны на  $t$ -распределении Стьюдента вместо нормального распределения; эти особые случаи обсуждаются в разделе 3.6. В конце главы (в разделе 3.7) обсуждаются выборочные корреляции и диаграммы рассеяния.

### 3.1. Оценка среднего значения генеральной совокупности

Предположим, что вы хотите знать среднее значение  $Y$  (т.е.  $\mu_Y$ ) в генеральной совокупности, например средние доходы женщин среди недавно закончивших колледж выпускников. Естественный способ оценить это среднее – вычислить выборочное среднее  $\bar{Y}$  из выборки  $n$  независимых и одинаково распределенных наблюдений (i.i.d.),  $Y_1, \dots, Y_n$  (вспомним, что  $Y_1, \dots, Y_n$  является i.i.d., если выборка сформирована простым случайным образом). Этот раздел рассматривает оценку  $\mu_Y$  и свойства  $\bar{Y}$  как оценки  $\mu_Y$ .

#### Оценки и их свойства

**Оценка.** Выборочное среднее  $\bar{Y}$  является естественным способом оценить  $\mu_Y$ , но не единственным. Например, для того чтобы оценить  $\mu_Y$  другим способом, можно просто взять его первое наблюдение  $Y_1$ . Обе эти оценки среднего значения,  $\bar{Y}$  и  $Y_1$ , являются функциями от данных, которые мы используем для оценки  $\mu_Y$ . Используя терминологию, представленную во вставке «Основные понятия 3.1», мы называем их оценками  $\mu_Y$ . Если для оценки  $\mu_Y$  использовать повторные выборки, то  $\bar{Y}$  и  $Y_1$  будут принимать различные значения в зависимости от используемой для оценки среднего выборки. Таким образом, обе оценки  $\bar{Y}$  и  $Y_1$  имеют выборочное распределение. Фактически существует множество оценок  $\mu_Y$ , из которых  $\bar{Y}$  и  $Y_1$  – только два примера.

**Оценки<sup>1</sup>**

*Оценка (an estimator)* – функция от результатов наблюдения (выборки), отобранных случайным образом из генеральной совокупности. *Оценка (an estimate)* – численное значение оценки, полученной по данным из конкретной случайной выборки. Это численное значение (оценка – *an estimate*) является случайной величиной, поскольку получена на основе случайного выделения выборки. В то же время оценка (*an estimator*) не является случайной.

**ОСНОВНЫЕ ПОНЯТИЯ**

3.1

Существует много возможных оценок. Так что же заставляет одну оценку быть «лучше» другой? Поскольку оценки являются случайными величинами, этот вопрос может быть сформулирован более точно: каковы желательные характеристики выборочного распределения оценки? В целом мы хотели бы иметь оценку, которая дает значение настолько близкое к истинному, насколько возможно. Другими словами, мы хотели бы, чтобы выборочное распределение оценки было столь плотно сосредоточено на неизвестном значении, насколько это возможно. Эта потребность приводит к трем конкретным желательным характеристикам оценки: несмещеннность (отсутствие смещения), состоятельность и эффективность.

**Несмешенность.** Предположим, что вы вычисляете оценку много раз по повторяющимся случайным выборкам. В этом случае разумно надеяться, что в среднем вы будете получать правильный ответ. Таким образом, желательным свойством оценки является то, что в среднем она равна  $\mu_Y$ ; если это так, оценка считается несмешенной.

Сформулируем понятие несмешенности математически. Пусть  $\hat{\mu}_Y$  обозначает некоторую оценку  $\mu_Y$ , например  $\bar{Y}$  или  $Y_1$ . Оценка  $\hat{\mu}_Y$  является несмешенной, если  $E(\hat{\mu}_Y) = \mu_Y$ , где  $E(\hat{\mu}_Y)$  – математическое ожидание выборочного распределения  $\hat{\mu}_Y$ ; в противном случае  $\hat{\mu}_Y$  является смещенной.

**Состоятельность.** Другое желательное свойство оценки  $\hat{\mu}_Y$  заключается в том, что в случае наличия выборки большого размера неопределенность относительно значения  $\mu_Y$ , возникающая из-за случайных изменений в выборке, очень мала. Формулируя более точно, скажем, что желательным свойством  $\hat{\mu}_Y$  является то, что вероятность нахождения этой оценки внутри малого интервала истинного значения  $\mu_Y$  приближается к единице при росте выборки, то есть  $\hat{\mu}_Y$  является состоятельной оценкой  $\mu_Y$  (см. вставку «Основные понятия 2.6»).

**Дисперсия и эффективность.** Предположим, что у вас есть две оценки,  $\hat{\mu}_Y$  и  $\tilde{\mu}_Y$ , являющиеся несмешенными. Какую оценку вы могли бы выбрать? Один из способов сделать это заключается в выборе оценки с самым плотно расположенным выборочным распределением. Это предлагает выбор между  $\hat{\mu}_Y$  и  $\tilde{\mu}_Y$ .

<sup>1</sup> В английском языке существует два слова, которые на русский язык переводятся одинаково: «оценка». Это *an estimator* и *an estimate*. Различия между этими понятиями представлены во вставке «Основные понятия 3.1», в которой для лучшего понимания в скобках оставлены английские эквиваленты. – Примеч. науч. ред. перевода.

на основе величин их дисперсий: мы выбираем оценку с наименьшей дисперсией. Если  $\hat{\mu}_y$  имеет меньшую дисперсию, чем  $\tilde{\mu}_y$ , то  $\hat{\mu}_y$  считается более эффективной, чем  $\tilde{\mu}_y$ . Термин «эффективность» означает, что если  $\hat{\mu}_y$  имеет меньшую дисперсию, чем  $\tilde{\mu}_y$ , то эта оценка использует информацию, получаемую из данных, более эффективно, чем  $\tilde{\mu}_y$ .

## ОСНОВНЫЕ ПОНЯТИЯ 3.2

### Смещение, состоятельность и эффективность

Пусть  $\hat{\mu}_y$  является оценкой  $\mu_y$ . Тогда:

- *Смещением*  $\hat{\mu}_y$  называется величина  $E(\hat{\mu}_y) - \mu_y$ .
- $\hat{\mu}_y$  есть *несмешенная оценка*  $\mu_y$ , если  $E(\hat{\mu}_y) = \mu_y$ .
- $\hat{\mu}_y$  есть *состоятельная оценка*  $\mu_y$ , если  $\hat{\mu}_y \xrightarrow{P} \mu_y$ .
- Пусть  $\tilde{\mu}_y$  — другая оценка  $\mu_y$  и предположим, что и  $\hat{\mu}_y$  и  $\tilde{\mu}_y$  являются несмешенными. Тогда  $\hat{\mu}_y$  называется более *эффективной*, чем  $\tilde{\mu}_y$ , если  $\text{var}(\hat{\mu}_y) < \text{var}(\tilde{\mu}_y)$ .

Определения понятий *смещения, состоятельности и эффективности* приведены во вставке «Основные понятия 3.2».

## Свойства $\bar{Y}$

Какими свойствами обладает  $\bar{Y}$  как оценка  $\mu_y$ , когда тестируется по критериям смещенности, состоятельности и эффективности?

**Смещение и состоятельность.** Выборочное распределение  $\bar{Y}$  уже было рассмотрено в разделах 2.5 и 2.6. Как показано в разделе 2.5,  $E(\bar{Y}) = \mu_y$ , поэтому  $\bar{Y}$  является несмешенной оценкой. Аналогично закон больших чисел (см. вставку «Основные понятия 2.6») утверждает, что  $\bar{Y} \xrightarrow{P} \mu_y$ ; то есть  $\bar{Y}$  является состоятельной.

**Эффективность.** Что можно сказать об эффективности? Поскольку эффективность влечет за собой сравнение оценок, мы должны специфицировать оценку или оценки, с которыми  $\bar{Y}$  должна быть сопоставима.

Мы начинаем со сравнения эффективности  $\bar{Y}$  с оценкой  $Y_1$ . Поскольку  $Y_1, \dots, Y_n$  является i.i.d., среднее выборочного распределения  $Y_1$  есть  $E(Y_1) = \mu_y$ ; таким образом,  $Y_1$  является несмешенной оценкой  $\mu_y$ . Ее дисперсия есть  $\text{var}(Y_1) = \sigma_y^2$ . Из раздела 2.5 дисперсия  $\bar{Y}$  — это  $\sigma_y^2 / n$ . Таким образом, для  $n \geq 2$  дисперсия  $\bar{Y}$  меньше, чем дисперсия  $Y_1$ ; то есть  $\bar{Y}$  более эффективная оценка, чем  $Y_1$ , так что, согласно критерию эффективности,  $\bar{Y}$  должна использоваться вместо  $Y_1$ . Оценка  $Y_1$ , очевидно, является плохой. Возникает естественный вопрос: зачем вы потратили столько сил для того, чтобы собрать выборку, состоящую из  $n$  наблюдений, чтобы затем выбросить все наблюдения, кроме первого? С этой точки зрения концепция эффективности предлагает формальный способ показать, что  $\bar{Y}$  является более желательной оценкой, чем  $Y_1$ .

**Эффективность  $\bar{Y}$ :  $\bar{Y}$  является наилучшей линейной несмешенной оценкой (BLUE)**

Пусть  $\hat{\mu}_y$  – оценка  $\mu_y$ , то есть  $\hat{\mu}_y$  является взвешенным средним  $Y_1, \dots, Y_n$  или  $\hat{\mu}_y = (1/n) \sum_{i=1}^n a_i Y_i$ , где  $a_1, \dots, a_n$  – неслучайные константы. Если  $\hat{\mu}_y$  является несмешенной оценкой среднего, то  $\text{var}(\bar{Y}) < \text{var}(\hat{\mu}_y)$  кроме случая, когда  $\hat{\mu}_y = \bar{Y}$ . Таким образом,  $\bar{Y}$  является наилучшей линейной несмешенной оценкой (Best Linear Unbiased Estimator, BLUE); то есть  $\bar{Y}$  является самой эффективной оценкой  $\hat{\mu}_y$  среди всех несмешенных оценок, являющихся средневзвешенными значениями  $Y_1, \dots, Y_n$ .

**ОСНОВНЫЕ ПОНЯТИЯ**  
**3.3**

Что можно сказать о не столь очевидно плохой оценке, чем та, что мы рассмотрели выше? Рассмотрим взвешенное среднее, в котором наблюдения взвешены с чередующимися весами, равными  $\frac{1}{2}$  и  $\frac{3}{2}$ :

$$\tilde{Y} = \frac{1}{n} \left( \frac{1}{2} Y_1 + \frac{3}{2} Y_2 + \frac{1}{2} Y_3 + \frac{3}{2} Y_4 + \dots + \frac{1}{2} Y_{n-1} + \frac{3}{2} Y_n \right), \quad (3.1)$$

где для удобства число наблюдений  $n$  предполагается четным. Математическое ожидание  $\tilde{Y}$  равно  $\mu_y$ , и его дисперсия равна  $\text{var}(\tilde{Y}) = 1,25\sigma_y^2 / n$  (см. упражнение 3.11). Таким образом,  $\tilde{Y}$  является несмешенной оценкой среднего, и поскольку  $\text{var}(\tilde{Y}) \rightarrow 0$  при  $n \rightarrow \infty$ , оценка  $\tilde{Y}$  является состоятельной оценкой. Однако  $\tilde{Y}$  имеет дисперсию большую, чем  $\bar{Y}$ . Таким образом,  $\bar{Y}$  более эффективная оценка среднего, чем  $\tilde{Y}$ .

Оценки  $\bar{Y}$ ,  $\tilde{Y}$  и  $\hat{\mu}_y$  имеют похожую математическую структуру: они являются взвешенными средними значений  $Y_1, \dots, Y_n$ . Сравнение эффективности оценок, проведенное выше, показывает, что взвешенные средние  $\tilde{Y}$  и  $\hat{\mu}_y$  имеют большую дисперсию, чем  $\bar{Y}$ . Фактически эти выводы отражают более общий результат:  $\bar{Y}$  – самая эффективная оценка среди всех несмешенных оценок, являющихся взвешенными средними  $Y_1, \dots, Y_n$ . Иначе говоря,  $\bar{Y}$  является наилучшей линейной несмешенной оценкой (Best Linear Unbiased Estimator – **BLUE**); то есть она – самая эффективная (наилучшая) оценка среди всех оценок, являющихся несмешенными и линейными функциями от  $Y_1, \dots, Y_n$ . Этот результат сформулирован во вставке «Основные понятия 3.3» и доказан в главе 5.

$\bar{Y}$  является оценкой наименьших квадратов  $\mu_y$ . Выборочное среднее  $\bar{Y}$  предлагает наилучшую подгонку к данным в том смысле, что среднее значение квадратов разностей между наблюдениями и  $\bar{Y}$  является наименьшим среди всех возможных оценок.

Рассмотрим проблему нахождения такой оценки  $m$ , которая минимизирует сумму:

$$\sum_{i=1}^n (Y_i - m)^2 \quad (3.2)$$

и является мерой общего квадрата разрыва или расстояния между оценкой  $m$  и точками выборки. Поскольку  $m$  является оценкой  $E(Y)$ , вы можете думать о ней как о прогнозе значения  $Y_i$ , поэтому о разрыве (остатке)  $Y_i - m$  можно думать как об ошибке предсказания (прогноза) или МНК-остатке. Тогда о сумме квадратов остатков в выражении (3.2) можно думать как о сумме квадратов ошибок прогнозов.

Оценка  $m$ , которая минимизирует сумму квадратов разрывов  $Y_i - m$  в выражении (3.2), называется *оценкой наименьших квадратов*. Можно представить, как решается задача наименьших квадратов, используя метод проб и ошибок: значения  $m$  перебираются до тех пор, пока вы не найдете то значение, которое сделает выражение (3.2) настолько малым, насколько возможно. Но можно показать, используя алгебру или дифференциальное исчисление (см. приложение 3.2), что, выбирая  $m = \bar{Y}$ , вы минимизируете сумму квадратов остатков в выражении (3.2), так что  $\bar{Y}$  является оценкой наименьших квадратов  $\mu_Y$ .

### **Важность случайного выбора**

Мы предположили, что  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами, как если бы они были получены простым случайным выбором. Это предположение важно, потому что неслучайный отбор может привести к смещеннности  $\bar{Y}$ . Пусть для того чтобы оценить ежемесячный уровень безработицы, статистическое агентство опрашивает респондентов по следующей схеме: интервьюеры опрашивают взрослых людей трудоспособного возраста, сидящих в городских парках в 10:00 утра во вторую среду месяца. Поскольку многие работающие люди работают в это время (не сидят в парке!), безработные являются чрезмерно представленными в выборке, и оценка уровня безработицы, основанная на такой схеме опроса, будет смещена. Это смещение возникает потому, что при такой схеме выбора безработные представлены избыточно по сравнению с генеральной совокупностью. Рассмотренный пример является искусственным, в то время как вставка «Лэндон выигрывает!» дает реальный пример смещения, полученного в результате неслучайного отбора.



### **Лэндон выигрывает!**

Незадолго до президентских выборов в США в 1936 году *Literary Gazette* опубликовала результаты опроса, из которых следовало, что Аль М. Лэндон победит действующего президента Франклина Д. Рузельта подавляющим числом голосов – 57% против 43%. *Gazette* оказалась права: выборы действительно завершились подавляющим числом голосов, но ошиблась с победителем – Рузельт выиграл, собрав 59% голосов против 41%!

Как *Gazette* могла совершить такую большую ошибку? *Gazette* использовала выборку из респондентов, составленную из данных телефонных справочников и базы данных регистрации автомобилей. Но в 1936 году многие семьи не имели автомобилей или телефонов. А те, которые имели, были, как правило, богаче остальных и, более вероятно, были республиканцами. Поскольку телефонный опрос не был случайной выборкой из всей генеральной совокупности (населения), а вместо этого в ней не были в достаточной мере представлены демократы, оценка была построена на выборке, изначально смещённой в сторону присутствия в ней республиканцев, в результате чего *Gazette* и совершила столь досадную ошибку.

Как вы думаете, могут ли исследования, проведенные через Интернет, иметь похожие проблемы со смещённостью?



Следовательно, важно разработать такие схемы исследований, которые минимизировали бы возможное смещение самой выборки. Приложение 3.1 включает в себя обсуждение того, что на самом деле делает Бюро статистики труда США, когда проводит текущее обследование населения (Current Population Survey, CPS), которое используется для оценки ежемесячного уровня безработицы США.

### **3.2. Тестирование гипотез о среднем значении генеральной совокупности**

Многие гипотезы о мире вокруг нас могут быть сформулированы в виде вопросов, на которые можно ответить «да» или «нет». Равна ли средняя почасовая заработка недавно закончивших колледж США выпускников 20 долл. в час? Является ли средняя заработка платы одинаковой для мужчин и женщин среди выпускников колледжей? Оба вопроса отражают конкретные гипотезы о распределении зарплат в генеральной совокупности. Статистическая задача заключается в том, чтобы ответить на эти вопросы на основании имеющейся выборки. В данном разделе описываются методы *тестирования гипотез* о среднем значении генеральной совокупности. (Равна ли средняя почасовая зарплата генеральной совокупности 20 долл.?) Методы тестирования гипотез о равенстве средних двух генеральных совокупностей (является ли средняя зарплата одинаковой для женщин и мужчин?) рассматриваются в разделе 3.4.

#### ***Нулевая и альтернативная гипотезы***

Отправной точкой при тестировании статистических гипотез является необходимость специфицировать проверяемую (тестируемую) гипотезу, называемую *нулевой гипотезой*. При тестировании гипотез мы используем данные выборки, чтобы сравнить нулевую гипотезу со второй гипотезой, которая называется *альтернативной* и имеет место, если нулевая гипотеза неверна.

При тестировании гипотезы о среднем значении генеральной совокупности нулевая гипотеза заключается в том, что генеральное среднее  $E(Y)$  равно некоторому конкретному значению, обозначаемому  $\mu_{Y,0}$ . Нулевая гипотеза обозначается как  $H_0$  и формально записывается так:

$$H_0 : E(Y) = \mu_{Y,0}. \quad (3.3)$$

Например, предположение о том, что в среднем в генеральной совокупности выпускники колледжей зарабатывают 20 долл. в час, представляет собой нулевую гипотезу о распределении почасовой зарплаты в генеральной совокупности. Говоря математически, если  $Y$  – почасовая заработка плата случайно выбранных недавно закончивших колледж выпускников, тогда нулевой гипотезой является  $E(Y) = 20$ , то есть в выражении (3.3)  $\mu_{Y,0} = 20$ .

Альтернативная гипотеза специфицирует то, что является верным, если нулевая гипотеза неверна. Самой общей альтернативной гипотезой является гипотеза о том, что  $E(Y) \neq \mu_{Y,0}$ , которая называется *двухсторонней альтернативной гипотезой*, поскольку она допускает, что  $E(Y)$  может быть или меньше, или больше, чем  $\mu_{Y,0}$ . Двухсторонняя альтернативная гипотеза записывается так:

$$H_1 : E(Y) \neq \mu_{Y,0}. \quad (3.4)$$

Односторонние альтернативы также возможны, и они обсуждаются далее в этом разделе.

Стоящая перед статистиком проблема заключается в том, чтобы, используя случайную выборку данных, принять (не отвергнуть) нулевую гипотезу  $H_0$  или отвергнуть ее в пользу альтернативной гипотезы  $H_1$ . Если нулевая гипотеза принимается (не отвергается), это не означает, что статистик утверждает, будто это правда; скорее, нулевая гипотеза принимается приблизительно. И мы понимаем, что она может быть отвергнута позже на основе дополнительных свидетельств. По этой причине статистическая проверка гипотез должна ориентироваться на тот факт, что мы можем отвергнуть нулевую гипотезу или же мы не в состоянии сделать это.

### ***p-значение***

В любой выборке выборочное среднее  $\bar{Y}$  редко будет в точности равно гипотетическому значению  $\mu_{Y,0}$ . Различия между  $\bar{Y}$  и  $\mu_{Y,0}$  могут возникать, потому что истинное среднее фактически не равно  $\mu_{Y,0}$  (нулевая гипотеза неверна) или потому что истинное среднее равно  $\mu_{Y,0}$  (нулевая гипотеза верна), но  $\bar{Y}$  отличается от  $\mu_{Y,0}$  из-за случайности выборки. И важно понимать, что эти две ситуации невозможно различить между собой. Хотя выборка данных не может предоставить убедительных доказательств о нулевой гипотезе, возможно выполнить вероятностный расчет, который позволяет тестировать нулевую гипотезу таким способом, который объясняет неопределенность выборки. Этот расчет включает в себя использование данных для вычисления *p-значения* нулевой гипотезы.

*p-значение* или *вероятность значимости* – это минимальная вероятность отвержения нулевой гипотезы на основе имеющейся выборки в предположе-

нии, что она (нулевая гипотеза) верна<sup>1</sup>. Иными словами,  $p$ -значение представляет собой вероятность того, что выборочное среднее  $\bar{Y}$  попадет в хвост распределения случайной величины при верной нулевой гипотезе.

Предположим, например, что в вашей выборке, составленной из выпускников колледжей, средняя зарплата составляет 22,64 долл. в час. Тогда  $p$ -значение есть вероятность того, что это выборочное среднее  $\bar{Y}$  отличается от 20 долл. (т.е. от генерального среднего при нулевой гипотезе), при этом предполагается, что нулевая гипотеза верна. Если  $p$ -значение мало, скажем 0,5 %, тогда очень маловероятно, что выборка получена при верной нулевой гипотезе; таким образом, разумно заключить, что нулевая гипотеза неверна. С другой стороны, если  $p$ -значение велико, скажем, 40 %, тогда достаточно правдоподобным выглядит то, что если нулевая гипотеза верна, то наблюдаемое выборочное среднее, равное 22,64 долл., могло возникнуть только из-за случайности выборки; соответственно, свидетельство против нулевой гипотезы слабо в вероятностном смысле, и разумно не отвергнуть нулевую гипотезу.

Сформулируем определение  $p$ -значения математически. Пусть  $\bar{Y}^{act}$  – значение выборочного среднего, вычисленное по имеющейся выборке, и пусть  $\Pr_{H_0}$  – вероятность, вычисленная в предположении нулевой гипотезы (т.е. вероятность, вычисленная при предположении, что  $E(Y_i) = \mu_{Y,0}$ ). Тогда  $p$ -значение – это

$$p\text{-value} = \Pr_{H_0} \left[ |\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}| \right]. \quad (3.5)$$

То есть  $p$ -значение – это область в хвосте распределения  $\bar{Y}$  при нулевой гипотезе за пределами  $|\bar{Y}^{act} - \mu_{Y,0}|$ . Если  $p$ -значение велико, тогда наблюдаемое значение  $\bar{Y}^{act}$  соответствует нулевой гипотезе, но если  $p$ -значение мало, это не так.

Чтобы вычислить  $p$ -значение, необходимо знать выборочное распределение  $\bar{Y}$  при нулевой гипотезе. Как обсуждалось в разделе 2.6, это распределение является сложным, когда размер выборки мал. Однако согласно центральной предельной теореме, когда размер выборки большой, выборочное распределение  $\bar{Y}$  хорошо аппроксимируется нормальным распределением. При нулевой гипотезе среднее нормального распределения – это  $\mu_{Y,0}$ , так что при нулевой гипотезе  $\bar{Y}$  распределено как  $N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$ , где  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$ . Эта аппроксимация нормальным распределением для больших выборок делает возможным вычисление  $p$ -значения без необходимости знания распределения генеральной совокупности  $Y$ . Детали вычисления, однако, зависят от того, является ли  $\sigma_Y^2$  известной.

### **Вычисление $p$ -значения, когда $\sigma_Y$ известна**

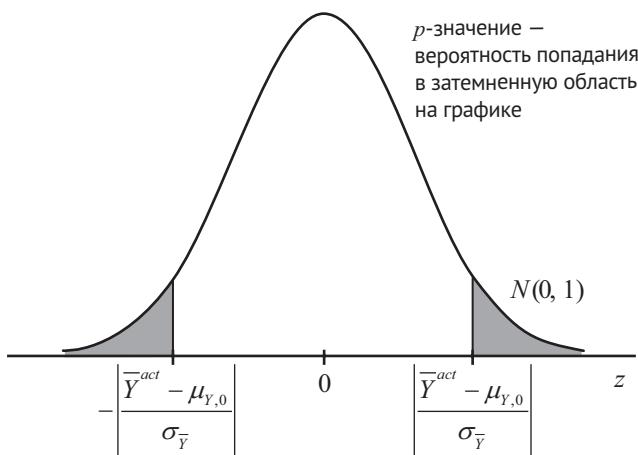
Вычисление  $p$ -значения при известной  $\sigma_Y$  показано на рисунке 3.1. Если размер выборки велик, то при нулевой гипотезе выборочное распределение  $\bar{Y}$  есть  $N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$ , где  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$ . Таким образом, при нулевой гипотезе стандартизованное среднее значение  $\bar{Y}$ , то есть  $(\bar{Y} - \mu_{Y,0}) / \sigma_{\bar{Y}}$ , имеет стандартное

<sup>1</sup>  $P$ -значение часто определяют как вероятность совершения ошибки первого рода, о чем будет сказано позже. – Примеч. науч. ред. перевода.

нормальное распределение.  $P$ -значение – это вероятность получения выборочного среднего значения ( $\bar{Y}$ ), более отдаленного от  $\mu_{Y,0}$ , чем  $\bar{Y}^{act}$  при нулевой гипотезе, или, что эквивалентно, это вероятность получения  $(\bar{Y} - \mu_{Y,0}) / \sigma_{\bar{Y}}$  большего, чем  $(\bar{Y}^{act} - \mu_{Y,0}) / \sigma_{\bar{Y}}$  по абсолютному значению. Эта вероятность равна площади затемненной части под графиком функции плотности вероятности, показанной на рисунке 3.1. Записывая математически, вероятность попадания в затемненную область на рисунке 3.1 (т.е.  $p$ -значение) есть

$$p\text{-value} = \Pr_{H_0} \left( \left| \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \right| > \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \right| \right) = 2\Phi \left( - \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \right| \right), \quad (3.6)$$

где  $\Phi$  – кумулятивная функция стандартного нормального распределения. То есть  $p$ -значение – это область в хвостах стандартного нормального распределения за пределами  $\pm(\bar{Y}^{act} - \mu_{Y,0}) / \sigma_{\bar{Y}}$ .



**Рисунок 3.1. Вычисление  $p$ -значения**

$p$ -значение – это вероятность того, что значение выборочного среднего  $\bar{Y}$  отлично от  $\mu_{Y,0}$ , по крайней мере, на величину большую, чем  $\bar{Y}^{act}$ . На больших выборках  $\bar{Y}$  распределено как  $N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$  в предположении нулевой гипотезы, поэтому  $(\bar{Y} - \mu_{Y,0}) / \sigma_{\bar{Y}}$  распределено как  $N(0, 1)$ . Таким образом,  $p$ -значение есть вероятность попадания в затемненный хвост стандартного нормального распределения за пределами  $\pm |(\bar{Y}^{act} - \mu_{Y,0}) / \sigma_{\bar{Y}}|$ .

Формула для  $p$ -значения в уравнении (3.6) зависит от дисперсии распределения генеральной совокупности  $\sigma^2$ . На практике эта дисперсия обычно неизвестна. (За исключением случая, когда  $Y_i$  бинарная случайная величина, так что имеет распределение Бернуlli, в котором дисперсия определяется нулевой гипотезой; см. уравнение (2.7) и упражнение 3.2.) Поскольку в общем случае  $\sigma^2$  должна быть оценена до вычисления  $p$ -значения, мы сейчас обратимся к проблеме оценивания  $\sigma^2$ .

### Выборочная дисперсия, выборочное стандартное отклонение и стандартная ошибка

Выборочная дисперсия  $s_y^2$  является оценкой дисперсии генеральной совокупности  $\sigma^2$ , выборочное стандартное отклонение  $s_y$  является оценкой стандартного отклонения генеральной совокупности  $\sigma_y$  и стандартная ошибка вы-

борочного среднего  $\bar{Y}$  является оценкой стандартного отклонения выборочного распределения  $\bar{Y}$ .

**Выборочная дисперсия и стандартное отклонение.** Выборочная дисперсия  $s_y^2$ :

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (3.7)$$

Выборочное стандартное отклонение  $s_y$  равно квадратному корню из выборочной дисперсии.

Формула для выборочной дисперсии очень похожа на формулу для дисперсии генеральной совокупности. Дисперсия генеральной совокупности  $E(Y - \mu_y)^2$  есть среднее значение  $(Y - \mu_y)^2$  в генеральной совокупности. Точно также выборочная дисперсия является средним значением  $(Y_i - \bar{Y})^2$ ,  $i = 1, \dots, n$  с двумя модификациями. Во-первых,  $\mu_y$  заменяется на  $\bar{Y}$ , а во-вторых, среднее делится на  $n-1$  вместо  $n$ .

### Стандартная ошибка $\bar{Y}$

Стандартная ошибка  $\bar{Y}$  — это оценка стандартного отклонения  $\bar{Y}$ . Стандартная ошибка  $\bar{Y}$  обозначается  $SE(\bar{Y})$  или  $\hat{\sigma}_{\bar{Y}}$ . Если  $Y_1, \dots, Y_n$  являются i.i.d., то

$$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s_y / \sqrt{n}. \quad (3.8)$$

## ОСНОВНЫЕ ПОНЯТИЯ

3.4

Причина первой модификации — замена  $\mu_y$  на  $\bar{Y}$  — заключается в том, что  $\mu_y$  является неизвестным и, таким образом, должно быть оценено; естественной оценкой  $\mu_y$  является  $\bar{Y}$ . Причина второй модификации — деление на  $n-1$  вместо  $n$  — следует из того, что оценка  $\mu_y$ , используя  $\bar{Y}$ , приводит к смещению вниз в  $(Y_i - \bar{Y})^2$ . Более точно, как показано в упражнении 3.18,  $E[(Y_i - \bar{Y})^2] = [(n-1)/n]\sigma_y^2$ . Таким образом,  $E \sum_{i=1}^n (Y_i - \bar{Y})^2 = nE[(Y_i - \bar{Y})^2] = (n-1)\sigma_y^2$ . Разделив выражение в уравнении (3.7) на  $n-1$  вместо  $n$ , мы корректируем небольшое смещение вниз, и в результате  $s_y^2$  является несмещенной оценкой выборочной дисперсии.

Деление на  $n-1$  вместо  $n$  в уравнении (3.7) называется коррекцией на *степени свободы*: при оценке выборочного среднего мы используем часть имеющейся информации, то есть расходуем одну «степень свободы», так что остается  $n-1$  степень свободы.

**Состоительность выборочной дисперсии.** Выборочная дисперсия является состоятельной оценкой дисперсии генеральной совокупности:

$$s_y^2 \rightarrow \sigma_y^2. \quad (3.9)$$

Другими словами, выборочная дисперсия близка к дисперсии генеральной совокупности с высокой вероятностью, когда  $n$  велико.

Результат, представленный в уравнении (3.9), доказан в приложении 3.3 при предположении о том, что  $Y_1, \dots, Y_n$  являются i.i.d., и  $Y_i$  имеет конечный четвертый момент, то есть  $E(Y_i^4) < \infty$ . На интуитивном уровне причина состоятельности  $s_y^2$  заключается в том, что она является выборочным средним, так что  $s_y^2$  удовлетворяет закону больших чисел. Но для того чтобы  $s_y^2$  удовлетворяла закону больших чисел, сформулированному во вставке «Основные понятия 2.6», случайная величина  $(Y_i - \bar{Y})^2$  должна иметь конечную дисперсию, что в свою очередь означает, что  $E(Y_i^4)$  должно быть конечно; другими словами, случайная величина  $Y_i$  должна иметь конечный четвертый момент.

**Стандартная ошибка  $\bar{Y}$ .** Поскольку стандартное отклонение выборочного распределения  $\bar{Y}$  есть  $\sigma_{\bar{Y}} = \sigma_y / \sqrt{n}$ , уравнение (3.9) обосновывает использование  $s_y / \sqrt{n}$  в качестве оценки для  $\sigma_{\bar{Y}}$ . Оценка для  $\sigma_{\bar{Y}}$ ,  $s_y / \sqrt{n}$  называется *стандартной ошибкой  $\bar{Y}$*  и обозначается  $SE(\bar{Y})$  или  $\hat{\sigma}_{\bar{Y}}$  (знак « $\wedge$ »<sup>1</sup> над символом означает, что он является оценкой  $\sigma_{\bar{Y}}$ ). Понятие «стандартная ошибка  $\bar{Y}$ » рассмотрено во вставке «Основные понятия 3.4».

Когда случайные величины  $Y_1, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами из распределения Бернуlli с вероятностью успеха  $p$ , формула для дисперсии  $\bar{Y}$  упрощается до  $p(1-p)/n$  (см. упражнение 3.2). Формула для стандартной ошибки также имеет простую форму, которая зависит от  $\bar{Y}$  и  $n$ :  $SE(\bar{Y}) = \sqrt{\bar{Y}(1-\bar{Y})/n}$ .

### Вычисление $p$ -значения, когда $\sigma_y$ неизвестна

Так как  $s_y^2$  является состоятельной оценкой  $\sigma_y^2$ ,  $p$ -значение может быть вычислено с помощью замены  $\sigma_{\bar{Y}}$  из уравнения (3.6) на стандартную ошибку  $SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}}$ . То есть когда  $\sigma_y$  неизвестно и случайные величины  $Y_1, \dots, Y_n$  являются i.i.d.,  $p$ -значение вычисляется с использованием формулы:

$$p\text{-value} = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}\right|\right). \quad (3.10)$$

### *t*-статистика

Стандартизованное выборочное среднее  $(\bar{Y} - \mu_{Y,0}) / SE(\bar{Y})$  играет центральную роль в тестировании статистических гипотез и имеет специальное название *–t-статистика* или *t-отношение*:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}. \quad (3.11)$$

В общем случае *тестовая статистика* – это статистика, используемая для проверки гипотез. *t*-статистика является важным примером тестовой статистики.

Распределение *t*-статистики на больших выборках. Когда  $n$  велико, выборочная дисперсия  $s_y^2$  близка к дисперсии генеральной совокупности  $\sigma_y^2$  с высокой вероятностью. Таким образом, распределение *t*-статистики приблизительно та-

<sup>1</sup> В русскоязычном варианте этот символ обычно называют крышкой и говорят «сигма с крышкой» и так далее. – Примеч. науч. ред. перевода.

кое же, как и распределение  $(\bar{Y} - \mu_{Y,0}) / \sigma_{\bar{Y}}$ , которое, в свою очередь, хорошо аппроксимируется стандартным нормальным распределением, когда  $n$  велико, что следует из центральной предельной теоремы (вставка «Основные понятия 2.7»). Соответственно при нулевой гипотезе:

$$t \text{ приблизительно распределена как } N(0,1) \text{ для больших } n. \quad (3.12)$$

Формула для  $p$ -значения в уравнении (3.10) может быть переписана в терминах  $t$ -статистики. Пусть  $t^{act}$  обозначает значение рассчитанной по данным  $t$ -статистики:

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}. \quad (3.13)$$

Тогда когда  $n$  велико,  $p$ -значение может быть вычислено с использованием

$$p\text{-value} = 2\Phi(-|t^{act}|). \quad (3.14)$$

В качестве гипотетического примера предположим, что выборка из  $n = 200$  недавно закончивших колледж выпускников используется для проверки нулевой гипотезы о том, что средняя заработка плата  $E(Y)$  составляет 20 долл. в час. Средняя зарплата по выборке составляет  $\bar{Y}^{act} = 22,64$ , а выборочное стандартное отклонение составляет  $s_y = \$18,14$ . Тогда стандартная ошибка  $\bar{Y}$  есть  $s_y / \sqrt{n} = 18,14 / \sqrt{200} = 1,28$ . Значение  $t$ -статистики есть  $t^{act} = (22,64 - 20) / 1,28 = 2,06$ . Из таблицы 1 приложения следует, что  $p$ -значение равно  $2\Phi(-2,06) = 0,039$ , или 3,9 %. То есть предполагая верной нулевую гипотезу, получаем, что вероятность того, что выборочное среднее отличается от предполагаемого при нулевой гипотезе значения, равна 3,9 %.

**Тестирование гипотез для определенного уровня значимости.** При проведении статистической проверки гипотез вы можете совершить два типа ошибок: вы можете некорректно отвергнуть нулевую гипотезу, когда она верна, или можете не отвергнуть нулевую гипотезу, когда она ложна. Тестирование гипотез может быть выполнено без вычисления  $p$ -значений, если вы готовы указать заранее вероятность, которую вы можете допустить для совершения ошибки первого рода – то есть некорректного отвержения нулевой гипотезы, когда она верна. Если вы выбираете предварительно специфицированную вероятность отвержения нулевой гипотезы, когда она верна (например 5 %), тогда вы будете отвергать нулевую гипотезу, если и только если  $p$ -значение меньше, чем 0,05. Такой подход придает нулевой гипотезе некоторое привилегированное положение, но во многих практических ситуациях он является адекватным.

**Тестирование гипотез, используя фиксированный уровень значимости.** Предположим, было решено, что нулевая гипотеза будет отвергнута, если  $p$ -значение меньше 5 %. Поскольку площадь под хвостом графика функции плотности нормального распределения за пределами  $\pm 1,96$  составляет 5 %, это дает простое правило:

$$\text{Отвергать } H_0, \text{ если } |t^{act}| > 1,96. \quad (3.15)$$

То есть следует отвергать нулевую гипотезу, если абсолютное значение  $t$ -статистики, вычисленное по выборке, больше, чем 1,96. Если  $n$  достаточно

велико, тогда в условиях нулевой гипотезы  $t$ -статистика имеет распределение  $N(0,1)$ . Таким образом, вероятность ошибочного отверждения нулевой гипотезы (отверждения нулевой гипотезы, когда она фактически верна) составляет 5 %.

**ОСНОВНЫЕ  
ПОНЯТИЯ**

**3.5**

**Терминология в тестировании гипотез**

Статистическая проверка гипотез может привести к двум типам ошибок: к *ошибке I рода*, при которой нулевая гипотеза отвергается, когда фактически верна, и к *ошибке II рода*, при которой нулевая гипотеза не отвергается, когда она фактически ложна. Предварительно специфицированная вероятность отверждения нулевой гипотезы, когда она верна — то есть предварительно специфицированная вероятность ошибки I рода — есть *уровень значимости* теста. *Критическое значение* тестовой статистики — это значение статистики, для которой тест отвергает нулевую гипотезу при данном уровне значимости. Множество значений тестовой статистики, для которых тест точно отвергает нулевую гипотезу, называется *областью отверждения*, а множество значений тестовой статистики, для которых он не отвергает нулевую гипотезу, называется *областью принятия (не отверждения)*. Вероятность того что тест фактически некорректно отвергает нулевую гипотезу, когда она верна, называется *размером теста*, а вероятность того что тест корректно отвергает нулевую гипотезу, когда верна альтернативная, называется *мощностью теста*.

*p*-значение является наименьшим уровнем значимости, при котором можно отвергнуть нулевую гипотезу.

Терминология, принятая при тестировании статистических гипотез, имеет некоторые специфические особенности и приведена во вставке «Основные понятия 3.5». Уровень значимости теста в уравнении (3.15) составляет 5 %, критическое значение этого двухстороннего теста есть 1,96, и область отверждения является значениями  $t$ -статистики за пределами  $\pm 1,96$ . Если тест отвергает на 5 %-м уровне значимости, среднее генеральной совокупности  $\mu_Y$  называют статистически значимо отличной от  $\mu_{Y,0}$  на 5 %-м уровне значимости.

Тестирование гипотез, используя предварительно специфицированный уровень значимости, не требует вычисления *p*-значений. В предыдущем примере тестирования гипотезы о том, что средняя зарплата недавно закончивших колледж выпускников составляет 20 долл. в час,  $t$ -статистика составила 2,06. Это значение превышает 1,96, поэтому гипотеза отвергается на 5 %-м уровне значимости. Хотя проверка нулевой гипотезы на 5 %-м уровне значимости менее

затратна с вычислительной точки зрения, информация о том, отвергается ли нулевая гипотеза для специфицированного уровня значимости, меньше, чем в случае с  $p$ -значением.

**Какой уровень значимости вы должны использовать на практике?** Во многих случаях статистики и эконометристы используют 5 %-й уровень значимости. Если вы тестируете много статистических гипотез на 5 %-м уровне значимости, вы можете некорректно отвергнуть нулевую гипотезу в среднем один раз из 20 случаев. Иногда более консервативный уровень значимости может быть более адекватным. Например, судебные дела иногда включают в себя статистические свидетельства, и нулевая гипотеза заключается в том, что подсудимый невиновен; тогда хотелось бы быть уверенным, что отвержение нулевой гипотезы (т.е. вывод о том, что обвиняемый виновен) является не только результатом изменчивости случайной выборки. В некоторых юридических ситуациях используется уровень значимости 1 % или даже 0,1 %, чтобы избежать этого вида ошибки. Аналогично, если государственный орган рассматривает разрешение на продажу новых лекарств, очень консервативный стандарт может быть более подходящим, чтобы потребители были уверены, что лекарства, доступные на рынке, действительно работают.

Тест является консервативным в том смысле, что использует очень низкий уровень значимости, но при этом имеет издержки: чем ниже уровень значимости, тем более высоко критическое значение, и становится сложнее отвергнуть нулевую гипотезу, если она ложна. Фактически самая консервативная вещь, которую можно сделать, – это никогда не отвергать нулевую гипотезу, но если вы это делаете, у вас нет необходимости обращать внимание на любые статистические факты, для того чтобы никогда не изменить свое мнение! Чем ниже уровень значимости, тем ниже мощность теста. Многие экономические и политические практики могут быть менее консервативными, чем судебное дело, так что 5 %-й уровень значимости часто рассматривается как разумный компромисс.

**Тестирование гипотезы**  $E(Y) = \mu_{Y,0}$  **против альтернативы**

$$E(Y) \neq \mu_{Y,0}$$

- 1) Вычислите стандартную ошибку  $\bar{Y}$ ,  $SE(\bar{Y})$  [уравнение (3.8)].
- 2) Вычислите  $t$ -статистику [уравнение (3.13)].
- 3) Вычислите  $p$ -значение [уравнение (3.13)]. Отвергайте нулевую гипотезу на 5 %-м уровне значимости, если  $p$ -значение меньше, чем 0,05 (эквивалентно, если  $|t^{act}| > 1,96$ ).

### ОСНОВНЫЕ ПОНЯТИЯ

3.6

Во вставке «Основные понятия 3.6» приводится схема тестирования гипотезы о генеральном среднем против двухсторонней альтернативы.

## Односторонние альтернативы

В некоторых обстоятельствах альтернативная гипотеза может заключаться в том, что среднее превышает  $\mu_{Y,0}$ . Например, хочется надеяться, что образование помогает на рынке труда, поэтому соответствующая альтернатива для нулевой гипотезы о том, что доходы являются одинаковыми для выпускников колледжей и тех, кто колледж не оканчивал, заключается не только в том, что их доходы отличаются, но и в том, что выпускники зарабатывают больше невыпускников. Это называется *односторонней альтернативной гипотезой*, которая может быть записана так:

$$H_1 : E(Y) > \mu_{Y,0}. \quad (3.16)$$

Общий подход к вычислению  $p$ -значения и тестирования гипотез для односторонних альтернатив тот же самый, что и для двухсторонних альтернатив, с тем отличием, что только высокие положительные значения  $t$ -статистики отвергают нулевую гипотезу, а не значения, которые являются высокими по абсолютной величине. В частности, для проверки односторонней гипотезы в уравнении (3.16) строится  $t$ -статистика в уравнении (3.13).  $p$ -значением является область под графиком функции плотности стандартного нормального распределения справа от вычисленной  $t$ -статистики. То есть  $p$ -значение, основанное на  $N(0,1)$  — аппроксимации распределения  $t$ -статистики, — имеет вид:

$$p\text{-value} = \Pr_{H_0}(Z > t^{act}) = 1 - \Phi(t^{act}). \quad (3.17)$$

Критическое значение для одностороннего теста с 5 %-м уровнем значимости составляет 1,64. Область отверждения для этого теста — это все значения  $t$ -статистики, превышающие 1,64.

Односторонняя гипотеза в уравнении (3.16) соответствует значению  $\mu_Y$ , превышающему  $\mu_{Y,0}$ . Если вместо этого альтернативная гипотеза состоит в том, что  $E(Y) < \mu_{Y,0}$ , то схема остается той же, за исключением изменения знаков; например, 5 %-я область отверждения состоит из значений  $t$ -статистики, меньших чем -1,64.

### 3.3. Доверительные интервалы для среднего генеральной совокупности

Из-за ошибки случайного отбора невозможно узнать точное значение среднего генеральной совокупности, используя только информацию о выборке. Однако возможно использовать данные из случайной выборки, чтобы построить множество значений, содержащих истинное среднее генеральной совокупности  $\mu_Y$  с определенной заданной вероятностью. Такое множество называется *доверительным множеством*, а предварительно специфицированная вероятность того, что  $\mu_Y$  содержится в этом множестве, называется *уровнем доверия*. Доверительное множество для  $\mu_Y$  содержит все возможные значения среднего

между нижней и верхней границами, так что доверительное множество – это интервал, называемый *доверительным интервалом*.

Рассмотрим один из способов построения 95 %-го доверительного множества для среднего генеральной совокупности. Начните с выбора некоторого произвольного значения для среднего; назовем его  $\mu_{Y,0}$ . Тестируйте нулевую гипотезу отом, что  $\mu_Y = \mu_{Y,0}$  против альтернативы отом, что  $\mu_Y \neq \mu_{Y,0}$ , вычисляя *t*-статистику; если она меньше, чем 1,96, это гипотетическое значение  $\mu_{Y,0}$  не отвергается на 5 %-м уровне значимости, и запишите это неотвергающееся значение  $\mu_{Y,0}$ . Теперь выберите другое произвольное значение  $\mu_{Y,0}$  и тестируйте его. Если вы не можете отвергнуть его, запишите это значение в ваш список. Повторяйте процедуру снова и снова для всех возможных значений генерального среднего. Продолжая этот процесс, получим множество всех значений генерального среднего, которые не могут быть отвергнуты на 5 %-м уровне значимости двухсторонним тестом.

Полученный список является полезным, поскольку он обобщает множество гипотез, которые вы можете или не можете отвергнуть (на 5 %-м уровне значимости) на основе имеющихся данных: если кто-то подходит к вам с конкретным вопросом о равенстве генерального среднего, вы можете сказать ему, отвергается ли его гипотеза или нет, просто взглянув на соответствующее число в вашем списке. Немного порассуждав, можно увидеть, что это множество значений обладает замечательным свойством: вероятность того, что оно содержит истинное значение среднего генеральной совокупности, составляет 95 %.

Эти рассуждения выглядят следующим образом. Предположим, что истинное значение  $\mu_Y = 21,5$  (хотя мы не знаем этого). Тогда  $\bar{Y}$  имеет нормальное распределение, центрированное в 21,5, и *t*-статистика, тестирующая нулевую гипотезу  $\mu_Y = 21,5$ , имеет  $N(0,1)$ -распределение. Таким образом, если  $n$  велико, вероятность отвержения нулевой гипотезы  $\mu_Y = 21,5$  на 5 %-м уровне значимости составляет 5 %. Но так как вы тестировали все возможные значения генерального среднего при построении вашего множества, то вы, в частности, тестировали истинное значение  $\mu_Y = 21,5$ . В 95 % всех выборок вы будете корректно принимать 21,5; это означает, что в 95 % всех выборок ваш список будет содержать действительное значение  $\mu_Y$ . Таким образом, значения в вашем списке составляют 95 %-е доверительное множество для  $\mu_Y$ .

Этот метод построения доверительного множества непрактичен, поскольку он требует от вас проверить все возможные значения  $\mu_Y$  в качестве нулевой гипотезы. К счастью, существует более легкий подход. Согласно формуле для *t*-статистики в уравнении (3.13), гипотеза о равенстве выборочного среднего пробному значению  $\mu_{Y,0}$  отвергается на 5 %-м уровне значимости, если оно отклоняется от  $\bar{Y}$  больше, чем на стандартную ошибку выборочного среднего  $\bar{Y}$ , умноженную на 1,96. Таким образом, множество значений  $\mu_Y$ , которые не отвергаются на 5 %-м уровне значимости, состоит из значений, отклоняющихся от  $\bar{Y}$  на величину, не большую чем  $\pm 1,96SE(\bar{Y})$ . То есть 95 %-й доверительный интервал для  $\mu_Y$  есть  $\bar{Y} - 1,96SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1,96SE(\bar{Y})$ . Во вставке «Основные понятия 3.7» описан этот подход.

**ОСНОВНЫЕ ПОНЯТИЯ****3.7****Доверительные интервалы для генерального среднего**

95 %-й двухсторонний доверительный интервал для  $\mu_y$  — это интервал, построенный так, чтобы он содержал истинное значение  $\mu_y$  в 95 % всех возможных случайных выборках. Когда размер выборки  $n$  велик, 95, 90 и 99 %-е доверительные интервалы для  $\mu_y$  есть

$$95\text{-}\text{й доверительный интервал для } \mu_y = \{\bar{Y} \pm 1,96SE(\bar{Y})\}.$$

$$90\text{-}\text{й доверительный интервал для } \mu_y = \{\bar{Y} \pm 1,64SE(\bar{Y})\}.$$

$$99\text{-}\text{й доверительный интервал для } \mu_y = \{\bar{Y} \pm 2,58SE(\bar{Y})\}.$$

Как пример, рассмотрим задачу построения 95 %-го доверительного интервала для средней почасовой зарплаты недавно закончивших колледж выпускников, используя гипотетическую случайную выборку из 200 недавно закончивших колледж выпускников, где  $\bar{Y} = \$22,64$  и  $SE(\bar{Y}) = 1,28$ . Тогда 95 %-й доверительный интервал для средней почасовой зарплаты составит  $22,64 \pm 1,96 \times 1,28 = 22,64 \pm 2,51 = [\$20,13; \$25,15]$ .

Обсуждение до сих пор было сосредоточено на построении двухсторонних доверительных интервалов. Вместо этого можно было бы построить односторонний доверительный интервал в качестве множества значений  $\mu_y$ , которые не могут быть отвергнуты односторонним тестом. Хотя односторонние доверительные интервалы применяются в некоторых отраслях статистики, они являются редкостью в прикладном эконометрическом анализе.

**Вероятность попадания.** Вероятность попадания среднего значения генеральной совокупности в доверительный интервал — это вероятность, с которой доверительный интервал содержит истинное среднее генеральной совокупности, рассчитанная по всем возможным случайным выборкам.

### 3.4. Сравнение средних значений различных генеральных совокупностей

Существуют ли различия в средних заработных платах мужчин и женщин, недавно закончивших колледж? Этот вопрос касается сравнения средних значений двух различных генеральных совокупностей. В этом разделе описывается, как тестировать гипотезы и как строить доверительные интервалы для разности средних двух различных генеральных совокупностей.

#### Тестирование гипотез для разности между двумя средними

Проиллюстрируем *тест для разности между двумя средними*. Пусть  $\mu_w$  будет средней почасовой заработной платой в генеральной совокупности женщин, недавно закончивших колледж, и пусть  $\mu_m$  будет средним генеральной совокупности для недавно закончивших колледж мужчин. Рассмотрим нулевую гипоте-

зу о том, что средние зарплаты для этих двух генеральных совокупностей отличаются на определенную величину, скажем,  $d_0$ . Тогда нулевая гипотеза и двухсторонняя альтернативная гипотеза имеют вид:

$$H_0 : \mu_m - \mu_w = d_0 \text{ против } H_1 : \mu_m - \mu_w \neq d_0. \quad (3.18)$$

Нулевая гипотеза о том, что мужчины и женщины в этих генеральных совокупностях имеют одинаковые средние зарплаты, соответствует  $H_0$  в уравнении (3.18) с  $d_0 = 0$ .

Поскольку эти средние генеральных совокупностей неизвестны, они должны оцениваться по выборкам мужчин и женщин. Предположим, что мы имеем выборки  $n_m$  мужчин и  $n_w$  женщин, выбранных случайно из соответствующих генеральных совокупностей. Пусть выборочная средняя зарплата составляет  $\bar{Y}_m$  для мужчин и  $\bar{Y}_w$  для женщин. Тогда оценкой  $\mu_m - \mu_w$  является  $\bar{Y}_m - \bar{Y}_w$ .

Чтобы тестировать нулевую гипотезу о том, что  $\mu_m - \mu_w = d_0$ , используя  $\bar{Y}_m - \bar{Y}_w$ , мы должны знать распределение  $\bar{Y}_m - \bar{Y}_w$ . Вспомним, что  $\bar{Y}_m$ , согласно центральной предельной теореме, приблизительно распределена как  $N(\mu_m, \sigma_m^2 / n_m)$ , где  $\sigma_m^2$  – дисперсия зарплат генеральной совокупности мужчин. Аналогично  $\bar{Y}_w$  приблизительно распределено как  $N(\mu_w, \sigma_w^2 / n_w)$ , где  $\sigma_w^2$  – дисперсия зарплат генеральной совокупности для женщин. Вспомним также из раздела 2.4, что взвешенное среднее двух нормально распределенных случайных величин является нормально распределенной случайной величиной. Поскольку  $\bar{Y}_m$  и  $\bar{Y}_w$  строятся, используя случайные выборки из различных генеральных совокупностей, они являются независимыми случайными величинами. Таким образом,  $\bar{Y}_m - \bar{Y}_w$  распределено как  $N[\mu_m - \mu_w, (\sigma_m^2 / n_m) + (\sigma_w^2 / n_w)]$ .

Если  $\sigma_m^2$  и  $\sigma_w^2$  известны, тогда это распределение, будучи приближенно нормальным, может использоваться для вычисления  $p$ -значений для тестиования нулевой гипотезы о том, что  $\mu_m - \mu_w = d_0$ . На практике, однако, эти дисперсии генеральных совокупностей обычно неизвестны, так что они должны быть оценены. Как и ранее, они могут быть оценены с использованием выборочных дисперсий  $s_m^2$  и  $s_w^2$ , где  $s_m^2$  определено как уравнение (3.7), за исключением того, что статистика вычислена только для мужчин в выборке, и  $s_w^2$  определяется аналогично для женщин. Таким образом, стандартная ошибка  $\bar{Y}_m - \bar{Y}_w$  – это:

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}. \quad (3.19)$$

Для упрощенной версии уравнения (3.19), когда  $Y$  является бернуlliевской случайной величиной, см. упражнение 3.15.

Тогда  $t$ -статистика для тестиования нулевой гипотезы о разности генеральных средних строится аналогично  $t$ -статистике для тестиования гипотезы о единственном среднем генеральной совокупности путем вычитания гипотетического значения  $\mu_m - \mu_w$ , тестируемого при нулевой гипотезе, из оценки  $\bar{Y}_m - \bar{Y}_w$  и деления результата на стандартную ошибку  $\bar{Y}_m - \bar{Y}_w$ :

$$t = \frac{(\bar{Y}_m - \bar{Y}_w) - d_0}{SE(\bar{Y}_m - \bar{Y}_w)}. \quad (3.20)$$

Если  $n_m$  и  $n_w$  велики, тогда  $t$ -статистика имеет стандартное нормальное распределение.

Поскольку  $t$ -статистика в уравнении (3.20) имеет стандартное нормальное распределение, когда  $n_m$  и  $n_w$  велики,  $p$ -значение двухстороннего теста вычисляется точно также, как было в рассмотренном выше случае единственной генеральной совокупности. То есть соответствующее  $p$ -значение вычисляется с использованием уравнения (3.14).

Для проведения теста с предварительно специфицированным уровнем значимости просто вычислите  $t$ -статистику в уравнении (3.20) и сравните ее с соответствующим критическим значением. Например, нулевая гипотеза отвергается на 5 %-м уровне значимости, если абсолютное значение  $t$ -статистики превышает 1,96.

Если альтернатива односторонняя, а не двухсторонняя (т.е. если альтернатива заключается в том, что  $\mu_m - \mu_w > d_0$ ), тогда тест модифицируется, как описано в разделе 3.2. В этом случае  $p$ -значение вычисляется с использованием уравнения (3.17), и нулевая гипотеза отвергается на 5 %-м уровне значимости, если  $t > 1,64$ .

### ***Доверительные интервалы для разности между двумя средними генеральных совокупностей***

Метод построения доверительных интервалов, описанный в разделе 3.3, обобщается на случай доверительного интервала для разности между средними  $d = \mu_m - \mu_w$ . Поскольку гипотетическое значение  $d_0$  отвергается на 5 %-м уровне значимости, если  $|t| > 1,96$ ,  $d_0$  будет находиться в доверительном множестве, если  $|t| \leq 1,96$ . Но выполнение неравенства  $|t| \leq 1,96$  означает, что оцененная разность  $\bar{Y}_m - \bar{Y}_w$  отличается от  $d_0$  на величину, меньшую чем 1,96 стандартных ошибки разности выборочных средних. Таким образом, 95 %-й двухсторонний доверительный интервал для  $d$  состоит из значений  $d$ , отклоняющихся от разности выборочных средних на величину  $\pm 1,96$  стандартных ошибки  $\bar{Y}_m - \bar{Y}_w$  или меньше:

95 %-й доверительный интервал для  $d = \mu_m - \mu_w$  имеет вид:

$$(\bar{Y}_m - \bar{Y}_w) \pm 1,96SE(\bar{Y}_m - \bar{Y}_w). \quad (3.21)$$

Вставка «Гендерный разрыв в заработных платах выпускников колледжей в Соединенных Штатах» содержит результаты эмпирических исследований гендерных различий в доходах американских выпускников колледжей.

### ***3.5. Оценка причинных эффектов при помощи разности средних значений, используя экспериментальные данные***

Вспомним из раздела 1.2, что случайный управляемый эксперимент случайным образом выбирает субъекты (индивидуумов или, в более общем случае, организаций) из интересующей генеральной совокупности, затем случайным образом распределяет их либо в экспериментальную группу, которая получает

экспериментальное воздействие, либо в контрольную группу, которая не получает воздействия. Разница между выборочными средними экспериментальной и контрольной групп является оценкой влияния условий эксперимента (наличия или отсутствия воздействия).

### **Причинный эффект как разность условных математических ожиданий**

Причинный эффект условий эксперимента – это ожидаемое влияние на исход эксперимента интересующего нас воздействия, измеряемое в идеальном случайном управляемом эксперименте. Это влияние может быть выражено как разность двух условных математических ожиданий. Более конкретно, *причинный эффект* влияния уровня  $x$  на случайную величину  $Y$  есть разность в условных математических ожиданиях:  $E(Y|X=x) - E(Y|X=0)$ , где  $E(Y|X=x)$  является ожидаемым значением  $Y$  для исследуемой группы (которая получает воздействие уровня  $X = x$ ) в идеальном случайном управляемом эксперименте, а  $E(Y|X=0)$  – это ожидаемое значение  $Y$  в контрольной группе (которая получает уровень воздействия  $X = 0$ ). В контексте экспериментов причинный эффект также называется *эффектом воздействия (условий эксперимента)*. Если существует только два уровня воздействия (т.е. если воздействие является бинарным), то мы можем считать  $X = 0$  для контрольной группы и  $X = 1$  для исследуемой группы. Если воздействие является бинарным, тогда причинный эффект (т.е. эффект воздействия) – это  $E(Y|X=1) - E(Y|X=0)$  в идеальном случайном управляемом эксперименте.

### **Оценка причинного эффекта при помощи разности средних значений**

Если воздействие в случайном управляемом эксперименте является бинарным, то причинный эффект может быть оценен разностью выборочных средних исходов в экспериментальной и контрольной группах. Гипотеза о том, что воздействие неэффективно, эквивалентна гипотезе о том, что два средних одинаковы, что может быть проверено с использованием  $t$ -статистики для сравнения двух средних, заданной в уравнении (3.20). 95 %-й доверительный интервал для разности средних двух групп есть 95 %-й доверительный интервал для причинного эффекта. Таким образом, 95 %-й доверительный интервал для причинного эффекта может быть построен с использованием уравнения (3.21).

Хорошо продуманный, хорошо управляемый эксперимент может дать убедительные оценки причинного эффекта. По этой причине случайные управляемые эксперименты обычно проводятся в таких областях, как медицина. В экономике, однако, эксперименты, как правило, дороги, сложны в управлении и, в некоторых случаях, этически сомнительны, поэтому они все еще редки. По этой причине эконометристы иногда изучают «естественные эксперименты», также называемые квазиэксперименты, в которых некоторое событие, не связанное с воздействием или характеристикой субъекта, имеет назначение различных видов воздействия на различные субъекты, как если бы они были частью случайного управляемого эксперимента. Вставка «Новый способ увеличения пенсионных

накоплений» представляет собой пример такого квазиэксперимента, который привел к некоторым неожиданным выводам.



### ***Гендерный разрыв в заработных plataх выпускников колледжей в Соединенных Штатах***

Вставка «Распределение заработных плат в США в 2008 году» из главы 2 показывает, что в среднем выпускники колледжей мужского пола зарабатывают больше, чем выпускники колледжей женского пола. Каковы последние тенденции в этом «гендерном разрыве» в доходах? Социальные нормы и законы, регулирующие половую дискриминацию на рабочем месте, существенно изменились в Соединенных Штатах. Является ли гендерный разрыв в зарплатах выпускников колледжей стабильным и сокращается ли с течением времени?

Таблица 3.1 дает оценки почасовой зарплаты для 25–34-летних людей, работающих полный рабочий день и имеющих высшее образование, в Соединенных Штатах в 1992, 1996, 2000, 2004 и 2008 годах на основе данных из текущего обследования населения. Зарплаты для 1992, 1996, 2000 и 2004 годов были скорректированы на инфляцию в ценах 2008 года с использованием индекса потребительских цен (ИПЦ)<sup>1</sup>. В 2008 году средняя почасовая зарплата 1832 опрошенных мужчин составляла 24,98 долл., и стандартное отклонение зарплат мужчин было 11,78 долл. Средняя почасовая зарплата в 2008 году для 1871 опрошенных женщин составляла 20,87 долл., а стандартное отклонение — 9,66 долл. Таким образом, оценка гендерного разрыва в зарплатах для 2008 года составляла 4,11 долл. ( $= \$24,98 - \$20,87$ ) со стандартным отклонением 0,35 долл. ( $= \sqrt{11,78^2 / 1838 + 9,66^2 / 1871}$ ). 95%-й доверительный интервал для гендерного разрыва в зарплатах в 2008 году составляет  $4,11 \pm 1,96 \times 0,35 = (\$3,41; \$4,80)$ .

Из результатов, представленных в таблице 3.1, можно сделать четыре вывода. Во-первых, гендерный разрыв велик. Почасовой разрыв в 4,11 долл. может выглядеть не слишком большим, но за год он добавляет 8220 долл. с учетом 40-часовой рабочей недели и 50 оплачиваемых недель в году. Во-вторых, с 1992 до 2008 год оцененный гендерный разрыв возрастил на 0,87 долл. в час в реальном выражении, это возрастание незначимо статистически на 5%-м уровне значимости (упражнение 3.17). В-третьих, разрыв велик, если он измеряется в процентах: согласно оценкам, представленным в таблице 3.1, в 2008 году женщины зарабатывали в час на 16% меньше, чем мужчины (4,11 долл./24,98 долл.), немного больше, чем разрыв в 14%, имеющий место в 1992 году (3,22 долл./23,27 долл.). В-четвертых, гендерный разрыв меньше для молодых выпускников (группа, рассмотренная в таблице 3.1), чем для всех выпускников колледжей (рассмотренных в таблице 2.4). Как показано в таблице 2.4, средняя зарплата для всех женщин, закончивших колледж, работающих полный рабочий день, в 2008 году составляла 23,93 долл., в то время как для мужчин средняя зарплата была 30,97 долл., которая соответствует гендерному разрыву в 23% [ $= (30,97 - 23,93) / 30,97$ ] среди всех работающих полный рабочий день выпускников колледжей.

Эмпирический анализ показывает, что «гендерный разрыв» в почасовой зарплате высок и был достаточно стабильным (или, возможно, несколько увеличи-

вался) в недавнем прошлом. Анализ не говорит нам, однако, почему этот разрыв высок. Возрастает ли он из-за половой дискриминации на рынке труда? Означает ли он различие в квалификации, опыте и образовании между мужчинами и женщинами? Означает ли он различие в выборе работы? Или существуют некоторые другие причины? Мы вернемся к этим вопросам, когда у нас в руках будут инструменты анализа множественной регрессии, рассматриваемые во второй части.

Таблица 3.1

**Изменения почасовых зарплат работающих выпускников колледжей  
25–34 лет в США с 1992 по 2008 год в долларах 2008 года**

Год	Мужчины			Женщины			Разница между мужчинами и женщинами		95%-й доверительный интервал для $d$
	$\bar{Y}_m$	$s_m$	$n_m$	$\bar{Y}_w$	$s_w$	$n_w$	$\bar{Y}_m - \bar{Y}_w$	$SE(\bar{Y}_m - \bar{Y}_w)$	
1992	23,27	10,17	1594	20,05	7,87	1368	3,22**	0,33	2,58–3,88
1996	22,48	10,10	1379	18,98	7,95	1230	3,50**	0,35	2,80–4,19
2000	24,88	11,60	1303	20,74	9,36	1181	4,14**	0,42	3,32–4,97
2004	25,12	12,01	1894	21,02	9,36	1735	4,10**	0,36	3,40–4,80
2008	24,98	11,78	1838	20,87	9,66	1871	4,11**	0,35	3,41–4,80

*Примечание:* Эти оценки вычислены на основе данных обо всех работающих полный рабочий день 25–34-летних работников из текущего обследования населения, проведенного в марте следующего года (например, данные для 2008 года собраны в марте 2009 года). «\*\*» означают, что разность значимо отличается от нуля на 1%-м уровне значимости

<sup>1</sup> Из-за инфляции доллар в 1992 году стоил больше, чем доллар в 2008, в том смысле что на 1 долл. в 1992 году можно было купить больше товаров и услуг, чем на 1 доллар в 2008. Таким образом, зарплаты в 1992 году не могут непосредственно сравниваться с зарплатами в 2008 без корректировки на инфляцию. Один из способов сделать это заключается в использовании ИПЦ, меры цен «рыночной корзины» потребительских товаров и услуг, построенного Бюро статистики труда. За 16 лет, с 1992 по 2008, цена потребительской корзины по ИПЦ выросла на 53,4%, другими словами, корзина товаров и услуг по ИПЦ, которая стоила 100 долл. в 1992 году, стала стоить 153,40 долл. в 2008. Для того чтобы сопоставить зарплаты в 1992 и 2008 годах в таблице 3.1, зарплаты 1992 года завышены в соответствии с ростом стоимости потребительской корзины, то есть путем умножения доходов 1992 года на 1,534, чтобы перевести их в «доллары 2008 года».



### 3.6. Использование $t$ -статистики при малом размере выборки

В разделах 3.2–3.5  $t$ -статистика в сочетании с критическими значениями из стандартного нормального распределения используется для тестирования гипотез и построения доверительных интервалов. Использование стандартного нормального распределения оправдано центральной предельной теоремой, которая применяется, если размер выборки велик. Если размер выборки мал, стандартное нормальное распределение может дать плохую аппроксимацию распределения  $t$ -статистики. Если, однако, само распределение генеральной совокупности нормально, тогда точное распределение (т.е. распределение

на конечных выборках; см. раздел 2.6)  $t$ -статистики для тестирования гипотезы о среднем единственной генеральной совокупности является  $t$ -распределением Стьюдента с  $n-1$  степенью свободы, и критические значения могут быть взяты из  $t$ -распределения Стьюдента.

### ***t*-статистика и *t*-распределение Стьюдента**

***t*-статистика для тестирования гипотезы о среднем.** Рассмотрим  $t$ -статистику, используемую для того, чтобы проверить гипотезу о том, что среднее  $Y$  есть  $\mu_{Y,0}$ , используя данные  $Y_1, \dots, Y_n$ . Формула для этой статистики задается уравнением (3.11), где стандартная ошибка выборочного среднего  $\bar{Y}$  задается уравнением (3.8). Подстановка второго выражения в первое приводит к формуле для  $t$ -статистики:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{\sqrt{s_Y^2 / n}}, \quad (3.22)$$

где  $s_Y^2$  задается уравнением (3.7).

Как обсуждалось в разделе 3.2, при выполнении общих условий  $t$ -статистика имеет стандартное нормальное распределение, если размер выборки велик и нулевая гипотеза верна [см. уравнение (3.12)]. Хотя аппроксимация стандартным нормальным распределением является надежной для широкого класса распределений случайной величины  $Y$  при больших  $n$ , она может быть плохой при малых  $n$ . Точное распределение  $t$ -статистики зависит от распределения случайной величины  $Y$  и может быть очень сложным. Существует, однако, один частный случай, в котором точное распределение  $t$ -статистики является относительно простым: если случайная величина  $Y$  распределена нормально, то  $t$ -статистика в уравнении (3.22) имеет  $t$ -распределение Стьюдента с  $n-1$  степенью свободы.

Чтобы проверить этот результат, вспомним из раздела 2.4, что  $t$ -распределение Стьюдента с  $n-1$  степенью свободы определяется как распределение  $Z / \sqrt{W / (n-1)}$ , где  $Z$  – случайная величина со стандартным нормальным распределением,  $W$  – случайная величина с хи-квадрат распределением с  $n-1$  степенью свободы, а  $Z$  и  $W$  независимо распределены. Когда  $Y_1, \dots, Y_n$  есть i.i.d. и распределение генеральной совокупности  $Y$  есть  $N(\mu_Y, \sigma_Y^2)$ , то  $t$ -статистика может быть записана как такое отношение. Более точно, пусть  $Z = (\bar{Y} - \mu_{Y,0}) / \sqrt{\sigma_Y^2 / n}$  и пусть  $W = (n-1)s_Y^2 / \sigma_Y^2$ ; тогда алгебраические преобразования<sup>1</sup> показывают, что  $t$ -статистика в уравнении (3.22) может быть записана как  $t = Z / \sqrt{W / (n-1)}$ . Вспомним из раздела 2.4, что если  $Y_1, \dots, Y_n$  являются i.i.d. и распределение генеральной совокупности  $Y$  является  $N(\mu_Y, \sigma_Y^2)$ , то выборочное распределение  $\bar{Y}$  является точным  $N(\mu_Y, \sigma_Y^2 / n)$  для всех  $n$ ; таким образом, если нулевая гипотеза  $\mu_Y = \mu_{Y,0}$  корректна, то  $Z = (\bar{Y} - \mu_{Y,0}) / \sqrt{\sigma_Y^2 / n}$  имеет стандартное нормальное распределение для всех  $n$ . В дополнение  $W = (n-1)s_Y^2 / \sigma_Y^2$

<sup>1</sup> Требуемое выражение получается умножением и делением на  $\sqrt{\sigma_Y^2}$  и соответствующими преобразованиями:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{\sqrt{s_Y^2 / n}} = \frac{(\bar{Y} - \mu_{Y,0})}{\sqrt{\sigma_Y^2 / n}} \div \sqrt{\frac{s_Y^2}{\sigma_Y^2}} = \frac{(\bar{Y} - \mu_{Y,0})}{\sqrt{\sigma_Y^2 / n}} \div \sqrt{\frac{(n-1)s_Y^2 / \sigma_Y^2}{n-1}} = Z \div \sqrt{W / (n-1)}.$$

имеет распределение  $\chi^2_{n-1}$  для всех  $n$  и  $\bar{Y}$  и  $s_y^2$  независимо распределены. Отсюда следует, что если распределение генеральной совокупности  $Y$  является нормальным, тогда при нулевой гипотезе  $t$ -статистика, заданная в уравнении (3.22), имеет точное  $t$ -распределение Стьюдента с  $n-1$  степенью свободы.

Если распределение генеральной совокупности является нормальным, то критические значения из  $t$ -распределения Стьюдента могут быть использованы для проверки гипотез и построения доверительных интервалов. В качестве примера рассмотрим гипотетическую задачу, в которой  $t^{act} = 2,15$  и  $n = 20$ , так что количество степеней свободы  $n-1=19$ . Из таблицы 2 приложения следует, что 5 %-е двухстороннее критическое значение для распределения  $t_{19}$  есть 2,09. Поскольку  $t$ -статистика больше по абсолютному значению, чем критическое значение ( $2,15 > 2,09$ ), нулевая гипотеза будет отвергаться на 5 %-м уровне значимости против двухсторонней альтернативы. 95 %-й доверительный интервал для  $\mu_y$ , построенный с использованием распределения  $t_{19}$ , будет равен  $\bar{Y} \pm 2,09SE(\bar{Y})$ . Этот доверительный интервал несколько более широкий, чем доверительный интервал, построенный с использованием критических значений стандартного нормального распределения.

***t*-статистика для тестирования гипотезы о разности средних.**  $t$ -статистика для тестирования гипотезы о разности двух средних, заданная в уравнении (3.20), не распределена согласно  $t$ -распределению Стьюдента, даже если распределение генеральной совокупности  $Y$  является нормальным.  $t$ -распределение Стьюдента не применяется здесь, потому что оценка дисперсии, используемая для вычисления стандартной ошибки в уравнении (3.19), не приводит к знаменателю в  $t$ -статистике с распределением хи-квадрат.

Модифицированная версия  $t$ -статистики для разности средних, основанная на другой формуле стандартной ошибки – формула стандартной ошибки для «пула» или формула объединенной стандартной ошибки<sup>1</sup>, – имеет точное  $t$ -распределение Стьюдента, когда  $Y$  нормально распределено; однако формула стандартного распределения для пула применяется только в специальных случаях, когда две группы имеют одинаковую дисперсию, или в сообразных выборках, когда группы имеют одинаковое число наблюдений (упражнение 3.21). Примем обозначения из уравнения (3.19), так чтобы две группы были обозначены как  $m$  и  $w$ . Оценка объединенной дисперсии ( $s_{pooled}^2$ ) разности средних равна:

$$s_{pooled}^2 = \frac{1}{n_m + n_w - 2} \left[ \sum_{\substack{i=1 \\ \text{в группе } m}}^{n_m} (\bar{Y}_i - \bar{Y}_m)^2 + \sum_{\substack{i=1 \\ \text{в группе } w}}^{n_w} (\bar{Y}_i - \bar{Y}_w)^2 \right], \quad (3.23)$$

где первое суммирование выполняется для наблюдений в группе  $m$ , а второе суммирование – для наблюдений в группе  $w$ . Стандартная ошибка для пула (объединения) разности средних есть  $SE_{pooled}(\bar{Y}_m - \bar{Y}_w) = s_{pooled} \times \sqrt{1/n_m + 1/n_w}$ , а соответствующая  $t$ -статистика вычисляется с использованием уравнения (3.20), где стандартная ошибка является стандартной ошибкой для пула, то есть  $SE_{pooled}(\bar{Y}_m - \bar{Y}_w)$ .

<sup>1</sup> В данном случае мы используем такое название – формула стандартной ошибки для «пула» или формула объединенной стандартной ошибки, – потому что для расчета стандартной ошибки мы используем обе непересекающиеся выборки, то есть весь их пул или объединение. – Примеч. науч. ред. перевода.

Если распределение генеральной совокупности  $Y$  в группе  $m$  имеет вид  $N(\mu_m, \sigma_m^2)$ , а распределение генеральной совокупности  $Y$  в группе  $w - N(\mu_w, \sigma_w^2)$ , и если дисперсии двух групп одинаковы (т.е.  $\sigma_m^2 = \sigma_w^2$ ), тогда в условиях нулевой гипотезы  $t$ -статистика, вычисленная с использованием стандартной ошибки, имеет  $t$ -распределение Стьюдента с  $n_m + n_w - 2$  степенями свободы.



### **Новый способ увеличения пенсионных накоплений**

Многие экономисты считают, что пенсионные накопления людей недостаточны. Традиционные методы стимулирования пенсионных накоплений сосредоточены на финансовых стимулах, но был всплеск интереса к нетрадиционным способам стимулирования пенсионных накоплений.

Важное исследование опубликовано в 2001 году. В нем Бриджит Мэдриан (Brigitte Madrian) и Деннис Ши (Dennis Shea) рассмотрели один такой нетрадиционный метод стимулирования пенсионных накоплений. Многие фирмы предлагают сбалансированный способ пенсионных накоплений, в рамках которого они полностью или частично обеспечивают соответствие будущих пенсионных выплат тем взносам, которые делают наемные работники, участвующие в программе. Пенсионные взносы в таких планах, называемых после внесения соответствующих изменений в налоговое законодательство США планами 401 (k), всегда являются добровольными и необязательными. Однако в некоторых фирмах сотрудники автоматически зачисляются в план, хотя они могут и отказаться, а в других фирмах сотрудники зачисляются, только если они подтвердили свое участие в программе. Согласно обычным экономическим моделям поведения, метод регистрации – отказ или согласие – не должен иметь значения: рациональный работник вычисляет оптимальное действие, а затем принимает его. Но Мэдриан и Ши задаются вопросом, может ли в традиционной экономике быть не так? Влияет ли *способ включения* в программу пенсионных накоплений на охват работников этой программой?

Для оценки влияния метода регистрации Мэдриан и Ши изучили большую фирму, которая изменила значение, применяемое по умолчанию, для своей программы 401 (k) от неучастия к участию. Они сравнивали две группы работников: нанятых за год до изменения, которым зачисления не производились автоматически (но они могли согласиться на участие в программе), и нанятых в течение года после изменения, которым зачисления делались автоматически (но они могли отказаться от участия в программе). Финансовые условия программы остались прежними, и Мэдриан и Ши не нашли систематических отличий между работниками, нанятыми до и после изменения. Таким образом, с точки зрения эконометристов, изменение играло роль случайного воздействия, и причинный эффект изменения мог быть оценен как разность средних между двумя группами.

Мэдриан и Ши обнаружили, что правило включения в программу участия по умолчанию повлекло большие различия в уровне охвата: уровень охвата для группы тех, кого включали в программу по предварительному согласию (контрольная группа), был равен 37,4% ( $n = 4249$ ), тогда как уровень охвата для исследуемой

группы (для тех, кого включали в программу автоматически) был равен 85,9% ( $n = 5801$ ). Оценка эффекта воздействия составила 48,5% (= 85,9% – 37,4%). Поскольку рассматриваемая выборка велика, 95%-й доверительный интервал (вычисляемый в упражнении 3.15) для эффекта воздействия довольно узок, от 46,8 до 50,2%.

Как выбор опции автоматического включения в программу участия мог иметь такое значение? Может быть, работники считали, что размышления об участии в программе влекут слишком много издержек, или, может быть, они просто не хотели думать о старении. Ни одно из объяснений не является экономически рациональным – но оба согласуются с расширяющейся областью «поведенческой экономики», и оба могли бы привести к принятию опции регистрации участия в программе по умолчанию.

Это исследование сделало важный практический вклад. В августе 2006 года Конгресс США принял закон о пенсионной защите, который (среди прочих) призвал фирмы предлагать план 401 (k), в котором регистрация выполнялась бы по умолчанию. Эконометрические результаты Мэдриан и Ши и другие исследования заняли видное место в числе доводов в пользу этой части законодательства.

Для того чтобы лучше ознакомиться с поведенческой экономикой и тем, как устроены планы пенсионных накоплений, см.: Bernartzi, Thaler (2007) и Beshears, Choi, Laibson, Madrian (2008).



Недостаток использования оценки дисперсии в пуле ( $s_{pooled}^2$ ) заключается в том, что она применяется только в том случае, если дисперсии двух генеральных совокупностей совпадают (предполагая, что  $n_m \neq n_w$ ). Если дисперсии генеральных совокупностей различные, оценка дисперсии в пуле смещена и несостоительна. Если дисперсии генеральных совокупностей различные, но используется формула для дисперсии в пуле, распределение  $t$ -статистики в пуле при нулевой гипотезе не является распределением Стьюдента, даже если данные нормально распределены; фактически оно не будет нормальным даже в больших выборках. Поэтому стандартные ошибки и  $t$ -статистики в пуле не должны использоваться, если отсутствует веская причина верить в то, что дисперсии генеральных совокупностей одинаковые.

### **Использование $t$ -распределения Стьюдента на практике**

Для задачи тестирования среднего значения случайной величины  $Y$   $t$ -распределение Стьюдента применяется, если распределение генеральной совокупности  $Y$  является нормальным. Для экономических переменных, однако, нормальное распределение является исключением (для примера см. вставки «Распределение заработных плат в США в 2008 году» и «Черный день Уолл-стрит» в главе 2). Даже если лежащие в основе данные не распределены нормально, нормальная аппроксимация распределения  $t$ -статистики правомерна при большом размере выборки. Поэтому статистическая проверка – тестирование гипотез и построение доверительных интервалов – о среднем распределении должно основываться на нормальной аппроксимации на больших выборках.

Когда сравниваются два средних, любая экономическая причина для двух групп, имеющих различные средние, обычно предполагает, что две группы также могут иметь различные дисперсии. Соответственно, формула для стандартной ошибки в пуле не является корректной, и корректная формула для стандартной ошибки, которая допускает различные дисперсии в группах, приведена в уравнении (3.19). Даже если распределение генеральной совокупности нормально,  $t$ -статистика, вычисленная с использованием формулы стандартной ошибки из уравнения (3.19), не имеет  $t$ -распределения Стьюдента. Поэтому на практике статистическая проверка гипотезы о разности средних должна быть основана на уравнении (3.19), использованном в сочетании со стандартной нормальной аппроксимацией для больших выборок.

Хотя  $t$ -распределение Стьюдента редко применяется в экономике, некоторые программные пакеты используют  $t$ -распределение Стьюдента для вычисления  $p$ -значений и доверительных интервалов. На практике это не представляет проблемы, поскольку разница между  $t$ -распределением Стьюдента и стандартным нормальным распределением незначительна, если размер выборки большой. Для  $n > 15$  различие в  $p$ -значениях, вычисленных с использованием  $t$ -распределения Стьюдента и стандартного нормального распределения, никогда не превышает 0,01; для  $n > 80$  различие никогда не превышает 0,002. В большинстве современных приложений и во всех приложениях в этом учебнике размер выборки составляет от сотен до тысяч, что достаточно велико для того, чтобы различие между  $t$ -распределением Стьюдента и стандартным нормальным распределением было незначительным.

### **3.7. Диаграммы рассеяния, выборочная ковариация и выборочная корреляция**

Какая взаимосвязь между возрастом и зарплатами? Этот вопрос, как и многие другие, связывает одну переменную,  $X$  (возраст), с другой,  $Y$  (зарплаты). В данном разделе рассматриваются три способа отразить отношения между переменными: диаграмма рассеяния, выборочная ковариация и выборочный коэффициент корреляции.

#### **Диаграмма рассеяния**

Диаграмма рассеяния – это график  $n$  наблюдений  $X_i$  и  $Y_i$ , где каждое наблюдение представлено точкой  $(X_i, Y_i)$ . Например, рисунок 3.2 является диаграммой рассеяния возраста ( $X$ ) и почасовой зарплаты ( $Y$ ) в выборке из 200 менеджеров в отрасли информационных технологий в марте 2009 года (данные CPS). Каждая точка на рисунке 3.2 соответствует паре  $(X, Y)$  для одного наблюдения. Например, одному из работников в этой выборке 40 лет и он зарабатывает 35,78 долл. в час; этот возраст работника и зарплата указаны выделенной точкой на рисунке 3.2. Диаграмма рассеяния показывает положительную связь между возрастом и зарплатами в выборке: работники старшего возраста, как правило, зарабатывают больше, чем молодые. Это отношение неточно, однако и зарплаты не могут предсказываться точно, используя только возраст человека.

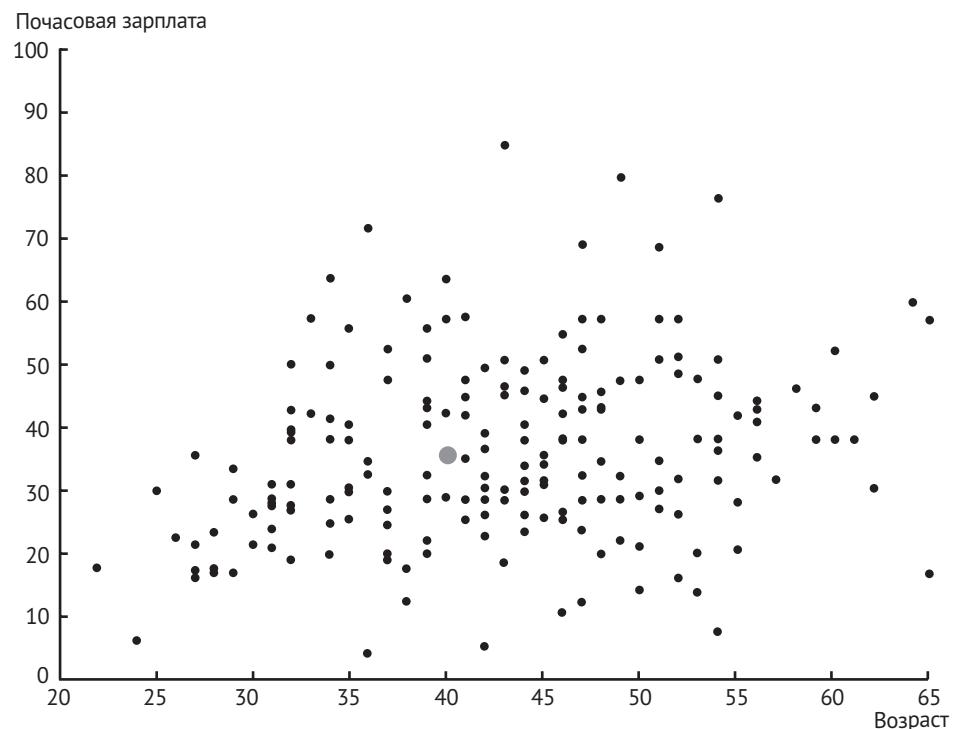
## Выборочные ковариация и корреляция

Понятия ковариации и корреляции случайных величин введены в разделе 2.3 как две характеристики совместного распределения вероятностей случайных величин  $X$  и  $Y$ . Поскольку распределение генеральной совокупности неизвестно, на практике мы не знаем ковариацию и корреляцию генеральной совокупности. Ковариация и корреляция генеральной совокупности могут, однако, быть оценены, если рассматривать случайную выборку из  $n$  элементов генеральной совокупности и собирать данные в  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ .

Выборочные ковариация и корреляция являются оценками ковариации и корреляции генеральной совокупности. Как и оценки, обсуждаемые ранее в этой главе, они вычисляются путем замены генерального среднего (математическое ожидание) на выборочное среднее. *Выборочная ковариация*, обозначаемая  $s_{XY}$ , равна:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}). \quad (3.24)$$

Как и для случая выборочной дисперсии, среднее в уравнении (3.24) вычислено путем деления на  $n-1$  вместо  $n$ ; здесь различие также происходит из-за использования  $\bar{X}$  и  $\bar{Y}$ , оценивающих соответствующие генеральные средние. При больших  $n$  не имеет значения, на что мы делим: на  $n$  или на  $n-1$ .



**Рисунок 3.2. Диаграмма рассеяния почасовой зарплаты в зависимости от возраста**

Каждая точка на графике представляет возраст и среднюю зарплату одного из 200 работников в выборке. Выделенная серым точка соответствует 40-летнему работнику, который зарабатывает 35,78 долл. в час. Данные для менеджеров компьютеров и информационных систем с марта 2009 года (данные CPS).

*Выборочный коэффициент корреляции* или *выборочная корреляция*, обозначаемые  $r_{XY}$ , представляют собой отношение выборочной ковариации к произведению выборочных стандартных отклонений:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \quad (3.25)$$

Выборочная корреляция измеряет силу линейной связи между  $X$  и  $Y$  в выборке из  $n$  наблюдений. Как и корреляция генеральной совокупности, выборочная корреляция безразмерна и лежит между  $-1$  и  $1$ :  $|r_{XY}| \leq 1$ .

Выборочная корреляция равна  $1$ , если  $X_i = Y_i$  для всех  $i$  и равна  $-1$ , если  $X_i = -Y_i$  для всех  $i$ . В общем случае корреляция равна  $\pm 1$ , если диаграмма рассеяния является прямой линией. Если линия наклонена вверх, тогда существует положительная связь между  $X$  и  $Y$ , и корреляция равна  $1$ . Если наклон линии вниз, тогда существует отрицательная связь и корреляция равна  $-1$ . Чем ближе диаграмма рассеяния к прямой линии, тем ближе корреляция к  $\pm 1$ . Высокий коэффициент корреляции не обязательно означает, что линия имеет крутой наклон; скорее, это означает, что точки на диаграмме рассеяния находятся очень близко к прямой линии.

**Состоятельность выборочных ковариации и корреляции.** Как и выборочная дисперсия, выборочная ковариация является состоятельной оценкой ковариации, то есть:

$$s_{XY} \xrightarrow{P} \sigma_{XY}. \quad (3.26)$$

Другими словами, на больших выборках выборочная ковариация с высокой вероятностью близка к ковариации генеральной совокупности.

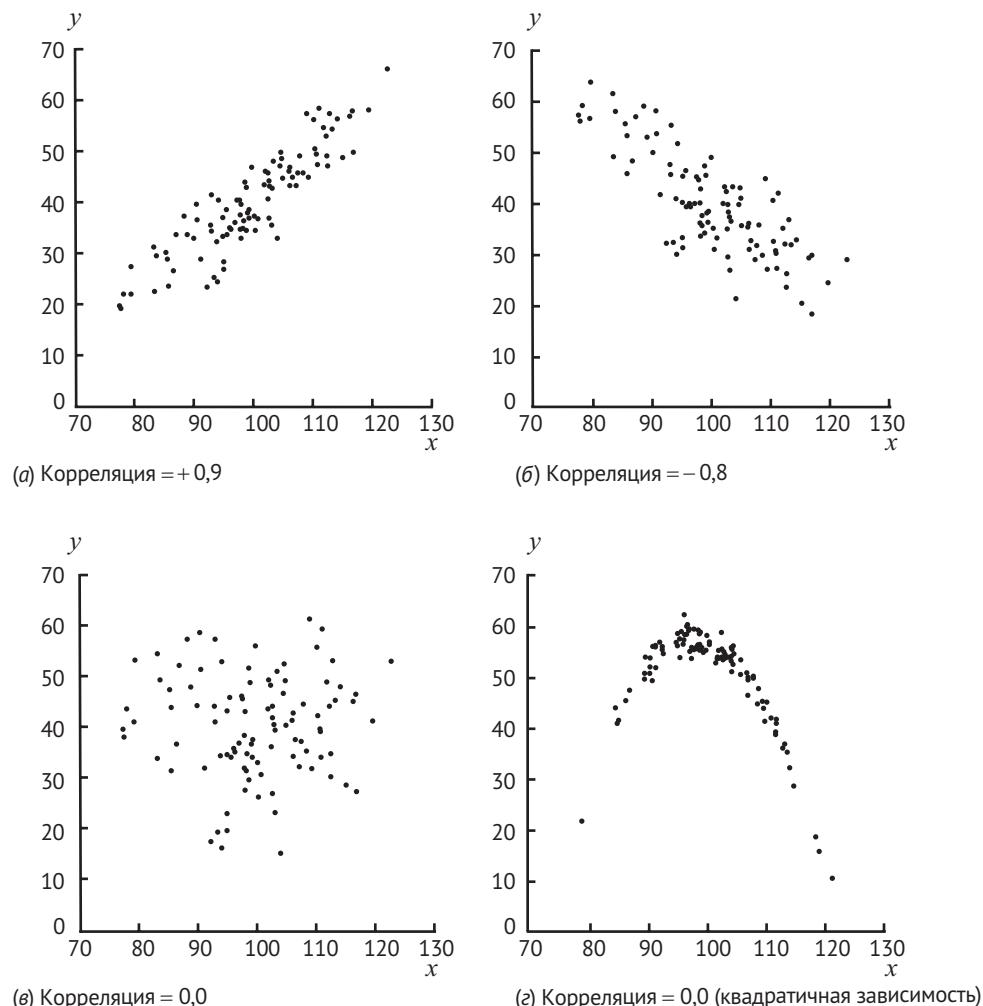
Доказательство результата в уравнении (3.26) при предположениях о том, что  $(X_i, Y_i)$  являются i.i.d. и  $X_i$  и  $Y_i$  имеют конечные четвертые моменты, аналогично доказательству того, что выборочная ковариация состоятельна, приведенному в приложении 3.3, и дается как упражнение 3.20.

Поскольку выборочная дисперсия и выборочная ковариация представляют собой состоятельные оценки, выборочный коэффициент корреляции также является состоятельной оценкой корреляции, то есть  $r_{XY} \xrightarrow{P} \text{corr}(X_i, Y_i)$ .

**Пример.** В качестве примера рассмотрим данные по возрасту и зарплате, приведенные на рисунке 3.2. Для этих 200 работников выборочное стандартное отклонение возраста составило  $s_A = 9,07$  лет и выборочное стандартное отклонение зарплат составило  $s_E = 14,37$  долл. в час. Ковариация между возрастом и зарплатами составила  $s_{AE} = 33,16$  (единиц лет  $\times$  долл. в час, не интерпретируемых легко). Таким образом, коэффициент корреляции равен  $r_{AE} = 33,16 / (9,07 \times 14,37) = 0,25$ , или  $25\%$ . Корреляция  $0,25$  означает, что существует положительная связь между возрастом и зарплатами, но, как видно на диаграмме рассеяния, это соотношение далеко от совершенства.

Чтобы проверить, что корреляция не зависит от единиц измерений, предположим, что зарплаты приводятся в центах, в этом случае выборочное стандартное отклонение зарплат составит 1437 долл. в час, а выборочная ковариация между возрастом и зарплатами будет 3316 (единиц лет  $\times$  центов в час); тогда корреляция составит  $3316 / (9,07 \times 1437) = 0,25$ , или  $25\%$ .

На рисунке 3.3 приведены дополнительные примеры диаграмм рассеяния и корреляции. Рисунок 3.3а показывает сильную положительную линейную связь между переменными, и выборочная корреляция равна 0,9. Рисунок 3.3б показывает сильную отрицательную связь с выборочным коэффициентом корреляции, равным -0,8. На рисунке 3.3в приведена диаграмма рассеяния с отсутствием свидетельства связи, и выборочная корреляция равна нулю. Рисунок 3.3г показывает ясную связь: при увеличении  $X$  —  $Y$  сначала увеличивается, но затем уменьшается. Несмотря на эту заметную связь между  $X$  и  $Y$ , выборочная корреляция равна нулю; причина заключается в том, что для этих данных малые значения  $Y$  связаны как с большими, так и с малыми значениями  $X$ .



**Рисунок 3.3. Диаграмма рассеяния для четырех гипотетических наборов данных**

Диаграммы рассеяния на рисунках 3.3а и 3.3б показывают сильную линейную связь между  $X$  и  $Y$ . На рисунке 3.3в  $X$  не зависит от  $Y$ , и две переменные не коррелированы. На рисунке 3.3г две переменные также не коррелированы, хотя они имеют нелинейную связь.

Этот последний пример подчеркивает важный момент: коэффициент корреляции является мерой линейной связи. Связь между случайными величинами на рисунке 3.3г существует, но она не является линейной.

## Выходы

1. Выборочное среднее  $\bar{Y}$  является оценкой среднего значения генеральной совокупности  $\mu_Y$ . Когда  $Y_1, \dots, Y_n$  являются i.i.d.,
  - а) выборочное распределение  $\bar{Y}$  имеет среднее  $\mu_Y$  и дисперсию  $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$ ;
  - б)  $\bar{Y}$  является несмешенной оценкой генерального среднего;
  - в) по закону больших чисел  $\bar{Y}$  является состоятельной оценкой;
  - г) по центральной предельной теореме  $\bar{Y}$  имеет приблизительно нормальное выборочное распределение, если размер выборки велик.
2.  $t$ -статистика используется для тестирования нулевой гипотезы о том, что среднее генеральной совокупности принимает определенное значение. Если  $n$  велико,  $t$ -статистика имеет стандартное нормальное выборочное распределение в предположении, что нулевая гипотеза верна.
3.  $t$ -статистика может быть использована для вычисления  $p$ -значений, связанных с нулевой гипотезой. Малое  $p$ -значение можно рассматривать как свидетельство того, что нулевая гипотеза неверна.
4. 95 %-й доверительный интервал для  $\mu_Y$  – это интервал, построенный так, чтобы он содержал истинное значение  $\mu_Y$  в 95 % всех возможных выборок.
5. Тестирование гипотез и построение доверительных интервалов для различия средних двух генеральных совокупностей концептуально похожи на тестирование гипотез и построение доверительных интервалов для среднего единственной генеральной совокупности.
6. Выборочный коэффициент корреляции – это оценка корреляции генеральной совокупности, измеряющая линейную связь между двумя переменными, то есть того, насколько сильно их диаграмма рассеяния приближается к прямой линии.

## Основные понятия

Оценка (estimator) (с. 69).

Оценка (estimate) (с. 69).

Смещение, состоятельность и эффективность (с. 70).

BLUE (наилучшая линейная несмешенная оценка) (с. 71).

Оценка наименьших квадратов (с. 72).

Тестирование гипотез (с. 73).

Нулевая гипотеза (с. 73).

Альтернативная гипотеза (с. 73).

Двухсторонняя альтернативная гипотеза (с. 74).

$p$ -значение (вероятность значимости) (с. 74).

Выборочная дисперсия (с. 77).

Выборочное стандартное отклонение (с. 77).

Степени свободы (с. 77).  
Стандартная ошибка  $\bar{Y}$  (с. 78).  
 $t$ -статистика ( $t$ -отношение) (с. 78).  
Тестовая статистика (с. 78).  
Ошибка I рода (с. 80).  
Ошибка II рода (с. 80).  
Уровень значимости (с. 80).  
Критическое значение (с. 80).  
Область отвержения (с. 80).  
Область принятия (с. 80).  
Размер теста (с. 80).  
Мощность теста (с. 80).  
Односторонняя альтернативная гипотеза (с. 82).  
Доверительное множество (с. 82).  
Уровень доверия (с. 82).  
Доверительный интервал (с. 83).  
Вероятность попадания (с. 84).  
Тест для разности между двумя средними (с. 84).  
Причинный эффект (с. 87).  
Эффект воздействия (условий эксперимента) (с. 87).  
Диаграмма рассеяния (с. 94).  
Выборочная ковариация (с. 95).  
Выборочный коэффициент корреляции (выборочная корреляция) (с. 96).

### **Вопросы для повторения и закрепления основных понятий**

- 3.1. Объясните различие между выборочным средним  $\bar{Y}$  и средним генеральной совокупности.
- 3.2. Объясните различие между оценкой (an estimator) и оценкой (an estimate). Приведите пример каждой из них.
- 3.3. Распределение генеральной совокупности имеет среднее 10 и дисперсию 16. Определите среднее и дисперсию  $\bar{Y}$  для случая i.i.d. выборки из этой генеральной совокупности для (а)  $n = 10$ ; (б)  $n = 100$ ; (в)  $n = 1000$ . Соотнесите ваши ответы с законом больших чисел.
- 3.4. Какую роль играет центральная предельная теорема в статистической проверке гипотез? В построении доверительных интервалов?
- 3.5. В чем различие между нулевой и альтернативной гипотезой? между размером теста (критерия), уровнем значимости и мощностью? между односторонней альтернативной гипотезой и двухсторонней альтернативной гипотезой?
- 3.6. Почему доверительный интервал содержит больше информации, чем результат единственной проверки гипотезы?
- 3.7. Объясните, почему оценка разности средних, применяемая к данным из случайного управляемого эксперимента, является оценкой эффекта воздействия (условий эксперимента)?

- 3.8. Нарисуйте схематически гипотетическую диаграмму рассеяния выборки, состоящей из 10 наблюдений, для двух случайных величин с корреляцией в генеральной совокупности, равной (a) 1,0; (б) -1,0; (в) 0,9; (г) -0,5; (д) 0,0.

## Упражнения

- 3.1. Известно, что в генеральной совокупности  $\mu_Y = 100$  и  $\sigma_Y^2 = 43$ . Используйте центральную предельную теорему, чтобы ответить на следующие вопросы:
- В случайной выборке размера  $n = 100$  найдите  $\Pr(\bar{Y} < 101)$ .
  - В случайной выборке размера  $n = 64$  найдите  $\Pr(101 < \bar{Y} < 103)$ .
  - В случайной выборке размера  $n = 165$  найдите  $\Pr(\bar{Y} > 98)$ .
- 3.2. Пусть  $Y$  – случайная величина Бернулли с вероятностью успеха  $\Pr(Y = 1) = p$ , и пусть случайные  $Y_1, \dots, Y_n$  есть i.i.d. из этого распределения. Пусть  $\hat{p}$  – доля успехов (т.е. число единиц) в этой выборке.
- Покажите, что  $\hat{p} = \bar{Y}$ .
  - Покажите, что  $\hat{p}$  является несмешенной оценкой  $p$ .
  - Покажите, что  $\text{var}(\hat{p}) = p(1-p)/n$ .
- 3.3. В ходе опроса из 400 потенциальных избирателей 215 ответили, что они будут голосовать за действующего президента, а 185 ответили, что будут голосовать за его соперника. Пусть  $p$  обозначает долю всех потенциальных избирателей, которые предпочитают действующего президента на время опроса, и пусть  $\hat{p}$  – доля респондентов, которые предпочитают действующего президента.
- Используйте результаты опроса для оценки  $p$ .
  - Используйте оценку дисперсии  $\hat{p}, \hat{p}(1-\hat{p})/n$ , чтобы вычислить стандартную ошибку вашей оценки.
  - Каково  $p$ -значение для теста  $H_0 : p = 0,5$  против  $H_1 : p \neq 0,5$ ?
  - Каково  $p$ -значение для теста  $H_0 : p = 0,5$  против  $H_1 : p > 0,5$ ?
  - Почему результаты в (в) и (г) отличаются?
  - Содержит ли опрос статистически значимое свидетельство того, что действующий президент опередил соперника на момент опроса? Объясните.
- 3.4. Используйте данные в упражнении 3.3:
- Постройте 95 %-й доверительный интервал для  $p$ .
  - Постройте 99 %-й доверительный интервал для  $p$ .
  - Почему интервал в (б) более широкий, чем интервал в (а)?
  - Без выполнения каких-либо дополнительных вычислений проверьте гипотезу о том, что  $H_0 : p = 0,50$  против  $H_1 : p \neq 0,50$  на 5 %-м уровне значимости.
- 3.5. Проводится опрос 1055 зарегистрированных избирателей, и их просят выбрать между кандидатом  $A$  и кандидатом  $B$ . Пусть  $p$  обозначает долю голосующих в генеральной совокупности, которые предпочитают кандидата  $A$ , и пусть  $\hat{p}$  обозначает долю голосующих в выборке, которые предпочитают кандидата  $A$ .

- a) Вы заинтересованы в тестировании конкурирующих гипотез  $H_0 : p = 0,5$  против  $H_1$ . Предположим, что вы решаете отвергнуть  $H_0$ , если  $|\hat{p} - 0,5| > 0,02$ .
- (i) Какой размер этого теста?
  - (ii) Вычислите мощность этого теста, если  $p = 0,53$ .
- б) В опросе  $\hat{p} = 0,54$ .
- (i) Протестируйте гипотезу о том, что  $H_0 : p = 0,5$  против  $H_1 : p \neq 0,5$ , используя 5 %-й уровень значимости.
  - (ii) Протестируйте гипотезу о том, что  $H_0 : p = 0,5$  против  $H_1 : p > 0,5$ , используя 5 %-й уровень значимости.
  - (iii) Постройте 95 %-й доверительный интервал для  $p$ .
  - (iv) Постройте 99 %-й доверительный интервал для  $p$ .
  - (v) Постройте 50 %-й доверительный интервал для  $p$ .
- в) Предположим, что опрос проводился 20 раз, используя независимо выбранных голосующих в каждом опросе. Для каждого из этих 20 опросов построен 95 %-й доверительный интервал для  $p$ .
- (i) Какова вероятность того, что истинное значение  $p$  содержится во всех 20 доверительных интервалах?
  - (ii) Сколько этих доверительных интервалов, как вы ожидаете, содержат истинное значение  $p$ ?
- г) На жаргоне опроса, «погрешность» – это  $1,96 \times SE(\hat{p})$ ; то есть это половина длины 95 %-го доверительного интервала. Предположим, что вы хотите создать опрос, который имел погрешность не более 1 %. То есть вы хотите, чтобы  $\Pr(|\hat{p} - p| > 0,01) \leq 0,05$ . Насколько большим должно быть  $n$ , если опрос использует простую случайную выборку?
- 3.6. Пусть  $Y_1, \dots, Y_n$  есть i.i.d., выбранные из распределения со средним значением  $\mu$ . Тест  $H_0 : \mu = 5$  против  $H_1 : \mu \neq 5$ , используя обычную  $t$ -статистику, приводит к  $p$ -значению, равному 0,03.
- а) Содержит ли 95 %-й доверительный интервал  $\mu = 5$ ? Объясните.
  - б) Можете ли вы определить, содержит ли  $\mu = 6$  в доверительном интервале? Объясните.
- 3.7. В заданной генеральной совокупности 11 % потенциальных избирателей – афроамериканцы. Опрос, проведенный на основе простой случайной выборки из 600 стационарных телефонных номеров, находит 8 % афроамериканцев. Можно ли считать, что результаты обзора смещены? Объясните.
- 3.8. Известны результаты новой версии теста SAT 1000 случайно выбранных старшеклассников. Выборочное среднее экзаменационной оценки составляет 1110 и выборочное стандартное отклонение – 123. Постройте 95 %-й доверительный интервал для среднего экзаменационной оценки генеральной совокупности старшеклассников.
- 3.9. Предположим, что завод по производству лампочек производит лампы со средней продолжительностью работы, равной 2000 часов, и стандартным отклонением в 200 часов. Изобретатель утверждает, что разработал улучшенный процесс, который производит лампы с более длинной

средней продолжительностью работы и тем же самым стандартным отклонением. Директор завода случайно выбирает 100 новых ламп. Она сказала, что поверит в заявление изобретателя, если выборочное среднее продолжительности работы ламп будет больше, чем 2100 часов; в противном случае она заключит, что новый процесс не лучше, чем старый. Пусть  $\mu$  обозначает среднее нового процесса. Рассмотрим нулевую и альтернативную гипотезы  $H_0 : \mu = 2000$  против  $H_1 : \mu > 2000$ .

- a) Какой размер тестовой процедуры директора завода?
  - b) Предположим, что новый процесс фактически лучше и средняя продолжительность работы лампы равна 2150 часов. Какова мощность тестовой процедуры директора завода?
  - c) Какая тестовая процедура должна использоваться директором завода, если она хочет получить размер теста, равный 5 %?
- 3.10. Предположим, что новый стандартизированный тест дан 100 случайно выбранным студентам третьего курса в Нью-Джерси. Выборочная средняя оценка за тест составила 58 пунктов, а выборочное стандартное отклонение  $s_y = 8$  пунктов.
- a) Авторы планируют проводить тест для всех студентов третьего курса в Нью-Джерси. Постройте 95 %-й доверительный интервал для средней оценки всех студентов третьего курса Нью-Джерси.
  - b) Предположим, что тот же самый тест дан 200 случайно выбранным студентам третьего курса из Айовы, получаемое выборочное среднее равно 62 пунктам, а выборочное стандартное отклонение – 11 пунктов. Постройте 90 %-й доверительный интервал для разности средних оценок между студентами из Айовы и Нью-Джерси.
  - c) Можете ли вы заключить с высокой степенью доверия, что средние генеральных совокупностей для студентов из Айовы и Нью-Джерси различны? (Какова стандартная ошибка разности двух выборочных средних? Каково  $p$ -значение теста об отсутствии различия в средних против некоторого отличия?)
- 3.11. Рассмотрим оценку  $\tilde{Y}$ , определенную в уравнении (3.1). Покажите, что (a)  $E(\tilde{Y}) = \mu_y$  и (b)  $\text{var}(\tilde{Y}) = 1,25\sigma_y^2 / n$ .
- 3.12. Для исследования гендерной дискриминации в фирмах случайно выбрана выборка из 100 мужчин и 64 женщин с похожими должностями. Информация об их месячных зарплатах представлена в таблице.

	Средняя зарплата ( $\bar{Y}$ )	Стандартное отклонение ( $s_y$ )	$n$
Мужчины	3100 долл.	200 долл.	100
Женщины	2900 долл.	320 долл.	64

- a) Что говорят эти данные по поводу гендерных различий в фирме? Представляют ли они статистически значимое свидетельство того, что средние зарплаты мужчин и женщин различны? (Чтобы ответить на этот вопрос, во-первых, сформулируйте нулевую и альтернативную гипотезы; во-вторых, вычислите соответствующую  $t$ -статистику;

в-третьих, вычислите  $p$ -значение, связанное с  $t$ -статистикой, и, наконец, используйте  $p$ -значение для ответа на вопрос.)

- б) Свидетельствуют ли эти данные о том, что в фирме есть гендерная дискриминация в политике оплаты труда? Объясните.
- 3.13. Данные по результатам тестов в пятом классе (чтение и математика) для 420 школьных округов в Калифорнии приводят к выборочному среднему  $\bar{Y} = 646,2$  и стандартному отклонению  $s_Y = 19,5$ .
- а) Постройте 95 %-й доверительный интервал для среднего результата теста в генеральной совокупности.
- б) После того как округ был разделен на округа с малыми ( $< 20$  учеников на учителя) и большими классами ( $\geq 20$  учеников на учителя), были получены следующие результаты:

Размер класса	Средняя оценка ( $\bar{Y}$ )	Стандартное отклонение ( $s_Y$ )	$n$
Малый	657,4	19,4	238
Большой	650,0	17,9	182

Есть ли статистически значимое свидетельство того, что округа с малыми классами имеют более высокие средние результаты? Объясните.

- 3.14. Значения высоты в дюймах ( $X$ ) и веса в фунтах ( $Y$ ) записаны по выборке из 300 студентов колледжей мужского пола. Результирующая сводная статистика показывает:  $\bar{X} = 70,5$  дюймов,  $\bar{Y} = 158$  фунтов,  $s_X = 11,8$  дюймов,  $s_Y = 14,2$  фунтов,  $s_{XY} = 21,73$  дюймов  $\times$  фунты и  $r_{XY} = 0,85$ . Конвертируйте эти статистики в метрическую систему (метры и килограммы).

- 3.15. Пусть  $Y_a$  и  $Y_b$  обозначают случайные величины Бернулли из двух различных генеральных совокупностей, обозначенных  $a$  и  $b$ . Предположим, что  $E(Y_a) = p_a$  и  $E(Y_b) = p_b$ . Случайная выборка размера  $n_a$  выбрана из генеральной совокупности  $a$  с выборочным средним  $\hat{p}_a$ , и случайная выборка размера  $n_b$  выбрана из генеральной совокупности  $b$  с выборочным средним  $\hat{p}_b$ . Предположим, что выборка из генеральной совокупности  $a$  не зависит от выборки из генеральной совокупности  $b$ .

а) Покажите, что  $E(\hat{p}_a) = p_a$  и  $\text{var}(\hat{p}_a) = p_a(1-p_a)/n_a$ . Покажите, что  $E(\hat{p}_b) = p_b$  и  $\text{var}(\hat{p}_b) = p_b(1-p_b)/n_b$ .

б) Покажите, что  $\text{var}(\hat{p}_a - \hat{p}_b) = \frac{p_a(1-p_a)}{n_a} + \frac{p_b(1-p_b)}{n_b}$ . (Подсказка: вспомните, что выборки независимы.)

в) Предположим, что  $n_a$  и  $n_b$  велики. Покажите, что 95 %-й доверительный

интервал для  $p_a - p_b$  задается как  $\hat{p}_a - \hat{p}_b \pm 1,96 \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}$ .

Как построить 90 %-й доверительный интервал для  $p_a - p_b$ ?

г) Прочитайте вставку «Новый способ увеличения пенсионных накоплений» из раздела 3.5. Пусть генеральная совокупность  $a$  обозначает исследуемую группу, и генеральная совокупность  $b$  обозначает

контрольную группу. Постройте 95 %-й доверительный интервал для эффекта воздействия, то есть  $p_a - p_b$ .

- 3.16. Известно, что результаты стандартного школьного теста в США имеют среднее, равное 1000. Тест проходят 453 случайно выбранных школьников во Флориде; и в этой выборке среднее составляет 1013, а стандартное отклонение ( $s$ ) – 108.
- Постройте 95 %-й доверительный интервал для среднего результата теста школьников Флориды.
  - Можно ли говорить о наличии статистического свидетельства того, что школьники из Флориды выполняют тест иначе, чем другие школьники в Соединенных Штатах?
  - Другие 503 школьника были случайно выбраны во Флориде. Им был прочитан 3-часовой подготовительный курс до выполнения теста. Их средняя оценка составила 1019 со стандартным отклонением 95.
    - Постройте 95 %-й доверительный интервал для изменения в средней оценке, связанной с подготовительным курсом.
    - Есть ли статистическое свидетельство того, что подготовительный курс помог?
  - Первым 453 студентам также прочитали подготовительный курс, а затем попросили пройти тест во второй раз. Среднее изменение их оценки составило 9 пунктов, и стандартное отклонение изменилось на 60 пунктов.
    - Постройте 95 %-й доверительный интервал для изменения в средней оценке.
    - Есть ли статистически значимое свидетельство того, что студенты выполняют тест лучше со второй попытки после подготовительного курса?
    - Студенты могут выполнить тест лучше со второй попытки из-за подготовительного курса или из-за своего полученного опыта при первой попытке. Опишите эксперимент, который будет количественной оценкой этих двух эффектов.
- 3.17. Прочтите вставку «Гендерный разрыв в заработных платах выпускников колледжей в США» из раздела 3.5.
- Постройте 95 %-й доверительный интервал для изменения в средних почасовых зарплатах мужчин между 1992 и 2008 годами.
  - Постройте 95 %-й доверительный интервал для изменения в средних почасовых зарплатах женщин между 1992 и 2008 годами.
  - Постройте 95 %-й доверительный интервал для изменения в гендерном разрыве средних почасовых зарплат между 1992 и 2008 годами.  
*(Подсказка:  $\bar{Y}_{m,1992} - \bar{Y}_{w,1992}$  не зависит от  $\bar{Y}_{m,2008} - \bar{Y}_{w,2008}$ .)*
- 3.18. Это упражнение показывает, что выборочная дисперсия является несмещенной оценкой дисперсии генеральной совокупности, когда  $Y_1, \dots, Y_n$  являются i.i.d. со средним  $\mu_Y$  и дисперсией  $\sigma_Y^2$ .

- a) Используйте уравнение (2.31), чтобы показать, что  $E[(Y_i - \bar{Y})^2] = \text{var}(Y_i) - 2\text{cov}(Y_i, \bar{Y}) + \text{var}(\bar{Y})$ .
- б) Используйте уравнение (2.33), чтобы показать, что  $\text{cov}(Y_i, \bar{Y}) = \sigma_Y^2 / n$ .
- в) Используйте результаты в (a) и (б), чтобы показать, что  $E(s_Y^2) = \sigma_Y^2$ .
- 3.19. а)  $\bar{Y}$  – несмешенная оценка  $\mu_y$ . Является ли  $\bar{Y}^2$  несмешенной оценкой  $s_y^2$ ?
- б)  $\bar{Y}$  – состоятельная оценка  $\mu_y$ . Является ли  $\bar{Y}^2$  состоятельной оценкой  $s_y^2$ ?
- 3.20. Предположим, что  $(X_i, Y_i)$  есть i.i.d. с конечным четвертым моментом. Докажите, что выборочная ковариация является состоятельной оценкой ковариации генеральной совокупности, то есть  $s_{XY} \xrightarrow{P} \sigma_{XY}$ , где  $s_{XY}$  определено в уравнении (3.24) (Подсказка: используйте стратегию приложения 3.3 и неравенство Коши–Шварца<sup>1</sup>).
- 3.21. Покажите, что стандартная ошибка в пуле  $[SE_{\text{pooled}}(\bar{Y}_m - \bar{Y}_w)]$ , заданная в уравнении (3.23), равна обычной стандартной ошибке для разности средних в уравнении (3.19), если размер двух групп одинаковый ( $n_m = n_w$ ).

### Компьютерные упражнения

На веб-сайте книги [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) вы найдете файл с данными **CPS92\_0**, который содержит расширенную версию базы данных, использованной в таблице 3.1 книги для 1992 и 2008 годов. Она содержит данные о 25–34-летних работниках с полным рабочим днем, занятых в течение всего года и имеющих диплом средней школы или В.А./В.С. в качестве диплома о высшем образовании. Детальное описание дано в файле **CPS92\_08\_Description**, доступном на веб-сайте. Используйте эти данные, чтобы ответить на следующие вопросы:

- а) Вычислите выборочное среднее для средней почасовой зарплаты (AHE – average hourly earnings) в 1992 и 2008 годах. Постройте 95 %-й доверительный интервал для средних генеральных совокупностей AHE в 1992 и 2008 годах и изменения между 1992 и 2008 годами.
- б) В 2008 году значение индекса потребительских цен (ИПЦ) было 215,2. В 1992 году значение ИПЦ составляло 140,3. Повторите пункт (а), но используйте меру АНЕ в реальных долларах 2008 года (S2008); то есть скорректируйте данные 1992 года на инфляцию, которая произошла между 1992 и 2008 годами.
- в) Если вы интересуетесь изменением покупательной способности работников с 1992 по 2008 год, будете ли вы использовать результаты из пунктов (а) или (б)? Объясните.
- г) Используйте данные 2008 года для построения 95 %-го доверительного интервала для среднего значения АНЕ для выпускников средней школы. Постройте 95 %-й доверительный интервал для среднего значения АНЕ для выпускников колледжей. Постройте 95 %-й доверительный интервал для разности между двумя средними.

<sup>1</sup> В русскоязычной литературе это неравенство называется неравенством Коши–Буняковского. – Примеч. науч. ред. перевода.

- д) Повторите (г), используя данные 1992 года, выраженные в ценах 2008 года.
- е) Возрастают ли реальные (скорректированные на инфляцию) зарплаты выпускников средней школы с 1992 по 2008 год? Объясните. А реальные зарплаты для выпускников колледжей? Возрастает ли разрыв между зарплатами выпускников средней школы и колледжей? Объясните это, используя соответствующие оценки, доверительные интервалы и тестовые статистики.
- ж) Таблица 3.1 представляет информацию о гендерном разрыве для выпускников колледжей. Подготовьте такую же таблицу для выпускников средней школы, используя данные 1992 и 2008 годов. Есть ли существенные различия между результатами для выпускников средней школы и колледжей?

## Приложения

### *Приложение 3.1. Текущее обследование населения США*

Каждый месяц Бюро статистики труда Департамента труда США проводит текущее обследование населения (CPS – Current Population Survey), которое представляет данные о характеристиках рабочей силы, включая уровень занятости, безработицы и зарплат. Более чем 50 тыс. домохозяйств США опрашиваются каждый месяц. Выборка делается случайным выбором из базы данных адресов самой последней десятилетней переписи, дополненная данными о новых домохозяйствах, появившихся после последней переписи. Точная схема случайного выбора является довольно сложной (сначала небольшие географические районы выбираются случайным образом, а затем домохозяйства в этих районах выбираются случайным образом); подробную информацию можно найти в справочнике по статистике труда (Handbook of Labor Statistics) и на веб-сайте Бюро статистики труда ([www.bls.gov](http://www.bls.gov)).

Опрос, проведенный в марте каждого года, является более подробным, чем за другие месяцы, и включает вопросы о заработке в течение предыдущего года. Статистика в таблицах 2.4 и 3.1 вычислена с использованием мартовского опроса. Данные по зарплатам CPS для работающих полный рабочий день определены для тех, кто работал более чем 35 часов в неделю в течение не менее чем 48 недель в предыдущем году.

### *Приложение 3.2. Два доказательства того, что $\bar{Y}$ является оценкой наименьших квадратов $\mu_Y$*

В этом приложении представлены два доказательства того, что  $\bar{Y}$  минимизирует сумму квадратов МНК-остатков в уравнении (3.2), то есть что  $\bar{Y}$  является оценкой наименьших квадратов  $E(Y)$ : одно из доказательств использует дифференциальное исчисление, а другое – нет.

**Доказательство, использующее дифференциальное исчисление**

Чтобы минимизировать сумму квадратов МНК-остатков, возьмем производную и положим ее равной нулю:

$$\frac{d}{dm} \sum_{i=1}^n (Y_i - m)^2 = -2 \sum_{i=1}^n (Y_i - m) = -2 \sum_{i=1}^n Y_i + 2nm = 0. \quad (3.27)$$

Решение последнего уравнения для  $m$  показывает, что  $\sum_{i=1}^n (Y_i - m)^2$  минимизируется, когда  $m = \bar{Y}$ .

**Доказательство, не использующее дифференциальное исчисление**

В этом случае стратегия заключается в том, чтобы показать, что разность между оценкой наименьших квадратов и  $\bar{Y}$  должна быть равной нулю, из чего следует, что  $\bar{Y}$  является оценкой наименьших квадратов. Пусть  $d = \bar{Y} - m$ , так что  $m = \bar{Y} - d$ . Тогда  $(Y_i - m)^2 = (Y_i - [\bar{Y} - d])^2 = ([Y_i - \bar{Y}] + d)^2 = (Y_i - \bar{Y})^2 + 2d(Y_i - \bar{Y}) + d^2$ . Таким образом, сумма квадратов МНК-остатков [уравнение (3.2)] имеет вид:

$$\sum_{i=1}^n (Y_i - m)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2d \sum_{i=1}^n (Y_i - \bar{Y}) + nd^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 + nd^2, \quad (3.28)$$

где переход ко второму равенству использует тот факт, что  $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$ . Поскольку оба слагаемых в последней строке уравнения (3.28) неотрицательны и так как первое слагаемое не зависит от  $d$ ,  $\sum_{i=1}^n (Y_i - m)^2$  минимизируется, выбирай  $d$  таким образом, чтобы сделать второе слагаемое  $nd^2$  настолько малым, насколько возможно. Следовательно, минимум достигается, если  $d = 0$ , то есть  $m = \bar{Y}$ , так что  $\bar{Y}$  является оценкой наименьших квадратов  $E(Y)$ .

**Приложение 3.3. Доказательство состоятельности выборочной дисперсии**

В этом приложении закон больших чисел используется для доказательства того, что выборочная дисперсия  $s_y^2$  является состоятельной оценкой дисперсии генеральной совокупности  $\sigma_y^2$ , как утверждается в уравнении (3.9), когда  $Y_1, \dots, Y_n$  являются i.i.d. и  $E(Y_i^4) < \infty$ .

Во-первых, добавим и вычтем  $\mu_Y$ , чтобы записать:  $(Y_i - \bar{Y})^2 = [(Y_i - \mu_Y) - (\bar{Y} - \mu_Y)]^2 = (Y_i - \mu_Y)^2 - 2(Y_i - \mu_Y)(\bar{Y} - \mu_Y) + (\bar{Y} - \mu_Y)^2$ . Заменяя это выражение для  $(Y_i - \bar{Y})^2$  на определение  $s_y^2$  [уравнение (3.7)], получаем:

$$\begin{aligned} s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \\ &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \mu_Y)^2 - \frac{2}{n-1} \sum_{i=1}^n (Y_i - \mu_Y)(\bar{Y} - \mu_Y) + \frac{1}{n-1} \sum_{i=1}^n (\bar{Y} - \mu_Y)^2 = \end{aligned}$$

$$= \left( \frac{n}{n-1} \right) \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \mu_Y)^2 \right] - \left( \frac{n}{n-1} \right) (\bar{Y} - \mu_Y)^2, \quad (3.29)$$

где последнее равенство следует из определения  $\bar{Y}$  [которое предполагает, что  $\sum_{i=1}^n (Y_i - \mu_Y)^2 = n(\bar{Y} - \mu_Y)^2$ ] и после приведения подобных в выражении.

Теперь можно применить закон больших чисел к двум слагаемым в последней строке уравнения (3.29). Определим  $W_i = (Y_i - \mu_Y)^2$ . Сейчас  $E(W_i) = \sigma_Y^2$  (по определению дисперсии). Поскольку случайные величины  $Y_1, \dots, Y_n$  являются i.i.d., случайные величины  $W_1, \dots, W_n$  также i.i.d. В дополнение к этому  $E(W_i^2) = E[(Y_i - \mu_Y)^4] < \infty$ , поскольку, по предположению,  $E(Y_i^4) < \infty$ . Таким образом,  $W_1, \dots, W_n$  являются i.i.d. и  $\text{var}(W_i) < \infty$ , поэтому  $\bar{W}$  удовлетворяет условиям закона больших чисел, сформулированным во вставке «Основные понятия 2.6», и  $\bar{W} \xrightarrow{p} E(W_i)$ . Но  $\bar{W} = (1/n) \sum_{i=1}^n (Y_i - \mu_Y)^2$  и  $E(W_i) = \sigma_Y^2$ , так что  $(1/n) \sum_{i=1}^n (Y_i - \mu_Y)^2 \xrightarrow{p} \sigma_Y^2$ . Также  $n/(n-1) \rightarrow 1$ , поэтому первое слагаемое в уравнении (3.29) сходится по вероятности к  $\sigma_Y^2$ . Поскольку  $\bar{Y} \xrightarrow{p} \mu_Y$ ,  $(\bar{Y} - \mu_Y)^2 \xrightarrow{p} 0$ , так что второе слагаемое сходится по вероятности к нулю. Объединяя эти результаты, получаем, что  $s_Y^2 \xrightarrow{p} \sigma_Y^2$ .

Часть II

ОСНОВЫ  
РЕГРЕССИОННОГО  
АНАЛИЗА



# Глава 4. Парная линейная регрессия

Государство повысило штрафы за вождение в пьяном виде: как эта мера повлияет на смертность на дорогах? В начальных школах некоего округа уменьшается количество детей в классах: как это повлияет на результаты стандартных тестов? Вы окончили еще один курс в колледже: повлияет ли это каким-либо образом на вашу будущую зарплату?

Во всех трех рассмотренных ситуациях мы задаемся вопросом о влиянии изменения одной переменной  $X$  (штрафы за вождение в пьяном виде, уменьшение количества детей в классе или число лет обучения в колледже), на другую переменную  $Y$  (смертность на дорогах, оценка за тест или зарплата).

В данной главе мы вводим понятие парной линейной модели регрессии, которая связывает между собой переменные  $X$  и  $Y$ . Такая модель предполагает наличие линейной связи между  $X$  и  $Y$ , в которой коэффициент при переменной  $X$  характеризует то, как изменение  $X$  на 1 единицу влияет на  $Y$ . Так же как среднее  $Y$  является неизвестной характеристикой распределения генеральной совокупности  $Y$ , угловой коэффициент линии регрессии  $X$  на  $Y$  является неизвестной характеристикой совместного распределения  $X$  и  $Y$ . Задачей эконометрики является оценить этот угловой коэффициент, то есть оценить, каким образом меняется  $Y$ , если  $X$  изменится на единицу, используя выборку этих двух переменных.

В настоящей главе мы описываем методы оценки этого углового коэффициента, используя случайную выборку данных  $X$  и  $Y$ . Например, на основе данных о количестве детей в классе и результатов тестов в различных школьных округах мы показываем, как оценить ожидаемое влияние от уменьшения на одного ученика числа учеников в классе на результаты итоговых тестов. Угловой коэффициент и константа в линии регрессии  $X$  и  $Y$  могут быть оценены методом, называемым методом наименьших квадратов (МНК).

## 4.1. Модель парной линейной регрессии

Окружной школьный инспектор по начальным школам должен решить, стоит ли увеличивать число учителей в начальной школе, и поэтому она хочет, чтобы ей дали совет. Если она наймет больше учителей, то уменьшит число учеников в классе на одного учителя (отношение «ученик – учитель», соотношение учеников и учителей). Перед ней дилемма. С одной стороны, родители хотят, чтобы в классах было меньше учеников и их дети получали бы больше внимания. С другой стороны, если она наймет больше учителей и, как следствие, для этого потребуются большие траты, это вряд ли понравится тем, кто оплачивает счета! Поэтому она хочет, чтобы вы ответили ей на вопрос: как повлияет на успеваемость в начальной школе уменьшение числа учеников в классах?

Во многих школьных округах успеваемость учеников оценивается при помощи стандартных тестов, и то, сколько будут зарабатывать окружные школьные инспекторы и прочие администраторы, может зависеть, в частности, и от результатов тестов учащихся. Поэтому мы уточняем вопрос окружного школьного инспектора: если она уменьшит в среднем число учеников в классе на два, то как это повлияет на результаты стандартных тестов в ее округе?

Чтобы получить точный ответ на этот вопрос, необходимо сформулировать количественное утверждение об изменении. Если окружной школьный инспектор изменяет число учащихся в классе на определенное количество, какое она должна ожидать *изменение* в результатах тестов? Мы можем записать математическое отношение, используя греческую букву бета,  $\beta_{ClassSize}$ , где индекс *ClassSize* (т.е. размер класса – число учеников в классе) выделяет эффект изменения числа учеников в классе среди других эффектов. Таким образом:

$$\beta_{ClassSize} = \frac{\text{изменение в } TestScore}{\text{изменение в } ClassSize} = \frac{\Delta TestScore}{\Delta ClassSize}, \quad (4.1)$$

где греческая буква  $\Delta$  (дельта) обозначает «изменение в». То есть  $\beta_{ClassSize}$  является изменением в оценке за тест (*TestScore*), которое является результатом изменения размера класса, деленное на изменение размера класса.

Если бы вы знали  $\beta_{ClassSize}$ , вы могли бы сказать окружному школьному инспектору, что уменьшение размера класса на одного ученика изменило бы оценки за тест в ее школьном округе на  $\beta_{ClassSize}$ . Вы могли бы также ответить на актуальный для окружного школьного инспектора вопрос, который касается изменения размера класса на двух учеников в классе. Для этого изменим уравнение (4.1) так, чтобы оно приняло вид

$$\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize. \quad (4.2)$$

Предположим, что  $\beta_{ClassSize} = -0,6$ . Тогда уменьшение размера класса на двух учеников в классе привело бы к прогнозируемому изменению оценок тестов на  $(-0,6) \times (-2) = 1,2$ , то есть вы бы могли сказать, что оценки за тест возросли бы на 1,2 пункта как результат уменьшения размеров класса на двух учеников в классе.

Уравнение (4.1) представляет собой определение углового коэффициента прямой линии, соотносящей оценки за тест и число учеников в классе. Эта прямая линия может быть записана как

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize, \quad (4.3)$$

где  $\beta_0$  является константой, и, как указано выше,  $\beta_{ClassSize}$  является угловым коэффициентом. Из уравнения (4.3) следует, что если бы вы знали  $\beta_0$  и  $\beta_{ClassSize}$ , вы бы смогли не только определить изменение оценок за тесты в округе, связанное с изменением числа учеников в классах, но и могли предсказать среднюю оценку за сам тест для класса конкретного размера.

Когда вы показываете уравнение (4.3) окружному школьному инспектору, она говорит, что в предлагаемой вами формулировке что-то неверно. Она указывает на то, что число учеников в классе является лишь одним из многих аспектов начального образования и что два округа с одинаковыми размерами классов будут иметь раз-

личные оценки за тесты по многим причинам. Например, один из округов может иметь лучших учителей или может использовать лучшие учебники. Два округа с со-поставимыми размерами классов, учителями и учебниками все еще могут очень сильно различаться составом учеников; возможно, в одном округе живет больше иммигрантов (и, таким образом, меньшее носителей английского языка) или больше детей из обеспеченных семей. Наконец, она указывает, что даже если два округа одинаковы по всем этим параметрам, результаты тестов в них все равно могут существенно различаться из-за случайных причин, связанных с конкретным состоянием отдельных учеников в день проведения теста. Конечно, она права; по всем этим причинам уравнение (4.3) нельзя рассматривать как равенство, верное для каждого ребенка в каждом округе. Вместо этого его следует рассматривать как некоторое утверждение о связи, которая выполняется в *среднем* для школьников в округе.

Для того чтобы линейное соотношение выполнялось для *каждого* округа, оно должно содержать и другие факторы, влияющие на результаты тестов, в том числе уникальные характеристики каждого округа (например, насколько хороши учителя в округе, из каких семей ученики или насколько удачным был день тестирования для ученика). Тогда один из подходов заключается в том, чтобы перечислить самые важные факторы, влияющие на качество образования, и ввести их в уравнение (4.3) (мы вернемся к этой идеи в главе 6). Однако сейчас мы просто объединим все эти «другие факторы» (*other factors*) вместе и запишем соотношение для заданного округа как

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + other\ factors \quad (4.4)$$

Таким образом, оценка за тест для округа записана в терминах одной компоненты,  $\beta_0 + \beta_{ClassSize} \times ClassSize$ , которая представляет средний эффект влияния числа учеников в классе на оценки в генеральной совокупности школьных округов, и второй компоненты, которая представляет все другие факторы.

Несмотря на то что мы сосредоточили свое внимание на оценках за стандартный тест и размере класса, идея, выраженная в уравнении (4.4), является более общей, и поэтому она полезна для введения более общих обозначений. Предположим, что у вас есть выборка из  $n$  округов. Пусть  $Y_i$  – средняя оценка за тест в  $i$ -м округе, а  $X_i$  – средний размер класса в  $i$ -м округе, и пусть  $u_i$  обозначает другие факторы, влияющие на оценку теста в  $i$ -м округе. Тогда уравнение (4.4) может быть переписано в более общем случае следующим образом:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4.5)$$

для каждого округа (т.е.  $i = 1, \dots, n$ ), где  $\beta_0$  является константой, а  $\beta_1$  – угловым коэффициентом линии регрессии. [Более общее обозначение  $\beta_1$  используется для углового коэффициента в уравнении (4.5) вместо  $\beta_{ClassSize}$ , поскольку это уравнение записано в терминах общей переменной  $X_i$ .]

Уравнение (4.5) – это модель парной линейной регрессии, в которой  $Y$  является *зависимой переменной*, а  $X$  является *независимой переменной*, или *регрессором*.

Правая часть уравнения (4.5),  $\beta_0 + \beta_1 X_i$ , называется *линией регрессии генеральной совокупности* или *функцией регрессии генеральной совокупности*

(теоретической функцией регрессии). Оно является соотношением, верным для  $Y$  и  $X$  в среднем по генеральной совокупности. Таким образом, если бы вы знали значение  $X$ , то, согласно этой теоретической линии регрессии, вы могли бы утверждать, что значение зависимой переменной  $Y$  будет  $\beta_0 + \beta_1 X$ .

Константа  $\beta_0$  и угловой коэффициент  $\beta_1$  – это коэффициенты теоретической линии регрессии, также известные как параметры теоретической линии регрессии. Угловой коэффициент  $\beta_1$  характеризует изменение в  $Y$ , связанное с изменением  $X$  на единицу. Константа – это значение линии регрессии генеральной совокупности, когда  $X = 0$ ; это точка, в которой теоретическая линия регрессии пересекает ось  $Y$ . В некоторых эконометрических приложениях константа имеет значимую экономическую интерпретацию. В других приложениях константа не имеет реального смысла; например, когда  $X$  является числом учеников в классе, строго говоря, константа представляет собой значение оценки за тест, когда в классе нет учеников! Если у константы отсутствует реальный смысл, лучше всего думать о ней математически, как о коэффициенте, который определяет уровень линии регрессии.

Компонента  $u_i$  в уравнении (4.5) называется ошибкой. Она включает в себя все факторы, отвечающие за разность между  $i$ -й средней оценкой за тест по округу и значением, предсказанным теоретической линией регрессии. Ошибка содержит все те факторы, кроме  $X$ , которые определяют значение зависимой переменной  $Y$  для конкретного наблюдения  $i$ . В примере про число учеников в классе эти факторы включают все уникальные особенности  $i$ -го округа, которые влияют на то, как учащиеся напишут тест, включая то, насколько хорошие учителя работают в этом округе, из каких семей происходят ученики, сопутствует ли им удача во время экзамена и даже любые ошибки, которые могут возникнуть во время проверки теста.

Основные термины, используемые в модели парной линейной, приведены во вставке «Основные понятия 4.1».

## ОСНОВНЫЕ ПОНЯТИЯ 4.1

### Модель парной линейной регрессии: терминология

Модель парной линейной регрессии имеет вид:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

где индекс  $i$  пробегает по всем наблюдениям,  $i = 1, \dots, n$ ;

$Y_i$  – зависимая переменная, регressируемая переменная или просто переменная слева;

$X_i$  – независимая переменная, объясняющая переменная, regressor или просто переменная справа;

$\beta_0 + \beta_1 X_i$  – линия теоретической регрессии, линия регрессии генеральной совокупности, или функция регрессии генеральной совокупности;

$\beta_0$  – константа линии теоретической регрессии;

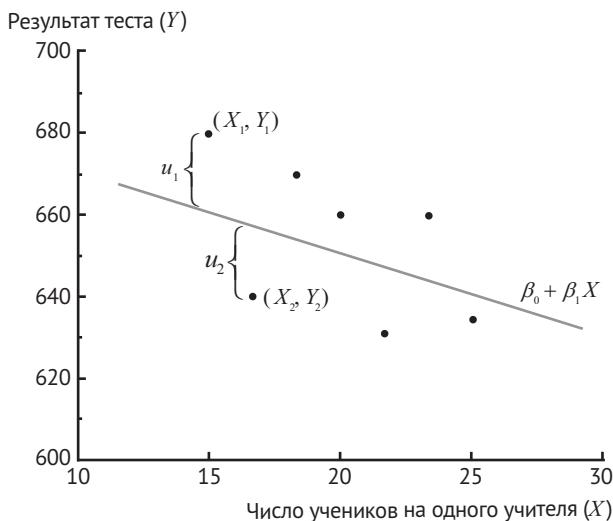
$\beta_1$  – угловой коэффициент линии теоретической регрессии; и

$u_i$  – ошибка.

На рисунке 4.1 приведен пример модели парной линейной регрессии для семи гипотетических наблюдений результатов теста ( $Y$ ) и числа учеников в классе ( $X$ ). Линия регрессии генеральной совокупности – это прямая линия  $\beta_0 + \beta_1 X$ . Линия теоретической регрессии имеет отрицательный наклон ( $\beta_1 < 0$ ), что означает, что в округах с более низким отношением «ученик – учитель» (с маленькими классами) результаты тестов выше. Константа  $\beta_0$  – это значение, в котором линия теоретической регрессии пересекает ось  $Y$ , но, как было замечено ранее, это не несет никакого экономического смысла в рассматриваемом примере.

Из-за неучтенных факторов, влияющих на результаты тестов, гипотетические наблюдения на рисунке 4.1 не попадают точно на линию теоретической регрессии. Например, значение  $Y$  для округа № 1, то есть  $Y_1$ , выше линии теоретической регрессии. Это означает, что оценки теста в округе № 1 были выше, чем предсказано линией теоретической регрессии, поэтому ошибка  $u_1$  для этого округа положительна. В противоположность этому значение  $Y_2$  для округа № 2 лежит ниже линии теоретической регрессии, поэтому оценки теста для этого округа оказались хуже, чем показывает линия теоретической регрессии и ошибка  $u_2 < 0$ .

Вернемся теперь к вашей проблеме как советнику окружного школьного инспектора: каковы наши ожидания влияния уменьшения числа учеников в классах на два (т.е. отношения «ученик – учитель») на результаты тестов? Ответ является легким: ожидаемое изменение равно  $(-2) \times \beta_{ClassSize}$ . Но остается неизвестным ответ на один вопрос: каково значение  $\beta_{ClassSize}$ ?



**Рисунок 4.1. Диаграмма рассеяния результатов тестов относительно числа учеников в классе (гипотетические данные)**

На диаграмме рассеяния показаны гипотетические наблюдения для семи школьных округов. Линия теоретической регрессии – это  $\beta_0 + \beta_1 X_i$ . Расстояние от  $i$ -й точки до линии теоретической регрессии по вертикали равно  $Y_i - (\beta_0 + \beta_1 X_i)$  и является ошибкой  $u_i$  для  $i$ -го наблюдения.

## 4.2. Оценка коэффициентов в модели парной линейной регрессии

В практических ситуациях, в таких как рассмотренный в разделе 4.1 пример, константа  $\beta_0$  и угловой коэффициент  $\beta_1$  линии теоретической регрессии неизвестны. Следовательно, мы должны использовать эмпирические данные, чтобы оценить неизвестные нам коэффициенты линии теоретической регрессии.

Проблема оценки коэффициентов теоретической регрессии аналогична другим проблемам, с которыми вы уже сталкивались ранее в статистике. Предположим, например, что вы хотите сравнить средние зарплаты мужчин и женщин, недавно окончивших колледж. Несмотря на то что генеральное среднее неизвестно, мы можем оценить его, используя случайную выборку мужчин и женщин, являющихся выпускниками колледжей. Тогда естественной оценкой неизвестного среднего генеральной совокупности для женщин, например, будет средняя зарплата женщин – выпускниц колледжей в этой выборке.

Аналогичная идея возникает при необходимости оценить модель парной линейной регрессии. Мы не знаем теоретического значения коэффициента  $\beta_{ClassSize}$  (т.е. значения, которое принимает этот коэффициент в генеральной совокупности), являющегося угловым коэффициентом неизвестной линии теоретической регрессии, связывающей  $X$  (число учеников в классе) и  $Y$  (результаты тестов). Но точно так же, как мы могли бы узнать о среднем генеральной совокупности, используя выборку данных из генеральной совокупности, мы можем узнати и о теоретическом угловом коэффициенте  $\beta_{ClassSize}$ , используя выборку данных.

Данные, которые используются в этом разделе, представляют собой информацию о результатах тестов и размерах классов в 1999 году в 420 школьных округах Калифорнии, в которых расположены школы, обучающие детей с подготовительного до восьмого класса. Оценка за тест представляет собой среднюю по округу оценку за тесты по чтению и математике для пятиклассников. Численная характеристика размера класса может быть определена различными способами. Мы используем одну из самых распространенных мер: число учащихся в округе, деленное на число учителей этого округа, то есть среднее по округу отношение «ученик – учитель».

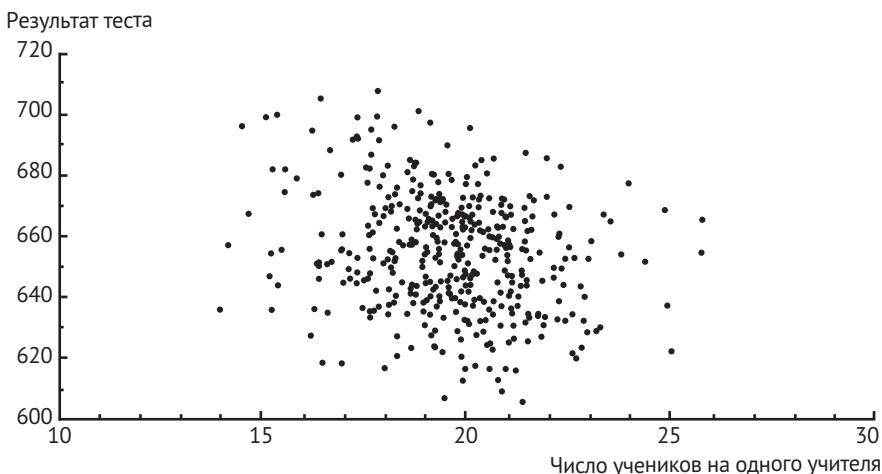
В таблице 4.1 приведены основные статистические характеристики рассматриваемой выборки данных. Выборочное среднее значение отношения «ученик – учитель» равно 19,6 учеников на одного учителя, а выборочное стандартное отклонение составляет 1,9 ученика на одного учителя. 10%-й процентиль распределения отношения «ученик – учитель» равен 17,3 (т.е. только в 10% округов отношение «ученик – учитель» ниже, чем 17,3), в то время как округ из 90%-го перцентиля имеет отношение «ученик – учитель», равное 21,9.

Таблица 4.1

**Основные статистические характеристики распределения отношения числа учеников, приходящихся на одного учителя, и результатов тестов пятиклассников в 420 школьных округах Калифорнии в 1999 году**

	Выборочное среднее	Выборочное стандартное отклонение	Процентили						
			10 %	25 %	40 %	50 % медиана	60 %	75 %	90 %
Число учеников на одного учителя	19,6	1,9	17,3	18,6	19,3	19,7	20,1	20,9	21,9
Оценка за тест	654,2	19,1	630,4	640,0	649,1	654,5	659,4	666,7	679,1

Диаграмма рассеяния этих 420 наблюдений результатов тестов и отношения «ученик – учитель» представлена на рисунке 4.2. Выборочная корреляция между рассматриваемыми показателями равна  $-0,23$ , что указывает на слабую отрицательную связь между двумя переменными. Несмотря на то что ученики из больших классов в данной выборке, как правило, имеют более низкие результаты тестов, существуют и другие факторы, влияющие на эти результаты, которые не позволяют наблюдениям оказаться точно на прямой линии.



**Рисунок 4.2. Диаграмма рассеяния результатов тестов и соотношения числа учеников в классе на одного учителя (данные школьных округов Калифорнии)**

Данные по 420 школьным округам Калифорнии. На диаграмме рассеяния видна слабая отрицательная зависимость между отношением «ученик – учитель» и результатами тестов: выборочная корреляция составляет  $-0,23$ .

Несмотря на такую низкую корреляцию, если бы было можно как-то построить прямую линию по этим данным, то коэффициент ее наклона был бы оценкой  $\beta_{ClassSize}$ , полученной по этим данным. Можно, например, начертить такую линию, взяв в руки карандаш и линейку, на глаз. Этот метод очень прост, но в то же время он очень ненаучен, и у разных людей получатся различные оцененные ими линии.

Тогда возникает вопрос: как же следует выбирать нужную линию среди многих возможных? Наиболее распространенным методом подгонки линии регрессии является метод «наименьших квадратов», примененный к имеющимся данным. То есть следует использовать обычные оценки метода наименьших квадратов (МНК).

### **Метод наименьших квадратов**

МНК-оценка выбирает коэффициенты регрессии так, чтобы оцененная линия регрессии была близка настолько, насколько возможно к наблюдаемым данным, где близость измеряется суммой квадратов ошибок, которые возникают при предсказании  $Y$  при заданных  $X$ .

Как обсуждалось в разделе 3.1, выборочное среднее  $\bar{Y}$  представляет собой оценку наименьших квадратов генерального среднего  $E(Y)$ , то есть  $\bar{Y}$  минимизирует полную сумму квадратов оцененных ошибок  $\sum_{i=1}^n (Y_i - m)^2$  среди всех возможных оценок  $m$  [см. выражение (3.2)].

МНК-оценка обобщает эту идею на случай модели парной линейной регрессии. Пусть  $b_0$  и  $b_1$  – некоторые оценки  $\beta_0$  и  $\beta_1$ . Тогда линия регрессии, основанная на этих оценках, имеет вид  $b_0 + b_1 X$ , поэтому значение  $Y_i$ , подобранное таким образом, имеет вид  $b_0 + b_1 X_i$ . Следовательно, ошибка (остаток), сделанная в предсказании  $i$ -го наблюдения, это разность  $Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$ . Сумма квадратов остатков модели по всем  $n$  наблюдениям равна:

$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2. \quad (4.6)$$

Сумма квадратов остатков для модели линейной регрессии в выражении (4.6) является обобщением суммы квадратов ошибок для задачи оценивания среднего в выражении (3.2). Действительно, если регрессоры в модели отсутствуют, тогда  $b_1$  не входит в выражение (4.6), и две задачи становятся идентичными, за исключением различных обозначений [ $m$  – в выражении (3.2) и  $b_0$  – в выражении (4.6)]. Подобно тому что существует единственная оценка  $\bar{Y}$ , минимизирующая выражение (3.2), существует и единственная пара оценок  $\beta_0$  и  $\beta_1$ , которые минимизируют выражение (4.6).

Оценки константы и углового коэффициента, которые минимизируют сумму квадратов остатков в выражении (4.6), называются *оценками метода наименьших квадратов (МНК-оценками)*  $\hat{\beta}_0$  и  $\hat{\beta}_1$ .

Метод наименьших квадратов имеет собственные специальные обозначения и терминологию. МНК-оценка  $\beta_0$  обозначается  $\hat{\beta}_0$ , и МНК-оценка  $\beta_1$  обозначается  $\hat{\beta}_1$ . МНК-оценка линии регрессии, также называемая линией выборочной ре-

грессии или функцией выборочной регрессии, это прямая линия, построенная с использованием МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$ . Предсказанное значение<sup>1</sup>  $\hat{Y}_i$ , основанное на МНК-оценке линии регрессии, имеет вид  $\hat{\beta}_0 + \hat{\beta}_1 X_i$ . Остаток для  $i$ -го наблюдения есть разность между  $Y_i$  и предсказанным значением:  $\hat{u}_i = Y_i - \hat{Y}_i$ .

### МНК-оценка, предсказанные значения и остатки

МНК-оценка углового коэффициента  $\beta_1$  и константы  $\beta_0$  равны:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \quad (4.7)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.8)$$

Предсказанные МНК-значения  $\hat{Y}_i$  и остатки  $\hat{u}_i$  равны:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n; \quad (4.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n. \quad (4.10)$$

Оцененная константа ( $\hat{\beta}_0$ ), угловой коэффициент ( $\hat{\beta}_1$ ) и остаток ( $\hat{u}_i$ ) вычисляются по выборке с  $n$  наблюдениями  $X_i$  и  $Y_i$ ,  $i = 1, \dots, n$ . Они являются оценками неизвестных параметров теоретической регрессии: константы ( $\beta_0$ ), углового коэффициента ( $\beta_1$ ) и ошибки ( $u_i$ ).

## ОСНОВНЫЕ ПОНЯТИЯ

4.2

МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  можно было бы вычислить, перебирая различные значения  $b_0$  и  $b_1$  тех пор, пока вы не найдете те, которые минимизируют полную сумму квадратов остатков в выражении (4.6); они являются оценками наименьших квадратов. Однако такой метод весьма утомителен. К счастью, у нас есть формулы, полученные при минимизации выражения (4.6) путем дифференциального исчисления, что сильно упрощает вычисление МНК-оценок.

Формулы для вычисления МНК-оценок и основные термины перечислены во вставке «Основные понятия 4.2». Эти формулы реализуются практически во всех статистических программных пакетах и выводятся в приложении 4.2.

### МНК-оценки зависимости между результатами тестов и соотношением учеников и учителей

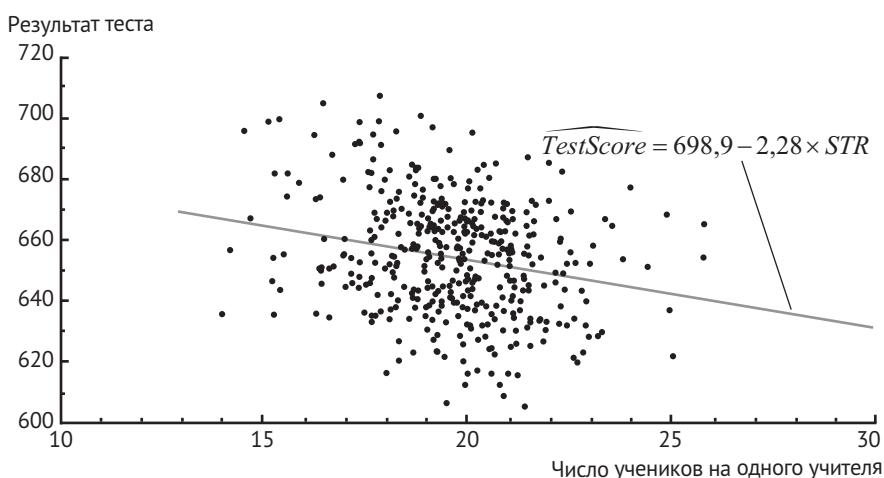
Когда МНК используется, чтобы оценить линейную зависимость отношения «ученик – учитель» к оценкам за тест на основе 420 наблюдений на рисунке 4.2, оцененный коэффициент наклона составляет  $-2,28$  и оцененная константа составляет  $698,9$ . Соответственно МНК-регрессионная линия для этих 420 наблюдений имеет вид:

$$\widehat{\text{TestScore}} = 698,9 - 2,28 \times \text{STR}, \quad (4.11)$$

<sup>1</sup> Русскоязычные аналоги английского понятия predicted value многочисленны – предсказанное, оцененное, подогнанное, прогнозное, рассчитанное или аппроксимированное значение. – Примеч. науч. ред. перевода.

где  $TestScore$  – это средняя оценка по тестам в округе, а  $STR$  – отношение «ученик – учитель». Знак « $\hat{}$ » над  $TestScore$  в уравнении (4.11) указывает, что это предсказанное значение, основанное на МНК-регрессии. На рисунке 4.3 изображена МНК-регрессия и она наложена на диаграмму рассеяния данных, которую мы уже видели ранее на рисунке 4.2.

Угловой коэффициент, равный  $-2,28$ , означает, что увеличение отношения «ученик – учитель» на единицу вызывает в среднем снижение оценок за тест в округе на  $2,28$  пункта. Уменьшение отношения «ученик – учитель» на два ученика на одного учителя влечет за собой в среднем улучшение результатов тестов на  $4,56$  пункта  $[-2 \times (-2,28)]$ . Отрицательный угловой коэффициент указывает на то, что большее количество учеников на одного учителя (большие классы) связано с худшими результатами за тест.



**Рисунок 4.3. Оценка линии регрессии для данных по Калифорнии**

Оцененная линия регрессии показывает отрицательную связь между результатами тестов и числом учеников в классе, приходящимся на одного учителя. Если относительный размер класса уменьшается на одного ученика, то оцененная регрессия предсказывает, что оценки за тест увеличиваются на  $2,28$  пункта.

Теперь можно предсказать результат теста при заданном значении количества учеников на одного учителя в классе. Например, для округа с  $20$  учениками на одного учителя предсказанная оценка за тест равна  $698,9 - 2,28 \times 20 = 653,3$ . Конечно, это предсказание не является точным из-за ряда других факторов, которые влияют на качество образования в округе. Но по линии регрессии действительно можно сделать предсказание (МНК-предсказание), основанное на знании соотношения учеников и учителей, о том, какие результаты теста были бы в этом районе при отсутствии других влияющих на них факторов.

Является ли эта оценка углового коэффициента большой или нет? Чтобы ответить на этот вопрос, мы возвращаемся к задаче окружного школьного инспектора. Вспомним, что она рассматривает возможность найти достаточного количества учителей, чтобы уменьшить относительное число учеников на одного учителя на  $2$ . Предположим, что ее округ является медианным среди округов Калифорнии. Из таблицы 4.1 видим, что медианное значение отношения «уче-

ник – учитель» составляет 19,7, а медианная оценка за тест составляет 654,5. Сокращение двух учеников в классе с 19,7 до 17,7 передвинет отношение «ученик – учитель» в ее округе с 50-го процентиля до значения очень близкого к 10-му процентилю. Это довольно большое изменение, и для этого необходимо нанять много новых учителей. Как такое снижение числа учеников на одного учителя повлияет на результаты тестов?

Согласно уравнению (4.11), сокращение соотношения «ученик – учитель» на 2 предсказывает увеличение результатов тестов приблизительно на 4,6 пункта; если средний результат тестов в ее округе (на медиане) равен 654,5, то можно предсказать его рост до 659,1. Велико ли такое улучшение или мало? Согласно таблице 4.1, это улучшение передвинет рассматриваемый округ с медианного до почти 60-го процентиля. Таким образом, сокращение числа учеников на одного учителя переместит в категорию, близкую к 10% округов с наименьшим соотношением числа учеников и учителей, но потенциальные результаты тестов улучшатся не так сильно: по этому показателю округ переместится из медианного в 60-й процентиль. Следовательно, по результатам оценки уменьшение размеров классов в округе приведет к улучшению результатов обучения, и, возможно, стоит это сделать в зависимости от бюджетной ситуации в округе, но такое решение не является панацеей.

Что делать, если окружной школьный инспектор размышляет о гораздо более радикальных изменениях, таких как сокращение количества учеников на одного учителя с 20 до 5? К сожалению, оценки из уравнения (4.11) будут не очень полезны для нее. Данная регрессия была оценена с использованием данных, представленных на рисунке 4.2, и, как видно из рисунка, наименьшее возможное отношение «ученик – учитель» равно 14. Имеющиеся данные не содержат информации о результатах тестов в округах с очень маленькими классами, таким образом, эти данные сами по себе не являются надежной основой для оценки влияния радикального перехода к такому крайне низкому отношению «ученик – учитель».

### **Почему используется МНК-оценка?**

Существуют и практические, и теоретические причины использовать МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$ . Поскольку МНК является доминирующим методом на практике, он стал общим языком для регрессионного анализа экономики, финансов (см. вставку «Бета-акции») и социальных наук в целом. Представление результатов с использованием МНК (или его варианты, обсуждаемые позже в этой книге) означает, что вы «говорите на таком же языке», как и другие экономисты и статистики. Формулы МНК встроены практически во все программные статистические пакеты, что делает МНК простым в использовании.



#### **Бета-акции**

Фундаментальная идея современных финансовых заключается в том, что инвестору нужен финансовый стимул, чтобы идти на риск. Иначе говоря, ожидаемая

доходность\* рисковых инвестиций  $R$  должна превышать доходность безопасных или безрисковых инвестиций  $R_f$ . Таким образом, ожидаемая избыточная доходность  $R - R_f$  рисковых инвестиций, как и владение ценными бумагами в компании, должна быть положительной.

На первый взгляд может показаться, что риск владения акцией следует измерять ее дисперсией. Большая часть этого риска, однако, может быть уменьшена покупкой других акций и формированием «портфеля» акций – другими словами, за счет диверсификации ваших финансовых активов. Это означает, что правильный способ измерения риска акции это не ее дисперсия, а ее ковариация с рынком.

Модель оценки финансовых активов (CAPM) формализует эту идею. Согласно CAPM, ожидаемое повышение доходности актива пропорционально ожидаемой доходности портфеля всех доступных активов («рыночный портфель»). То есть CAPM показывает, что

$$R - R_f = \beta(R_m - R_f), \quad (4.12)$$

где  $R_m$  – ожидаемая доходность рыночного портфеля, а  $\beta$  – коэффициент теоретической регрессии  $R - R_f$  на  $R_m - R_f$ . На практике безрисковая доходность часто берется равной ставке процента по краткосрочным облигациям правительства США. Согласно CAPM, акция с  $\beta < 1$  имеет меньший риск, чем рыночный портфель и, следовательно, имеет более низкую избыточную доходность, чем рыночный портфель. В противоположность этому акция с  $\beta > 1$  более рисковая, чем рыночный портфель и, таким образом, имеет более высокую ожидаемую доходность.

Бета-акция стала рабочей лошадкой инвестиционной отрасли, и вы можете узнать ее оцененные значения для акций приблизительно сотни компаний на веб-сайтах инвестиционных компаний. Эти беты обычно оцениваются с помощью МНК-регрессии фактической избыточной доходности акции против фактической избыточной доходности рыночного портфеля.

В таблице ниже приведены оцененные значения беты для акций семи американских компаний. Производители с низким риском потребительских товаров, такие как Kellogg, имеют акции с низкими бетами; более рисковые акции имеют высокие беты.

Wal-Mart (розничный продавец со скидками)	0,3
Kellogg (сухие завтраки)	0,5
Waste Management (размещение отходов)	0,6
Verizon (телефонные коммуникации)	0,6
Microsoft (программное обеспечение)	1,0
Best Buy (розничное электронное оборудование)	1,3
Bank of America (банк)	2,4

\* Доходность по инвестициям представляет собой изменение цены актива плюс любой платеж (дивиденд) по инвестиции как процент от их начальной цены. Например, акция, купленная 1 января за 100 долл., по которой затем платится 2,50 долл. дивидендов в течение года, и проданная 31 декабря за 105 долл., будет иметь доходность  $R = [(\$105 - \$100) + \$2,50] / \$100 = 7,5\%$ .



МНК-оценки имеют, кроме того, желаемые теоретические свойства. Они аналогичны свойствам  $\bar{Y}$ , изученным в разделе 3.1 как оценки генерального среднего. При предположениях, введенных в разделе 4.4, МНК-оценка является несмещенной и состоятельной. МНК-оценка также эффективна среди определенного класса несмешанных оценок; однако результат, касающийся эффективности, выполняется при некоторых дополнительных специальных условиях, и дальнейшее обсуждение этого результата откладывается до раздела 5.5.

### 4.3. Критерии качества приближения данных моделью

Имея оцененную линейную регрессию, вы могли бы задаться вопросом о том, насколько хорошо линия регрессии описывает данные. Какую долю дисперсии зависимой переменной (большую или малую) объясняет регрессор? Плотно ли располагаются около линии регрессии наблюдения или нет?

$R^2$  и стандартная ошибка регрессии измеряют, как хорошо МНК-регрессия аппроксимирует (подгоняет) данные.  $R^2$  принимает значения между 0 и 1 и измеряет долю дисперсии  $Y_i$ , которая объясняется  $X_i$ . Стандартная ошибка регрессии показывает, насколько, как правило, далеки  $Y_i$  от предсказанного значения.

#### $R^2$

$R^2$  регрессии<sup>1</sup> равен доле выборочной дисперсии  $Y_i$ , объясненной (или предсказанной)  $X_i$ . Определения предсказанного значения и остатка (см. вставку «Основные понятия 4.2») позволяют нам записать зависимую переменную  $Y_i$  как сумму предсказанного значения  $\hat{Y}_i$  плюс остаток  $\hat{u}_i$ :

$$Y_i = \hat{Y}_i + \hat{u}_i \quad (4.13)$$

В этом обозначении  $R^2$  является отношением выборочной дисперсии  $\hat{Y}_i$  к выборочной дисперсии  $Y_i$ .

Математически  $R^2$  может быть записан как отношение объясненной суммы квадратов к полной сумме квадратов. *Объясненная сумма квадратов (ESS)* – это сумма квадратов отклонений предсказанных значений  $Y_i$ , то есть  $\hat{Y}_i$  от их среднего, а *полная сумма квадратов (TSS)* есть сумма квадратов отклонений  $Y_i$  от их среднего:

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2; \quad (4.14)$$

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (4.15)$$

В уравнении (4.14) используется тот факт, что выборочное среднее МНК-оценки  $\hat{Y}_i$  равно  $\bar{Y}$  (доказано в приложении 4.3).

---

<sup>1</sup>  $R^2$  регрессии довольно часто называют коэффициентом детерминации регрессии. – Примеч. науч. ред. перевода.

Тогда  $R^2$  записывается как отношение объясненной суммы квадратов к полной сумме квадратов:

$$R^2 = \frac{ESS}{TSS}. \quad (4.16)$$

С другой стороны,  $R^2$  может быть записан в терминах доли дисперсии  $Y_i$ , не объясненной  $X_i$ . Сумма квадратов остатков или  $SSR$  – это сумма квадратов МНК-остатков регрессии:

$$SSR = \sum_{i=1}^n \hat{u}_i^2. \quad (4.17)$$

В приложении 4.3 показано, что выполняется соотношение  $TSS = ESS + SSR$ . Таким образом,  $R^2$  также может быть выражен как разность единицы и отношения суммы квадратов остатков к полной сумме квадратов:

$$R^2 = 1 - \frac{SSR}{TSS}. \quad (4.18)$$

Наконец,  $R^2$  регрессии  $Y$  на единственный регрессор  $X$  равен квадрату выборочного коэффициента корреляции между  $Y$  и  $X$ .

$R^2$  может принимать значения в интервале от 0 до 1. Если  $\hat{\beta}_1 = 0$ , тогда  $X_i$  не объясняет ничего из дисперсии  $Y_i$  и предсказанное по регрессии значение  $Y_i$  равно выборочному среднему  $\bar{Y}_i$ . В этом случае объясненная сумма квадратов равна нулю, а сумма квадратов остатков равна полной сумме квадратов; таким образом,  $R^2$  также равен нулю. В отличие от этого, если  $X_i$  полностью объясняет дисперсию  $Y_i$ , тогда  $Y_i = \hat{Y}_i$  для всех  $i$ , и любой остаток равен нулю (т.е.  $\hat{u}_i = 0$ ), поэтому  $ESS = TSS$  и  $R^2 = 1$ . В общем случае  $R^2$  не дает крайние значения 0 или 1, а находится где-то между ними. Если  $R^2$  близок к единице, это указывает на то, что регрессор хорошо предсказывает  $Y_i$ . Если же, наоборот,  $R^2$  близок к нулю, это свидетельство в пользу того, что регрессор не очень хороший с точки зрения предсказания  $Y_i$ .

## Стандартная ошибка регрессии

Стандартная ошибка регрессии ( $SER$ ) это оценка стандартного отклонения ошибок регрессии  $u_i$ . Единицы измерения  $u_i$  и  $Y_i$  одинаковые, поэтому  $SER$  является мерой разброса наблюдений около линии регрессии, измеряемой в тех же единицах, что и зависимая переменная. Например, если зависимая переменная измеряется в долларах, то  $SER$  измеряет величину обычных отклонений от линии регрессии, то есть величину обычной ошибки регрессии – в долларах.

Поскольку ошибки регрессии  $u_1, \dots, u_n$  ненаблюдаются,  $SER$  вычисляется с использованием выборочных аналогов, МНК-остатков  $\hat{u}_1, \dots, \hat{u}_n$ . Формула для  $SER$  записывается в таком виде:

$$SER = s_{\hat{u}}, \text{ где } s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}, \quad (4.19)$$

где в формуле для  $s_{\hat{u}}^2$  используется тот факт, что выборочное среднее МНК-остатков равно нулю (см. приложение 4.3).

Формула для  $SER$  в уравнении (4.19) аналогична формуле для выборочного стандартного отклонения  $Y$ , заданного в уравнении (3.7) в разделе 3.2, за исключением того, что  $Y_i - \bar{Y}$  в уравнении (3.7) заменяется на  $\hat{u}_i$ , и знаменатель в уравнении (3.7) равен  $n-1$ , тогда как здесь он равен  $n-2$ . Причина такой корректировки (т.е. использования знаменателя  $n-2$  вместо  $n$ ) здесь та же самая, что в случае использования знаменателя  $n-1$  в уравнении (3.7): он корректирует небольшое смещение вниз из-за потери степеней свободы, возникающее из-за необходимости оценки двух коэффициентов. Это называется коррекцией на «степени свободы», поскольку два коэффициента оценены ( $\beta_0$  и  $\beta_1$ ), две «степени свободы» в данных были потеряны, поэтому знаменатель дроби равен  $n-2$ . (Математика, лежащая в основе этого утверждения, обсуждается в разделе 5.6.) Когда  $n$  велико, различие между делением на  $n$ ,  $n-1$  или  $n-2$  незначительно.

### **Пример: применение к данным по результатам тестов**

Уравнение (4.11) – это линия регрессии, оцененная с использованием данных по результатам школьных тестов в Калифорнии, связывающая оценку по стандартному тесту (*TestScore*) и соотношение учеников и учителей (*STR*). Коэффициент детерминации  $R^2$  этой регрессии составляет 0,051, или 5,1 %, а  $SER$  равна 18,6.

Коэффициент детерминации  $R^2$ , равный 0,051, означает, что регрессор *STR* объясняет 5,1 % дисперсии зависимой переменной *TestScore*. График линии регрессии на рисунке 4.3 накладывается на диаграмму рассеяния данных *TestScore* и *STR*. Как показывает диаграмма рассеяния, соотношение учеников и учителей объясняет некоторую вариацию в оценках за тест, но большая доля дисперсии остается необъясненной.

Стандартная ошибка регрессии (*SER*), равная 18,6, означает, что стандартное отклонение остатков регрессии равно 18,6 и измеряется в тех же единицах, что и результаты по стандартному тесту. Поскольку стандартное отклонение измеряет разброс данных, стандартная ошибка регрессии, равная 18,6, означает, что существует большой разброс на диаграмме рассеяния на рисунке 4.3 около линии регрессии, измеренный в единицах измерения зависимой переменной. Этот большой разброс означает, что предсказания результатов тестов, сделанные на основе только соотношения учеников и учителей для этого района, будут часто ошибаться на большую величину.

Как мы должны интерпретировать такой низкий  $R^2$  и высокую *SER*? Тот факт, что  $R^2$  регрессии является низким (и *SER* высокой), сам по себе не означает, что эта регрессия «хорошая» или «плохая». Низкий  $R^2$  говорит нам о том, что на результаты тестов влияют и другие важные факторы. Эти факторы могут включать территориальные различия между учениками, качественные различия между школами, не связанные с соотношением учеником и учителей в них,

или удачу, сопутствующую ученикам во время проведения теста. Низкий  $R^2$  и высокая  $SER$  не могут сказать нам о том, какие это факторы, но они показывают, что соотношение учеников и учителей объясняет лишь небольшую часть дисперсии показателя результатов тестов в имеющихся данных.

#### 4.4. Предположения метода наименьших квадратов

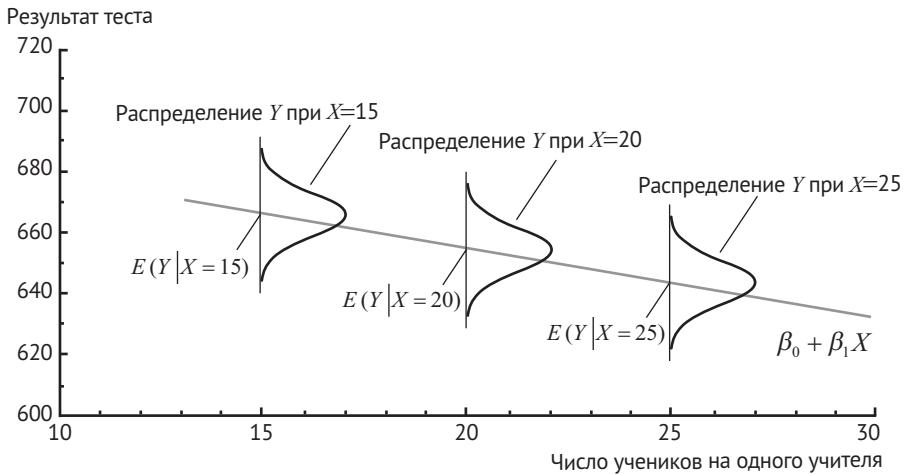
В этом разделе описаны три основных предположения модели линейной регрессии, при которых МНК дает хорошую оценку неизвестных регрессионных коэффициентов  $\beta_0$  и  $\beta_1$ . На первый взгляд эти предположения могут показаться абстрактными. Тем не менее каждое из них имеет естественную интерпретацию, и понимание этих предположений является существенным для понимания того, когда МНК будет – или не будет – давать хорошие оценки коэффициентов регрессии.

##### **Предположение № 1: условное распределение $u_i$ , относительно $X_i$ , имеет нулевое среднее**

Первое из трех *предположений метода наименьших квадратов* заключается в том, что условное распределение  $u_i$  относительно  $X_i$  имеет нулевое среднее. Это предположение является формальным математическим утверждением о других «факторах», содержащихся в  $u_i$ , и утверждает, что эти другие факторы не связаны с  $X_i$  в том смысле что при заданном значении  $X_i$  среднее значение распределения этих других факторов равно нулю.

Графическая иллюстрация этого предположения приведена на рисунке 4.4. Теоретическая регрессия – это отношение, которое выполняется в среднем между размером класса и результатами тестов в генеральной совокупности, а компонента ошибок  $u_i$  представляет другие факторы, из-за которых оценки за тесты в данном округе отличаются от предсказанных значений, основанных на линии теоретической регрессии. Как показано на рисунке 4.4, для заданного значения соотношения числа учеников и учителей, скажем, 20 учеников на одного учителя, иногда эти другие факторы положительно влияют на результаты тестов, и тогда  $u_i > 0$ , а иногда – отрицательно ( $u_i < 0$ ), но в среднем по всей генеральной совокупности предсказанные значения правильны. Другими словами, при заданном  $X_i = 20$  среднее значение распределения  $u_i$  равно нулю. На рисунке 4.4 показано распределение  $u_i$ , центрированное на линии теоретической регрессии в  $X_i = 20$  и, в более общем случае, для других значений  $X_i$ , равных  $x$ . Иначе говоря, распределение  $u_i$  относительно  $X_i = x$  имеет нулевое среднее; говоря математически,  $E(u_i | X_i = x) = 0$ , или, в более простом обозначении,  $E(u_i | X_i) = 0$ .

Как показано на рисунке 4.4, предположение, о том, что  $E(u_i | X_i) = 0$ , эквивалентно предположению о том, что линия теоретической регрессии есть условное среднее  $Y_i$  при заданном  $X_i$  (математическое доказательство этого утверждения остается в качестве упражнения 4.6).



**Рисунок 4.4. Условное распределение вероятности и линия теоретической регрессии**

Рисунок показывает условную вероятность результатов тестов по округам Калифорнии для классов размером 15, 20 и 25 учеников. Среднее значение условного распределения результатов тестов при заданном соотношении учеников и учителей  $E(Y|X)$  – это линия теоретической регрессии. При заданном значении  $X$  показатель  $Y$  распределен около линии регрессии, и ошибки  $\varepsilon = Y - (\beta_0 + \beta_1 X)$  имеют условное среднее, равное нулю для всех значений  $X$ .

**Условное среднее ошибки и в случайном управляемом эксперименте.** В случайном контролируемом эксперименте субъекты отбираются случайным образом в исследуемую группу ( $X = 1$ ) или контрольную группу ( $X = 0$ ). Случайный отбор обычно выполняется при помощи компьютерной программы, которая использует информацию о субъектах и гарантирует, что  $X$  распределен независимо от всех персональных характеристик субъекта. Случайный отбор гарантирует независимость  $X$  и  $\varepsilon$ , что в свою очередь означает, что условное среднее  $\varepsilon$  относительно  $X$  равно нулю.

В наблюдаемых данных  $X$  отбирается в эксперименте неслучайно. Лучшее, на что мы можем надеяться вместо этого, это то, что  $X$  будто бы случайно отбирается, в том смысле что  $E(\varepsilon_i | X_i) = 0$ . Выполняется ли это предположение в конкретном эмпирическом приложении с конкретными наблюдаемыми данными, является вопросом, требующим тщательного размышления, и мы будем возвращаться к этому вопросу неоднократно.

**Корреляция и условное среднее.** Вспомним из раздела 2.3, что если условное среднее одной случайной величины относительно другой случайной величины равно нулю, то две случайные величины имеют нулевую ковариацию и, таким образом, некоррелированы [уравнение (2.27)]. Значит, из предположения о том, что условное среднее  $E(\varepsilon_i | X_i) = 0$ , следует, что  $X_i$  и  $\varepsilon_i$  не коррелированы или  $\text{corr}(X_i, \varepsilon_i) = 0$ . Поскольку корреляция есть мера линейной взаимосвязи, это не значит, что верно и обратное; даже если  $X_i$  и  $\varepsilon_i$  не коррелированы, условное среднее  $\varepsilon_i$  относительно  $X_i$  может быть ненулевым. Однако если  $X_i$  и  $\varepsilon_i$  коррелированы, то  $E(\varepsilon_i | X_i)$  в этом случае будет ненулевым. Это обстоятельство часто бывает удобным при обсуждении предположения об условном среднем в терминах возможной корреляции между  $X_i$  и  $\varepsilon_i$ . Если  $X_i$  и  $\varepsilon_i$  коррелированы, тогда предположение об условном среднем нарушено.

### **Предположение № 2: $(X_i, Y_i), i=1, \dots, n$ , независимые и одинаково распределенные**

Второе предположение метода наименьших квадратов заключается в том, что  $(X_i, Y_i), i=1, \dots, n$  независимы и одинаково распределены (i.i.d.). Как обсуждалось в разделе 2.5 (см. вставку «Основные понятия 2.5»), это предположение является утверждением о способе формирования выборки. Если наблюдения отобраны простым случайным образом из единственной большой генеральной совокупности, тогда  $(X_i, Y_i), i=1, \dots, n$  являются i.i.d. Например, пусть  $X$  – это возраст работника, а  $Y$  – его или ее зарплата, и представим, что мы выбираем человека случайным образом из генеральной совокупности всех работников. Этот случайно выбранный человек будет иметь определенный возраст и зарплату (т.е.  $X$  и  $Y$  будут давать определенные значения). Если выборка из  $n$  работников извлечена из этой генеральной совокупности, тогда  $(X_i, Y_i), i=1, \dots, n$  обязательно имеют одинаковое распределение. Если они выбраны случайно, они также распределены независимо от наблюдения к наблюдению; то есть они i.i.d.

Предположение об i.i.d. является разумным для многих схем сбора данных. Например, данные обследования населения со случайно выбранным подмножеством всей генеральной совокупности (всего населения) обычно можно рассматривать как i.i.d.

Тем не менее не все схемы выбора предоставляют i.i.d. наблюдения  $(X_i, Y_i)$ . Один из примеров: когда значение  $X$  не выбирается случайно из генеральной совокупности, а задается исследователем в рамках эксперимента. Например, предположим, что садовод хочет изучить влияние различных методов органической прополки ( $X$ ) на производство помидоров ( $Y$ ) и, соответственно, выращивает помидоры на различных участках, используя различные органические методы прополки. Если он выбирает технику (уровень  $X$ ), которая будет использоваться на  $i$ -м участке и применяет ее технику на  $i$ -м участке для всех повторений эксперимента, тогда значение  $X_i$  не изменяется от одной выборки к другой. Таким образом,  $X_i$  не является случайным (хотя результат  $Y_i$  случайный), поэтому такая схема выбора – не i.i.d. Результаты, представленные в этой главе и разработанные для i.i.d. регрессоров, также верны для случая неслучайных регрессоров. Случай неслучайного регрессора, однако, достаточно специфический. Например, в современных экспериментальных исследованиях уровень  $X$  на различных участках выбирается при помощи компьютерного генератора случайных чисел, тем самым обходятся любые возможные отклонения садовода от него (он может использовать свой любимый способ прополки для помидоров на самом солнечном участке). В таких условиях уровень  $X$  также является случайнym, а  $(X_i, Y_i)$  являются i.i.d.

Другой пример не i.i.d. выборки – это ситуация, когда наблюдения относятся к одному и тому же объекту в разные моменты времени. Например, у нас могут иметься данные по уровню запасов ( $Y$ ) какой-либо фирмы и процентной ставки, под которую фирма может занимать деньги ( $X$ ), и эти данные собирались в те-

чение 30 лет ежеквартально. Это пример временного ряда, и основная особенность данных такого типа заключается в том, что наблюдения, расположенные близко друг к другу, не являются независимыми, а, как правило, коррелированы друг с другом; если процентные ставки низки сейчас, они, скорее всего, будут низкими в следующем квартале. Этот пример корреляции нарушает часть «независимость» в определении i.i.d. Такие данные приводят к множеству осложнений, которые лучше всего решать после разработки основных инструментов регрессионного анализа.

### **Предположение № 3: большие выбросы маловероятны**

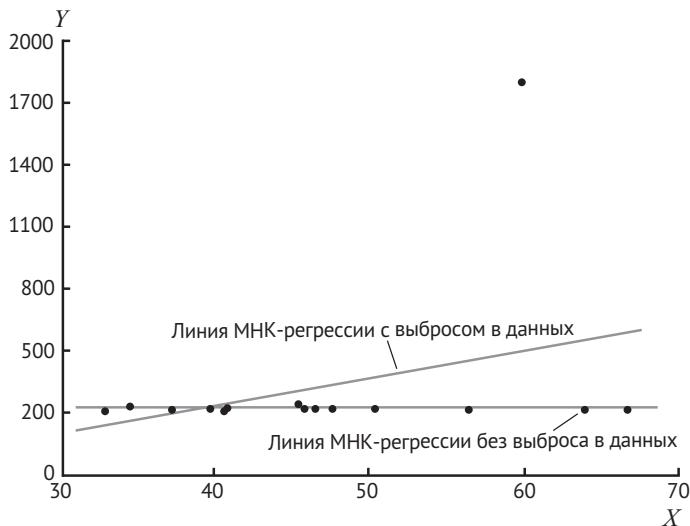
Третье предположение метода наименьших квадратов заключается в том, что большие выбросы, то есть тот факт, что наблюдения со значениями  $X_i$  или  $Y_i$  или обоих находятся далеко за пределами обычного диапазона данных, – маловероятны. Большие выбросы могут сделать результаты оценки МНК-регрессии странными. Эта потенциальная чувствительность МНК к экстремальным выбросам показана на рисунке 4.5 для гипотетических данных.

В данной книге предположение о том, что большие выбросы маловероятны, математически оформлено в предположении, что  $X$  и  $Y$  имеют конечный четвертый момент:  $0 < E(X_i^4) < \infty$  и  $0 < E(Y_i^4) < \infty$ . По-другому это предположение можно сформулировать в терминах эксцесса: мы говорим о том, что эксцесс  $X$  и  $Y$  конечен.

Предположение о конечности эксцесса вводится для того, чтобы математически обосновать аппроксимацию на больших выборках для распределения МНК-статистик. Мы встречались с этим предположением в главе 3, когда обсуждали состоятельность выборочной дисперсии. Более точно, уравнение (3.9) показывает, что выборочная дисперсия  $s_Y^2$  является состоятельной оценкой теоретической дисперсии  $\sigma_Y^2$  ( $s_Y^2 \xrightarrow{P} \sigma_Y^2$ ). Если  $Y_1, \dots, Y_n$  являются i.i.d. и четвертый момент  $Y$  конечен, то закон больших чисел, сформулированный во вставке

«Основные понятия 2.6», применяется к среднему  $\frac{1}{n} \sum_{i=1}^n (Y_i - \mu_Y)^2$ , и это есть основной шаг, используемый в приложении 3.3 при доказательстве того, что  $s_Y^2$  является состоятельной.

Одним из источников больших выбросов являются ошибки при вводе данных, такие как типографические ошибки или некорректное использование различных единиц для различных наблюдений. Например, представим себе сбор данных о росте студентов в метрах, но со случайной записью о росте одного из них в сантиметрах. Один из способов найти выбросы – начертить график по имеющимся данным. Если вы решите, что выброс происходит из-за ошибки ввода данных, то вы можете или исправить ошибки, или, если это невозможно, удалить наблюдение из набора данных.



**Рисунок 4.5. Чувствительность МНК к большим выбросам**

В рассматриваемом гипотетическом наборе данных присутствует один выброс: регрессионная линия МНК, оцененная с выбросом, показывает сильную положительную взаимосвязь между  $X$  и  $Y$ , но регрессионная линия МНК, оцененная без выброса, не показывает взаимосвязи.

Если оставить в стороне ошибки ввода данных, предположение о конечном экспесссе является правдоподобным во многих экономических приложениях. Число учеников в классе ограничено физической вместимостью класса; лучшее, что вы можете сделать в стандартизированном тесте, это ответить правильно на все вопросы, а худшее – ответить неправильно на каждый из них. Поскольку количество учеников в классе и результаты теста конечны, то они обязательно имеют конечный экспессс. В целом у всех широко используемых распределений, таких как нормальное распределение, первые четыре момента конечны. Тем не менее существуют распределения, имеющие бесконечные четвертые моменты, и это обстоятельство исключает их из рассмотрения.

### Использование предположений наименьших квадратов

Три предположения метода наименьших квадратов для модели линейной регрессии перечислены во вставке «Основные понятия 4.3». Предположения метода наименьших квадратов играют двойную роль, и мы будем возвращаться к ним неоднократно на протяжении всего учебника.

## ОСНОВНЫЕ ПОНЯТИЯ 4.3

### Предположения метода наименьших квадратов

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots, n,$$

где

1. Ошибка имеет нулевое условное среднее при заданном  $X_i$ :  
 $E(u_i | X_i) = 0$ .
2.  $(X_i, Y_i), i = 1, \dots, n$  — независимые и одинаково распределенные (i.i.d.), извлеченные из их совместного распределения.
3. Большие выбросы маловероятны:  $X_i$  и  $Y_i$  имеют ненулевой конечный четвертый момент.

Эти предположения выполняют, в первую очередь, математическую роль: если они выполняются, то, как показано в следующем разделе, в больших выборках МНК-оценка имеет выборочное распределение, являющееся нормальным. В свою очередь это нормальное распределение в больших выборках позволяет нам разработать методы проверки гипотез и построения доверительных интервалов с помощью МНК-оценок.

Вторая роль рассматриваемых предположений заключается в возможности формулирования трудностей при оценке МНК-регрессии, возникающих из-за нарушения этих предположений. Как мы увидим, первое предположение метода наименьших квадратов является самым важным для рассмотрения его на практике. Одна из причин, почему первое предположение метода наименьших квадратов может не выполняться на практике, обсуждается в главе 6, а дополнительные причины обсуждаются в разделе 9.2.

Также важно рассмотреть, выполняется ли второе предположение применительно к реальным данным. Хотя оно является правдоподобным для многих межобъектных выборок, предположение независимости является маловероятным для временных рядов. Таким образом, регрессионные методы, разработанные при выполнении второго предположения, требуют модификации для некоторых приложений к данным, имеющим структуру временных рядов.

Третье предположение служит напоминанием о том, что МНК, как и выборочное среднее, может быть чувствителен к большим выбросам. Если ваша выборка содержит большие выбросы, вы должны тщательно их изучить, чтобы убедиться, что эти наблюдения правильно записаны и действительно принадлежат этой выборке.

## 4.5. Выборочное распределение МНК-оценки

Поскольку МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  вычисляются по данным случайной выборки, сами оценки являются случайными величинами с вероятностным распределением – выборочным распределением, описывающим значения, которые они могли бы принимать на различных возможных случайных выборках. В данном разделе рассмотрены эти выборочные распределения. На малых выборках эти распределения сложные, но на больших выборках они аппроксимируются нормальным по центральной предельной теореме.

### Выборочное распределение МНК-оценки

*Описание выборочного распределения  $\bar{Y}$ .* Вспомним обсуждение в разделах 2.5 и 2.6, касающееся распределения выборочного среднего  $\bar{Y}$ , являющегося оценкой неизвестного теоретического среднего  $Y, \mu_Y$ . Поскольку  $\bar{Y}$  вычисляется с использованием случайной выборки,  $\bar{Y}$  является случайной величиной, которая принимает различные значения в зависимости от имеющейся выборки; вероятность этих различных значений описывается в выборочном распределении. Хотя выборочное распределение  $\bar{Y}$  может быть сложным при небольшом размере выборки, мы можем сделать определенные утверждения о нем, которые выполняются для всех  $n$ . В частности, выборочное среднее равно  $\mu_Y$ , то есть  $E(\bar{Y}) = \mu_Y$ , поэтому  $\bar{Y}$  является

несмешенной оценкой  $\mu_y$ . Если  $n$  велико, то мы можем сказать о выборочном распределении больше. В частности, центральная предельная теорема (раздел 2.6) говорит нам о том, что это распределение приблизительно нормальное.

**Выборочное распределение  $\hat{\beta}_0$  и  $\hat{\beta}_1$ .** Все эти идеи переносятся на МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  неизвестных константы  $\beta_0$  и углового коэффициента  $\beta_1$  линии теоретической регрессии. Поскольку МНК-оценка вычисляется с использованием случайной выборки,  $\hat{\beta}_0$  и  $\hat{\beta}_1$  являются случайными величинами, которые принимают различные значения от одной выборки к другой; вероятность этих различных значений описывается в их выборочных распределениях.

Выборочное распределение  $\hat{\beta}_0$  и  $\hat{\beta}_1$  может быть сложным при малых размерах выборки, но, несмотря на это, мы можем сделать определенные выводы о них, которые верны для всех  $n$ . В частности, выборочные средние распределений  $\hat{\beta}_0$  и  $\hat{\beta}_1$  есть  $\beta_0$  и  $\beta_1$ . Другими словами, в предположениях метода наименьших квадратов, сформулированных во вставке «Основные понятия 4.3»,

$$E(\hat{\beta}_0) = \beta_0 \text{ и } E(\hat{\beta}_1) = \beta_1; \quad (4.20)$$

то есть  $\hat{\beta}_0$  и  $\hat{\beta}_1$  являются несмешенными оценками  $\beta_0$  и  $\beta_1$ . Доказательство того что  $\hat{\beta}_1$  является несмешенной оценкой, приведено в приложении 4.3, а доказательство того что  $\hat{\beta}_0$  не смешена, остается в качестве упражнения 4.7.

Если выборка достаточно велика, по центральной предельной теореме выборочное распределение  $\hat{\beta}_0$  и  $\hat{\beta}_1$  хорошо аппроксимируется двумерным нормальным распределением (раздел 2.4). Это предполагает, что безусловные распределения  $\hat{\beta}_0$  и  $\hat{\beta}_1$  являются нормальными на больших выборках.

В качестве аргумента можно обратиться к центральной предельной теореме. Технически центральная предельная теорема касается распределения среднего (как  $\bar{Y}$ ). Если вы изучите числитель в выражении (4.7) для  $\hat{\beta}_1$ , то увидите, что он также является некоторым типом среднего значения — не простого среднего, как  $\bar{Y}$ , но среднего значения произведения  $(X_i - \bar{X})(Y_i - \bar{Y})$ . Как указывается далее в приложении 4.3, центральная предельная теорема применяется к этому среднему так же, как и к более простому среднему  $\bar{Y}$ , из чего можно сделать вывод об асимптотической нормальности (о нормальности распределения в больших выборках) распределения  $\hat{\beta}_1$ .

### Распределение на больших выборках $\hat{\beta}_0$ и $\hat{\beta}_1$

Если предположения метода наименьших квадратов из вставки «Основные понятия 4.3» выполняются, то на больших выборках  $\hat{\beta}_0$  и  $\hat{\beta}_1$  имеют совместное нормальное распределение.  $\hat{\beta}_1$  распределен асимптотически (в больших выборках) нормально как  $N(\beta_1, \sigma_{\beta_1}^2)$ , где дисперсия этого распределения  $\sigma_{\beta_1}^2$  равна:

$$\sigma_{\beta_1}^2 = \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{\left[\text{var}(X_i)\right]^2}. \quad (4.21)$$

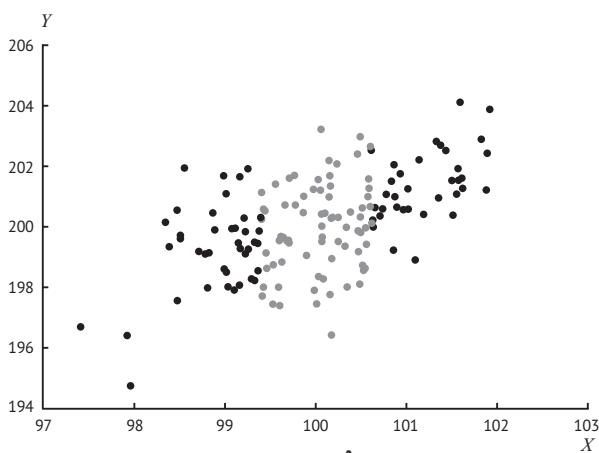
Асимптотически (на больших выборках) нормальное распределение  $\hat{\beta}_0$  имеет вид:  $N(\beta_0, \sigma_{\beta_0}^2)$ ,

$$\text{где } \sigma_{\beta_0}^2 = \frac{1}{n} \frac{\text{var}(H_i u_i)}{\left[E(H_i^2)\right]^2}, \text{ а } H_i = 1 - \left[ \frac{\mu_X}{E(X_i^2)} \right] X_i. \quad (4.22)$$

Нормальная аппроксимация распределения МНК-оценки в больших выборках описана во вставке «Основные понятия 4.4». (В приложении 4.3 излагается математический вывод этих формул.) Соответственно, на практике возникает вопрос о том, насколько большим должен быть размер выборки  $n$  для того, чтобы эти аппроксимации были верны. В разделе 2.6 мы предложили, что  $n=100$  достаточно велико для того, чтобы выборочное распределение  $\bar{Y}$  хорошо приближалось нормальным распределением, а иногда достаточно и меньших  $n$ . Этот критерий переносится на более сложные ситуации, возникающие в регрессионном анализе. Практически во всех современных эконометрических приложениях  $n > 100$ , поэтому мы будем рассматривать нормальную аппроксимацию распределения МНК-оценки как надежную, если нет веских причин думать иначе.

Результаты, представленные во вставке «Основные понятия 4.4», предполагают, что МНК-оценка состоятельна, то есть при больших размерах выборки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  будут близки к теоретическим коэффициентам  $\beta_0$  и  $\beta_1$  с высокой вероятностью. Это происходит потому, что дисперсии  $\sigma_{\beta_0}^2$  и  $\sigma_{\beta_1}^2$  оценок стремятся к нулю при возрастании  $n$  ( $n$  появляется в знаменателе формулы для дисперсии), поэтому распределение МНК-оценок будет плотно сосредоточено около их средних  $\beta_0$  и  $\beta_1$  при больших  $n$ .

Другим следствием из распределений, представленных во вставке «Основные понятия 4.4», является то, что в общем случае чем больше дисперсия  $X_i$ , тем меньше дисперсия  $\hat{\beta}_1$  оценки  $\hat{\beta}_1$ . Математически это следует потому, что дисперсия  $\hat{\beta}_1$  в уравнении (4.21) обратно пропорциональна квадрату дисперсии  $X_i$ : чем больше  $\text{var}(X_i)$ , тем больше знаменатель в уравнении (4.21), поэтому меньше  $\sigma_{\beta_1}^2$ . Чтобы лучше понять, почему это так, посмотрим на рисунок 4.6, на котором представлена диаграмма рассеяния 150 искусственно сгенерированных точек  $(X, Y)$ . Точки, указанные серым цветом, – это 75 наблюдений, близких к  $\bar{X}$ . Предположим, что вас попросили провести как можно более точно линию либо через серые точки, либо через черные – как бы вы это сделали? Было бы легче провести четкую линию через черные точки, дисперсии у которых больше, чем у серых. Кроме того, чем больше дисперсия  $X$ , тем точнее  $\hat{\beta}_1$ .



**Рисунок 4.6. Дисперсия  $\hat{\beta}_1$  и дисперсия  $X$**

Серые точки представляют множество  $X_i$  с малой дисперсией. Чёрные точки представляют множество  $X_i$  с большой дисперсией. Линия регрессии может быть оценена более точно по чёрным точкам, а не по серым точкам.

Распределения из вставки «Основные понятия 4.4» также предполагают, что чем меньше дисперсия ошибок  $u_i$ , тем меньше дисперсия  $\sigma_{\beta_1}^2$ . Это следует из уравнения (4.21), поскольку  $u_i$  входит в числитель  $\sigma_{\beta_1}^2$ , но не в знаменатель: если бы все  $u_i$  были меньше наполовину, но  $X$  не изменились бы, тогда  $\sigma_{\beta_1}$  была бы меньше на половину, а  $\sigma_{\beta_1}^2$  была бы меньше на четверть (упражнение 4.13). Говоря менее формально с математической точки зрения, если ошибки меньше (при фиксированных  $X$ ), тогда данные будут более плотно рассеиваться около линии регрессии генеральной совокупности, поэтому ее угловой коэффициент будет оцениваться более точно.

Нормальная аппроксимация выборочного распределения  $\hat{\beta}_0$  и  $\hat{\beta}_1$  является мощным инструментом. Зная ее, мы можем разработать методы, которые позволят нам делать выводы о теоретических значениях коэффициентов регрессии, используя информацию только о выборке данных.

## 4.6. Заключение

Эта глава посвящена применению метода наименьших квадратов для получения оценок константы и углового коэффициента линии теоретической регрессии с использованием выборки из  $n$  наблюдений зависимой переменной  $Y$  и единственного регрессора  $X$ . Существует много способов провести прямую линию через диаграмму рассеяния, но МНК-оценки линии регрессии имеют несколько достоинств. Если предположения метода наименьших квадратов выполняются, тогда МНК-оценки углового коэффициента и константы несмещенные, состоятельные и имеют выборочное распределение с дисперсией, которая обратно пропорциональна размеру выборки  $n$ . Кроме того, если  $n$  велико, тогда выборочное распределение МНК-оценок является нормальным.

Эти важные свойства выборочного распределения МНК-оценок выполняются при трех предположениях метода наименьших квадратов.

Первое предположение заключается в том, что условное среднее компоненты ошибок в модели линейной регрессии относительно  $X$  равно нулю. Это предположение подразумевает, что МНК-оценка является несмещенной.

Второе предположение говорит о том, что  $(X_i, Y_i)$  является i.i.d., как в случае, когда данные собираются простым случайным образом. Из этого предположения следуют формулы, приведенные во вставке «Основные понятия 4.4» для дисперсий выборочных распределений МНК-оценок.

Третье предположение заключается в том, что большие выбросы маловероятны. Говоря более формально,  $X$  и  $Y$  имеют конечный четвертый момент (конечный эксцесс). Причина введения этого предположения состоит в том, что МНК может давать ненадежные оценки, если в выборке присутствуют большие выбросы. Взятые вместе, эти три предположения метода наименьших квадратов подразумевают, что МНК-оценка распределена асимптотически нормально, как описано во вставке «Основные понятия 4.4».

Результаты этой главы описывают выборочное распределение МНК-оценки. Сами по себе эти результаты недостаточны, чтобы проверить гипотезу о значе-

ния  $\beta_1$  или построить доверительный интервал для  $\beta_1$ . Для этого необходимо оценить стандартное отклонение выборочного распределения, то есть стандартную ошибку МНК-оценки. Этот шаг – переход от выборочного распределения  $\hat{\beta}_1$  к оценке ее стандартной ошибки, тестированию гипотез и построению доверительных интервалов – будет сделан в следующей главе.

## Выводы

- Линия теоретической регрессии  $\beta_0 + \beta_1 X$  представляет собой среднее значение  $Y$  как функцию от значения  $X$ . Угловой коэффициент  $\beta_1$  является ожидаемым изменением  $Y$ , связанным с изменением  $X$  на единицу. Константа  $\beta_0$  определяет уровень (или высоту) линии регрессии (т.е. точку, в которой она пересекает ось  $Y$ ). Во вставке «Основные понятия 4.1» перечислены основные термины теоретической модели парной линейной регрессии.
- Линия теоретической регрессии может быть оценена с использованием выборки наблюдений  $(X_i, Y_i)$ ,  $i = 1, \dots, n$  методом наименьших квадратов. МНК-оценки константы и углового коэффициента регрессии обозначаются  $\hat{\beta}_0$  и  $\hat{\beta}_1$ .
- Коэффициент детерминации  $R^2$  и стандартная ошибка регрессии ( $SER$ ) являются мерами близости значения  $Y_i$  к оцененной линии регрессии.  $R^2$  находится в интервале между нулем и единицей. Если он близок к единице, это указывает на то, что  $Y_i$ -е близки к линии регрессии. Стандартная ошибка регрессии является оценкой стандартного отклонения ошибок теоретической регрессии.
- Существует три ключевых предположения для модели линейной регрессии: (1) регрессионные ошибки  $u_i$  имеют нулевое условное среднее относительно регрессора  $X_i$ ; (2) выборка наблюдений является i.i.d. случайно отобранный из генеральной совокупности и (3) большие выбросы маловероятны. Если эти предположения выполняются, МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  являются (1) несмещенными, (2) состоятельными и (3) нормально распределенными в больших выборках.

## Основные понятия

Модель парной линейной регрессии (с. 113).

Зависимая переменная (с. 113).

Независимая переменная (с. 113).

Регрессор (с. 113).

Линия теоретической регрессии (с. 113).

Функция теоретической регрессии (с. 113).

Теоретическая константа (с. 114).

Теоретический угловой коэффициент (с. 114).

Теоретические коэффициенты (с. 114).

Параметры (с. 114).

Ошибка (с. 114).

Метод наименьших квадратов (МНК) (с. 118).

Оценки (с. 118).

Линия МНК-регрессии (с. 118).

Линия выборочной регрессии (с. 118, 119).

Функция выборочной регрессии (с. 119).

Предсказанные значения (с. 119).

Остаток (с. 119).

$R^2$  регрессии (с. 123).

Объясненная сумма квадратов (ESS) (с. 123).

Полная сумма квадратов (TSS) (с. 123).

Сумма квадратов остатков (SSR) (с. 124).

Стандартная ошибка регрессии (SER) (с. 124).

Предположения метода наименьших квадратов (с. 126).

### **Вопросы для повторения и закрепления основных понятий**

- 4.1. Объясните различие между  $\hat{\beta}_1$  и  $\beta_1$ , между остатком  $\hat{u}_i$  и ошибкой регрессии  $u_i$  и между предсказанным МНК-значением  $\hat{Y}_i$  и  $E(Y_i | X_i)$ .
- 4.2. Для каждого предположения метода наименьших квадратов приведите пример, в котором предположения справедливы, затем приведите пример, в котором предположения не выполняются.
- 4.3. Сделайте набросок гипотетической диаграммы рассеяния данных для оцененной регрессии с  $R^2 = 0,9$ . Сделайте набросок гипотетической диаграммы рассеяния данных для регрессии с  $R^2 = 0,5$ .

### **Упражнения**

- 4.1. Предположим, что исследователь, используя данные о размере класса (CS) и средних оценках за тест по 100 третьим классам, оценивает МНК-регрессию:

$$\widehat{\text{TestScore}} = 520,4 - 5,82 \times CS, R^2 = 0,08, SER = 11,5.$$

- a) В классе 22 ученика. Какое предсказанное значение средней оценки за тест будет получено для этого класса, исходя из этих оценок?
  - b) В предыдущем году в классе было 19 учеников, а в этом году их стало 23. Какое предсказанное изменение средней оценки за тест можно получить для такого изменения числа учеников в классе?
  - c) Средний по выборке размер класса среди рассматриваемых 100 классов равен 21,4. Чему равна оценка среднего значения результатов тестов среди этих 100 классов? (Подсказка: посмотрите формулы МНК-оценок.)
  - d) Чему равно выборочное стандартное отклонение оценок за тест среди 100 классов? (Подсказка: просмотрите формулы для  $R^2$  и SER.)
- 4.2. Предположим, что мы рассматриваем выборку данных по росту и весу 200 двадцатилетних мужчин из некоторой генеральной совокупности.

Регрессия показателя, характеризующего вес, на показатель роста дает оценки:

$$\widehat{Weight} = -99,41 + 3,94 \times Height, R^2 = 0,81, SER = 10,2,$$

где вес (*Weight*) измеряется в фунтах, а рост (*Height*) измеряется в дюймах.

- a) Чему равно предсказанное по этой регрессии значение веса мужчины, если его рост равен 70 дюймам? 65 дюймам? 74 дюймам?
  - б) Допустим, мужчина продолжает расти и растет на 1,5 дюйма в течение года. Чему равно предсказанное по регрессии изменение его веса?
  - в) Предположим, что мы поменяли единицы измерения показателей веса и роста, и вместо фунтов и дюймов измеряем теперь их в сантиметрах и килограммах. Как изменятся представленные оценки в этом случае? (Пересчитайте все оцененные коэффициенты,  $R^2$  и  $SER$  для новых единиц измерения.)
- 4.3. Регрессия средней недельной зарплаты (*AWE*, измеренной в долларах) от возраста (*Age*, измеренного в годах) на основе случайной выборки 25–65-летних людей с высшим образованием и работающих полный рабочий день приводит к следующему:
- $$\widehat{AWE} = 696,7 + 9,6 \times Age, R^2 = 0,023, SER = 624,1.$$
- а) Объясните значение коэффициентов 696,7 и 9,6.
  - б) Стандартная ошибка регрессии (*SER*) равна 624,1. Каковы единицы измерения *SER*? (Доллары? Годы? Или *SER* безразмерна?)
  - в) Коэффициент детерминации  $R^2$  регрессии равен 0,023. Каковы единицы измерения для  $R^2$ ? (Доллары? Годы? Или  $R^2$  безразмерен?)
  - г) Чему равно предсказанное по регрессии значение зарплаты для 25-летнего работника? Для 45-летнего работника?
  - д) Будет ли регрессия давать надежное предсказание зарплаты для 99-летнего работника? Почему да или почему нет?
  - е) Учитывая ваши знания о распределении зарплат, насколько вероятно, что распределение ошибок регрессии является нормальным? (Подсказка: как вы думаете, распределение зарплат является симметричным или скошенным? Каково наименьшее значение зарплаты и согласуется ли это с нормальным распределением?)
  - ж) Средний возраст работника в выборке составляет 41,6 лет. Каково среднее значение *AWE* в выборке? (Подсказка: посмотрите вставку «Основные понятия 4.2».)
- 4.4. Прочтите вставку «Бета-акции» в разделе 4.2.
- а) Предположим, что значение  $\beta$  больше, чем 1, для определенной акции. Покажите, что дисперсия  $(R - R_f)$  для этой акции больше, чем дисперсия  $(R_m - R_f)$ .
  - б) Предположим, что значение  $\beta$  меньше, чем 1, для определенной акции. Возможно ли, что дисперсия  $(R - R_f)$  для этой акции больше, чем дисперсия  $(R_m - R_f)$ ? (Подсказка: не забудьте ошибку регрессии.)

- в) В данном году норма доходности на трехмесячные казначейские векселя составляет 3,5 %, а норма доходности на крупный диверсифицированный портфель акций (S&P 500) составляет 7,3 %. Для каждой компании, записанной в таблице из вставки, используйте оцененное значение  $\beta$ , чтобы оценить ожидаемую норму доходности акции.
- 4.5. Профессор решает провести эксперимент, чтобы измерить, как влияет эффект нехватки времени на финальную экзаменационную оценку. Он дает каждому из 400 студентов в его курсе одинаковый итоговый экзамен, но некоторые студенты получают 90 минут, чтобы закончить экзамен, в то время как другие получают 120 минут. Студенты распределяются по группам с разным временем на написание экзаменационной работы случайным образом, основываясь на бросании монеты. Пусть  $Y_i$  обозначает число результата экзамена  $i$ -го студента ( $0 \leq Y_i \leq 100$ ), а  $X_i$  обозначает количество времени, за которое студент закончил экзамен ( $X_i = 90$  или  $120$ ). Рассмотрим модель парной линейной регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$ .
- а) Объясните, что представляет компонента  $u_i$ . Почему у разных студентов различные значения  $u_i$ ?
  - б) Объясните, почему  $E(u_i | X_i) = 0$  для этой модели регрессии.
  - в) Выполняются ли другие предположения из вставки «Основные понятия 4.3»?
  - г) Оцененная регрессия имеет вид:  $\hat{Y} = 49 + 0,24X_i$ .
    - (i) Вычислите предсказанное на основе регрессии значение средней оценки студентов к группе, которая писала экзамен 90 минут. Повторите вычисления для случая 120 минут и 150 минут.
    - (ii) Вычислите предсказанный по регрессии выигрыш в оценке для студента, который получает дополнительные 10 минут на экзамене.
- 4.6. Покажите, что первое предположение метода наименьших квадратов, то есть  $E(u_i | X_i) = 0$ , предполагает, что  $E(Y_i | X_i) = \beta_0 + \beta_1 X_i$ .
- 4.7. Покажите, что  $\hat{\beta}_0$  является несмешенной оценкой  $\beta_0$ . (Подсказка: используйте тот факт, что  $\hat{\beta}_1$  является несмешенной оценкой, как показано в приложении 4.3.)
- 4.8. Предположим, что все предположения МНК-регрессии из вставки «Основные понятия 4.3» выполнены, за исключением того, что первое предположение заменяется на предположение  $E(u_i | X_i) = 2$ . Какая часть выводов из вставки «Основные понятия 4.4» останется верной? Каковы изменения? Почему? ( $\hat{\beta}_1$  асимптотически нормально распределена со средним значением и дисперсией, заданными во вставке «Основные понятия 4.4»? Что можно сказать о  $\hat{\beta}_0$ ?)
- 4.9. а) Известно, что оценка коэффициента наклона в модели парной линейной регрессии равна  $\hat{\beta}_1 = 0$ . Покажите, что  $R^2 = 0$ .
- б) Теперь известно, что  $R^2 = 0$ . Предполагает ли это, что  $\hat{\beta}_1 = 0$ ?

- 4.10. Предположим, что  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , где  $(X_i, u_i)$  является i.i.d., и  $X_i$  является случайной величиной Бернули с  $\Pr(X=1)=0,20$ . Когда  $X=1$ ,  $u_i$  имеет вид  $N(0,4)$ ; когда  $X=0$ ,  $u_i$  имеет вид  $N(0,1)$ .
- Покажите, что предположения метода наименьших квадратов из вставки «Основные понятия 4.3» выполнены.
  - Выполните выражение для дисперсии  $\hat{\beta}_1$  в случае больших выборок.  
[Подсказка: оцените все составляющие выражения (4.21).]
- 4.11. Рассмотрим модель парной линейной регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$ .
- Предположим, вы знаете, что  $\beta_0 = 0$ . Выполните формулу оценки наименьших квадратов для  $\beta_1$ .
  - Предположим, вы знаете, что  $\beta_0 = 4$ . Выполните формулу оценки наименьших квадратов для  $\beta_1$ .
- 4.12. а) Покажите, что коэффициент детерминации  $R^2$  в регрессии  $Y$  на  $X$  – это квадрат значения выборочного коэффициента корреляции между  $X$  и  $Y$ . То есть покажите, что  $R^2 = r_{XY}^2$ .
- б) Покажите, что  $R^2$  регрессии  $Y$  на  $X$  тот же самый, что и  $R^2$  в регрессии  $X$  на  $Y$ .
- в) Покажите, что  $\hat{\beta}_1 = r_{XY} (s_Y / s_X)$ , где  $r_{XY}$  – выборочная корреляция между  $X$  и  $Y$ , а  $s_Y$  и  $s_X$  – выборочные стандартные отклонения  $X$  и  $Y$ .
- 4.13. Предположим, что  $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$ , где  $\kappa$  является ненулевой константой, а  $(Y_i, X_i)$  удовлетворяет трем предположениям метода наименьших квадратов. Покажите, что дисперсию  $\hat{\beta}_1$  в больших выборках можно вычислить по формуле  $\sigma_{\hat{\beta}_1}^2 = \kappa^2 \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{\text{var}(X_i)^2}$ . [Подсказка: это выражение равно дисперсии, заданной в уравнении (4.21) и умноженной на  $\kappa^2$ .]
- 4.14. Покажите, что линия выборочной регрессии проходит через точку  $(\bar{X}, \bar{Y})$ .

## Компьютерные упражнения

- E4.1. На веб-сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) вы найдете файл с данными **CPS08**, который содержит расширенную версию множества данных, использованных в таблице 3.1 для 2008 года. Он содержит данные по работникам в возрасте от 25 до 34 лет, работающим полный рабочий в течение года и имеющим диплом о высшем образовании или В.А./В.С степени. Детальный комментарий к данным дан в файле **CPS08\_Description**, также доступном на веб-сайте. (Это те же самые данные, как в **CPS92\_08**, но ограниченные 2008 годом.) В этом упражнении вы будете исследовать взаимосвязь между возрастом работника и его зарплатой. (Как правило, старшие работники имеют больший опыт работы, что влечет за собой более высокие производительность труда и зарплату.)

- a) Оцените регрессию средней почасовой зарплаты (*AHE*) на возраст (*Age*). Чему равна оцененная константа? А чему равен оцененный угловой коэффициент? Используйте оцененную регрессию, чтобы ответить на вопрос: как сильно увеличится зарплата при увеличении возраста на один год?
- b) Боб является 26-летним работником. Предскажите зарплату Боба, используя оцененную регрессию. Алексис является 30-летним работником. Предскажите зарплату Алексис, используя оцененную регрессию.
- c) Учитывает ли возраст большую долю в дисперсии зарплат в выборке? Объясните.

E4.2. На веб-сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) вы найдете файл с данными **TeachingRating**, который содержит данные о том, как студенты оценивают качество курсов, о характеристиках этих курсов, а также о характеристиках профессоров для 463 курсов, читаемых в университете штата Техас в Остине (University of Texas at Austin<sup>1</sup>). Детальное описание массива данных приведено в файле **TeachingRatings\_Description** и также доступно на веб-сайте. Одной из характеристик является индекс «красоты» профессора, оцененный группой из шести судей. В этом упражнении вы будете исследовать, как оценка курса связана с индексом «красоты» профессора.

- a) Постройте диаграмму рассеяния средней оценки курса (*Course\_Eval*) и индекса «красоты» профессора (*Beauty*). Можно ли увидеть зависимость между рассматриваемыми показателями?
- b) Оцените регрессию средней оценки курса (*Course\_Eval*) от индекса «красоты» профессора (*Beauty*). Чему равна оценка константы? А чему равна оценка углового коэффициента? Объясните, почему оцененная константа равна выборочному среднему показателя *Course\_Eval*. (Подсказка: чему равно выборочное среднее показателя *Beauty*?)
- c) Значение индекса «красоты» профессора Уотсона равно среднему значению переменной *Beauty*, в то время как значение индекса «красоты» профессора Стока превышает значение *Beauty* на одно стандартное отклонение. Предскажите оценки курсов профессоров Стока и Уотсона.
- d) Прокомментируйте величину углового коэффициента регрессии. Оцененный эффект влияния переменной *Beauty* на переменную *Course\_Eval* большой или маленький? Объясните, что означает «большой» или «маленький».
- e) Объясняет ли переменная *Beauty* большую долю дисперсии в оценках курсов? Объясните.

E4.3. На веб-сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) вы найдете файл с данными **CollegeDistance**, который содержит дан-

---

<sup>1</sup> Эти данные были представлены профессором Дэниэлом Хамермешем (Daniel Hamermesh) из университета штата Техас в Остине и использованы в его совместной работе с Эмми Паркер (Amy Parker): «Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity», *Economics of Education Review*, August 2005, 24 (4): 369–376.

ные случайной выборки старшеклассников, опрошенных сначала в 1980 году, а потом еще раз в 1986 году. В этом упражнении необходимо использовать эти данные, чтобы исследовать отношение между числом полных лет обучения для молодых людей и расстоянием от средней школы каждого студента до ближайшего колледжа, в котором можно получить степень бакалавра. (Расстояние до колледжа снижает затраты на образование, так что студенты, которые живут ближе к колледжу, должны, в среднем, окончить больше курсов колледжа через большее количество лет.) Детальное описание дано в файле **CollegeDistance\_Description**, также доступном на веб-сайте<sup>1</sup>.

- a) Оцените регрессию числа полных лет обучения (*ED*) от расстояния до ближайшего колледжа (*Dist*), где *Dist* измеряется в десятках милях. (Например, *Dist* = 2 означает, что расстояние составляет 20 миль.) Чему равна оценка константы? А чему углового коэффициента? Используйте оцененную регрессию, чтобы ответить на вопрос: насколько изменяется среднее значение лет, необходимых для окончания школы, когда колледж построен ближе к школе, в которую ходят ученики?
  - б) Средняя школа Боба находилась в 20 милях от ближайшего колледжа. Предскажите количество лет, которые потратит Боб для окончания образования, используя оцененную регрессию. Как бы это предсказание изменилось, если бы Боб жил в 10 милях от ближайшего колледжа?
  - в) Расстояние до колледжа объясняет большую долю дисперсии в уровне образования студентов выборки? Объясните.
  - г) Чему равно значение стандартной ошибки регрессии? В каких единицах выражается стандартная ошибка (метры, граммы, годы, доллары, центы или что-нибудь еще)?
- E4.4. На веб-сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) вы найдете файл с данными **Growth**, который содержит данные о средних темпах роста для 65 стран с 1960 по 1995 год наряду с переменными, которые потенциально связаны с ростом. Детальное описание дано в файле **Growth\_Description**, также доступном на веб-сайте. В этом упражнении необходимо исследовать отношение между ростом и торговлей<sup>2</sup>.
- а) Постройте диаграмму рассеяния среднегодового темпа роста (*Growth*) и средней доли торговли (*TradeShare*). Существует ли, по вашему мнению, взаимосвязь между переменными?

<sup>1</sup> Данные предоставлены профессором Сесилией Руз (Cecilia Rouse) из Принстонского университета и использованы в ее работе «Democratization or Diversion? The Effect of Community Colleges on Educational Attainment» // Journal of Business and Economic Statistics, April 1995. 12 (2): 217–224.

<sup>2</sup> Данные предоставлены профессором Россом Левайном (Ross Levine) из Браунского университета (Brown University) и использованы в его работе с Торстеном Беком (Thorsten Beck) и Норманом Лойзой (Norman Loayza): «Finance and the Sources of Growth» // Journal of Financial Economics. 2000. 58: 261–300.

- б) Доля торговли одной из стран, Мальты, сильно превышает доли торговли других стран. Найдите Мальту на диаграмме рассеяния. Полагаете ли вы, что Мальта является выбросом?
- в) Используя все наблюдения, оцените регрессию переменной *Growth* на переменную *TradeShare*. Чему равен оцененный угловой коэффициент? Какова оценка константы? Используйте регрессию, чтобы предсказать темпы роста для страны с долей торговли 0,5 и для страны с долей торговли, равной 1,0.
- г) Оцените ту же самую регрессию, исключая данные по Мальте. Ответьте на тот же самый вопрос, что и в пункте в.
- д) Где находится Мальта? Почему доля торговли на Мальте настолько высока? Должна ли Мальта включаться или исключаться из анализа?

## Приложения

### Приложение 4.1. База данных по результатам тестов в Калифорнии

Данные о результатах стандартизованных тестов в Калифорнии и связанных характеристиках содержат информацию о результатах тестов, характеристиках калифорнийских школ и об особенностях школьников (таких как характеристики их семей и т.д.). Данные, используемые здесь, доступны с 1999 года для всех 420 К-6 и К-8 округов Калифорнии. Оценки за тест представляют собой средние значения оценок по чтению и математике по Стэнфордскому тесту (Stanford 9 Achievement Test), который является стандартизованным тестом для пятиклассников. Школьные характеристики (в среднем по всему школьному округу) включают регистрационные данные, количество учителей (измеряемое в эквиваленте полного рабочего дня), количество компьютеров в классе, а также расходы на одного ученика. Отношение «ученик – учитель», используемое здесь, – это количество учащихся в округе, деленное на количество учителей (в эквиваленте полного рабочего дня). Характеристики индивидуальных особенностей учеников также усредняются по округу. Эти переменные включают процент учеников, которые задействованы в программе CalWorks (ранее AFDC), процент учеников, которые имеют право на покупку обедов по сниженной цене, и процент учеников, которые изучают английский (т.е. тех учеников, для которых английский язык является вторым). Все эти данные были получены от департамента образования штата Калифорния ([www.cde.ca.gov](http://www.cde.ca.gov)).

### Приложение 4.2. Вывод МНК-оценок

В этом приложении мы используем дифференциальное исчисление, чтобы вывести формулы МНК-оценок, приведенных во вставке «Основные понятия 4.2». Чтобы минимизировать сумму квадратов ошибок предсказания  $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$  [уравнение (4.6)], возьмем сначала частные производные по  $b_0$  и  $b_1$ :

$$\frac{\partial}{\partial b_0} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \quad (4.23)$$

$$\text{и } \frac{\partial}{\partial b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) X_i. \quad (4.24)$$

МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  – это значения коэффициентов  $b_0$  и  $b_1$ , которые минимизируют сумму  $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$ , или, эквивалентно, значения коэффициентов  $b_0$  и  $b_1$ , для которых производные в уравнениях (4.23) и (4.24) равны нулю. Соответственно приравнивая эти производные к нулю, приводя подобные и деля на  $n$ , получаем, что МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  должны удовлетворять двум равенствам:

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \quad (4.25)$$

$$\text{и } \frac{1}{n} \sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n X_i^2 = 0. \quad (4.26)$$

Решая эти уравнения относительно  $\hat{\beta}_0$  и  $\hat{\beta}_1$ , получаем:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \bar{Y}}{\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (4.27)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.28)$$

Выражения (4.27) и (4.28) совпадают с формулами МНК-оценок  $\hat{\beta}_0$  и  $\hat{\beta}_1$ , приведенными во вставке «Основные понятия 4.2»; формула  $\hat{\beta}_1 = s_{XY} / s_X^2$  может быть получена делением числителя и знаменателя в выражении (4.27) на  $n-1$ .

### Приложение 4.3. Выборочное распределение МНК-оценки

В этом приложении мы покажем, что МНК-оценка  $\hat{\beta}_1$  является несмещенной и асимптотически нормальной, как указано во вставке «Основные понятия 4.4».

#### Представление $\hat{\beta}_1$ в терминах регрессоров и ошибок

Представим выражение для  $\hat{\beta}_1$  в терминах регрессоров и ошибок. Поскольку  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , то  $Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + u_i - \bar{u}$ , поэтому числитель в формуле для  $\hat{\beta}_1$  в выражении (4.27) равен:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n (X_i - \bar{X})[\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})] = \\ &= \beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}). \end{aligned} \quad (4.29)$$

Имеем, что  $\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^n (X_i - \bar{X})u_i - \sum_{i=1}^n (X_i - \bar{X})\bar{u} = \sum_{i=1}^n (X_i - \bar{X})u_i$ , где последнее равенство следует из определения  $\bar{X}$ , которое предполагает,

что  $\sum_{i=1}^n (X_i - \bar{X})\bar{u} = \left[ \sum_{i=1}^n X_i - n\bar{X} \right] \bar{u} = 0$ . Подставляя  $\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^n (X_i - \bar{X})u_i$  в последнюю часть выражения (4.29), получаем, что  $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X})u_i$ . Подставляя это выражение, в свою очередь, в формулу (4.27) для  $\hat{\beta}_1$ , получаем:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})u_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (4.30)$$

### **Доказательство несмещенности $\hat{\beta}_1$**

Математическое ожидание  $\hat{\beta}_1$  можно вычислить, взяв математическое ожидание от обеих частей выражения (4.30). Таким образом:

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 + E\left[\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})u_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}\right] = \\ &= \beta_1 + E\left[\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})E(u_i | X_1, \dots, X_n)}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}\right] = \beta_1, \end{aligned} \quad (4.31)$$

где второе равенство в выражении (4.31) следует из закона повторного математического ожидания (раздел 2.3). По второму предположению метода наименьших квадратов, ошибка  $u_i$  распределена независимо от  $X$  для всех наблюдений, кроме  $i$ , поэтому  $E(u_i | X_1, \dots, X_n) = E(u_i | X_i)$ . По первому предположению метода наименьших квадратов, однако,  $E(u_i | X_i) = 0$ . Следовательно, условное математическое ожидание в больших скобках во второй строке выражения (4.31) равно нулю, так что  $E(\hat{\beta}_1 - \beta_1 | X_1, \dots, X_n) = 0$ . Или, эквивалентно,  $E(\hat{\beta}_1 | X_1, \dots, X_n) = \beta_1$ ; то есть  $\hat{\beta}_1$  – условно несмешенная оценка при заданных  $X_1, \dots, X_n$ . По закону повторного математического ожидания,  $E(\hat{\beta}_1 - \beta_1) = E[E(\hat{\beta}_1 - \beta_1 | X_1, \dots, X_n)] = 0$ , поэтому  $E(\hat{\beta}_1) = \beta_1$ ; то есть  $\hat{\beta}_1$  – несмешенная оценка.

### **Асимптотическая нормальность МНК-оценки**

Нормальная аппроксимация на больших выборках предельного распределения  $\hat{\beta}_1$  (вставка «Основные понятия 4.4») может быть получена путем рассмотрения поведения последней компоненты выражения (4.30).

Рассмотрим сначала числитель этой компоненты. Поскольку  $\bar{X}$  состоятельная оценка, то в больших выборках  $\bar{X}$  примерно равно  $\mu_x$ . Таким образом,

в больших выборках числитель дроби в выражении (4.30) является выборочным средним  $\bar{v}$ , где  $v_i = (X_i - \mu_X)u_i$ . По первому предположению метода наименьших квадратов,  $v_i$  имеет среднее значение, равное нулю. По второму предположению метода наименьших квадратов,  $v_i$  является i.i.d. Дисперсия  $v_i$  имеет вид  $\sigma_v^2 = \text{var}[(X_i - \mu_X)u_i]$ , и она, по третьему предположению метода наименьших квадратов, ненулевая и конечная. Следовательно,  $\bar{v}$  удовлетворяет всем требованиям центральной предельной теоремы (вставка «Основные понятия 2.7»). Тогда  $\bar{v}/\sigma_{\bar{v}}$  в больших выборках распределено как  $N(0,1)$ , где  $\sigma_{\bar{v}}^2 = \sigma_v^2/n$ . Таким образом, распределение  $\bar{v}$  хорошо аппроксимируется распределением  $N(0, \sigma_v^2/n)$ .

Далее рассмотрим знаменатель выражения (4.30); это выборочная дисперсия  $X$  (за исключением деления на  $n$ , а не на  $n-1$ , что несущественно, если  $n$  велико). Как обсуждалось в разделе 3.2 [уравнение (3.8)], выборочная дисперсия является состоятельной оценкой дисперсии генеральной совокупности, поэтому в больших выборках она сколько угодно близка к дисперсии  $X$  в генеральной совокупности.

Объединяя эти два результата, имеем, что в больших выборках  $\hat{\beta}_1 - \beta_1 \cong \bar{v}/\text{var}(X_i)$ , поэтому выборочное распределение  $\hat{\beta}_1$  в больших выборках имеет вид  $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$ , где  $\sigma_{\hat{\beta}_1}^2 = \text{var}(\bar{v})/\text{var}(X_i)^2 = \text{var}[(X_i - \mu_X)u_i]/\{n[\text{var}(X_i)]^2\}$ , что совпадает с выражением (4.21).

### **Некоторые дополнительные алгебраические факты об МНК**

МНК-остатки и предсказанные значения удовлетворяют выражениям:

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i = 0; \quad (4.32)$$

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \bar{Y}; \quad (4.33)$$

$$\sum_{i=1}^n \hat{u}_i X_i = 0 \text{ и } s_{\hat{u}X} = 0 \text{ и } (4.34)$$

$$TSS = SSR + ESS. \quad (4.35)$$

Выражения (4.32) – (4.35) говорят, что выборочное среднее МНК-остатков равно нулю; выборочное среднее предсказанных МНК-значений равно  $\bar{Y}$ ; выборочная ковариация  $s_{\hat{u}X}$  между МНК-остатками и регрессорами равна нулю; и полная сумма квадратов остатков равна сумме квадратов остатков и объясненной сумме квадратов [ESS, TSS и SSR определены в выражениях (4.14), (4.15) и (4.17)].

Чтобы проверить корректность равенства (4.32), заметим, что определение  $\hat{\beta}_0$  позволяет нам записать МНК-остатки как  $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$ , таким образом,

$$\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X}).$$

Но определение  $\bar{Y}$  и  $\bar{X}$  предполагает, что  $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$  и  $\sum_{i=1}^n (X_i - \bar{X}) = 0$ , поэтому  $\sum_{i=1}^n \hat{u}_i = 0$ .

Для проверки равенства (4.33) заметим, что  $Y_i = \hat{Y}_i + \hat{u}_i$ , поэтому  $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n \hat{Y}_i$ , где второе равенство является следствием выражения (4.32).

Для доказательства корректности выражения (4.34) заметим, что равенство  $\sum_{i=1}^n \hat{u}_i = 0$  предполагает  $\sum_{i=1}^n \hat{u}_i X_i = \sum_{i=1}^n \hat{u}_i (X_i - \bar{X})$ , поэтому

$$\begin{aligned} \sum_{i=1}^n \hat{u}_i X_i &= \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] (X_i - \bar{X}) = \\ &= \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})^2 = 0, \end{aligned} \quad (4.36)$$

где последнее равенство в выражении (4.36) получено с использованием формулы для  $\hat{\beta}_1$  из выражения (4.27). Этот результат, объединенный с предыдущим результатом, предполагает, что  $s_{\hat{u}X} = 0$ .

Уравнение (4.35) следует из предыдущего результата и ряда алгебраических преобразований:

$$\begin{aligned} TSS &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \\ &= SSR + ESS + 2 \sum_{i=1}^n \hat{u}_i \hat{Y}_i = SSR + ESS, \end{aligned} \quad (4.37)$$

где последнее равенство следует из

$$\sum_{i=1}^n \hat{u}_i \hat{Y}_i = \sum_{i=1}^n \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n \hat{u}_i X_i = 0$$

по предыдущему результату.

# Глава 5. Парная регрессия: проверка гипотез и доверительные интервалы

В этой главе мы продолжим изучение парной (простой) регрессии. В главе 4 было рассмотрено, как МНК-оценка  $\hat{\beta}_1$  коэффициента наклона  $\beta_1$  отличается от одной выборки к другой, то есть  $\hat{\beta}_1$  имеет выборочное распределение. В этой главе мы покажем, как знание этого распределения может быть использовано для того, чтобы сделать выводы о  $\beta_1$ , которые учитывают неопределенность, возникающую из-за случайности выборки. Точкой отсчета можно считать стандартную ошибку оценки МНК, которая измеряет разброс выборочного распределения  $\hat{\beta}_1$ . В разделе 5.1 рассматривается формула для этой стандартной ошибки (и стандартной ошибки МНК-оценки свободного члена) и показано, как использовать  $\hat{\beta}_1$  и ее стандартную ошибку для тестирования гипотез. В разделе 5.2 объясняется, как построить доверительный интервал для  $\beta_1$ . В разделе 5.3 обсуждается специальный случай: бинарная объясняющая переменная.

В разделах 5.1–5.3 предполагается выполнение трех предпосылок метода наименьших квадратов, рассмотренного в главе 4. Если, помимо этого, выполнены некоторые более существенные условия, то могут быть получены некоторые более значимые результаты относительно распределения оценки МНК. Одно из этих условий – гомоскедастичность ошибок – понятие, рассмотренное в разделе 5.4. В разделе 5.5 приводится теорема Гаусса–Маркова, которая утверждает, что при некоторых условиях МНК-оценка является эффективной (имеет наименьшую дисперсию) среди определенного класса оценок. В разделе 5.6 обсуждается распределение оценки МНК при условии, что генеральная совокупность случайных ошибок регрессии распределена нормально.

## 5.1. Проверка гипотез о коэффициентах регрессии

Ваш клиент, чиновник из министерства образования, звонит вам со следующей проблемой. У нее в офисе находится рассерженный налогоплательщик, который утверждает, что сокращение числа учеников в классе не будет способствовать повышению результатов обучения (тестов), поэтому дальнейшее уменьшение классов является пустой тратой денег.

Как утверждает налогоплательщик, количество учеников в классе никак не влияет на результаты тестов. Требование налогоплательщика можно перефразировать, используя язык регрессионного анализа. Так как исследуется влияние количества учеников в классе  $\beta_{ClassSize}$  на результаты тестов, налогоплательщик заявляет, что линия теоретической регрессии плоская, то есть

коэффициент наклона  $\beta_{ClassSize}$  линии теоретической регрессии равен нулю. Чиновник спрашивает, отличается ли этот наклон от нуля в нашей выборке из 420 наблюдений по школьным округам Калифорнии? Можем ли мы отвергнуть гипотезу налогоплательщика о том, что  $\beta_{ClassSize} = 0$  или должны ли мы принять ее, по крайней мере предварительно, в ожидании дальнейших новых подтверждений?

Этот раздел обсуждает проверку гипотез о наклоне  $\beta_1$  или свободного члена  $\beta_0$  линии теоретической регрессии. Мы начинаем с детального рассмотрения двухсторонних тестов о наклоне  $\beta_1$ , потом обсуждаем односторонние тесты и тестирование гипотез относительно свободного члена  $\beta_0$ .

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**5.1**

### Общая форма $t$ -статистики

Обычно  $t$ -статистика имеет вид:

$$t = \frac{\text{оценка} - \text{предполагаемое при нулевой гипотезе значение}}{\text{стандартная ошибка оценки}}. \quad (5.1)$$

## Двухсторонние гипотезы относительно $\beta_1$

Основной подход к тестированию гипотез о коэффициенте  $\beta_1$  совпадает с тестированием гипотез о генеральном среднем, поэтому мы начнем с краткого обзора процедуры тестирования гипотез о генеральном среднем.

**Тестирование гипотез о генеральном среднем.** Вспомните из раздела 3.2, что нулевая гипотеза о том, что среднее значение случайной величины  $Y$  равно определенному значению  $\mu_{Y,0}$ , может быть записана как  $H_0 : E(Y) = \mu_{Y,0}$ , а двухсторонняя альтернативная гипотеза имеет вид:  $H_1 : E(Y) \neq \mu_{Y,0}$ .

Тестирование нулевой гипотезы  $H_0$  против двухсторонней альтернативной гипотезы проводится в три шага, которые описаны во вставке «Основные понятия 3.6». Первый шаг состоит в том, чтобы рассчитать стандартную ошибку  $\bar{Y}$ , то есть  $SE(\bar{Y})$ , которая является оценкой стандартного отклонения выборочного распределения случайной величины  $\bar{Y}$ . Второй шаг — расчет  $t$ -статистики, которая имеет общую форму, представленную во вставке «Основные понятия 5.1»; в нашем же случае  $t$ -статистика имеет вид:  $t = (\bar{Y} - \mu_{Y,0}) / SE(\bar{Y})$ .

Третий шаг состоит в расчете  $p$ -значения, то есть наименьшего уровня значимости, при котором нулевая гипотеза могла бы быть отвергнута на основании наблюдаемой тестовой статистики; или, эквивалентно,  $p$ -значение — это вероятность получения статистики посредством случайной выборочной вариации, по крайней мере, настолько отличной от значения в нулевой гипотезе, насколько отлична наблюдаемая статистика в предположении о правильности нулевой гипотезы (вставка «Основные понятия 3.5»). Так как при нулевой гипотезе в больших выборках  $t$ -статистика имеет стандартное нормальное распределение,  $p$ -значение для проверки двухсторонней гипотезы имеет вид:  $2\Phi(-|t^{act}|)$ ,

где  $t^{act}$  – значение рассчитанной  $t$ -статистики и  $\Phi$  – функция стандартного нормального распределения из таблицы 1 приложения. Иначе говоря, первый шаг может быть заменен простым сравнением  $t$ -статистики и критического значения, подходящего для тестирования с желаемым уровнем значимости. Например, двухсторонний тест с 5 %-м уровнем значимости отклонил бы нулевую гипотезу, если  $|t^{act}| > 1,96$ . В этом случае считается, что генеральное среднее статистически значимо отличается от проверяемого значения на 5 %-м уровне значимости.

**Тестирование гипотез о коэффициенте наклона  $\beta_1$ .** На теоретическом уровне важнейшее свойство, лежащее в основе рассмотренной выше процедуры тестирования гипотезы о генеральном среднем, состоит в том, что в больших выборках выборочное распределение  $\bar{Y}$  приблизительно нормально. Так как  $\beta_1$  в больших выборках также имеет нормальное выборочное распределение, гипотезы об истинном значении коэффициента наклона  $\beta_1$  могут быть проверены с использованием того же общего подхода.

Нулевая и альтернативная гипотезы должны быть точно сформулированы до того, как они могут быть протестираны. Гипотеза налогоплательщика состоит в том, что  $\beta_{ClassSize} = 0$ . Говоря более общими словами, при нулевой гипотезе истинный коэффициент наклона  $\beta_1$  (т.е. коэффициент наклона истинной регрессии) принимает некоторое конкретное значение,  $\beta_{1,0}$ . При альтернативной гипотезе  $\beta_1$  не равно  $\beta_{1,0}$ . Таким образом, нулевая и двухсторонняя альтернативная гипотезы имеют вид:

$$H_0 : \beta_1 = \beta_{1,0} \text{ против } H_1 : \beta_1 \neq \beta_{1,0}. \quad (5.2)$$

Чтобы проверить нулевую гипотезу  $H_0$ , мы следуем тем же трем шагам, что и для генерального среднего.

На первом шаге нужно вычислить *стандартную ошибку*  $\hat{\beta}_1$ ,  $SE(\hat{\beta}_1)$ . Стандартная ошибка  $\hat{\beta}_1$  является оценкой для  $\sigma_{\hat{\beta}_1}$ , стандартного отклонения выборочного распределения  $\hat{\beta}_1$ . Точнее:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}, \quad (5.3)$$

где

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2}. \quad (5.4)$$

Оценка дисперсии в уравнении (5.4) обсуждается в приложении 5.1. Хотя формула для  $\hat{\sigma}_{\hat{\beta}_1}^2$  сложна, на практике стандартная ошибка рассчитывается при помощи стандартных эконометрических пакетов, поэтому осуществляется достаточно легко.

Второй шаг – расчет  $t$ -статистики:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}. \quad (5.5)$$

## ОСНОВНЫЕ ПОНЯТИЯ

### 5.2

#### Тестирование гипотезы $\beta_1 = \beta_{1,0}$ против альтернативной $\beta_1 \neq \beta_{1,0}$

1. Вычислите стандартную ошибку  $\hat{\beta}_1, SE(\hat{\beta}_1)$  [уравнение (5.3)].
  2. Вычислите  $t$ -статистику [уравнение (5.5)].
  3. Вычислите  $p$ -значение [уравнение (5.7)]. Нулевая гипотеза отвергается на 5%-м уровне значимости, если  $p$ -значение меньше, чем 0,05, или, что эквивалентно, если  $|t^{act}| > 1,96$ .
- Стандартная ошибка и (обычно)  $t$ -статистика и  $p$ -значение, проверяющие гипотезу  $\beta_1 = 0$ , рассчитываются автоматически при помощи стандартных эконометрических пакетов.

Третий шаг заключается в расчете  $p$ -значения, вероятности наблюдения значения  $\hat{\beta}_1$ , по крайней мере, настолько же отличного от  $\beta_{1,0}$ , как рассчитанная оценка  $(\hat{\beta}_1^{act})$ , в предположении, что нулевая гипотеза верна. Математически это выражается следующим образом:

$$\begin{aligned} p\text{-value} &= \Pr_{H_0} \left[ \left| \hat{\beta}_1 - \beta_{1,0} \right| > \left| \hat{\beta}_1^{act} - \beta_{1,0} \right| \right] = \\ &= \Pr_{H_0} \left[ \left| \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right| > \left| \frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)} \right| \right] = \Pr_{H_0} (|t| > |t^{act}|), \end{aligned} \quad (5.6)$$

где  $\Pr_{H_0}$  – вероятность, рассчитанная при нулевой гипотезе, второе равенство получается после деления на  $SE(\hat{\beta}_1)$ , а  $t^{act}$  – значение рассчитанной  $t$ -статистики. Так как  $\hat{\beta}_1$  распределено в больших выборках приблизительно нормально, при нулевой гипотезе  $t$ -статистика приблизительно распределена как стандартная нормальная величина, так что в больших выборках:

$$p\text{-value} = \Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|), \quad (5.7)$$

где  $p$ -значение, меньшее 5 %, свидетельствует против нулевой гипотезы в том смысле, что при нулевой гипотезе вероятность получения значения  $\hat{\beta}_1$ , по крайней мере, настолько же далекого от нуля, как наблюдаемое, меньше 5 %. Если так, нулевая гипотеза отклоняется на 5%-м уровне значимости.

С другой стороны, гипотеза может быть проверена на 5%-м уровне значимости просто сравнением значения  $t$ -статистики с  $\pm 1,96$ , то есть с критическим значением для двухстороннего теста, и тогда нулевая гипотеза отклоняется на 5%-м уровне значимости, если  $|t^{act}| > 1,96$ .

Эти этапы отмечены во вставке «Основные понятия 5.2».

**Представление регрессионных уравнений и пример.** МНК-регрессия результатов тестов школьников от соотношения числа школьников и учителей, описанная в уравнении (4.11), дала оценки  $\hat{\beta}_0 = 698,9$  и  $\hat{\beta}_1 = -2,28$ . Стандартные ошибки этих оценок равны:  $SE(\hat{\beta}_0) = 10,4$  и  $SE(\hat{\beta}_1) = 0,52$ .

Вследствие важности стандартных ошибок принято включать их в отчет об оцененных коэффициентах МНК. Компактный способ включать стандартные ошибки в отчет – разместить их в скобках внизу соответствующих коэффициентов МНК-регрессии:

$$\widehat{TestScore} = 698,9 - 2,28 \times STR, R^2 = 0,051, SER = 18,6. \quad (5.8)$$

(10,4)      (0,52)

Уравнение (5.8) также содержит информацию о  $R^2$  и стандартной ошибке оцененной регрессии ( $SER$ ). Поэтому уравнение (5.8) содержит оценки параметров линии регрессии, оценки неопределенности коэффициента наклона и свободного члена (стандартные ошибки) и две меры качества подгонки регрессии  $R^2$  и  $SER$ . Это общепринятый формат для представления регрессионного уравнения, и он будет использоваться в оставшейся части этой книги.

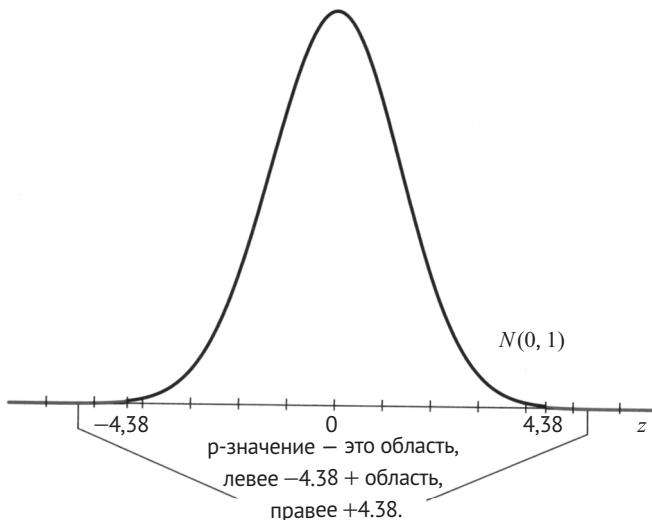
Предположим, вы хотите проверить нулевую гипотезу о том, что в уравнении (5.8) коэффициент наклона  $\beta_1$  равен нулю на 5 %-м уровне значимости. Для этого рассчитайте  $t$ -статистику и сравните ее с числом 1,96, то есть двухсторонним 5 %-м критическим значением, взятым из стандартного нормального распределения.  $t$ -статистика строится при помощи подстановки гипотетического значения  $\beta_1$  при нулевой гипотезе (в нашем случае оно равно нулю), оцененного коэффициента наклона и стандартной ошибки из уравнения (5.8) в общую формулу (5.5); результат имеет вид  $t^{act} = (2,28 - 0) / 0,52 = -4,38$ . Эта  $t$ -статистика превышает (по абсолютному значению) двухстороннее 5 %-е критическое значение 1,96, так что нулевая гипотеза отвергается в пользу двухсторонней альтернативы на 5 %-м уровне значимости.

С другой стороны, мы можем рассчитать  $p$ -значение, соответствующее  $t^{act} = -4,38$ . Эта вероятность – площадь под хвостами графика функции плотности стандартного нормального распределения, как показано на рисунке 5.1. Эта вероятность очень мала, примерно 0,00001, или 0,001 %. То есть если нулевая гипотеза  $\beta_{ClassSize} = 0$  выполняется, вероятность получения значения  $\hat{\beta}_1$  настолько же далекого от нуля, как значение, которое мы на самом деле наблюдаем, очень мала, меньше чем 0,001 %. Так как этот случай маловероятен, имеет смысл заключить, что нулевая гипотеза ложна.

### **Односторонние гипотезы о $\beta_1$**

До сих пор дискуссия была сконцентрирована на тестировании гипотезы о том, что  $\beta_1 = \beta_{1,0}$ , против гипотезы о том, что  $\beta_1 \neq \beta_{1,0}$ . Это проверка двухсторонней гипотезы, так как при альтернативной гипотезе  $\beta_1$  может быть либо меньше, либо больше, чем  $\beta_{1,0}$ . Однако иногда допустимо использовать проверку односторонней гипотезы. Например, в рассматриваемом нами примере о влиянии соотношения школьников и учителей на результаты тестов школьников многие люди думают, что меньшие по размеру классы обеспечивают лучшие условия для обучения. При такой гипотезе коэффициент  $\beta_1$  отрицателен: меньшее число учеников в классе приводит к более высоким оценкам. Следо-

вательно, имеет смысл проверить нулевую гипотезу о том, что  $\beta_1 = 0$  (никакого влияния нет), против односторонней альтернативной гипотезы о том, что  $\beta_1 < 0$ .



**Рисунок 5.1. Расчет  $p$ -значения для двухстороннего теста, когда  $t^{act} = -4,38$**

$p$ -значение для двухстороннего теста – это вероятность того, что  $|Z| > |t^{act}|$ , где  $Z$  – стандартная нормальная величина и  $t^{act}$  – значение  $t$ -статистики, рассчитанное по выборке. Когда  $t^{act} = -4,38$ ,  $p$ -значение составляет только 0,000 01.

Для одностороннего теста нулевая и односторонняя альтернативная гипотезы имеют вид:

$$H_0 : \beta_1 = \beta_{1,0} \text{ против } H_1 : \beta_1 < \beta_{1,0} \text{ (односторонняя альтернатива)}, \quad (5.9)$$

где  $\beta_{1,0}$  – значение  $\beta_1$  при нулевой гипотезе (0 в нашем примере) и альтернативная гипотеза состоит в том, что  $\beta_1$  меньше, чем  $\beta_{1,0}$ . Если альтернативная гипотеза состоит в том, что  $\beta_1$  больше, чем  $\beta_{1,0}$ , неравенство в уравнении (5.9) меняется на обратное.

Так как нулевая гипотеза одинакова для одно- и двухстороннего теста,  $t$ -статистика рассчитывается аналогично. Единственное отличие между односторонним и двухсторонним тестом состоит в том, как вы интерпретируете  $t$ -статистику. Для односторонней альтернативы в уравнении (5.9) нулевая гипотеза отвергается против односторонней альтернативы для больших отрицательных, но не для больших положительных значений  $t$ -статистики: вместо того чтобы отвергнуть нулевую гипотезу при  $|t^{act}| > 1,96$ , она отвергается на 5%-м уровне значимости, если  $t^{act} < -1,645$ .

$p$ -значение для одностороннего теста получается из стандартного нормального распределения:

$$p\text{-value} = \Pr(Z < t^{act}) = \Phi(t^{act}) \quad (p\text{-значение для левостороннего теста}). \quad (5.10)$$

Если альтернативная гипотеза состоит в том, что  $\beta_1$  больше, чем  $\beta_{1,0}$ , неравенства в уравнениях (5.9) и (5.10) меняются на противоположные, так что

*p*-значение представляет собой вероятность для правостороннего теста, то есть  $\Pr(Z > t^{act})$ .

**Когда нужно использовать односторонний тест?** На практике односторонняя альтернативная гипотеза должна быть использована, только если для этого есть веская причина. Эта причина может следовать из экономической теории, предыдущих эмпирических исследований, или из того и другого.

Однако даже если мы внутренне чувствуем, что необходимая альтернативная гипотеза — односторонняя, на деле может оказаться, что это не обязательно так. Глубокие клинические исследования нового лекарства могут показать, что оно опасно из-за ранее не выявленных побочных эффектов. В нашем примере с числом учеников в классе приходит на ум шутка, часто рассказываемая на выпускных вечерах: секрет успеха университета — принять талантливых студентов и следить за тем, чтобы преподаватели держались от них подальше, причиняя им как можно меньше вреда. На практике такая двусмысленность часто заставляет эконометристов использовать двусторонние тесты.

**Приложение для примера с результатами тестов.** *t*-статистика, проверяющая гипотезу о том, что число учеников в классе не влияет на результаты тестов (так что  $\beta_{1,0} = 0$  в уравнении (5.9)), равна  $t^{act} = -4,38$ . Это значение меньше, чем  $-2,33$  (критическое значение для одностороннего теста на уровне значимости 1 %), так что нулевая гипотеза отвергается против односторонней альтернативы на 1 %-м уровне значимости. Фактически *p*-значение меньше, чем 0,000 6 %. Основываясь на этих данных, вы можете отвергнуть утверждение рассерженного налогоплательщика о том, что отрицательная оценка коэффициента наклона возникла только из-за случайной специфики выборки на 1 %-м уровне значимости.

### Тестирование гипотез о свободном члене $\beta_0$

Наша дискуссия была сосредоточена на тестировании гипотез о коэффициенте наклона  $\beta_1$ . Однако иногда возникает необходимость проверить гипотезы относительно свободного члена  $\beta_0$ . Нулевая гипотеза о свободном члене и ее двухсторонняя альтернатива имеют вид:

$$H_0 : \beta_0 = \beta_{0,0} \text{ против } H_1 : \beta_0 \neq \beta_{0,0} \text{ (двуихсторонняя альтернатива).} \quad (5.11)$$

Общая схема для проверки этой нулевой гипотезы состоит из тех же трех шагов, которые описаны во вставке «Основные понятия 5.2», но примененных к коэффициенту  $\beta_0$  (формула для стандартной ошибки  $\hat{\beta}_0$  представлена в приложении 5.1). Если альтернатива — односторонняя, этот подход меняется, как обсуждалось в предыдущем подразделе, для гипотез о коэффициенте наклона.

Проверки гипотез весьма полезны, если у вас в голове есть определенная нулевая гипотеза (как это было с нашим разъяренным налогоплательщиком). Возможность принять или отвергнуть эту нулевую гипотезу, основанную на статистических наблюдениях, предоставляет эффективный инструмент для того, чтобы справиться с неопределенностью, присущей выборке, которую мы используем, чтобы сделать вывод обо всей генеральной совокупности. Однако часто ни одна

гипотеза о коэффициентах регрессии не является доминирующей, и вместо этого нам хотелось бы знать диапазон значений коэффициентов, согласующихся с данными. Для этого требуется построить доверительный интервал.

## 5.2. Доверительные интервалы для коэффициентов регрессии

Так как любые статистические оценки коэффициента наклона  $\beta_1$  обязательно имеют неопределенность, связанную с имеющейся выборкой, мы не можем точно определить истинное значение  $\beta_1$ , основываясь на выборке. Однако возможно использовать МНК-оценку и ее стандартную ошибку для построения доверительного интервала для коэффициента наклона  $\beta_1$  или для свободного члена  $\beta_0$ .

**Доверительный интервал для  $\beta_1$ .** Напомним, что есть два эквивалентных определения 95 %-го доверительного интервала для коэффициента  $\beta_1$ . Согласно первому из них, это набор значений, равенство которым не может быть отвергнуто на 5 %-м уровне значимости в двухстороннем тесте. По второму определению – это интервал, в котором с 95 %-й вероятностью содержится истинное значение  $\beta_1$ , то есть для 95 % выборок, которые могут быть сделаны, доверительный интервал будет содержать истинное значение  $\beta_1$ . Так как этот интервал содержит истинное значение для 95 % всех выборок, говорят, что интервал имеет уровень доверия (или доверительную вероятность), равный 95 %.

Причина, по которой эти два определения эквивалентны, состоит в следующем. При проверке гипотезы на 5 %-м уровне значимости мы, по определению, будем отвергать нулевую гипотезу о  $\beta_1$ , если она верна, только для 5 % всех возможных выборок; то есть для 95 % возможных выборок нулевая гипотеза о  $\beta_1$  не будет отвергнута. Из того, что 95 %-й доверительный интервал (как было сказано в первом определении) – это набор всех значений  $\beta_1$ , которые не были отвергнуты на 5 %-м уровне значимости, следует, что истинное значение  $\beta_1$  будет содержаться в доверительном интервале в 95 % случаев.

Как и в случае доверительного интервала для генерального среднего (раздел 3.3), в принципе 95 %-й доверительный интервал может быть рассчитан путем тестирования всех возможных значений  $\beta_1$  (т.е. проверки нулевой гипотезы  $\beta_1 = \beta_{1,0}$  для всех значений  $\beta_{1,0}$ ) на 5 %-м уровне значимости с использованием  $t$ -статистики. Тогда 95 %-й доверительный интервал будет представлять собой набор всех значений  $\beta_1$ , для которых не отвергается нулевая гипотеза. Но такой расчет  $t$ -статистик для всех значений  $\beta_1$  может занять целую вечность.

Более легкий способ построить доверительный интервал – просто заметить, что тестируемое значение  $\beta_{1,0}$  будет отвергаться при помощи  $t$ -статистики всегда, когда  $\beta_{1,0}$  лежит вне диапазона  $\hat{\beta}_1 \pm 1,96SE(\hat{\beta}_1)$ . То есть 95 %-й доверительный интервал для коэффициента  $\beta_1$  – это интервал  $[\hat{\beta}_1 - 1,96SE(\hat{\beta}_1), \hat{\beta}_1 + 1,96SE(\hat{\beta}_1)]$ , что аналогично случаю построения доверительного интервала для генерального среднего.

Схема построения доверительного интервала для коэффициента  $\beta_1$  описана во вставке «Основные понятия 5.3».

**Доверительный интервал для  $\beta_1$** 

95 %-й двухсторонний доверительный интервал для коэффициента  $\beta_1$  – это интервал, который содержит истинное значение  $\beta_1$  с 95 %-й вероятностью, то есть истинное значение коэффициента  $\beta_1$  содержится в нем для 95 % всех случайных выборок. Эквивалентно, это набор значений коэффициентов  $\beta_1$ , равенство которым не может быть отвергнуто при помощи двухстороннего теста на уровне значимости 5 %. Когда размер выборки большой, доверительный интервал строится следующим образом:

$$\begin{aligned} \text{95 %-й доверительный интервал для } \beta_1 = \\ = [\hat{\beta}_1 - 1,96SE(\hat{\beta}_1); \hat{\beta}_1 + 1,96SE(\hat{\beta}_1)]. \end{aligned} \quad (5.12)$$

**ОСНОВНЫЕ ПОНЯТИЯ**

5.3

**Доверительный интервал для  $\beta_0$ .** 95 %-й доверительный интервал для коэффициента  $\beta_0$  строится так же, как указано во вставке «Основные понятия 5.3», с заменой  $\hat{\beta}_1$  и  $SE(\hat{\beta}_1)$  на  $\hat{\beta}_0$  и  $SE(\hat{\beta}_0)$ .

**Пример: влияние количества школьников в классе на результаты обучения.** МНК-регрессия зависимости результатов тестов от соотношения числа учеников и учителей, представленная в уравнении (5.8), дает оценку  $\hat{\beta}_1 = -2,28$  и  $SE(\hat{\beta}_1) = 0,52$ . Тогда двухсторонний 95 %-й доверительный интервал для коэффициента  $\beta_1$  равен  $\{-2,28 \pm 1,96 \times 0,52\}$  или  $-3,30 \leq \beta_1 \leq -1,26$ . Значение  $\beta_1 = 0$  не содержится в этом доверительном интервале, так что (как мы уже знаем из раздела 5.1) гипотеза о том, что  $\beta_1 = 0$ , может быть отвергнута на 5 %-м уровне значимости.

**Доверительные интервалы для оценки влияния изменений  $X$ .** 95 %-й доверительный интервал для  $\beta_1$  может быть использован для построения 95 %-го доверительного интервала для расчета того, как влияет изменение  $X$  на зависимую переменную.

Рассмотрим изменение  $X$  на заданную величину  $\Delta x$ . Тогда оцененное изменение  $Y$ , связанное с изменением  $X$ , равно  $\beta_1 \Delta x$ . Теоретический коэффициент наклона  $\beta_1$  неизвестен, но так как мы можем построить доверительный интервал для  $\beta_1$ , мы можем построить и доверительный интервал для оцененного эффекта  $\beta_1 \Delta x$ . Так как левая граница 95 %-го доверительного интервала для  $\beta_1$  равна  $\hat{\beta}_1 - 1,96SE(\hat{\beta}_1)$ , то оцененный при помощи этой оценки эффект от изменения  $\Delta x$  равен  $[\hat{\beta}_1 - 1,96SE(\hat{\beta}_1)] \Delta x$ . Правая граница доверительного интервала равна  $\hat{\beta}_1 + 1,96SE(\hat{\beta}_1)$ . Тогда оценка соответствующего эффекта будет  $[\hat{\beta}_1 + 1,96SE(\hat{\beta}_1)] \Delta x$ . Поэтому 95 %-й доверительный интервал для эффекта изменения  $X$  на величину  $\Delta x$  может быть записан так:

$$\begin{aligned} \text{95 %-й доверительный интервал для } \beta_1 \Delta x = \\ = [\hat{\beta}_1 \Delta x - 1,96SE(\hat{\beta}_1) \times \Delta x; \hat{\beta}_1 \Delta x + 1,96SE(\hat{\beta}_1) \times \Delta x]. \end{aligned} \quad (5.13)$$

Например, наш гипотетический руководитель рассматривает сокращение отношения количества учеников на одного учителя на 2. Так как 95 %-й доверительный интервал для  $\beta_1$  равен  $[-3,30, -1,26]$ , эффект уменьшения количества

учеников на одного учителя на 2 мог бы быть равен максимум  $-3,30 \times (-2) = 6,60$  и минимум  $-1,26 \times (-2) = 2,52$ . Поэтому ожидается, что уменьшение количества учеников на одного учителя на 2 увеличит результаты тестов на величину между 2,52 и 6,60 пунктами с 95 %-й вероятностью.

### 5.3. Регрессия с бинарной объясняющей переменной

До сих пор наше обсуждение было сосредоточено на случае, когда регрессор – непрерывная переменная. Регрессионный анализ может быть использован и в том случае, когда регрессор – бинарная величина, то есть он принимает только два значения, 0 или 1. Например,  $X$  может определять пол работника (т.е. быть равным 1, если это женщина, и нулю, если это мужчина), либо местонахождение школы может быть городом или селом ( $= 1$ , если город;  $= 0$ , если село), или размер класса может быть большим или маленьким ( $= 1$ , если он маленький;  $= 0$ , если большой). Бинарная переменная также называется *индикаторной переменной* или иногда *дамми-переменной*<sup>1</sup>.

#### Интерпретация коэффициентов регрессии

Механизм регрессии с бинарным регрессором такой же, как в случае, когда регрессор является непрерывной переменной. Однако интерпретация коэффициента  $\beta_1$  отличается, и выясняется, что регрессия с бинарной объясняющей переменной эквивалентна тестированию гипотезы о равенстве разности средних значений какой-либо величине, что было рассмотрено в разделе 3.4.

Чтобы убедиться в этом, представьте, что у вас есть переменная  $D_i$ , которая равна либо 0, либо 1 в зависимости от того, меньше ли 20 или нет соотношение числа учеников и учителей (отношение «ученик – учитель»):

$$D_i = \begin{cases} 1, & \text{если отношение «ученик – учитель»} < 20; \\ 0, & \text{если отношение «ученик – учитель»} \geq 20. \end{cases} \quad (5.14)$$

Теоретическая модель с  $D_i$  в качестве регрессора имеет вид:

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad i = 1, \dots, n. \quad (5.15)$$

Это то же самое, что и регрессионная модель с непрерывными объясняющими переменными  $X_i$ , кроме того, что сейчас регрессор – бинарная переменная  $D_i$ . Так как  $D_i$  не непрерывна, бесполезно думать о  $\beta_1$  как о коэффициенте наклона. Действительно, так как  $D_i$  может принимать только два значения, нет никакой «линии», так что бессмысленно говорить о наклоне. Поэтому мы не будем ссылаться на  $\beta_1$  как на коэффициент наклона в уравнении (5.15); вместо этого мы будем просто ссылаться на  $\beta_1$  как на коэффициент, являющийся множителем  $D_i$  в этой регрессии или, более кратко, *коэффициентом при  $D_i$* .

Если  $\beta_1$  в уравнении (5.15) не является коэффициентом наклона, то что это? Лучший способ интерпретировать  $\beta_0$  и  $\beta_1$  в регрессии с бинарной объясняю-

<sup>1</sup> В российских учебниках дамми-переменную нередко называют фиктивной переменной. – Примеч. науч. ред. перевода.

щей переменной – это рассмотреть одновременно два возможных случая,  $D_i = 0$  и  $D_i = 1$ . Если отношение «ученик – учитель» высокое, то  $D_i = 0$  и уравнение (5.15) принимает вид:

$$Y_i = \beta_0 + u_i \quad (D_i = 0). \quad (5.16)$$

Так как  $E(u_i | D_i = 0)$ , условное ожидание  $Y_i$  при  $D_i = 0$  имеет вид:  $E(Y_i | D_i = 0) = \beta_0$ , то есть,  $\beta_0$  – генеральное среднее значение результатов тестов в ситуации, когда отношение «ученик – учитель» высокое. Аналогично, когда  $D_i = 1$ :

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1). \quad (5.17)$$

Следовательно, когда  $D_i = 1$ ,  $E(Y_i | D_i = 1) = \beta_0 + \beta_1$ , то есть  $\beta_0 + \beta_1$  – генеральное среднее значение результатов тестов, когда отношение «ученик – учитель» небольшое.

Так как  $\beta_0 + \beta_1$  – генеральное среднее  $Y_i$ , когда  $D_i = 1$  и  $\beta_0$  – генеральное среднее  $Y_i$ , когда  $D_i = 0$ , разность  $(\beta_0 + \beta_1) - \beta_0 = \beta_1$  представляет собой разность между этими двумя средними. Другими словами,  $\beta_1$  – разность между условным математическим ожиданием  $Y_i$ , когда  $D_i = 1$  или когда  $D_i = 0$ , или  $\beta_1 = E(Y_i | D_i = 1) - E(Y_i | D_i = 0)$ . В примере с результатами тестов  $\beta_1$  – разность между средней оценкой по тестам в районах с маленьким числом учеников на одного учителя и средней оценкой по тестам в районах с большим числом учеников на одного учителя.

Так как  $\beta_1$  – разность между генеральными средними, имеет смысл тот факт, что МНК-оценки коэффициента  $\beta_1$  представляют собой разность между выборочными средними для  $Y_i$  в двух группах, и по сути это так.

**Проверка гипотез и доверительные интервалы.** Если два генеральных средних значения одинаковы, то  $\beta_1$  в уравнении (5.15) равно нулю. Поэтому нулевая гипотеза о том, что два генеральных средних одинаковы, может быть проверена против альтернативной гипотезы о том, что они различны, при помощи тестирования нулевой гипотезы  $\beta_1 = 0$  против альтернативной  $\beta_1 \neq 0$ . Эта гипотеза может быть проверена с использованием процедуры, описанной в разделе 5.1. В частности, нулевая гипотеза может быть отвергнута на 5 %-м уровне значимости против двухсторонней альтернативной гипотезы, когда МНК  $t$ -статистика  $t = \hat{\beta}_1 / SE(\hat{\beta}_1)$  превышает 1,96 по модулю. Аналогично, 95 %-й доверительный интервал для коэффициента  $\beta_1$ , построенный как  $\hat{\beta}_1 \pm 1,96SE(\hat{\beta}_1)$ , как описано в разделе 5.2, дает 95 %-й доверительный интервал для разности между двумя генеральными средними.

**Пример: влияние количества школьников в классе на результаты обучения.** В качестве примера рассмотрим регрессию результатов тестов на бинарную переменную  $D$ , отражающую соотношение числа учеников и учителей, описанную в уравнении (5.14), оцененную МНК с использованием 420 наблюдений и изображенную на рисунке 4.2:

$$\widehat{TestScore} = 650,0 + 7,4D, R^2 = 0,037, SER = 18,7, \quad (5.18)$$

где стандартные ошибки МНК-оценок для коэффициентов  $\beta_0$  и  $\beta_1$  представлены в скобках под оценками коэффициентов. Таким образом, средняя оценка за тест для подвыборки с числом школьников на одного учителя, большим или равным 20 (т.е. для которых  $D=0$ ), равна 650,0, и средняя оценка за тест для подвыборки с числом школьников на одного учителя, меньшим 20 (т.е.  $D=1$ ), равна  $650,0 + 7,4 = 657,4$ . Разность между выборочными средними результатами тестов для двух групп равна 7,4. Это МНК-оценка коэффициента  $\beta_1$  при бинарной переменной  $D$ .

Является ли разность между генеральными средними значениями результатов тестов в двух группах статистически значимо отличающейся от нуля на 5 %-м уровне значимости? Чтобы определить это, построим  $t$ -статистику для  $\beta_1$ :  $t = 7,4/1,8 = 4,04$ . Это значение превышает по абсолютному значению 1,96, так что гипотеза о том, что генеральные средние значения результатов тестов в районах с высоким и низким отношением «ученик – учитель» совпадают, может быть отвергнута на 5 %-м уровне значимости.

МНК-оценка и ее стандартная ошибка могут быть использованы для построения 95 %-го доверительного интервала для истинной разности в средних. Получаем  $7,4 \pm 1,96 \times 1,8 = (3,9; 10,9)$ . Этот доверительный интервал не включает значение  $\beta_1 = 0$ , так что (как мы знаем из предыдущего абзаца) гипотеза  $\beta_1 = 0$  может быть отвергнута на 5 %-м уровне значимости.

## 5.4. Гетероскедастичность и гомоскедастичность

Наше единственное предположение о распределении  $u_i$  относительно  $X_i$  состоит в том, что оно имеет нулевое среднее (первое предположение метода наименьших квадратов). Если, кроме того, дисперсия этого условного распределения не зависит от  $X_i$ , то считается, что ошибки гомоскедастичны. В данном разделе обсуждается понятие гомоскедастичности, ее теоретические предпосылки, упрощенные формулы для стандартных ошибок МНК-оценок, для случая, когда ошибки гомоскедастичны, и риск, который мы берем на себя, если используем эти упрощенные формулы на практике.

### Что такое гетероскедастичность и гомоскедастичность?

**Определения гетероскедастичности и гомоскедастичности.** Случайная ошибка  $u_i$  называется *гомоскедастичной*, если дисперсия условного распределения  $u_i$  относительно  $X_i$  постоянна для  $i = 1, \dots, n$  и, в частности, не зависит от  $X_i$ . В противном случае ошибка называется *гетероскедастичной*.

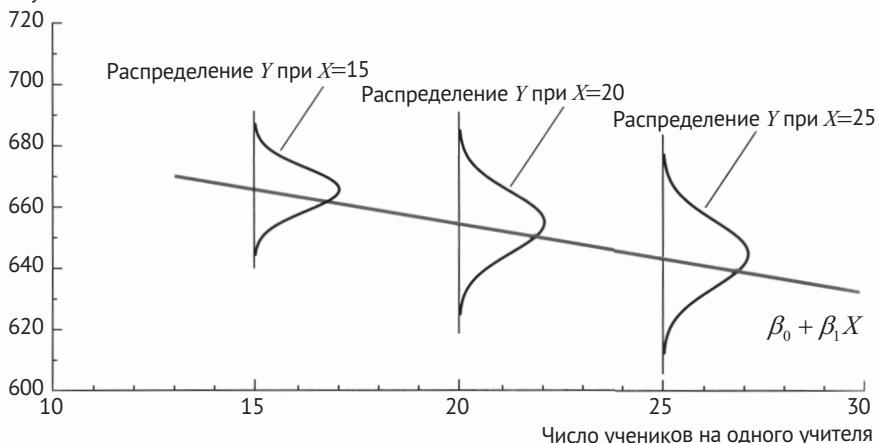
В качестве иллюстрации вернемся к рисунку 4.4. Распределение ошибок  $u_i$  показано для различных значений  $x$ . Так как это распределение рассмотрено для значений, равных 15, 20 и 25, то это условные распределения  $u_i$  относительно данных  $X_i = x$ . Как видно на рисунке, все условные распределения имеют одинаковый разброс; точнее, дисперсия распределений одна и та же для различных значений  $x$ . Таким образом, на рисунке 4.4 условная дисперсия  $u_i$

при  $X_i = x$  не зависит от  $x$ , то есть ошибки, изображенные на рисунке 4.4, – гомоскедастичны.

Напротив, рисунок 5.2 иллюстрирует случай, когда условное распределение  $u_i$  расширяется с ростом  $x$ . Для маленьких значений  $x$  это распределение имеет маленький разброс, но для больших значений  $x$  оно имеет больший разброс. Поэтому на рисунке 5.2 дисперсия  $u_i$  при  $X_i = x$  возрастает с ростом  $x$ , поэтому ошибки на рисунке 5.2 гетероскедастичны.

Определения гетероскедастичности и гомоскедастичности обобщаются во вставке «Основные понятия 5.4».

Результат теста



**Рисунок 5.2. Гетероскедастичность: пример**

Как и рисунок 4.4, этот рисунок показывает условное распределение результатов тестов для трех различных размеров класса. В отличие от рисунка 4.4, эти распределения становятся более широкими (имеют большую дисперсию) для классов больших размеров. Так как дисперсия распределения  $u_i$  при  $X_i = x$ , то есть  $\text{var}(u_i|X)$ , зависит от  $X$ , ошибка  $u_i$  гетероскедастична.

### Гетероскедастичность и гомоскедастичность

Ошибка  $u_i$  – гомоскедастична, если дисперсия условного распределения  $u_i$  при  $X_i = x$ ,  $\text{var}(u_i|X_i = x)$  – постоянна для  $i = 1, \dots, n$  и, в частности, не зависит от  $x$ . В противном случае ошибка гетероскедастична.

**ОСНОВНЫЕ ПОНЯТИЯ**  
**5.4**

**Пример.** Рассматриваемые понятия труднопроизносимы, а определения могут показаться абстрактными. Чтобы прояснить их на примере, мы отвлечемся от соотношения числа учеников на одного учителя в задаче моделирования показателя результатов тестов и вместо этого вернемся к примеру о различиях в зарплатах мужчин и женщин с высшим образованием, рассмотренном в главе 3 во вставке «Гендерный разрыв в заработных платах выпускников колледжей в Соединенных Штатах». Пусть  $MALE_i$  – бинарная переменная, которая

равна единице для выпускников мужского пола и нулю для выпускников женского пола. Модель регрессии зарплат выпускников колледжей с бинарной объясняющей переменной имеет вид:

$$Earnings_i = \beta_0 + \beta_1 MALE_i + u_i \quad (5.19)$$

для  $i=1, \dots, n$ . Так как регрессор является бинарной переменной,  $\beta_1$  представляет собой разность генеральных средних двух групп – в данном случае разность между средними зарплатами мужчин и женщин, имеющих высшее образование.

Из определения гомоскедастичности следует, что дисперсия  $u_i$  не зависит от регрессора. Здесь регрессором является переменная  $MALE_i$ , так что вопрос состоит в том, является ли дисперсия ошибки зависящей от нее. Другими словами, является ли дисперсия ошибки одинаковой для мужчин и женщин? Если да, ошибки гомоскедастичны, если нет, то гетероскедастичны.

Для того чтобы принять решение о том, является ли дисперсия  $u_i$  зависимой от зарплаты (в долл.)  $MALE_i$ , необходимо поразмышлять о том, что же на самом деле представляет собой ошибка регрессии. Для этого полезно записать уравнение (5.19) в виде двух отдельных уравнений: одного – для мужчин, другого – для женщин:

$$Earnings_i = \beta_0 + u_i \text{ (женщины) и} \quad (5.20)$$

$$Earnings_i = \beta_0 + \beta_1 + u_i \text{ (мужчины).} \quad (5.21)$$

Таким образом, мы видим, что для женщин  $u_i$  – отклонение зарплаты  $i$ -й женщины от генерального среднего для генеральной совокупности женщин ( $\beta_0$ ), а для мужчин  $u_i$  – отклонение зарплаты  $i$ -го мужчины от генерального среднего для генеральной совокупности мужчин ( $\beta_1 + \beta_0$ ). Отсюда следует, что утверждение «дисперсия  $u_i$  не зависит от  $MALE_i$ » эквивалентно утверждению «дисперсия зарплаты одинакова для мужчин и женщин». Другими словами, в этом примере ошибка гомоскедастична, если дисперсия распределения генеральной совокупности зарплат одинакова для мужчин и женщин; если же дисперсия различна, ошибка гетероскедастична.

### **Математические следствия гомоскедастичности**

**МНК-оценки остаются несмещенными и асимптотически нормальными.** Так как предположения метода наименьших квадратов, рассмотренные во вставке «Основные понятия 4.3», не устанавливают ограничений на условную дисперсию, они применимы как для общего случая гетероскедастичности, так и для частного случая гомоскедастичности. Следовательно, оценка МНК остается несмещенной и состоятельной, даже если ошибки гомоскедастичны. В дополнение, оценки МНК имеют выборочное распределение, которое нормально в больших выборках, даже если ошибки гомоскедастичны. Независимо от того, гомоскедастичны или гетероскедастичны ошибки, оценки МНК не смещены, состоятельны и асимптотически нормальны.

**Эффективность МНК-оценок для случая гомоскедастичных ошибок.** Если предположения метода наименьших квадратов из вставки «Основные поня-

тия 4.3» выполняются и ошибки гомоскедастичны, то МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$  эффективны среди всех линейных по  $Y_1, \dots, Y_n$  оценок, являющихся условно несмещенными относительно  $X_1, \dots, X_n$ . Этот результат, называемый теоремой Гаусса–Маркова, мы обсудим в разделе 5.5.

**Формула для дисперсии для случая гомоскедастичных ошибок.** Если ошибки гомоскедастичны, то формулы для дисперсий  $\hat{\beta}_0$  и  $\hat{\beta}_1$  из вставки «Основные понятия 4.4» упрощаются. Соответственно, если ошибки гомоскедастичны, то существуют специальные формулы, которые могут быть использованы для оценки стандартных ошибок  $\hat{\beta}_0$  и  $\hat{\beta}_1$ . Формула оценки стандартной ошибки  $\hat{\beta}_1$  в предположении гомоскедастичности ошибок регрессии, выведенная в приложении 5.1, имеет вид:  $SE(\hat{\beta}_1) = \sqrt{\tilde{\sigma}_{\hat{\beta}_1}^2}$ , где  $\tilde{\sigma}_{\hat{\beta}_1}^2$  – оценка дисперсии  $\hat{\beta}_1$  в случае гомоскедастичных ошибок регрессии:

$$\tilde{\sigma}_{\hat{\beta}_1}^2 = \frac{s_{\hat{u}}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad (5.22)$$

где  $s_{\hat{u}}^2$  дана в уравнении (4.19). Формула оценки стандартной ошибки  $\hat{\beta}_0$  для случая гомоскедастичных ошибок регрессии выводится в приложении 5.1. В частном случае, когда  $X$  – бинарная переменная, оценка дисперсии  $\hat{\beta}_1$  при предположении гомоскедастичности ошибок регрессии (т.е. квадрат стандартной ошибки  $\hat{\beta}_1$  при предположении гомоскедастичности ошибок регрессии) – это так называемая формула для объединенной дисперсии разности средних, представленная в уравнении (3.23).

Так как эти альтернативные формулы выведены для частного случая, когда ошибки гомоскедастичны и не применимы, если ошибки гетероскедастичны, они будут отнесены к формулам оценок дисперсии и стандартных ошибок МНК-оценок «только для гомоскедастичных ошибок регрессии». Как следует из этого названия, если ошибки регрессии гетероскедастичны, то стандартные ошибки для случая гомоскедастичных ошибок не применимы. В частности, если ошибки регрессии гетероскедастичны, то  $t$ -статистика, рассчитанная с использованием стандартных ошибок для случая гомоскедастичных ошибок регрессии, не имеет стандартного нормального распределения даже в больших выборках. Фактически правильные критические значения для возможности использования  $t$ -статистик, рассчитанных для случая гомоскедастичных ошибок регрессии, зависят от точной природы гетероскедастичности, поэтому эти критические значения не могут быть сведены в одну таблицу. Аналогично, если ошибки регрессии гетероскедастичны, но доверительный интервал построен как  $\pm 1,96$  от оценки стандартной ошибки оценки коэффициента регрессии, рассчитанной только для случая гомоскедастичных ошибок регрессии, то в общем случае вероятность того, что этот интервал содержит истинное значение коэффициента, не равна 95 % даже в больших выборках.

С другой стороны, так как гомоскедастичность – частный случай гетероскедастичности, оценки  $\hat{\sigma}_{\hat{\beta}_1}^2$  и  $\hat{\sigma}_{\hat{\beta}_0}^2$  дисперсий оценок  $\hat{\beta}_1$  и  $\hat{\beta}_0$ , представленных в выражениях (5.4) и (5.26), дают возможность для адекватной проверки статистических гипотез, независимо от того, гетероскедастичны или гомоскедастичны

ошибки регрессии. Поэтому проверка гипотез и доверительные интервалы, основанные на этих стандартных ошибках, адекватны, независимо от того, гетероскедастичны ошибки или нет. Так как стандартные ошибки, которые мы использовали до сих пор [т.е. те, которые рассчитаны по формулам (5.4) и (5.26)], дают возможность адекватно проверять статистические гипотезы независимо от того, гетероскедастичны ошибки или нет, и они называются *устойчивыми к гетероскедастичности стандартными ошибками*. Так как эти формулы были предложены в работах Эйкера (Eicker, 1967), Хьюбера (Huber, 1967) и Уайта (White, 1980), они называются стандартными ошибками Эйкера – Хьюбера – Уайта (Eicker – Huber – White).

### **Что это означает на практике?**

#### **Что более реалистично: гетероскедастичность или гомоскедастичность?**

Ответ на этот вопрос зависит от области применения. Разберем эти вопросы в рамках примера с гендерными различиями в зарплатах среди выпускников колледжей. Знание того, каким образом оплачивается труд людей вокруг нас, предоставляет нам некие подсказки о том, какие предположения более разумны. На протяжении многих лет и, в меньшей степени, сегодня женщины не находились на самых высокооплачиваемых должностях: всегда были низкооплачиваемые мужчины, но редко высокооплачиваемые женщины. Это говорит о том, что распределение доходов среди женщин уже, чем среди мужчин (см. вставку «Гендерный разрыв в заработных платах выпускников колледжей в Соединенных Штатах» в главе 3). Иными словами, дисперсия ошибки в уравнении (5.20) для женщин, скорее всего, меньше, чем дисперсия ошибки в уравнении (5.21) для мужчин. Таким образом, наличие верхней планки в зарплатах для женщин предполагает, что ошибка в модели регрессии с бинарной переменной, которая представлена уравнением (5.19), является гетероскедастичной. Пока нет веских причин для предположения обратного – и мы продолжаем считать, что таких причин нет, – имеет смысл рассматривать ошибку регрессии в этом примере как гетероскедастичную.



#### **Экономическая ценность дополнительного года обучения: гомоскедастичность или гетероскедастичность?**

В среднем работники с более высоким уровнем образования имеют более высокие доходы, чем работники с низким уровнем образования. Но если лучшие рабочие места, в основном, достаются выпускникам колледжей, также может быть, что разброс доходов больше для работников с более высоким уровнем образования. Будет ли распределение доходов иметь больший разброс с увеличением уровня образования?

Этот вопрос является эмпирическим, поэтому ответ на него требует анализа реальных данных. Рисунок 5.3 представляет собой диаграмму рассеяния почасовой заработной платы и количества лет обучения для выборки из 2989 занятых полный рабочий день работников 29 и 30 лет в Соединенных Штатах в 2008 году, число лет,

потраченных на получение образования которых колеблется от 6 до 18 лет. Данные взяты из текущего обследования населения, проведенного в марте 2009 года, которое было описано в приложении 3.1.

Рисунок 5.3 имеет две отличительные черты. Первое – это то, что среднее распределение доходов возрастает с увеличением числа лет обучения. Это увеличение отражено в МНК-регрессии:

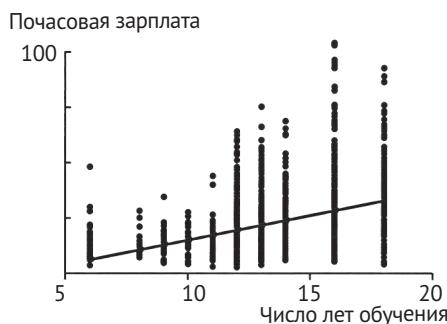
$$\widehat{\text{Earnings}} = -5,38 + 1,76 \text{ Years Education},$$

$$(1,05) \quad (0,08)$$

$$R^2 = 0,159, \text{ SER} = 9,50, \quad (5.23)$$

Линия этой регрессии построена на рисунке 5.3. Коэффициент 1,76 в МНК-регрессии означает, что в среднем почасовая заработная плата увеличится на 1,76 долл. за каждый дополнительный год обучения. 95%-й доверительный интервал для этого коэффициента равен  $1,76 \pm 1,96 \times 0,08$  или [1,60; 1,91].

Второй отличительной чертой рисунка 5.3 является то, что разброс распределения доходов увеличивается с числом лет, потраченных на образование. В то время как некоторые работники с высоким уровнем образованием имеют низкооплачиваемую работу, очень немногие работники с низким уровнем образования имеют высокооплачиваемую работу. Это может быть определено количественно, если мы посмотрим на разброс остатков вокруг линии МНК-регрессии. Для работников с десятилетним образованием стандартное отклонение остатков составляет 4,34 долл.; для работников с дипломом средней школы это стандартное отклонение равно 7,30 долл., а для работников с дипломом колледжа стандартное отклонение увеличивается до 12,25 долл. Поскольку эти стандартные отклонения различаются для разных уровней образования, дисперсия остатков в регрессии (5.23) зависит от значения регрессора (количество лет образования) – иными словами, ошибка регрессии гетероскедастична. В реальном мире далеко не все выпускники вузов будут получать 50 долл. в час к 29 годам, но некоторые из них будут, и работники только с десятью годами образования не имеют шансов найти такую работу.



**Рисунок 5.3. Диаграмма рассеяния почасовой зарплаты и числа лет обучения для 29- и 30-летних работников в США в 2008 году**

Почасовая оплата изображена относительно числа лет обучения для 2989 занятых полный рабочий день работников, достигших возраста 29–30 лет. Разброс вокруг линии регрессии увеличивается с ростом числа лет образования, что свидетельствует о том, что ошибки регрессии гетероскедастичны.



Как показывает пример моделирования доходов, гетероскедастичность возникает во многих эконометрических приложениях. В целом экономическая теория редко дает основания полагать, что ошибки гомоскедастичны. Поэтому разумно предположить, что ошибки могут быть гетероскедастичны, пока у вас нет убедительных причин верить обратному.

**Практические последствия.** С практической точки зрения основной вопрос во всей этой дискуссии заключается в том, следует ли использовать устойчивые к гетероскедастичности стандартные ошибки или можно использовать стандартные ошибки для случая гомоскедастичности. В связи с этим полезно представить себе вычисления обоими способами, а потом выбрать один из них. Если стандартные ошибки, вычисленные для случая гомоскедастичности ошибок регрессии, и устойчивые к гетероскедастичности стандартные ошибки одинаковы, то мы ничего не потеряли, использовав устойчивые к гетероскедастичности стандартные ошибки, однако если они отличаются, вы должны использовать более надежные, учитывающие наличие гетероскедастичности ошибок регрессии. Самое простое, что можно сделать, – это всегда использовать устойчивые к гетероскедастичности стандартные ошибки.

В силу исторических причин во многих программных пакетах по умолчанию рассчитываются стандартные ошибки для случая гомоскедастичных ошибок регрессии, и устанавливать опцию для расчета устойчивых к гетероскедастичности стандартных ошибок или нет – это выбор исследователя. Детали того, как это сделать, зависят от пакета программного обеспечения, который вы используете.

Во всех эмпирических примерах в этой книге мы используем устойчивые к гетероскедастичности стандартные ошибки, если не указано другое<sup>1</sup>.

## 5.5. Теоретические основы обычного метода наименьших квадратов\*

Как отмечалось в разделе 4.5, МНК-оценка регрессии является несмещенной, состоятельной, имеет дисперсию, которая обратно пропорциональна  $n$  и распределена асимптотически нормально. Кроме того, при определенных условиях МНК-оценка является более эффективной, чем некоторые другие кандидаты. В частности, если предположения наименьших квадратов выполняются и если ошибки гомоскедастичны, то МНК-оценка имеет наименьшую дисперсию из всех условно несмешанных оценок, которые являются линейными функциями  $Y_1, \dots, Y_n$ . В этом разделе описывается и обсуждается этот результат, который является следствием теоремы Гаусса–Маркова. Раздел заканчивается обсуждением альтернативных методов оценивания, которые более эффективны, чем МНК, когда условия теоремы Гаусса–Маркова не выполняются.

<sup>1</sup> В случае если эта книга используется в сочетании с другими учебниками, было бы полезно отметить, что некоторые учебники добавляют гомоскедастичность в список предположений наименьших квадратов. Однако, как уже обсуждалось, это дополнительное предположение не является необходимым для того, чтобы можно было использовать МНК для оценки регрессий до тех пор, пока используются устойчивые к гетероскедастичности стандартные ошибки.

\* Этот раздел является необязательным и не представлен в последующих главах.

## Линейные условно несмешенные оценки и теорема Гаусса–Маркова

Если три предположения метода наименьших квадратов (вставка «Основные понятия 4.3») выполняются и если ошибка гомоскедастична, то МНК-оценка имеет наименьшую условную относительно  $X_1, \dots, X_n$  дисперсию среди всех оценок в классе линейных условно несмешенных оценок. Иными словами, МНК-оценка является наилучшей линейной условно несмешенной оценкой, то есть является BLUE (Best Linear conditionally Unbiased Estimator). Этот результат является продолжением результата, обобщенного во вставке «Основные понятия 3.3», о том, что выборочное среднее  $\bar{Y}$  является эффективной оценкой выборочного среднего в классе всех оценок, которые не смешены и являются линейными функциями (средневзвешенными значениями)  $Y_1, \dots, Y_n$ .

**Линейные условно несмешенные оценки.** Класс линейных условно несмешенных оценок состоит из всех оценок  $\beta_1$ , которые являются линейными функциями  $Y_1, \dots, Y_n$  и не смешены условно относительно  $X_1, \dots, X_n$ . То есть если  $\tilde{\beta}_1$  является линейной оценкой, то она может быть записана так:

$$\tilde{\beta}_1 = \sum_{i=1}^n a_i Y_i, \quad (5.24)$$

где веса  $a_1, \dots, a_n$  могут зависеть от  $X_1, \dots, X_n$ , но не от  $Y_1, \dots, Y_n$ . Оценка  $\tilde{\beta}_1$  условно не смешена, если среднее ее условного выборочного распределения относительно  $X_1, \dots, X_n$  есть  $\beta_1$ . То есть оценка  $\tilde{\beta}_1$  условно не смешена, если

$$E(\tilde{\beta}_1 | X_1, \dots, X_n) = \beta_1. \quad (5.25)$$

Оценка  $\tilde{\beta}_1$  является линейной условно несмешенной оценкой, если она может быть записана в виде выражения (5.24) (линейность) и если выражение (5.25) имеет место (условная несмешенность). Как показано в приложении 5.2, МНК-оценка является линейной и условно несмешенной.

**Теорема Гаусса–Маркова.** Теорема Гаусса–Маркова утверждает, что при определенном наборе условий, известных как условия Гаусса–Маркова, МНК-оценка  $\hat{\beta}_1$  имеет наименьшую условную дисперсию, относительно  $X_1, \dots, X_n$  среди всех линейных условно несмешенных оценок  $\beta_1$ ; то есть МНК-оценка является BLUE. Условия Гаусса–Маркова, которые указаны в приложении 5.2, вытекают из трех предположений метода наименьших квадратов, а также предположения о том, что ошибки гомоскедастичны. Следовательно, если три предположения метода наименьших квадратов выполняются и ошибки гомоскедастичны, то МНК является BLUE. Теорема Гаусса–Маркова сформулирована во вставке «Основные понятия 5.5» и доказана в приложении 5.2.

### Теорема Гаусса–Маркова для $\hat{\beta}_1$

Если выполняются предположения метода наименьших квадратов из вставки «Основные понятия 4.3» и если ошибки регрессии гомоскедастичны, то МНК-оценка  $\hat{\beta}_1$  является наилучшей (эффективной) линейной условно несмешенной оценкой (BLUE).

**ОСНОВНЫЕ ПОНЯТИЯ**

5.5

**Ограничения теоремы Гаусса–Маркова.** Теорема Гаусса–Маркова дает теоретическое обоснование использования МНК. Однако теорема имеет два важных ограничения. Во–первых, ее условия могут не выполняться на практике. В частности, если ошибка гетероскедастична – как это часто бывает в экономических приложениях, – МНК–оценка больше не BLUE. Как уже говорилось в разделе 5.4, наличие гетероскедастичности не представляет угрозы для проверки гипотез, если мы используем устойчивые к гетероскедастичности оценки стандартных ошибок, но это означает, что МНК больше не является эффективной линейной условно несмешенной оценкой. Альтернативой является МНК–оценка при наличии гетероскедастичности известного вида, называемая взвешенной оценкой наименьших квадратов и обсуждаемая далее.

Второе ограничение теоремы Гаусса–Маркова состоит в том, что даже если условия теоремы выполнены, существуют другие возможные оценки, которые не являются линейными и условно несмешенными; но при некоторых условиях эти оценки будут более эффективными, чем МНК.

### **Отличные от МНК методы оценивания**

При определенных условиях некоторые методы оценивания дают оценки, которые являются более эффективными, чем оценки МНК.

**Взвешенный метод наименьших квадратов.** Если ошибки регрессии гетероскедастичны, то МНК больше не является BLUE. Если характер гетероскедастичности известен – особенно если условная дисперсия относительно  $X_i$  известна с точностью до постоянного коэффициента пропорциональности, – то можно построить оценку, которая имеет меньшую дисперсию, чем МНК–оценка. Этот метод, называемый *взвешенным методом наименьших квадратов* (WLS), взвешивает  $i$ –е наблюдение с весом, обратным квадратному корню из условной дисперсии  $u_i$  относительно  $X_i$ . В результате такого взвешивания ошибки в этой взвешенной регрессии являются гомоскедастичными, поэтому метод наименьших квадратов, примененный к взвешенным данным, является BLUE. Несмотря на теоретическую элегантность этого метода, практической проблемой при его использовании является то, что вы должны знать, как условная дисперсия  $u_i$  зависит от  $X_i$ , что редко известно в эконометрических приложениях. По этой причине взвешенный метод наименьших квадратов используется гораздо реже, чем МНК, и дальнейшее его обсуждение мы откладываем до главы 17.

**Метод наименьших абсолютных отклонений.** Как уже говорилось в разделе 4.3, МНК–оценка может быть чувствительна к выбросам. Если большие выбросы не являются редкостью, то существуют другие оценки, которые могут быть более эффективными, чем МНК, и соответствующая проверка гипотез на их основе может быть более надежной. Одной из таких оценок является оценка метода наименьших абсолютных отклонений (LAD), при которой коэффициенты регрессии  $\beta_0$  и  $\beta_1$  получаются путем минимизации суммы, аналогичной сумме (4.6), за исключением того, что вместо суммы квадратов остатков используется сумма модулей остатков. То есть оценки LAD для  $\beta_0$  и  $\beta_1$  – это значения  $b_0$  и  $b_1$ ,

которые минимизируют сумму  $\sum_{i=1}^n |Y_i - b_0 - b_1 X_i|$ . LAD-оценка является менее чувствительной к большим выбросам в  $u$ , чем МНК-оценка.

В экономических данных большие выбросы в  $u$  встречаются нечасто, поэтому использование LAD-оценки или других оценок с пониженной чувствительностью к выбросам является редкостью в приложениях. В связи с этим в дальнейшем мы сосредоточим свое внимание исключительно на изучении методов наименьших квадратов.

## 5.6. Использование $t$ -статистики регрессии в малых выборках\*

Когда размер выборки небольшой, точное распределение  $t$ -статистики является сложным и зависит от неизвестного выборочного распределения данных. Однако если выполняются все три предположения метода наименьших квадратов, ошибки регрессии гомоскедастичны и нормально распределены, то распределение МНК-оценки является нормальным, и  $t$ -статистика, рассчитанная при условии наличия гомоскедастичности, имеет распределение Стьюдента. Эти пять предположений — три предположения метода наименьших квадратов, предположения о том, что ошибки гомоскедастичны и нормально распределены — называются *предположениями модели нормальной линейной регрессии с гомоскедастичными ошибками*.

### $t$ -статистика и $t$ -распределение Стьюдента

Вспомним из раздела 2.4, что распределением Стьюдента с  $m$  степенями свободы называется распределение  $Z/\sqrt{W/m}$ , где  $Z$  является случайной величиной со стандартным нормальным распределением,  $W$  является случайной величиной с распределением хи-квадрат с  $m$  степенями свободы и  $Z$  и  $W$  являются независимыми. При нулевой гипотезе  $t$ -статистика, вычисленная с использованием стандартных ошибок при условии гомоскедастичности ошибок регрессии, может быть записана в указанном далее виде.

$t$ -статистика, проверяющая нулевую гипотезу  $\beta_1 = \beta_{1,0}$  и вычисленная в предположении гомоскедастичности ошибок регрессии, имеет вид:  $\tilde{t} = (\hat{\beta}_1 - \beta_{1,0})/\tilde{\sigma}_{\hat{\beta}_1}$ , где  $\tilde{\sigma}_{\hat{\beta}_1}$  определяется в уравнении (5.22). В предположениях модели нормальной линейной регрессии с гомоскедастичными ошибками  $Y$  имеет нормальное распределение, условное относительно  $X_1, \dots, X_n$ . Как обсуждается в разделе 5.5, МНК-оценка является взвешенным средним  $\bar{Y}_1, \dots, \bar{Y}_n$ , где веса зависят от  $X_1, \dots, X_n$  [см. выражение (5.32) в приложении 5.2]. Так как взвешенное среднее независимых нормальных случайных величин нормально распределено,  $\hat{\beta}_1$  имеет нормальное распределение, условное относительно  $X_1, \dots, X_n$ . Таким образом,  $(\hat{\beta}_1 - \beta_{1,0})$  имеет условное относительно  $X_1, \dots, X_n$  нормальное распределение

\* Этот раздел является необязательным и не используется в последующих главах.

в условиях нулевой гипотезы. В дополнение, (нормированная) оценка дисперсии при условии гомоскедастичности ошибок регрессии имеет хи-квадрат распределение с  $n-2$  степенями свободы, деленное на  $n-2$ , и  $\hat{\sigma}_{\beta_1}$  и  $\hat{\beta}_1$  независимо распределены. Следовательно,  $t$ -статистика, рассчитанная при условии гомоскедастичности ошибок регрессии, имеет распределение Стьюдента с  $n-2$  степенями свободы.

Этот результат тесно связан с результатом, который обсуждался в разделе 3.5 в рамках тестирования гипотезы о равенстве средних. В этой задаче, если два выборочных распределения нормальны и имеют одинаковые дисперсии и если  $t$ -статистика построена с использованием объединенной формулы для стандартной ошибки [выражение (3.23)], то (объединенная)  $t$ -статистика имеет  $t$ -распределение Стьюдента. Если  $X$  является бинарной переменной, стандартная ошибка для  $\hat{\beta}_1$  при условии гомоскедастичности ошибок регрессии упрощается до объединенной формулы для стандартной ошибки для разности средних. Отсюда следует, что результат раздела 3.5 является частным случаем этого результата, полученного при выполнении предположений модели нормальной линейной регрессии с гомоскедастичными ошибками: мы пришли к выводу, что  $t$ -статистика, рассчитанная при условии гомоскедастичности остатков регрессии, имеет  $t$ -распределение Стьюдента (см. упражнение 5.10).

### ***Использование $t$ -распределения Стьюдента на практике***

Если ошибки регрессии гомоскедастичны и нормально распределены и если используется  $t$ -статистика, рассчитанная в предположении гомоскедастичности остатков регрессии, то критические значения следует брать из распределения Стьюдента (таблица 2 приложения) вместо стандартного нормального распределения. Так как разница между распределением Стьюдента и нормальным распределением незначительна, если  $n$  умеренное или большое, то это различие имеет значение только при небольшом размере выборки.

В эконометрических приложениях очень редко есть основание полагать, что ошибки гомоскедастичны и нормально распределены. Размер выборки, как правило, большой, однако мы можем проверять гипотезы, как описано в разделах 5.1 и 5.2, то есть сначала необходимо вычислить устойчивые к гетероскедастичности стандартные ошибки, а затем с помощью стандартного нормального распределения вычислить  $p$ -значения для проверки гипотез и построения доверительных интервалов.

## **5.7. Заключение**

Вернемся ненадолго к задаче, с которой начиналась глава 4: окружной школьный инспектор рассматривает возможность найма дополнительных учителей, чтобы уменьшить соотношение числа учеников и учителей. Узнали ли мы что-нибудь, что инспектор могла бы счесть полезным для себя при принятии такого решения?

Оценки регрессии, полученные на основе 420 наблюдений за 1998 год из базы данных по результатам тестов в школьных округах Калифорнии, показали, что существует отрицательная связь между соотношением учеников и учи-

телей и результатами тестов: школьные округа с меньшими классами имеют более высокие показатели теста. Коэффициент умеренно велик с практической точки зрения: школьные округа, где число учеников на одного учителя на два меньше, в среднем имеют результаты тестов, которые на 4,6 балла выше. Это соответствует перемещению школьного округа из 50-го процентиля распределения тестовых баллов примерно в 60-й процентиль.

Коэффициент при соотношении учеников и учителей статистически значимо отличается от нуля на 5 %-м уровне значимости. Теоретический коэффициент наклона может быть равен нулю, и мы могли бы получить такую оценку посредством случайного изменения выборки. Тем не менее вероятность того, что мы не отвергнем гипотезу о равенстве нулю углового коэффициента регрессии, чрезвычайно мала и примерно равна 0,001 %. 95 %-й доверительный интервал для  $\beta_1$  равен  $-3,30 \leq \beta_1 \leq -1,26$ .

Этот результат дает нам возможность ответить на вопрос окружного школьного инспектора, хотя все еще остаются некоторые сомнительные моменты. Существует отрицательное отношение между соотношением учеников и учителей и результатами тестов, но обязательно ли эта связь служить основой для принятия решения окружным школьным инспектором? Школьные округа с более низким соотношением учеников и учителей в среднем показывают более высокие результаты при тестировании. Но значит ли это, что сокращение соотношения учеников и учителей, по сути, увеличит тестовые баллы?

На самом деле существует причина полагать, что этого может и не произойти. Найм большего числа учителей в конце концов стоит денег, поэтому более богатые школьные округа могут позволить себе меньшие классы. Но студенты в богатых школах также имеют и другие преимущества над своими бедными соседями, в том числе лучшие условия жизни, новые книги и более высоко оплачиваемых учителей. Кроме того, студенты в богатых школах, как правило, сами приходят из более состоятельных семей и, следовательно, имеют и другие преимущества, непосредственно не связанные с их школой. Например, в Калифорнии есть большое иммигрантское сообщество; эти иммигранты, как правило, беднее, чем в общей выборке, и, во многих случаях, их дети не являются носителями английского языка. Таким образом, может оказаться, что наше оцененное отрицательное отношение между результатами тестов и соотношением учеников и учителей является следствием наличия классов больших размеров в сочетании со многими другими факторами, которые, по сути, являются реальной причиной более низких результатов тестов.

Эти другие факторы, или «пропущенные переменные», могут означать, что результаты проведенного анализа, в основе которого лежит МНК, не имеют большого значения для окружного школьного инспектора. Действительно, это может ввести в заблуждение: изменение соотношения учеников и учителей само по себе не приведет к изменению этих других факторов, определяющих успеваемость ребенка в школе. Чтобы решить эту проблему, нам нужен метод, который позволяет изолировать влияние на результаты тестов изменения соотношения учеников и учителей, если считать эти другие факторы постоянными. Этот метод – множественный регрессионный анализ – рассматривается в главах 6 и 7.

## Выходы

1. Проверка гипотез для коэффициентов регрессии аналогична проверке гипотез для выборочного среднего: используйте  $t$ -статистику для расчета  $p$ -значения и либо отвергните нулевую гипотезу, либо нет. Как и доверительный интервал для выборочного среднего, 95 %-й доверительный интервал для коэффициента регрессии рассчитывается как оценка  $\pm 1,96$  стандартной ошибки.
2. Если  $X$  является бинарной переменной, регрессионная модель может быть использована для оценки и проверки гипотезы о разности между генеральными средними для группы с  $X=0$  и группы с  $X=1$ .
3. В общем случае ошибка регрессии  $U_i$  гетероскедастична, то есть условная дисперсия  $u_i$  относительно  $X_p$ ,  $\text{var}(u_i | X_i = x)$  зависит от  $x$ . Особый случай – когда ошибка регрессии гомоскедастична, то есть  $\text{var}(u_i | X_i = x)$  является постоянной. Стандартные ошибки оценок коэффициентов, вычисленные в предположении гомоскедастичности ошибок регрессии, не позволяют сделать достоверные статистические выводы, если на самом деле ошибки гетероскедастичны, в отличие от устойчивых к гетероскедастичности стандартных ошибок.
4. Если выполняются три предположения метода наименьших квадратов и если регрессионные ошибки гомоскедастичны, то вследствие теоремы Гаусса–Маркова МНК-оценки регрессии являются BLUE.
5. Если помимо выполнения трех предположений метода наименьших квадратов ошибки регрессии гомоскедастичны и распределены нормально, то МНК  $t$ -статистика, вычисленная в предположении гомоскедастичности ошибок регрессии, имеет распределение Стьюдента, если нулевая гипотеза верна. Разница между распределением Стьюдента и нормальным распределением ничтожно мала, если размер выборки умеренный или большой.

## Основные понятия

Нулевая гипотеза (с. 149).

Двухсторонняя альтернативная гипотеза (с. 149).

Стандартная ошибка  $\beta_1$  (с. 149).

$t$ -статистика (с. 149).

$p$ -значение (с. 150).

Доверительный интервал для  $\beta_1$  (с. 154).

Уровень доверия (доверительная вероятность) (с. 154).

Индикаторная переменная (с. 156).

Дамми-переменная (с. 156).

Множитель  $D_i$  (с. 156).

Коэффициент при  $D_i$  (с. 156).

Гетероскедастичность и гомоскедастичность (с. 158).

Оценка стандартной ошибки в предположении гомоскедастичности ошибок регрессии (с. 161).

Устойчивая к гетероскедастичности стандартная ошибка (с. 162).

Теорема Гаусса–Маркова (с. 165).

Наилучшая линейная несмешенная оценка (BLUE) (с. 165).

Взвешенный метод наименьших квадратов (с. 166).

Предположения модели нормальной линейной регрессии с гомоскедастичными ошибками (с. 167).

Условия Гаусса–Маркова (с. 178).

### **Вопросы для повторения и закрепления основных понятий**

- 5.1. Опишите процедуру для вычисления  $p$ -значения для проверки двухсторонней гипотезы  $H_0 : \mu_Y = 0$ , используя выборку независимых одинаково распределенных случайных величин  $Y_i, i=1, \dots, n$ . Опишите процедуру для вычисления  $p$ -значения для проверки двухсторонней гипотезы  $H_0 : \beta_1 = 0$  в модели регрессии, используя выборку независимых одинаково распределенных случайных величин  $(Y_i, X_i), i=1, \dots, n$ .
- 5.2. Объясните, как вы могли бы использовать модель регрессии для оценки гендерного разрыва в заработной плате мужчин и женщин? Каковы зависимые и независимые переменные?
- 5.3. Определите *гомоскедастичность* и *гетероскедастичность* ошибок регрессии. Приведите гипотетический эмпирический пример, в котором, как вы считаете, ошибки будут гетероскедастичны, и объясните свои рассуждения.

### **Упражнения**

- 5.1. Предположим, что исследователь, используя данные о размере класса ( $CS$ ) и среднем балле за тест в 100 третьих классах, оценивает МНК-регрессию так:

$$\widehat{TestScore} = 520,4 - 5,82 \times CS, \quad R^2 = 0,08, \quad SER = 11,5.$$

(20,4)                                   (2,21)

- a) Постройте 95 %-й доверительный интервал для коэффициента наклона регрессии  $\beta_1$ .
  - b) Вычислите  $p$ -значение для проверки двухсторонней нулевой гипотезы  $H_0 : \beta_1 = 0$ . Отвергнете ли вы нулевую гипотезу на уровне значимости 5 %? На уровне значимости 1 %?
  - c) Вычислите  $p$ -значение для проверки двусторонней нулевой гипотезы  $H_0 : \beta_1 = -5,6$ . Без дополнительных расчетов определите, содержится ли  $-5,6$  в 95 %-м доверительном интервале для  $\beta_1$ .
  - d) Постройте 99 %-й доверительный интервал для  $\beta_0$ .
- 5.2. Предположим, что исследователь, используя данные о заработной плате 250 случайно выбранных мужчин и 280 случайно выбранных женщин, оценивает МНК-регрессию так:

$$\widehat{Wage} = 12,52 + 2,12 \times Male, \quad R^2 = 0,06, \quad SER = 4,2,$$

(0,23)                                   (0,36)

где заработная плата (*Wage*) измеряется в долларах в час, и *Male* – бинарная переменная, равная единице для мужчин и нулю для женщин. Определите гендерный разрыв в заработных платах мужчин и женщин как разность в их средних доходах.

- Каков предполагаемый гендерный разрыв?
- Отличается ли рассчитанный гендерный разрыв значимо от нуля? (Вычислите *p*-значение для проверки нулевой гипотезы, которая состоит в том, что гендерного разрыва не существует).
- Постройте 95 %-й доверительный интервал для гендерного разрыва.
- Какова средняя заработная плата женщин в этом примере? А мужчин?
- Другой исследователь использует те же данные, но оценивает регрессию заработной платы (*Wages*) на *Female*, переменную, равную единице для женщин и нулю для мужчин. Каковы будут оценки этой регрессии?

$$\widehat{Wage} = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} \times Female, \quad R^2 = \underline{\hspace{2cm}}, \quad SER = \underline{\hspace{2cm}}.$$

- 5.3. Предположим, что у нас есть случайная выборка из генеральной совокупности двадцатилетних молодых людей, состоящая из 200 наблюдений, и известны их рост и вес. Регрессия зависимости веса от роста имеет вид:

$$\widehat{Weight} = -99,41 + 3,94 \times Height, \quad R^2 = 0,81, \quad SER = 10,2,$$

(2,15)                    (0,31)

где вес (*Weight*) измеряется в фунтах и рост (*Height*) измеряется в дюймах. Человек растет на 1,5 дюйма в год. Постройте 99 %-й доверительный интервал для изменения его веса.

- 5.4. Прочтите текст во вставке «Экономическая ценность дополнительного года обучения: гомоскедастичность или гетероскедастичность?» из раздела 5.4. Используйте регрессию (5.23), чтобы ответить на следующие вопросы:

- Случайно выбранный 30-летний работник сообщает, что его уровень образования составляет 16 лет. Какова ожидаемая средняя почасовая зарплата этого работника?
- Выпускник средней школы (12 лет обучения) рассматривает обучение в колледже для получения двухлетней степени. Насколько увеличится его средняя почасовая зарплата?
- Консультант средней школы говорит студенту, что выпускник колледжа в среднем зарабатывает на 10 долл. в час больше, чем выпускник средней школы. Согласуется ли это утверждение с оценками регрессии? Какой диапазон значений соответствует регрессии?

- 5.5. В 1980-х годах в штате Теннесси провели эксперимент, в котором воспитанники подготовительного класса начальной школы были случайным образом распределены в «обычные» и «маленькие» классы. В конце учебного года они сдавали стандартные экзамены (тесты). (В обычных классах обучалось около 24 детей, а в маленьких классах – около 15 детей).

Предположим, что в генеральной совокупности средний балл по стандартизованным тестам равен 925 со стандартным отклонением 75 баллов. Пусть  $SmallClass$  означает бинарную переменную, равную единице, если ребенок обучался в маленьком классе, и равную нулю в противном случае. Регрессия показателя результатов тестов ( $TestScore$ ) на эту бинарную переменную имеет вид:

$$\widehat{TestScore} = 918,4 + 13,9 \times SmallClass, \quad R^2 = 0,01, \quad SER = 74,6.$$

(1,6)                          (2,5)

- a) Являются ли результаты тестов в маленьких классах лучшими, чем в обычных? Насколько? Является ли этот эффект большим? Объясните.
  - b) Является ли предполагаемое влияние размера класса на результаты тестов статистически значимым? Выполните проверку на уровне значимости 5 %.
  - c) Постройте 99 %-й доверительный интервал для описания влияния бинарной переменной, характеризующей размер класса, на результат теста.
- 5.6. Рассмотрим регрессию из упражнения 5.5.
- a) Считаете ли вы, что ошибки регрессии являются гомоскедастичными? Объясните.
  - b)  $SE(\hat{\beta}_1)$  была вычислена с помощью уравнения (5.3). Предположим, что ошибки регрессии были гомоскедастичны: будет ли это влиять на доверительный интервал, построенный в упражнении 5.5 (в)? Объясните.
- 5.7. Предположим, что  $(Y_i, X_i)$  удовлетворяют предположениям из вставки «Основные понятия 4.3». По случайной выборке объема  $n=250$  оценена следующая регрессия:
- $$\hat{Y} = 5,4 + 3,2 \times X, \quad R^2 = 0,26, \quad SER = 6,2.$$
- (3,1)                          (1,5)
- a) Проверьте нулевую гипотезу  $H_0: \beta_1 = 0$  против альтернативной  $H_1: \beta_1 \neq 0$  на уровне значимости 5 %.
  - b) Постройте 95 %-й доверительный интервал для  $\beta_1$ .
  - c) Предположим, вы узнали, что  $Y_i$  и  $X_i$  были независимы. Удивились ли вы? Объясните, почему?
  - d) Предположим, что  $Y_i$  и  $X_i$  являются независимыми, мы рассмотрели много выборок размера  $n=250$ , оценили соответствующие им регрессии и ответили на пункты (a) и (b) упражнения. В какой доле выборок нулевая гипотеза из пункта (a) была бы отвергнута? В какой доле выборок значение  $\beta_1 = 0$  было бы включено в доверительный интервал из пункта (b)?
- 5.8. Предположим, что  $(Y_i, X_i)$  удовлетворяют предположениям из вставки «Основные понятия 4.3», и, кроме того,  $u_i$  является  $N(0, \sigma_u^2)$  и не зависит от  $X_i$ . По выборке из  $n=30$  наблюдений получаем:

$$\hat{Y} = 43,2 + 61,5 \times X, \quad R^2 = 0,54, \quad SER = 1,52,$$

(10,2)                          (7,4)

где числа в скобках – стандартные ошибки коэффициентов регрессии, рассчитанные при предположении гомоскедастичности остатков регрессии.

- Постройте 95 %-й доверительный интервал для  $\beta_0$ .
- Проверьте нулевую гипотезу  $H_0 : \beta_1 = 55$  против  $H_1 : \beta_1 \neq 55$  на уровне значимости 5 %.
- Проверьте нулевую гипотезу  $H_0 : \beta_1 = 55$  против  $H_1 : \beta_1 > 55$  на уровне значимости 5 %.

5.9. Рассмотрите модель линейной регрессии:

$$Y_i = \beta X_i + u_i,$$

где  $u_i$  и  $X_i$  удовлетворяют условиям из вставки «Основные понятия 4.3». Пусть  $\bar{\beta}$  – оценка коэффициента  $\beta$ , которая построена как  $\bar{\beta} = \bar{Y}/\bar{X}$ , где  $\bar{Y}$  и  $\bar{X}$  являются выборочными средними для  $Y_i$  и  $X_i$ , соответственно.

- Покажите, что  $\bar{\beta}$  является линейной функцией по  $Y_1, Y_2, \dots, Y_n$ .
- Покажите, что  $\bar{\beta}$  условно не смещена.

5.10. Пусть  $X_i$  обозначает бинарную переменную. Рассмотрим регрессию  $Y_i = \beta_0 + \beta_1 X_i + u_i$ . Пусть  $\bar{Y}_0$  означает выборочное среднее для наблюдений с  $X=0$  и  $\bar{Y}_1$  обозначает выборочное среднее для наблюдений с  $X=1$ . Покажите, что  $\hat{\beta}_0 = \bar{Y}_0$ ,  $\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_1$  и  $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$ .

5.11. Случайная выборка состоит из  $n_m = 120$  мужчин и  $n_w = 131$  женщин. Вы-

борочное среднее недельных заработков мужчин  $\bar{Y}_m = (1/n_m) \sum_{i=1}^{n_m} Y_{m,i}$

равно 523,10 долл., а выборочное стандартное отклонение

$s_m = \sqrt{\frac{1}{n_m - 1} \sum_{i=1}^{n_m} (Y_{m,i} - \bar{Y}_m)^2}$  составляет 68,1 долл. Соответствующие зна-

чения для женщин равны  $\bar{Y}_w = \$485,10$  и  $s_w = \$51,10$ . Пусть  $Women$  – фиктивная переменная, принимающая значение единица для женщин и нуль для мужчин. Предположим, что 251 наблюдение используются в регрессии  $Y_i = \beta_0 + \beta_1 Women_i + u_i$ . Найдите МНК-оценки коэффициентов  $\beta_0$  и  $\beta_1$  и их соответствующие стандартные ошибки.

5.12. Рассмотрим уравнение (4.22). Выведите формулу дисперсии  $\beta_0$  при условии гомоскедастичности ошибок регрессии [см. уравнение (5.28) в приложении 5.1].

5.13. Предположим, что  $(Y_i, X_i)$  удовлетворяют предположениям из вставки «Основные понятия 4.3», и, кроме того,  $u_i$  является  $N(0, \sigma_u^2)$  и не зависит от  $X_i$ .

- Является ли  $\hat{\beta}_1$  условно несмещенной?
- Является ли  $\hat{\beta}_1$  наилучшей линейной условно несмещенной оценкой  $\beta_1$ ?

- в) Как изменились бы ваши ответы в пунктах (а) и (б), если бы вы только предполагали, что  $(Y_i, X_i)$  удовлетворяют предположениям из вставки «Основные понятия 4.3» и  $\text{var}(u_i | X_i = x)$  постоянна?
- г) Как изменились бы ваши ответы на пункты (а) и (б), если бы вы только предполагали, что  $(Y_i, X_i)$  удовлетворяют предположениям из вставки «Основные понятия 4.3»?
- 5.14. Предположим, что  $Y_i = \beta X_i + u_i$ , где  $(u_i, X_i)$  удовлетворяет условиям Гаусса–Маркова, приведенным в (5.31).
- Выведите оценку наименьших квадратов коэффициента  $\beta$  и покажите, что она является линейной функцией по  $Y_1, Y_2, \dots, Y_n$ .
  - Покажите, что эта оценка является условно несмешенной.
  - Выведите условную дисперсию оценки.
  - Докажите, что оценка является BLUE.
- 5.15. Исследователь имеет две независимые выборки наблюдений  $(Y_i, X_i)$ . Для определенности предположим, что  $Y_i$  обозначает доходы,  $X_i$  обозначает число лет обучения и имеются независимые выборки мужчин и женщин. Запишите регрессию для мужчин как  $Y_{m,i} = \beta_{m,0} + \beta_{m,1}X_{m,i} + u_{m,i}$  и регрессию для женщин как  $Y_{w,i} = \beta_{w,0} + \beta_{w,1}X_{w,i} + u_{w,i}$ . Пусть  $\hat{\beta}_{m,1}$  обозначает МНК-оценку, построенную с использованием выборки мужчин, а  $\hat{\beta}_{w,1}$  – МНК-оценку, построенную из выборки женщин, и  $SE(\hat{\beta}_{m,1})$  и  $SE(\hat{\beta}_{w,1})$  обозначают соответствующие стандартные ошибки. Покажите, что стандартная ошибка для  $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$  равна  $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$ .

### **Компьютерные упражнения**

- E5.1. Используя базу данных **CPS08**, описанную в упражнении E4.1, оцените регрессию средней почасовой заработной платы (*AHE*) от возраста (*Age*) и выполните следующие упражнения:
- Является ли рассчитанный коэффициент наклона регрессии статистически значимым? То есть можете ли вы отвергнуть нулевую гипотезу  $H_0: \beta_1 = 0$  против двухсторонней альтернативной гипотезы на 10 %, 5 % или на 1 %-м уровнях значимости? Каково *p*-значение, соответствующее *t*-статистике коэффициента наклона?
  - Постройте 95 %-й доверительный интервал для коэффициента наклона.
  - Повторите (а), используя данные только для выпускников средней школы.
  - Повторите (а), используя данные только для выпускников колледжей.
  - Различно ли влияние возраста на доход для выпускников средних школ и для выпускников колледжей? Объясните. (Подсказка: см. упражнение 5.15.)

- E5.2. Используя базу данных **TeachingRatings**, описанную в упражнении E4.2, оцените регрессию средней оценки курса (*Course\_Eval*) от индекса «красоты» профессора (*Beauty*). Является ли рассчитанный коэффициент наклона регрессии статистически значимым? То есть можете ли вы отвергнуть нулевую гипотезу  $H_0 : \beta_1 = 0$  против двухсторонней альтернативной гипотезы на 10 %, 5 % или на 1 %-м уровнях значимости? Каково  $p$ -значение, соответствующее  $t$ -статистике коэффициента наклона?
- E5.3. Используя базу данных **CollegeDistance**, описанную в упражнении E4.3, оцените регрессию лет полного образования (*ED*) от расстояния до ближайшего колледжа (*Dist*) и выполните следующие упражнения:
- Является ли рассчитанный коэффициент наклона регрессии статистически значимым? То есть можете ли вы отвергнуть нулевую гипотезу  $H_0 : \beta_1 = 0$  против двухсторонней альтернативной гипотезы на 10 %, 5 % или на 1 %-м уровнях значимости? Каково  $p$ -значение, соответствующее  $t$ -статистике коэффициента наклона?
  - Постройте 95 %-й доверительный интервал для коэффициента наклона.
  - Оцените регрессию с использованием данных только для женщин и повторите (б).
  - Оцените регрессию с использованием данных только для мужчин и повторите (б).
  - Различно ли влияние расстояния до колледжа на количество лет, полученных на образование, для мужчин и для женщин? (Подсказка: см. упражнение 5.15.)

## Приложения

### **Приложение 5.1. Формулы для стандартных ошибок МНК-оценок**

В данном приложении рассматриваются формулы для стандартных ошибок МНК-оценок, которые были представлены в рамках предположений наименьших квадратов во вставке «Основные понятия 4.3», допускающих гетероскедастичность ошибок регрессии; это «устойчивые к гетероскедастичности» стандартные ошибки. Формулы для дисперсии МНК-оценок коэффициентов регрессии и соответствующих стандартных ошибок приводятся и для частного случая гомоскедастичности ошибок регрессии.

#### **Устойчивые к гетероскедастичности стандартные ошибки**

Оценка  $\hat{\sigma}_{\hat{\beta}}^2$ , определенная в уравнении (5.4), может быть получена заменой дисперсии генеральной совокупности в уравнении (4.21) соответствующими выборочными дисперсиями с последующей модификацией. Дисперсия в чис-

лителе уравнения (4.21) оценивается как  $\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2$ , где сумма делится на  $n-2$  (вместо  $n$ ) для корректировки на степени свободы, чтобы исправить смещение оценки вниз. Это делается аналогично коррекции на степени свободы, использованной в определении *SER* в разделе 4.3. Дисперсия в знаменателе оценивается как  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ . Заменяя  $\text{var}[(X_i - \mu_X)u_i]$  и  $\text{var}(X_i)$  в выражении (4.21) этими двумя оценками, получаем  $\hat{\sigma}_{\hat{\beta}_1}^2$  из выражения (5.4). Состоительность устойчивых к гетероскедастичности стандартных ошибок оценок коэффициентов регрессии обсуждается в разделе 17.3.

Оценка дисперсии  $\hat{\beta}_0$  равна:

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n \hat{H}_i^2 \hat{u}_i^2}{\left( \frac{1}{n} \sum_{i=1}^n \hat{H}_i^2 \right)^2}, \quad (5.26)$$

где  $\hat{H}_i = 1 - \left( \bar{X} / \left( \frac{1}{n} \sum_{i=1}^n X_i \right) \right) X_i$ . Стандартная ошибка  $\hat{\beta}_0$  равна  $SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}_{\hat{\beta}_0}^2}$ . Вывести формулу оценки  $\hat{\sigma}_{\hat{\beta}_0}^2$  можно так же, как и в случае с оценкой  $\hat{\sigma}_{\hat{\beta}_1}^2$ , заменяя генеральные средние на выборочные средние.

### **Дисперсия оценок коэффициентов для случая гомоскедастичных ошибок регрессии**

При выполнении условия гомоскедастичности ошибок регрессии условная дисперсия  $u_i$  относительно  $X_i$  является постоянной:  $\text{var}(u_i | X_i) = \sigma_u^2$ . Если ошибки регрессии гомоскедастичны, формулы из вставки «Основные понятия 4.4» упрощаются до

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_u^2}{n \sigma_X^2} \quad (5.27)$$

$$\text{и } \sigma_{\hat{\beta}_0}^2 = \frac{E(X_i^2)}{n \sigma_X^2} \sigma_u^2. \quad (5.28)$$

Чтобы вывести выражение (5.27), запишите числитель в уравнении (4.21) как  $\text{var}[(X_i - \mu_X)u_i] = E\left(\{(X_i - \mu_X)u_i - E[X_i - \mu_X]u_i\}^2\right) = E\left\{(X_i - \mu_X)u_i\right\}^2 = E\left[(X_i - \mu_X)^2 u_i^2\right] = E\left[(X_i - \mu_X)^2 \text{var}(u_i | X_i)\right]$ , где второе равенство следует из  $E[(X_i - \mu_X)u_i] = 0$  (по первому предположению метода наименьших квадратов), а последнее равенство следует из закона повторного математического

ожидания (см. раздел 2.3). Если  $u_i$  – гомоскедастична, то  $\text{var}(u_i | X_i) = \sigma_u^2$ , так что  $E[(X_i - \mu_X)^2 \text{var}(u_i | X_i)] = \sigma_u^2 E[(X_i - \mu_X)^2] = \sigma_u^2 \sigma_x^2$ . Тогда выражение (5.27) следует из подстановки последнего выражения в числитель уравнения (4.21) и последующего упрощения. Аналогичным образом получаем уравнение (5.28).

### **Оценки стандартных ошибок для случая гомоскедастичных ошибок регрессии**

Оценки стандартных ошибок для случая гомоскедастичных ошибок регрессии получаются заменой средних и дисперсий генеральной совокупности в выражениях (5.27) и (5.28) на выборочные средние и дисперсии и при оценке дисперсии  $u_i$  как квадрата  $SER$ . Тогда оценки дисперсии для случая гомоскедастичных ошибок регрессии имеют вид:

$$\tilde{\sigma}_{\hat{\beta}_1}^2 = \frac{s_{\hat{u}}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \text{ и} \quad (5.29)$$

$$\tilde{\sigma}_{\hat{\beta}_0}^2 = \frac{\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) s_{\hat{u}}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad (5.30)$$

где  $s_{\hat{u}}^2$  дано в уравнении (4.19). Оценки стандартных ошибок для случая гомоскедастичных ошибок регрессии равны квадратному корню из  $\tilde{\sigma}_{\hat{\beta}_0}^2$  и  $\tilde{\sigma}_{\hat{\beta}_1}^2$ .

### **Приложение 5.2. Условия Гаусса–Маркова и доказательство теоремы Гаусса–Маркова**

Как отмечалось в разделе 5.5, теорема Гаусса–Маркова утверждает, что если условия Гаусса–Маркова выполнены, то МНК-оценка является наилучшей (наиболее эффективной) линейной условно несмешанной оценкой (BLUE). В данном приложении мы сформулируем условия Гаусса–Маркова и покажем, что они вытекают из трех условий метода наименьших квадратов при наличии гомоскедастичности ошибок регрессии. Затем мы покажем, что оценка МНК является линейной и условно несмешанной. И, наконец, мы докажем саму теорему.

#### **Условия Гаусса–Маркова**

Три условия Гаусса–Маркова выглядят следующим образом:

- (i)  $E(u_i | X_1, \dots, X_n) = 0;$
- (ii)  $\text{var}(u_i | X_1, \dots, X_n) = \sigma_u^2, 0 < \sigma_u^2 < \infty;$
- (iii)  $E(u_i u_j | X_1, \dots, X_n) = 0, i \neq j,$

где условия выполняются для  $i, j = 1, \dots, n$ . Эти три условия, соответственно, утверждают, что  $u_i$  имеет нулевое среднее, что  $u_i$  имеет постоянную дисперсию и что ошибки не коррелированы для различных наблюдений, где все утверждения выполняются условно относительно всех наблюдаемых значений  $(X_1, \dots, X_n)$  случайной величины  $X$ .

Условия Гаусса–Маркова вытекают из трех предположений метода наименьших квадратов (вставка «Основные понятия 4.3»), а также дополнительного предположения о том, что ошибки регрессии гомоскедастичны. Поскольку наблюдения независимы и нормально распределены (предположение 2),  $E(u_i | X_1, \dots, X_n) = E(u_i | X_i)$  и по предположению 1  $E(u_i | X_i) = 0$ , таким образом, условие (i) выполняется. Аналогично, по предположению 2  $\text{var}(u_i | X_1, \dots, X_n) = \text{var}(u_i | X_i)$  и так как ошибки предполагаются гомоскедастичными,  $\text{var}(u_i | X_i) = \sigma_u^2$ , что является постоянной величиной. Предположение 3 (ненулевые конечные четвертые моменты) гарантирует, что  $0 < \sigma_u^2 < \infty$ , так что условие (ii) выполняется. Чтобы показать, что условие (iii) вытекает из предположений метода наименьших квадратов, заметим, что  $E(u_i u_j | X_1, \dots, X_n) = E(u_i u_j | X_i, X_j)$ , так как  $(X_i, Y_i)$  – независимые одинаково распределенные случайные величины по предположению 2. Предположение 2 также подразумевает, что  $E(u_i u_j | X_i, X_j) = E(u_i | X_i)E(u_j | X_j)$  для  $i \neq j$ , так как  $E(u_i | X_i) = 0$  для всех  $i$ , из чего следует, что  $E(u_i u_j | X_1, \dots, X_n) = 0$  для всех  $i \neq j$ , так что условие (iii) выполняется. Таким образом, из предположений метода наименьших квадратов из вставки «Основные понятия 4.3» и гомоскедастичности ошибок регрессии следуют условия Гаусса–Маркова, перечисленные в (5.31).

**МНК-оценка  $\hat{\beta}_1$  является линейной условно несмещенной**

Чтобы показать, что  $\hat{\beta}_1$  – линейна, сначала заметим, что, поскольку  $\sum_{i=1}^n (X_i - \bar{X}) = 0$  (по определению  $\bar{X}$ ),  $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y} \sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})Y_i$ . Подставляя этот результат в формулу для  $\hat{\beta}_1$  в выражении (4.7), получаем:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{j=1}^n (X_j - \bar{X})^2} = \sum_{i=1}^n \hat{a}_i Y_i, \text{ где } \hat{a}_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (5.32)$$

Так как веса  $\hat{a}_i$ ,  $i = 1, \dots, n$  в уравнении (5.32) зависят от  $X_1, \dots, X_n$ , но не от  $Y_1, \dots, Y_n$ , МНК-оценка  $\hat{\beta}_1$  – линейная оценка.

Если условия Гаусса–Маркова выполнены, оценка  $\hat{\beta}_1$  условно не смещена, дисперсия условного распределения  $\hat{\beta}_1$  относительно  $X_1, \dots, X_n$  имеет вид:

$$\text{var}(\hat{\beta}_1 | X_1, \dots, X_n) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (5.33)$$

Доказательство того, что оценка  $\hat{\beta}_1$  условно не смещена, было приведено ранее в приложении 4.3.

### **Доказательство теоремы Гаусса–Маркова**

Начнем с доказательства некоторых фактов, которые справедливы для всех линейных условно несмешенных оценок, то есть для всех оценок  $\tilde{\beta}_1$ , удовлетворяющих уравнениям (5.24) и (5.25). Подставляя  $Y_i = \beta_0 + \beta_1 X_i + u_i$  в  $\tilde{\beta}_1 = \sum_{i=1}^n a_i Y_i$  и группируя слагаемые, получаем:

$$\tilde{\beta}_1 = \beta_0 \left( \sum_{i=1}^n a_i \right) + \beta_1 \left( \sum_{i=1}^n a_i X_i \right) + \sum_{i=1}^n a_i u_i. \quad (5.34)$$

По первому условию Гаусса–Маркова,  $E\left(\sum_{i=1}^n a_i u_i | X_1, \dots, X_n\right) = \sum_{i=1}^n a_i E(u_i | X_1, \dots, X_n) = 0$ , тогда возьмем условное математическое ожидание от обеих частей равенства (5.34) и получим:  $E(\tilde{\beta}_1 | X_1, \dots, X_n) = \beta_0 \left( \sum_{i=1}^n a_i \right) + \beta_1 \left( \sum_{i=1}^n a_i X_i \right)$ . Так как, по предположению,  $\tilde{\beta}_1$  условно не смещена, она должна быть равна:  $\beta_0 \left( \sum_{i=1}^n a_i \right) + \beta_1 \left( \sum_{i=1}^n a_i X_i \right) = \beta_1$ , но для того чтобы это равенство выполнялось для любых значений  $\beta_0$  и  $\beta_1$  при условии, что  $\tilde{\beta}_1$  условно не смещена, должно выполняться следующее:

$$\sum_{i=1}^n a_i = 0 \text{ и } \sum_{i=1}^n a_i X_i = 1. \quad (5.35)$$

При выполнении условий Гаусса–Маркова формула условной дисперсии  $\tilde{\beta}_1$  относительно  $X_1, \dots, X_n$  довольно проста. Подставляя выражения (5.35) в выражение (5.34), получаем  $\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^n a_i u_i$ . Таким образом,  $\text{var}(\tilde{\beta}_1 | X_1, \dots, X_n) = \text{var}\left(\sum_{i=1}^n a_i u_i | X_1, \dots, X_n\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{cov}(u_i, u_j | X_1, \dots, X_n)$ . Тогда, применяя второе и третье условия теоремы Гаусса–Маркова, получаем, что ковариации (при  $i \neq j$ ) равны нулю, и выражение для условной дисперсии упрощается до вида:

$$\text{var}(\tilde{\beta}_1 | X_1, \dots, X_n) = \sigma_u^2 \sum_{i=1}^n a_i^2. \quad (5.36)$$

Заметим, что выражения (5.35) и (5.36) применимы к  $\hat{\beta}_1$  с весами  $a_i = \hat{a}_i$ , приведенными в выражении (5.32).

Покажем теперь, что два ограничения в выражении (5.35) и выражение для условной дисперсии в выражении (5.36) подразумевают, что условная диспер-

сия  $\tilde{\beta}_1$  больше условной дисперсии  $\hat{\beta}_1$  для любых  $\tilde{\beta}_1$ . Пусть  $a_i = \hat{a}_i + d_i$ , так что

$$\sum_{i=1}^n a_i^2 = \sum_{i=1}^n (\hat{a}_i + d_i)^2 = \sum_{i=1}^n \hat{a}_i^2 + 2 \sum_{i=1}^n \hat{a}_i d_i + \sum_{i=1}^n d_i^2.$$

Используя определение  $\hat{a}_i$  из (5.32), получаем:

$$\begin{aligned} \sum_{i=1}^n \hat{a}_i d_i &= \frac{\sum_{i=1}^n (X_i - \bar{X}) d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i d_i - \bar{X} \sum_{i=1}^n d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \\ &= \frac{\left( \sum_{i=1}^n a_i X_i - \sum_{i=1}^n \hat{a}_i X_i \right) - \bar{X} \left( \sum_{i=1}^n a_i - \sum_{i=1}^n \hat{a}_i \right)}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0, \end{aligned}$$

где предпоследнее равенство следует из  $d_i = a_i - \hat{a}_i$ , а окончательное равенство следует из уравнения (5.35) (которое выполняется как для  $a_i$ , так и для  $\hat{a}_i$ ). Таким образом,  $\sigma_u^2 \sum_{i=1}^n a_i^2 = \sigma_u^2 \sum_{i=1}^n \hat{a}_i^2 + \sigma_u^2 \sum_{i=1}^n d_i^2 = \text{var}(\beta_1 | X_1, \dots, X_n) + \sigma_u^2 \sum_{i=1}^n d_i^2$ ; подставляя

этот результат в уравнение (5.36), получаем:

$$\text{var}(\tilde{\beta}_1 | X_1, \dots, X_n) - \text{var}(\hat{\beta}_1 | X_1, \dots, X_n) = \sigma_u^2 \sum_{i=1}^n d_i^2. \quad (5.37)$$

Таким образом,  $\tilde{\beta}_1$  имеет большую условную дисперсию, чем  $\hat{\beta}_1$ , если  $d_i$  отлична от нуля для любых  $i=1, \dots, n$ . Но если  $d_i = 0$  для всех  $i$ , то  $a_i = \hat{a}_i$  и  $\tilde{\beta}_1 = \hat{\beta}_1$ , что доказывает, что МНК-оценка является BLUE.

### Теорема Гаусса–Маркова для неслучайных $X$

С незначительными изменениями в интерпретации теоремы Гаусса–Маркова также относится к неслучайным регрессорам, то есть к регрессорам, которые не меняют своих значений по повторяющимся выборкам. В частности, если второе предположение наименьших квадратов заменяется на предположение о том, что  $X_1, \dots, X_n$  являются неслучайными (постоянны по повторяющимся выборкам) и  $u_1, \dots, u_n$  – независимые одинаково распределенные случайные величины, то указанная выше формулировка и доказательство теоремы Гаусса–Маркова применимы напрямую, за исключением того, что все «условные относительно  $X_1, \dots, X_n$ » утверждения можно исключить, так как  $X_1, \dots, X_n$  принимают одни и те же значения от одной выборки к другой.

### Выборочное среднее – эффективная линейная оценка $E(Y)$

Применением теоремы Гаусса–Маркова является утверждение о том, что выборочное среднее  $\bar{Y}$  является самой эффективной линейной оценкой  $E(Y)$ ,

когда  $Y_1, \dots, Y_n$  – независимые одинаково распределенные случайные величины. Чтобы убедиться в этом, рассмотрим случай регрессии без  $X$ , так что единственный регрессор – константа  $X_{0i} = 1$ . МНК-оценка  $\hat{\beta}_0 = \bar{Y}$ . Отсюда следует, что при выполнении предположений Гаусса–Маркова  $\bar{Y}$  является BLUE. Обратите внимание, что требование Гаусса–Маркова о том, что ошибка гомоскедастична, автоматически удовлетворено в данном случае, когда нет регрессоров; отсюда следует, что  $\bar{Y}$  является BLUE, если  $Y_1, \dots, Y_n$  – независимые одинаково распределенные случайные величины. Этот результат уже был отмечен ранее во вставке «Основные понятия 3.3».

# **Глава 6. Множественная линейная регрессия**

Главу 5 мы завершили неоднозначно. Несмотря на то что школьные округа с меньшим числом учеников на одного учителя в классе, как правило, имеют более высокие результаты тестов в рассматриваемой базе данных по Калифорнии, возможно, ученики из районов с маленькими классами имеют и другие преимущества, которые помогают им достичь лучших результатов. Может ли это обстоятельство привести к ошибочным результатам при оценке регрессии, и если да, то что можно с этим сделать?

Пропущенные факторы, такие как персональные характеристики ученика, на самом деле могут привести к тому, что оценка метода наименьших квадратов (МНК) в рассматриваемом нами примере будет ошибочной или, точнее, смещённой. В данной главе мы вводим понятие «смещения из-за пропущенных переменных» и объясняем его, а также вводим в рассмотрение множественную линейную регрессию — метод, который может устраниТЬ смещение из-за пропущенных переменных. Ключевая идея множественной линейной регрессии заключается в том, что если у нас есть данные о пропущенных переменных, то мы можем включить их в качестве дополнительных регрессоров и тем самым оценить влияние одного регрессора (соотношения учеников и учителей), считая, что другие переменные (такие как персональные характеристики ученика) не изменяются.

В данной главе описывается способ оценки коэффициентов множественной линейной регрессии. Множественная линейная регрессия во многом похожа на парную линейную регрессию, которая была изучена в главах 4 и 5. Коэффициенты модели множественной линейной регрессии могут быть оценены на основе имеющейся выборки с помощью МНК; МНК-оценки в множественной регрессии являются случайными величинами, поскольку они зависят от данных из случайной выборки, и в больших выборках выборочное распределение МНК-оценок сходится к нормальному.

## **6.1. Смещение из-за пропущенных переменных**

Если мы сконцентрируем наше внимание только на соотношении учеников и учителей, то при оценке регрессии в главах 4 и 5 мы игнорируем несколько потенциально важных факторов, влияющих на результаты экзаменов, поскольку их влияние включается в ошибку регрессии. Пропущенные факторы включают в себя характеристики школы, такие как квалификация учителей и использование компьютеров, а также персональные характеристики ученика, такие

как особенности его семьи. Рассмотрим сначала пропущенные характеристики учеников, что особенно актуально для Калифорнии из-за большого числа иммигрантов среди населения: в школах округов очень много учеников, для которых английский язык не является родным, и они все еще его изучают.

Если мы игнорируем процент изучающих английский язык учеников в школьном округе, МНК-оценка коэффициента наклона в регрессии результатов тестов на соотношение учеников и учителей может быть смещенной, то есть выборочное среднее МНК-оценки может быть не равно истинному воздействию, который оказывает изменение на единицу соотношения числа учеников и учителей на результаты итоговых тестов. Обоснование этому следующее. Ученики, которые до сих пор изучают английский язык, могут хуже справляться со стандартными тестами, чем носители английского языка. Поэтому если в школьных округах с большими классами также имеется много учеников, по-прежнему изучающих английский язык, то МНК-регрессия тестовых баллов на отношение «ученик – учитель» может ошибочно найти корреляцию и оценить большой коэффициент, в то время как на самом деле истинный причинный эффект влияния на результаты тестов, возникающий из-за уменьшения размеров классов, очень мал или даже практически равен нулю. Таким образом, на основе анализа из глав 4 и 5 окружной школьный инспектор может нанять достаточное количество новых учителей, чтобы уменьшить соотношение учеников и учителей в 2 раза, но ее надежда на улучшение результатов тестов не оправдается, если истинный коэффициент мал или равен нулю.

Посмотрев на данные по Калифорнии, мы убеждаемся в том, что наша озабоченность имеет под собой основания. Корреляция между соотношением учеников и учителей и процентом детей, изучающих английский язык (школьники, которые не являются носителями английского языка и которые еще его не освоили) в школьном округе, равна 0,19. Эта небольшая, но положительная корреляция показывает, что школьные округа с большим количеством изучающих английский, как правило, имеют более высокое соотношение учеников и учителей (большие классы). Если бы отношение «ученик – учитель» не было связано с процентом изучающих английский язык, то было бы безопасно игнорировать знание английского языка в регрессии результатов тестов от соотношения учеников и учителей. Но так как соотношение учеников и учителей и процент изучающих английский язык коррелированы, вполне возможно, что МНК-коэффициент в регрессии зависимости результатов тестов от соотношения учеников и учителей отражает это влияние.

### ***Определение смещения из-за пропущенной переменной***

Если регрессор (соотношение учеников и учителей) коррелирован с переменной, которая была исключена из анализа (процент изучающих английский язык) и которая определяет, в частности, зависимую переменную (результат теста), то МНК-оценка будет иметь *смещение из-за пропущенной переменной*.

Смещение из-за пропущенной переменной возникает, если выполнены два условия: (1) если пропущенная переменная коррелирует с включенным регрес-

сором и (2) если пропущенная переменная является детерминантой зависимой переменной. Чтобы проиллюстрировать эти условия, рассмотрим три примера переменных, которые исключены из регрессии зависимости результатов тестов от соотношения учеников и учителей.

**Пример № 1: Процент изучающих английский язык.** Поскольку процент изучающих английский коррелирует с соотношением учеников и учителей, первое условие возникновения смещения из-за пропущенных переменных выполняется.

Вполне вероятно, что школьники, которые до сих пор изучают английский язык, будут хуже справляться со стандартными тестами, чем носители английского языка, и в этом случае процент изучающих английский язык является определяющим фактором результатов тестов и выполняется второе условие возникновения смещения из-за пропущенных переменных. Таким образом, МНК-оценка в регрессии зависимости результатов тестов от соотношения учеников и учителей может неверно отражать влияние этой пропущенной переменной, характеризующей долю изучающих английский язык. Следовательно, пропуск этой переменной (процента изучающих английский язык школьников) может привести к возникновению смещения из-за пропущенных переменных.

**Пример № 2: Время проведения теста.** Другая переменная, исключенная из анализа,— время суток, в которое проводится тест. Для этой пропущенной переменной вполне вероятно, что первое условие для возникновения смещения из-за пропущенной переменной не выполняется, но второе условие имеет место. Например, если время проведения теста колеблется от одного района к другому безотносительно размера класса, то время суток и размер класса будут не коррелированы, поэтому первое условие не выполняется. И наоборот, время проведения теста может повлиять на результаты теста (степень волнения из-за предстоящего экзамена меняется в течение школьного дня), поэтому второе условие выполняется. Однако поскольку в этом примере время, когда проводится тест, не коррелирует с соотношением учеников и учителей, эта переменная (соотношение учеников и учителей) не могла неправильно принять на себя эффект «времени проведения теста». Таким образом, пропуск времени проведения теста не приводит к смещению, возникающему из-за пропущенных переменных.

**Пример № 3: Парковочное пространство на одного ученика.** Другой пропущенной переменной является показатель, характеризующий парковочное пространство на одного ученика (площадь места парковки для учителя, деленная на число учеников). Эта переменная удовлетворяет первому, но не второму условию для возникновения смещения из-за пропущенной переменной. В частности, школы с большим количеством учителей на одного ученика, скорее всего, имеют больше мест парковок для учителей, так что первое условие будет выполнено. Тем не менее в предположении, что обучение происходит в классе, а не на стоянке, парковочное место не имеет прямого влияния на обучение, поэтому второе условие не выполняется. Так как парковочное место на одного ученика не является определяющим фактором для результатов тестов, пропуск этой переменной в регрессионном анализе не приводит к смещению из-за пропущенных переменных.

Понятие смещения вследствие пропущенных переменных обобщается во вставке «Основные понятия 6.1».

**Смещение из-за пропущенных переменных и первое предположение метода наименьших квадратов.** Смещение вследствие пропущенных переменных означает, что первое предположение метода наименьших квадратов –  $E(u_i | X_i) = 0$  из вставки «Основные понятия 4.3» – не выполняется. Чтобы понять, почему так происходит, вспомним, что ошибка регрессии  $u_i$  в парной модели линейной регрессии включает все факторы, кроме  $X_i$ , которые оказывают влияние на  $Y_i$ . Если один из этих факторов коррелирует с  $X_i$ , это означает, что ошибка регрессии (которая содержит этот фактор) коррелирована с  $X_i$ . Другими словами, если пропущенная переменная является определяющим  $Y_i$  фактором, то она содержится в ошибке, и если она коррелирует с  $X_i$ , то ошибка регрессии тоже коррелирует с  $X_i$ . Так как  $u_i$  и  $X_i$  коррелированы, условное среднее  $u_i$  относительно  $X_i$  не равно нулю. Эта корреляция, следовательно, нарушает первое предположение метода наименьших квадратов и имеет серьезные последствия: оценка МНК смещена. Это смещение не исчезает даже в больших выборках, и МНК-оценка несостоительна.

## ОСНОВНЫЕ ПОНЯТИЯ

### 6.1

#### Смещение из-за пропущенных переменных в модели парной линейной регрессии

Смещение вследствие пропущенных переменных – это смещение в МНК-оценке, которое появляется, если регрессор  $X$  коррелирован с пропущенной переменной. Для того чтобы возникало смещение вследствие пропущенных переменных, должны быть выполнены два условия:

1. Объясняющая переменная  $X$  коррелирована с пропущенной переменной.
2. Пропущенная переменная влияет на зависимую переменную  $Y$ .

#### Формула для расчета смещения из-за пропущенных переменных

Обсуждение возникновения смещения из-за пропущенных переменных из предыдущего раздела может быть математически formalизовано, если получить формулу для расчета этого смещения. Обозначим коэффициент корреляции между  $X_i$  и  $u_i$  как  $\text{corr}(X_i, u_i) = \rho_{Xu}$ . Предположим, что второе и третье предположения метода наименьших квадратов выполнены, но первое не выполняется, так что коэффициент корреляции  $\rho_{Xu}$  отличен от нуля. Тогда МНК-оценка имеет предел (выведенный в приложении 6.1):

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}. \quad (6.1)$$

То есть по мере увеличения размера выборки  $\hat{\beta}_1$  с высокой вероятностью приближается к  $\beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$ .

Формула (6.1) обобщает некоторые из идей о смещении вследствие пропущенных переменных, обсуждавшиеся выше:

1. Смещение из-за пропущенных переменных является проблемой вне зависимости от его величины. Так как  $\hat{\beta}_1$  не сходится по вероятности к истинному значению  $\beta_1$ ,  $\hat{\beta}_1$  смещена и несостоительна, то есть  $\hat{\beta}_1$  не является состоятельной оценкой  $\beta_1$ , если присутствует смещение из-за пропущенной переменной. Член  $\rho_{Xu} (\sigma_u / \sigma_X)$  в уравнении (6.1) равен величине смещения оценки  $\hat{\beta}_1$ , которое сохраняется даже в больших выборках.
2. Является ли это смещение большим или маленьким, на практике зависит от корреляции  $\rho_{Xu}$  между регрессором и ошибкой. Чем больше  $\rho_{Xu}$ , тем больше смещение.
3. Направление смещения  $\hat{\beta}_1$  зависит от того, коррелируют ли  $X$  и  $u$  положительно или отрицательно. Например, мы предположили, что процент студентов, изучающих английский язык, оказывает *отрицательное* влияние на результаты тестов по школьному округу (учащиеся, до сих пор изучающие английский язык, имеют более низкие баллы), так что процент изучающих английский язык входит в ошибку с отрицательным знаком. В наших данных доля изучающих английский *положительно* коррелирует с соотношением учеников и учителей (районы с большим количеством изучающих английский имеют большие классы). Таким образом, соотношение учеников и учителей ( $X$ ) будет *отрицательно* коррелировано с ошибкой ( $u$ ), так что  $\rho_{Xu} < 0$ , а коэффициент при соотношении учеников и учителей  $\hat{\beta}_1$  будет смещенным в сторону отрицательного значения. Иными словами, наличие небольшого процента изучающих английский связано как с *высокими* результатами экзаменов, так и с *низким* соотношением учеников и учителей, поэтому одной из причин, по которой МНК-оценка предполагает, что результаты тестов в маленьких классах выше, может быть тот факт, что в школьных округах с классами небольших размеров обучается меньше учеников, изучающих английский язык.



### **Эффект Моцарта: смещение из-за пропущенных переменных?**

Исследование, опубликованное в журнале *Nature* в 1993 году (Rauscher, Shaw and Ky, 1993), показало, что если слушать музыку Моцарта в течение 10–15 минут, то можно на какое-то время поднять свой IQ на 8 или 9 пунктов. Это исследование стало большой сенсацией: и политики, и родители увидели простой способ сделать своих детей умнее. Какое-то время в штате Джорджия для всех детей штата даже распространялись компакт-диски с классической музыкой.

Существуют ли свидетельства в пользу существования «эффекта Моцарта»?

Обзор десятков исследований показал, что те ученики средней школы, которые посещали дополнительные курсы по музыке или искусству в школе, на самом деле получали более высокие баллы на экзаменах по английскому языку и математике, чем те, кто таких курсов не посещал\*. Более пристальный взгляд на эти исследования, однако, показывает, что реальная причина для более высоких результатов экзаменов не очень сильно связана с этими курсами. Авторы обзора предполагают, что корреляция между успешной сдачей экзаменов и посещением курсов по искусству или музыке может возникнуть по ряду других причин. Например, ученики с хорошей успеваемостью могли бы иметь больше времени для факультативных курсов по музыке или быть более заинтересованными в этом, или школы, в которых музыка изучается более глубоко, просто могут быть лучше остальных школ.

В терминах регрессии оцененное соотношение между результатами тестов и посещением дополнительных курсов по музыке, по-видимому, имеет смещение из-за пропущенных переменных. Если мы опустим такие факторы, как врожденная одаренность школьника или рейтинг школы, то посещение дополнительных курсов по музыке может оказаться значимым фактором в регрессии, оценивающей его влияние на результаты экзаменов, когда на самом деле его нет.

Так есть ли эффект Моцарта? Один из способов выяснить это – провести случайный управляемый эксперимент. (Как говорилось в главе 4, случайные контролируемые эксперименты устраниют смещение из-за пропущенных переменных путем случайного включения участников в исследуемую и контрольную группу.) Рассмотренные вместе, случайные контролируемые эксперименты по изучению эффекта Моцарта не показали, что прослушивание Моцарта повышает IQ или общий результат экзамена. Но не вполне понятным причинам, однако, было показано, что прослушивание классической музыки действительно временно помогает в одной узкой области: складывание бумаги и визуализация образов. Так что когда вы в следующий раз будете готовиться к экзамену по оригами, послушайте немного Моцарта.

---

\* См. специальный выпуск журнала *Journal of Aesthetic Education*, 34 (осень/зима), вышедший в 2000 году, особенно статьи Эллен Винер, Моники Купер (Ellen Winner, Monica Cooper. P. 11–76) и Луиса Хетланда (Lois Hetland. P. 105–148).



## **Поможет ли решить проблему смещения из-за пропущенной переменной разбиение данных на подгруппы?**

Что вы можете сделать со смещением из-за пропущенных переменных? Наш окружной школьный инспектор рассматривает вопрос об увеличении количества учителей в своем школьном округе, но она никак не может контролировать долю иммигрантов в сообществе. В результате она заинтересована в понимании того, как влияет соотношение учеников и учителей на результаты тестов при условии постоянства других факторах, в том числе доли изучающих английский язык. Этот новый способ постановки вопроса наводит на мысль о том, что, вме-

сто того чтобы использовать данные по всем школьным округам, возможно, мы должны сосредоточить внимание на изучении округов с процентом изучающих английский язык, сравнимым с процентом по ее округу. Возникает вопрос: справляются ли ученики школ из этого подмножества и с меньшими классами со стандартными тестами лучше, чем остальные в этом подмножестве?

В таблице 6.1 приведены данные о взаимосвязи между размером класса и результатами тестов в районах с сопоставимым процентом изучающих английский язык. Районы делятся на восемь групп. Во-первых, районы разбиты на четыре категории, которые соответствуют квартилям распределения процента изучающих английский язык по районам. Во-вторых, в рамках каждой из этих четырех категорий районы разбиты на две группы в зависимости от того, является ли соотношение учеников и учителей маленьким ( $STR < 20$ ) или большим ( $STR \geq 20$ ).

Таблица 6.1

**Различия в результатах тестов для школьных округов Калифорнии с низкими и высокими отношениями «ученик – учитель» по проценту изучающих английский язык в районе**

	Соотношение учеников и учителей ( $STR < 20$ )		Соотношение учеников и учителей ( $STR \geq 20$ )		Разность оценок за тест, низкое $STR$ – высокое $STR$	
	Средняя оценка за тест	<i>n</i>	Средняя оценка за тест	<i>n</i>	Разность	<i>t</i> -статистика
Все районы	657,4	238	650,0	182	7,4	4,04
Процент изучающих английский язык школьников						
<1,9 %	664,5	76	665,4	27	-0,9	-0,30
1,9 – 8,8 %	665,2	64	661,8	44	3,3	1,13
8,8 – 23,0 %	654,9	54	649,7	50	5,2	1,72
>23,0 %	636,7	44	634,8	61	1,9	0,68

Первая строка в таблице 6.1 показывает общую разность в средних результатах тестов между районами с низким и высоким соотношением учеников и учителей, то есть разность оценок по тесту между этими двумя группами без разбиения их на группы по проценту изучающих английский язык. (Напомним, что эта разность ранее приводилась в регрессии (5.18) как МНК-оценка коэффициента  $D_i$  в регрессии  $TestScore$  на  $D_i$ , где  $D_i$  является бинарной объясняющей переменной, равная единице, если  $STR < 20$ , и нулю – в противном случае.) В полной выборке из 420 округов средний результат тестов в школьных округах с высоким соотношением учеников и учителей на 7,4 пункта превышает аналогичный результат в районах с низким соотношением; причем *t*-статистика равна 4,04, так что нулевая гипотеза о том, что средний балл по тесту одинаков в двух группах, отвергается на 1 %-м уровне значимости.

Последние четыре строки в таблице 6.1 показывают разность в результатах тестов между районами с низким и высоким соотношением учеников и учителей, с разбивкой по группам в зависимости от процента учеников, изучающих английский язык. Тут мы видим иную картину. Среди школьных округов

с наименьшим количеством изучающих английский (<1,9 %) средний результат тестов для тех 76 округов с низким соотношением учеников и учителей составляет 664,5 и среднее по 27 округам с высоким соотношением учеников и учителей равно 665,4. Таким образом, для школьных округов с наименьшим числом изучающих английский язык школьников результаты тестов в школах с низким соотношением учеников и учителей были в среднем всего на 0,9 пунктов ниже, чем в школах с высоким! Во второй подгруппе районы с низким соотношением учеников и учителей имели результаты тестов в среднем на 3,3 балла выше, чем районы с высоким соотношением учеников и учителей; этот разрыв составлял 5,2 балла в третьей подгруппе и лишь 1,9 баллов в подгруппе с округами с наибольшим количеством изучающих английский язык. Как только мы сделаем процент изучающих английский постоянным, разница в успеваемости между районами с высоким и низким отношением «ученик — учитель», пожалуй, составит половину (или меньше) от общей оценки в 7,4 балла.

На первый взгляд этот вывод может показаться странным. Как разность результатов тестов во всем массиве может быть в 2 раза больше разности результатов тестов в любой его подгруппе? Ответ заключается в том, что районы с наибольшим количеством изучающих английский язык учеников, как правило, имеют как более высокое сопротивление учеников и учителей, так и более низкие результаты тестов. Разность средних результатов тестов между школьными округами в подгруппах с самым низким и самым высоким процентами изучающих английский язык велика, она составляет около 30 пунктов. Районы с маленьким числом изучающих английский язык школьников, как правило, имеют более низкое сопротивление учеников и учителей: 74 % (76 из 103) округов в первой подгруппе изучающих английский язык имеют небольшие классы ( $STR < 20$ ), в то время как только 42 % (44 из 105) районов в подгруппе с самым большим количеством изучающих английский язык имеют небольшие классы. Таким образом, школьные округа с наибольшим процентом изучающих английский язык имеют как наименьшие оценки по тестам, так и более высокие сопротивления учеников и учителей, чем другие округа.

Проведенный анализ усиливает обеспокоенность окружного школьного инспектора относительно возможного смещения в результатах регрессионного анализа из-за пропущенных переменных. Если мы посмотрим на данные по подгруппам, то увидим, что различия в результатах тестов во второй части таблицы 6.1 улучшают простой анализ разности в средних в первой строке таблицы 6.1. Тем не менее этот анализ еще не представляет окружному школьному инспектору полезную информацию о влиянии на результаты тестов изменений в размерах классов, считая постоянной долю изучающих английский язык. Такая оценка может быть получена с помощью множественного регрессионного анализа.

## 6.2. Модель множественной регрессии

Модель множественной регрессии расширяет модель парной линейной регрессии, рассмотренную в главах 4 и 5, на случай дополнительных регрессоров. Эта модель позволяет оценить влияние на  $Y_i$  изменения одной переменной ( $X_{1i}$ ), считая другие регрессоры ( $X_{2i}$ ,  $X_{3i}$  и т.д.) постоянными. В задаче о размере класса

модель множественной регрессии дает возможность изолировать влияние на результаты тестов ( $Y_i$ ) соотношения учеников и учителей ( $X_{1i}$ ), предполагая постоянным процент учащихся в школьном округе, изучающих английский язык ( $X_{2i}$ ).

### **Линия теоретической регрессии**

Предположим, что есть только две независимые переменные,  $X_{1i}$  и  $X_{2i}$ . В модели множественной линейной регрессии среднее отношение между этими двумя независимыми переменными и зависимой переменной  $Y$  задается линейной функцией:

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad (6.2)$$

где  $E(Y_i | X_{1i} = x_1, X_{2i} = x_2)$  – условное ожидание  $Y_i$  при условии, что  $X_{1i} = x_1$  и  $X_{2i} = x_2$ . То есть, если соотношение учеников и учителей в  $i$ -м округе ( $X_{1i}$ ) равно некоторому значению  $x_1$  и процент изучающих английский язык в  $i$ -м округе ( $X_{2i}$ ) равен  $x_2$ , то условное математическое ожидание  $Y_i$  относительно соотношения учеников и учителей в школах округа и доли изучающих английский язык приведено в выражении (6.2).

Выражение (6.2) называется *линией теоретической регрессии* или *функцией теоретической регрессии* в модели множественной регрессии. Коэффициент  $\beta_0$  – константа или *свободный член*; коэффициент  $\beta_1$  – *угловой коэффициент при  $X_{1i}$*  или, проще говоря, *коэффициент при  $X_{1i}$*  и коэффициент  $\beta_2$  – *угловой коэффициент при  $X_{2i}$*  или, проще говоря, *коэффициент при  $X_{2i}$* . Одна или более независимых переменных в модели множественной регрессии иногда называются *контрольными переменными*.

Интерпретация коэффициента  $\beta_1$  в (6.2) отлична от той, что имела место, когда переменная  $X_{1i}$  была единственным регрессором: в выражении (6.2)  $\beta_1$  характеризует влияние на  $Y$  единичного изменения в  $X_1$ , считая  $X_2$  *постоянной* или *контролируя  $X_2$* .

Эта интерпретация коэффициента  $\beta_1$  следует из определения того, что ожидаемый эффект влияния на  $Y$  изменения в  $X_1, \Delta X_1$ , фиксируя  $X_2$ , представляет собой разность между ожидаемым значением  $Y$ , когда объясняющие переменные принимают значения  $X_1 + \Delta X_1$  и  $X_2$ , и ожидаемым значением  $Y$ , когда объясняющие переменные принимают значения  $X_1$  и  $X_2$ . Соответственно запишем функцию теоретической регрессии из уравнения (6.2) как  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$  и представим, что  $X_1$  изменяется на величину  $\Delta X_1$  без изменения  $X_2$ , то есть мы считаем  $X_2$  постоянной. Так как  $X_1$  изменилась,  $Y$  изменится на некоторое значение, скажем, на  $\Delta Y$ . После этого изменения новое значение  $Y$ ,  $Y + \Delta Y$  будет иметь вид:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2. \quad (6.3)$$

Выражение для  $\Delta Y$  в терминах  $\Delta X_{1i}$  можно получить, вычитая равенство  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$  из выражения (6.3), в результате чего получим, что  $\Delta Y = \beta_1 \Delta X_1$ .

То есть  $\beta_1 = \frac{\Delta Y}{\Delta X_1}$ , считая  $X_2$  постоянной. (6.4)

Коэффициент  $\beta_1$  показывает, как единичное изменение в  $X_1$  влияет на  $Y$  (ожидаемое изменение  $Y$ ), считая  $X_2$  фиксированным. Другой термин, используемый для описания  $\beta_1$ , – частный эффект влияния  $X_1$  на  $Y$  при постоянном втором регрессоре  $X_2$ .

Интерпретация свободного члена  $\beta_0$  в модели множественной регрессии похожа на интерпретацию свободного члена в модели парной регрессии: это ожидаемое значение  $Y_i$ , когда  $X_{1i}$  и  $X_{2i}$  равны нулю. Проще говоря, свободный член  $\beta_0$  определяет, насколько далеко от начала координат проходит линия теоретической регрессии.

### **Теоретическая модель множественной линейной регрессии**

Линия теоретической регрессии (6.2) отражает связь  $Y$  с  $X_1$  и  $X_2$ , которая сохраняется в среднем в генеральной совокупности. Так же, как и в случае регрессии с одной объясняющей переменной, это отношение не выполняется в точности, так как множество других факторов влияет на зависимую переменную. В дополнение к соотношению учеников и учителей и проценту школьников, по-прежнему изучающих английский язык, результаты тестов зависят, например, от характеристик школы, других характеристик ученика и удачи. Таким образом, функция теоретической регрессии в уравнении (6.2) должна быть дополнена включением этих дополнительных факторов.

Так же как и в случае парной регрессии, факторы, которые определяют  $Y_i$  в дополнение к  $X_{1i}$  и  $X_{2i}$ , включены в уравнение (6.2) в качестве «ошибки» регрессии  $u_i$ . Эта ошибка представляет собой отклонение отдельного наблюдения (результатов тестов в  $i$ -м школьном округе в нашем примере) от линии теоретической регрессии. Соответственно, мы имеем:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i=1, \dots, n, \quad (6.5)$$

где индекс  $i$  означает  $i$ -е наблюдение из  $n$  школьных округов (в нашем примере) в выборке.

Выражение (6.5) иногда называют *теоретической моделью множественной линейной регрессии* с двумя объясняющими переменными  $X_{1i}$  и  $X_{2i}$ .

В регрессии с бинарными объясняющими переменными может быть полезным рассмотреть  $\beta_0$  как коэффициент при регрессоре, который всегда равен 1; подумайте о  $\beta_0$  как о коэффициенте при переменной  $X_{0i}$ , где  $X_{0i}=1$  для  $i=1, \dots, n$ . Соответственно, модель теоретической множественной регрессии (6.5) может быть записана в другом виде, а именно:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{ где } X_{0i} = 1, \quad i = 1, \dots, n. \quad (6.6)$$

Переменная  $X_{0i}$  иногда называется *постоянным регрессором*, так как она принимает одни и те же значения, равные единице, для всех наблюдений. Аналогично, свободный член  $\beta_0$  иногда называется *постоянным членом регрессии*.

Два способа записи модели теоретической регрессии, выражения (6.5) и (6.6), эквивалентны.

До сих пор обсуждение было сосредоточено на случае одной дополнительной переменной  $X_2$ . На практике, однако, в модели парной линейной регрессии

может быть пропущено несколько факторов. Например, игнорирование показателя, характеризующего экономическое положение семьи ученика, может привести к смещению из-за пропущенной переменной, равно как и игнорирование доли изучающих английский язык. Это рассуждение приводит нас к рассмотрению модели с тремя регрессорами или, говоря обобщенно, к модели, которая включает в себя  $k$  регрессоров. Во вставке «Основные понятия 6.2» обобщается информация о модели множественной регрессии с  $k$  регрессорами  $X_{1i}, X_{2i}, \dots, X_{ki}$ .

Определения гомоскедастичности и гетероскедастичности в модели множественной регрессии являются расширениями соответствующих понятий в модели парной линейной регрессии. Ошибка  $u_i$  в модели множественной линейной регрессии называется *гомоскедастичной*, если дисперсия условного распределения  $u_i$  относительно  $X_{1i}, X_{2i}, \dots, X_{ki}$ ,  $\text{var}(u_i | X_{1i}, X_{2i}, \dots, X_{ki})$ , является постоянной для  $i=1, \dots, n$  и, следовательно, не зависит от значений  $X_{1i}, X_{2i}, \dots, X_{ki}$ . В противном случае ошибка называется *гетероскедастичной*.

Модель множественной линейной регрессии дает возможность ответить на вопрос окружного школьного инспектора о влиянии изменений в соотношении учеников и учителей на результаты тестов, считая постоянными другие факторы, которые находятся вне ее контроля. Эти факторы включают в себя не только процент изучающих английский язык, но и другие измеримые переменные, которые могут повлиять на успеваемость школьников, в том числе экономические характеристики их семей. Но чтобы быть полезным для окружного школьного инспектора с практической точки зрения, мы должны оценить неизвестные теоретические коэффициенты  $\beta_0, \beta_1, \dots, \beta_k$  теоретической модели множественной линейной регрессии на основе имеющейся выборки данных. К счастью, эти коэффициенты могут быть оценены с использованием метода наименьших квадратов.

### Модель множественной линейной регрессии

Множественная регрессионная модель:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i, \quad i=1, \dots, n, \quad (6.7)$$

где

- $Y_i$  –  $i$ -е наблюдение зависимой переменной;  $X_{1i}, X_{2i}, \dots, X_{ki}$  –  $i$ -е наблюдения каждого из  $k$  регрессоров и  $u_i$  – ошибка регрессии.
- Линия теоретической регрессии – отношение, которое выполняется между  $Y$  и  $X$ -ми в среднем в генеральной совокупности:  
 $E(Y|X_{1i} = x_1, X_{2i} = x_2, \dots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$ .
- $\beta_1$  – коэффициент наклона при переменной  $X_1$ ,  $\beta_2$  – коэффициент наклона при переменной  $X_2$  и так далее. Коэффициент  $\beta_1$  отражает ожидаемое изменение  $Y_i$ , являющееся результатом изменения  $X_{1i}$  на одну единицу, считая постоянными  $X_{2i}, \dots, X_{ki}$ . Коэффициенты при других  $X$  интерпретируются похожим образом.
- Свободный член  $\beta_0$  – это ожидаемое значение  $Y$  при условии, что все  $X$  равны нулю. Свободный член может быть представлен как коэффициент при регрессоре  $X_{0i}$ , который равен единице для всех  $i$ .

## ОСНОВНЫЕ ПОНЯТИЯ

6.2

### 6.3. МНК-оценка множественной линейной регрессии

В данном разделе мы описываем, как коэффициенты модели множественной линейной регрессии могут быть оценены с использованием МНК.

#### МНК-оценка

В разделе 4.2 показано, как оценить свободный член и коэффициент наклона в модели парной регрессии, применяя МНК к выборке наблюдений  $Y$  и  $X$ . Ключевая идея состоит в том, что эти коэффициенты могут быть оценены с помощью минимизации суммы квадратов остатков, то есть подбирая  $b_0$  и  $b_1$  так, что минимизируется  $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$ . Оценки, которые удовлетворяют этому условию, – это МНК-оценки коэффициентов  $\beta_0$  и  $\beta_1$ .

МНК также может быть использован для оценки коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$  в модели множественной линейной регрессии. Пусть  $b_0, b_1, \dots, b_k$  – оценки коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$ . Предсказанное значение  $\hat{Y}_i$ , рассчитанное с использованием этих оценок, имеет вид:  $b_0 + b_1 X_{1i} + \dots + b_k X_{ki}$ , и ошибка в предсказании  $\hat{Y}_i$  равна  $\hat{Y}_i - (b_0 + b_1 X_{1i} + \dots + b_k X_{ki}) = Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki}$ . Сумма квадратов ошибок предсказания по всем  $n$  наблюдениям равна:

$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2. \quad (6.8)$$

Сумма квадратов ошибок для линейной регрессионной модели в выражении (6.8) – расширение суммы квадратов ошибок, приведенной в уравнении (4.6) для парной линейной регрессионной модели.

Оценки коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$ , которые минимизируют сумму квадратов ошибок в выражении (6.8), называются *оценками наименьших квадратов* (МНК-оценками) коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$ . МНК-оценки обозначаются  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ .

Терминология МНК в модели множественной линейной регрессии та же самая, что и в парной модели линейной регрессии. *Линия МНК-регрессии* – это прямая, построенная с использованием МНК-оценок:  $\hat{Y}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k$ . Предсказанное значение  $\hat{Y}_i$  при  $X_{1i}, X_{2i}, \dots, X_{ki}$ , полученной по МНК-регрессии, имеет вид:  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki}$ . *МНК-остатки* для  $i$ -го наблюдения – это разность между  $Y_i$  и его *предсказанным по МНК значением*; то есть МНК-остатки равны  $\hat{u}_i = Y_i - \hat{Y}_i$ .

МНК-оценки могут быть вычислены методом проб и ошибок, то есть подбором различных значений  $b_0, b_1, \dots, b_k$  до тех пор, пока вы не убедитесь в том, что минимизировали полную сумму квадратов в выражении (6.8). Гораздо проще, однако, использовать явные формулы для оценок, которые получены с использованием методов математического анализа. Формулы для оценок в модели множественной регрессии аналогичны тем, которые приведены во вставке «Основные понятия 4.2» для модели парной регрессии. Эти формулы используются во всех современных статистических пакетах. В модели множественной

регрессии эти формулы лучше всего можно выразить и объяснить с помощью матричных обозначений, поэтому их представление мы откладываем до раздела 18.1.

Определение и терминология метода наименьших квадратов во множественной регрессии приведены во вставке «Основные понятия 6.3».

### МНК-оценки, предсказанные значения и остатки в модели множественной регрессии

МНК-оценки коэффициентов регрессии  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  – это значения  $b_0, b_1, \dots, b_k$ , минимизирующие сумму квадратов предсказанных ошибок  $\sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2$ . Предсказанные по МНК значения  $\hat{Y}_i$  и остатков  $\hat{u}_i$  имеют вид:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki}, \quad i=1, \dots, n, \text{ и} \quad (6.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i=1, \dots, n. \quad (6.10)$$

МНК-оценки коэффициентов  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  и остатки  $\hat{u}_i$  рассчитываются на основе выборки из  $n$  наблюдений  $(X_{1i}, X_{2i}, \dots, X_{ki}, Y_i)$ ,  $i=1, \dots, n$ . Это оценки неизвестных истинных коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$  и ошибки регрессии  $u_i$ .

### ОСНОВНЫЕ ПОНЯТИЯ

6.3

### Пример: результаты тестов и соотношение учеников и учителей

В разделе 4.2 мы использовали МНК для оценки свободного члена и коэффициента наклона регрессии для результатов тестов (*TestScore*) и соотношения учеников и учителей (*STR*) на основе выборки из 420 наблюдений для школьных округов Калифорнии; оцененная МНК-регрессия, приведенная в выражении (4.11), имеет вид:

$$\widehat{\text{TestScore}} = 698,9 - 2,28 \times \text{STR}. \quad (6.11)$$

Мы подозревали, что полученные оценки не соответствуют фактическому положению дел, так как соотношение учеников и учителей может включать в себя эффект наличия большого числа школьников, изучающих английский язык в районах с большими классами. То есть вполне возможно, что МНК-оценка смещена вследствие пропущенных переменных.

Сейчас мы можем решить эту проблему с помощью МНК и оценить множественную регрессию, в которую включены зависимая переменная – результаты тестов  $\hat{Y}_i$ , – и есть два регрессора: соотношение учеников и учителей  $X_{1i}$  и процент изучающих английский в школьном округе  $X_{2i}$  для 420 округов ( $i=1, \dots, 420$ ). Оцененная линия множественной МНК-регрессии имеет вид:

$$\widehat{\text{TestScore}} = 686,0 - 1,10 \times \text{STR} - 0,65 \times \text{PctEL}, \quad (6.12)$$

где  $PctEL$  – процент школьников, которые изучают английский язык, в школьном округе. МНК-оценка свободного члена ( $\hat{\beta}_0$ ) равна 686,0, МНК-оценка коэффициента при соотношении учеников и учителей ( $\hat{\beta}_1$ ) равна –1,10 и МНК-оценка коэффициента при проценте изучающих английский язык школьников ( $\hat{\beta}_2$ ) равна –0,65.

Оцененное влияние на результаты тестов изменения в соотношении учеников и учителей во множественной регрессии примерно вдвое меньше, чем в случае парной регрессии: в уравнении с одним регрессором [уравнение (6.11)] единичное снижение  $STR$  по оценкам увеличит результат тестов на 2,28 пункта, но во множественной регрессии [уравнение (6.12)] предполагается увеличение результатов тестов только на 1,10 пункта. Это различие возникает из-за того, что коэффициент при  $STR$  во множественной регрессии является следствием изменения в  $STR$ , фиксируя в качестве постоянной (или контролируемой) переменную  $PctEL$ , в то время как в парной регрессии регрессор  $PctEL$  не является постоянным.

Эти две оценки согласуются друг с другом, если мы сделаем вывод о наличии смещения из-за пропущенной переменной в оценках модели парной регрессии в уравнении (6.11). В разделе 6.1 мы видели, что школьные округа с высоким процентом изучающих английский язык, как правило, имеют не только низкие результаты тестов, но и высокое соотношение учеников и учителей. Если доля изучающих английский язык исключается из регрессии, сокращение соотношения учеников и учителей в классах, по оценкам, будет иметь большое влияние на результаты тестов, но эта оценка отражает как влияние изменения в самом соотношении учеников и учителей, так и пропущенный эффект от наличия меньшего числа школьников, изучающих английский язык в конкретном школьном округе.

Мы пришли к выводу о наличии смещения из-за пропущенной переменной в соотношении между результатами тестов и числом учеников в расчете на одного учителя двумя разными путями: при помощи сравнения данных о результатах тестов в различных группах (раздел 6.1) и при помощи оценки множественной регрессии [выражение (6.12)]. Из этих двух методов множественная регрессия имеет два важных преимущества. Во-первых, она дает количественную оценку влияния снижения соотношения учеников и учителей на единицу, что является ключевым фактором при принятии решения окружным школьным инспектором. Во-вторых, она легко расширяется до случая более чем двух регрессоров, так что множественная регрессия может быть использована для оценки влияния изменений в соотношении учеников и учителей при условии постоянства не только процента изучающих английский язык школьников, но и при условии постоянства других важных факторов.

Оставшаяся часть этой главы посвящена обсуждению МНК в приложении к модели множественной линейной регрессии. Многое из того, что мы узнали об МНК-оценке парной регрессии, переносится на множественную регрессию с малыми изменениями или вообще без изменений, поэтому мы сосредоточим внимание на том, что является новым в случае множественной регрессии. Начнем с обсуждения качества приближения данных моделью множественной линейной регрессии.

## 6.4. Качество приближения данных моделью множественной линейной регрессии

Три широко используемые статистики в модели множественной линейной регрессии – это стандартная ошибка регрессии (*SER*), коэффициент детерминации регрессии  $R^2$  и скорректированный коэффициент детерминации регрессии  $\bar{R}^2$  (также известный как  $\bar{R}^2$ ). Все три статистики оценивают, насколько хорошо МНК-оценка линии множественной регрессии описывает, или «подгоняет» (приближает), данные.

### Стандартная ошибка регрессии (*SER*)

Стандартная ошибка регрессии (*SER*) оценивает стандартное отклонение ошибки регрессии  $u_i$ . Таким образом, *SER* является мерой разброса переменной  $Y$  вокруг линии регрессии. В модели множественной регрессии *SER* имеет вид:

$$SER = s_{\hat{u}},$$

$$\text{где } s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-k-1} \quad (6.13)$$

и где  $SSR$  – сумма квадратов остатков,  $SSR = \sum_{i=1}^n \hat{u}_i^2$ .

Единственное отличие между определением из выражения (6.13) и определением *SER* в разделе 4.3 для модели парной регрессии состоит в том, что в нашем случае делитель равен  $n-k-1$ , а не  $n-2$ . В разделе 4.3 делитель  $n-2$  (а не  $n$ ) корректирует занижение оценки, которое имеет место, так как мы оцениваем два коэффициента (коэффициент наклона и свободный член регрессии). Таким образом, делитель  $n-k-1$  корректирует занижение оценки вследствие необходимости оценивать  $k+1$  коэффициент ( $k$  – коэффициент наклона и свободный член). Как и в разделе 4.3, использование  $n-k-1$  вместо  $n$  называется корректировкой на степени свободы. Если есть только один регрессор, то  $k=1$ , и формула в разделе 4.3 идентична уравнению (6.13). При больших  $n$  влияние корректировки на степени свободы незначительно.

### Коэффициент детерминации $R^2$

$R^2$  регрессии – это доля выборочной дисперсии  $Y_i$ , объясненная (или предсказанная) регрессорами. Это определение равносильно тому, что  $R^2$  равен разности между единицей и долей дисперсии  $Y_i$ , не объясненной регрессорами.

Математическое определение  $R^2$  такое же, как для случая парной регрессии:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \quad (6.14)$$

где объясненная сумма квадратов равна  $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$  и полная сумма квадратов равна  $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$ .

В множественной регрессии  $R^2$  увеличивается всякий раз, когда мы добавляем регрессоры, если только оцененный коэффициент при добавленном регрессоре не равен нулю. Чтобы убедиться в этом, представьте, что мы начали с одного регрессора, а затем добавили второй. При использовании МНК для оценки модели с двумя регрессорами МНК находит значения коэффициентов, которые минимизируют сумму квадратов остатков. Если МНК выбирает коэффициент при новом регрессоре, равный нулю, то  $SSR$  будет такой же, независимо от того, включена ли вторая переменная в регрессию. Но если МНК выбирает любое значение, отличное от нуля, то это значение должно уменьшить  $SSR$  по отношению к регрессии, которая исключает этот регрессор. На практике точное равенство нулю оцененного коэффициента встречается крайне редко, так что в целом  $SSR$  будет уменьшаться, когда мы добавляем новый регрессор. Но это означает, что  $R^2$  обычно увеличивается (и никогда не уменьшается), когда мы добавляем новый регрессор.

### **Скорректированный коэффициент детерминации $R^2$**

Поскольку  $R^2$  увеличивается при добавлении в регрессию новой переменной, увеличение  $R^2$  не означает, что включение этой переменной действительно улучшает качество подгонки модели. В этом смысле  $R^2$  дает завышенную оценку того, насколько хорошо регрессия объясняет данные. Один из способов исправить это – уменьшить  $R^2$  на некоторый фактор – получил название скорректированного коэффициента детерминации  $R^2$ , или  $\bar{R}^2$ .

Скорректированный  $R^2$ , или  $\bar{R}^2$ , представляет собой модифицированную версию  $R^2$ , который не обязательно увеличивается при добавлении нового регрессора.  $\bar{R}^2$  имеет вид:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS} = 1 - \frac{s_u^2}{s_Y^2}. \quad (6.15)$$

Разница между этой формулой и вторым определением  $R^2$  в выражении (6.14) состоит в том, что отношение суммы квадратов остатков к общей сумме квадратов умножается на множитель  $(n-1)/(n-k-1)$ . Как показывает второе выражение в уравнении (6.15), это означает, что скорректированный  $R^2$  равен разности между единицей и отношением выборочной дисперсии МНК-остатков [с корректировкой на степени свободы в уравнении (6.13)] к выборочной дисперсии  $Y$ .

Существуют три вещи, которые необходимо знать о коэффициенте детерминации  $\bar{R}^2$ . Во-первых,  $(n-1)/(n-k-1)$  всегда больше 1, так что  $\bar{R}^2$  всегда меньше, чем  $R^2$ .

Во-вторых, включение в модель дополнительного регрессора вызывает два противоположных эффекта в  $\bar{R}^2$ . С одной стороны,  $SSR$  уменьшается, что увеличивает  $\bar{R}^2$ . С другой стороны, множитель  $(n-1)/(n-k-1)$  увеличивается. Увеличится или уменьшится  $\bar{R}^2$  – зависит от того, какой из эффектов сильнее.

В-третьих,  $\bar{R}^2$  может быть отрицательным. Это происходит, когда все регрессоры, вместе взятые, уменьшают сумму квадратов остатков на такую маленькую величину, что это сокращение не в состоянии компенсировать влияние фактора  $(n - 1)/(n - k - 1)$ .

### **Пример: результаты тестов и соотношение учеников и учителей**

Уравнение (6.12) представляет собой оцененную линию множественной регрессии результатов тестов (*TestScore*) в зависимости от соотношения учеников и учителей (*STR*) и процента изучающих английский язык школьников (*PctEL*). Коэффициент детерминации  $R^2$  для этой линии регрессии равен  $R^2 = 0,426$ , скорректированный коэффициент детерминации  $\bar{R}^2$  равен  $\bar{R}^2 = 0,424$ , а стандартная ошибка регрессии составляет  $SER = 14,5$ .

Сравнивая эти характеристики качества приближения данных с аналогичными для регрессии, в которой *PctEL* исключена [уравнение (6.11)], видим, что включение *PctEL* в регрессию увеличивает  $R^2$  с 0,051 до 0,426. Когда единственный регрессор — *STR*, только небольшая часть изменений в *TestScore* объяснена; однако когда к регрессии добавляется *PctEL*, объясняется более двух пятых (42,6 %) от вариации оценок по тестам. В этом смысле включение доли изучающих английский язык существенно улучшает качество подгонки регрессии. Так как  $n$  велико и только два регрессора появляются в уравнении (6.12), разница между коэффициентом детерминации  $R^2$  и скорректированным коэффициентом детерминации  $\bar{R}^2$  очень мала —  $R^2 = 0,426$  по сравнению с  $\bar{R}^2 = 0,424$ .

*SER* для регрессии без учета *PctEL* составляет 18,6; это значение падает до 14,5, когда переменная *PctEL* включена в качестве второго регрессора. *SER* измеряется в тех же единицах измерения, что и зависимая переменная. Снижение *SER* говорит нам, что предсказываемые при помощи второй регрессии (с двумя объясняющими переменными) результаты тестов существенно более точны, чем если бы это было сделано с помощью оценок парной регрессии.

**Использование  $R^2$  и скорректированного  $R^2$ .** Скорректированный коэффициент детерминации  $\bar{R}^2$  является полезным, поскольку он количественно выражает информацию о том, в какой степени учитываются регрессоры или объясняется дисперсия зависимой переменной. Тем не менее сильное доверие  $\bar{R}^2$  (или  $R^2$ ) может привести в западню. На практике «максимизация  $\bar{R}^2$ » редко является ответом на любой экономически и статистически содержательный вопрос. Вместо этого решение о том, включать ли переменную во множественную регрессию, должно быть основано на том, позволяет ли включение переменной лучше оценить причинно-следственную связь. Мы вернемся к решению вопроса о том, какие переменные включать в регрессию, а какие не стоит, в главе 7. А для начала мы должны разработать методы количественной оценки выборочной неопределенности МНК-оценки. Отправной точкой для этого является расширение предположений метода наименьших квадратов из главы 4 для случая множественной линейной регрессии.

## 6.5. Предположения метода наименьших квадратов для множественной линейной регрессии

В модели множественной линейной регрессии присутствуют четыре предположения метода наименьших квадратов. Первые три – те же, что и в разделе 4.3 для модели парной линейной регрессии (вставка «Основные понятия 4.3»), расширены для случая нескольких объясняющих переменных и обсуждаются лишь кратко. Четвертое предположение является новым и поэтому обсуждается более детально.

### **Предположение № 1: условное распределение $u_i$ , относительно $X_{1i}, X_{2i}, \dots, X_{ki}$ имеет нулевое среднее**

Первое предположение метода наименьших квадратов состоит в том, что условное распределение  $u_i$  относительно  $X_{1i}, X_{2i}, \dots, X_{ki}$  имеет среднее, равное нулю. Эта предпосылка расширяет первое предположение метода наименьших квадратов для модели парной регрессии на случай нескольких объясняющих переменных. Это предположение означает, что иногда  $Y_i$  находится выше линии теоретической регрессии, иногда ниже ее, но в среднем по генеральной совокупности  $Y_i$  попадает на линию теоретической регрессии. Таким образом, для любых значений регрессоров ожидаемое значение  $u_i$  равно нулю. Как и в случае парной регрессии, это предположение является ключевым и гарантирует несмещенност МНК-оценок. К обсуждению проблемы смещения оценок вследствие пропуска существенных переменных в модели множественной регрессии мы вернемся в разделе 7.5.

### **Предположение № 2: $X_{1i}, X_{2i}, \dots, X_{ki}, Y_i, i=1, \dots, n$ являются независимыми и одинаково распределенными (i.i.d.) случайными величинами**

Второе предположение метода наименьших квадратов для случая множественной линейной регрессии состоит в том, что  $X_{1i}, \dots, X_{ki}, Y_i, i=1, \dots, n$  – являются независимыми и одинаково распределенными (i.i.d.) случайными величинами. Это предположение выполняется автоматически, если данные собраны с помощью простого случайного отбора. Комментарии об этом предположении, появившиеся в разделе 4.3 для модели с одним регрессором, также применимы к модели с несколькими объясняющими переменными.

### **Предположение № 3: большие выбросы маловероятны**

Третье предположение метода наименьших квадратов состоит в том, что большие выбросы, то есть наблюдения со значениями, находящимися далеко за пределами обычного диапазона данных, маловероятны. Это предположение служит напоминанием о том, что, как и в случае парной регрессии, МНК-оценки коэффициентов в модели множественной регрессии могут быть чувствительны к большим выбросам.

Предпосылка о том, что большие выбросы маловероятны, становится математически точной, если мы предположим, что  $X_{1i}, \dots, X_{ki}$ , и  $Y_i$  имеют ненулевые

и конечные четвертые моменты:  $0 < E(X_{li}^4) < \infty$ , ...,  $0 < E(X_{ki}^4) < \infty$  и  $0 < E(Y_i^4) < \infty$ . Можно переформулировать это предположение так: зависимая переменная и регрессоры имеют конечный эксцесс. Заметим, что это предположение используется для доказательства свойств выборочных статистик МНК-оценок регрессии в больших выборках.

### **Предположение № 4: отсутствие совершенной мультиколлинеарности**

Четвертое предположение является новым для модели множественной регрессии. Оно исключает неудобные ситуации, называемые совершенной мультиколлинеарностью, при возникновении которых невозможно вычислить МНК-оценки. Согласно ему, объясняющие переменные являются *совершенно мультиколлинеарными*, если один из регрессоров является точной линейной комбинацией (функцией) других регрессоров. Четвертое предположение наименьших квадратов состоит в том, что регрессоры не являются совершенно мультиколлинеарными.

Почему совершенная мультиколлинеарность делает невозможным вычисление МНК-оценок? Предположим, вы хотите оценить коэффициент при  $STR$  в регрессии  $TestScore_i$  на  $STR_i$  и  $PctEL_p$ , но вы ошибаетесь и случайно вводите  $STR_i$  во второй раз вместо  $PctEL_i$ , то есть вы пытаетесь оценить регрессию  $TestScore_i$  на  $STR_i$  и на  $STR_i$ . Это как раз случай совершенной мультиколлинеарности, потому что один из регрессоров (первое включение  $STR$ ) является линейной функцией от другого регрессора (второе включение  $STR$ ). В зависимости от того, как программный пакет обрабатывает совершенную мультиколлинеарность, если вы попытаетесь оценить эту регрессию, эконометрический пакет будет делать одно из двух: либо он уберет одну из включенных переменных  $STR$ , либо он не станет рассчитывать МНК-оценки и выдаст сообщение об ошибке. Математическая причина этого состоит в том, что совершенная мультиколлинеарность приводит к делению на ноль в формулах МНК-оценок коэффициентов.

На интуитивном уровне совершенная мультиколлинеарность является проблемой, потому что вы хотите найти ответ на нелогичный вопрос. Во множественной регрессии коэффициент при одном из регрессоров отражает величину изменения зависимой переменной как следствие изменений в рассматриваемом регрессоре при постоянстве других регрессоров. В гипотетической регрессии  $TestScore$  на  $STR$  и  $STR$  коэффициент при первом включении  $STR$  характеризует влияние на результаты тестов при изменении в  $STR$  при условии постоянства второй объясняющей переменной – тоже  $STR$ . Это не имеет смысла, и МНК не может оценить этот бесполезный частный эффект.

Чтобы решить проблему совершенной мультиколлинеарности в этой гипотетической регрессии, нужно просто исправить опечатку и заменить одно из включений  $STR$  переменной, которую вы изначально хотели включить. Этот пример типичен: когда совершенная мультиколлинеарность имеет место, она часто отражает логическую ошибку в выборе регрессоров или некоторых ранее неизвестных особенностей набора данных. В общем случае решением

проблемы совершенной мультиколлинеарности является изменение набора объясняющих переменных в регрессии.

Дополнительные примеры совершенной мультиколлинеарности мы рассмотрим в разделе 6.7, в котором также определим и обсудим проблему несовершенной мультиколлинеарности.

Все предположения метода наименьших квадратов для модели множественной линейной регрессии обобщены во вставке «Основные понятия 6.4».

## ОСНОВНЫЕ ПОНЯТИЯ 6.4

### Предположения метода наименьших квадратов в модели множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i, \quad i = 1, \dots, n,$$

где

1.  $u_i$  имеет нулевое условное среднее относительно  $X_{1i}, X_{2i}, \dots, X_{ki}$ ; то есть  $E(u_i | X_{1i}, X_{2i}, \dots, X_{ki}) = 0$ .
2.  $(X_{1i}, X_{2i}, \dots, X_{ki}, Y_i)$ ,  $i = 1, \dots, n$ , – независимые и одинаково распределенные (i.i.d.) случайные величины из их совместного распределения.
3. Большие выбросы маловероятны:  $X_{1i}, \dots, X_{ki}$  и  $Y_i$  имеют не-нулевые конечные четвертые моменты.
4. В модели отсутствует совершенная мультиколлинеарность.

## 6.6. Распределение МНК-оценок в модели множественной линейной регрессии

Поскольку данные, на основе которых мы получаем оценки модели, могут отличаться, в различных выборках мы получим различные МНК-оценки. Такая изменчивость оценок по всем возможным выборкам приводит к (выборочной) неопределенности, связанной с МНК-оценками коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$  в теоретической регрессии. Так же, как и в случае парной регрессии, эта изменчивость отражается в выборочном распределении МНК-оценок.

Мы помним из раздела 4.4, что при выполнении предположений метода наименьших квадратов МНК-оценки ( $\hat{\beta}_0$  и  $\hat{\beta}_1$ ) являются несмещенными и состоятельными оценками неизвестных коэффициентов ( $\beta_0$  и  $\beta_1$ ) в модели парной линейной регрессии. Кроме того, в больших выборках выборочное распределение  $\hat{\beta}_0$  и  $\hat{\beta}_1$  хорошо приближается двумерным нормальным распределением.

Эти результаты переносятся на случай модели множественной линейной регрессии. То есть, по предположениям метода наименьших квадратов из вставки «Основные понятия 6.4», МНК-оценки  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  являются несмещенными и состоятельными оценками коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$  в модели линейной множественной регрессии. В больших выборках совместное выборочное распределение  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  хорошо приближается многомерным нормальным рас-

пределением, которое является расширением двумерного нормального распределения до случая двух или более совместно распределенных нормальных случайных величин (раздел 2.4).

### Распределение $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ в больших выборках

Если выполняются все предположения метода наименьших квадратов для модели множественной линейной регрессии (вставка «Основные понятия 6.4»), то в больших выборках МНК-оценки  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  коэффициентов модели совместно нормально распределены, и оценка любого коэффициента ( $\hat{\beta}_j$ ) распределена как  $N(\beta_j, \sigma_{\hat{\beta}_j}^2)$ ,  $j = 0, \dots, k$ .

## ОСНОВНЫЕ ПОНЯТИЯ

6.5

Несмотря на то что алгебраические преобразования в случае модели множественной регрессии будут более сложными, центральная предельная теорема применима к МНК-оценкам множественной регрессионной модели по той же причине, по какой она применяется к  $\bar{Y}$  и МНК-оценкам модели парной регрессии: МНК-оценки  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  являются средними значениями оценок, полученных по различным случайным выборкам, и при достаточно большом размере выборки выборочное распределение этих средних становится нормальным. Так как многомерное нормальное распределение компактнее всего записывается математически при помощи обозначений матричной алгебры, запись выражений для совместного распределения МНК-оценок в явном виде откладывается до главы 18.

Во вставке «Основные понятия 6.5» обобщен результат о том, что в больших выборках совместное распределение МНК-оценок множественной регрессии хорошо приближается совместным нормальным распределением. В большинстве случаев МНК-оценки коррелированы; эта корреляция вытекает из корреляции между регрессорами. Совместное выборочное распределение МНК-оценок будет обсуждаться более подробно для случая модели с двумя объясняющими переменными и в предположении гомоскедастичности ошибок регрессии в приложении 6.2, а общий случай будет рассмотрен в разделе 18.2.

## 6.7. Мультиколлинеарность

Как отмечалось в разделе 6.5, совершенная мультиколлинеарность возникает, если один из регрессоров является точной линейной комбинацией других объясняющих переменных. В этом разделе приведены некоторые примеры совершенной мультиколлинеарности и обсуждается, как совершенная мультиколлинеарность может возникнуть и как ее можно избежать в регрессии с несколькими бинарными объясняющими переменными. Несовершенная мультиколлинеарность возникает, если один из регрессоров очень сильно, но не полностью, коррелирован с другими регрессорами. В отличие от совершенной мультиколлинеарности, несовершенная мультиколлинеарность не мешает

оценке регрессии и не подразумевает логическую проблему с выбором регрессоров. Тем не менее это означает, что один или несколько коэффициентов регрессии могут быть оценены неточно.

### **Примеры совершенной мультиколлинеарности**

Мы продолжаем обсуждение проблемы совершенной мультиколлинеарности из раздела 6.5, рассматривая три дополнительные гипотетические регрессии. В каждой из них третий регрессор добавляется к регрессии  $TestScore_i$  на  $STR_i$  и  $PctEL_i$  в уравнении (6.12).

**Пример № 1: Доля изучающих английский язык школьников.** Пусть переменная  $FracEL_i$  означает долю изучающих английский язык школьников в  $i$ -ом школьном округе и принимает значения между нулем и единицей. Если бы переменная  $FracEL_i$  была включена в качестве третьего регрессора в дополнение к  $STR_i$  и  $PctEL_i$ , то регрессоры были бы совершенно мультиколлинеарны. Причина этого состоит в том, что  $PctEL_i$  характеризует процент изучающих английский язык школьников, так что  $PctEL_i = 100 \times FracEL_i$  для каждого района. Таким образом, один из регрессоров ( $PctEL_i$ ) можно записать в виде точной линейной функции от другого регрессора ( $FracEL_i$ ).

Из-за этой совершенной мультиколлинеарности невозможно вычислить МНК-оценки регрессии  $TestScore_i$  на  $STR_i$ ,  $PctEL_i$  и  $FracEL_i$ . На интуитивном уровне понятно, что МНК не работает потому, что вы задаете вопрос о степени влияния единичного изменения в проценте изучающих английский язык при постоянной доле изучающих английский язык. Поскольку процент изучающих английский язык и доля изучающих английский менятся вместе согласно точной линейной зависимости, этот вопрос не имеет никакого смысла, и МНК не может на него ответить.

**Пример № 2: «Не очень маленькие» классы.** Пусть  $NVS_i$  – бинарная переменная, равная единице, если соотношение учеников и учителей в  $i$ -ом районе «не очень мало», а именно  $NVS_i$  равна единице, если  $STR_i \geq 12$  и равна нулю в противном случае. Эта регрессия также испытывает совершенную мультиколлинеарность, но линейная зависимость между регрессорами в этом случае не так очевидна, как в предыдущем примере. На самом деле в нашей выборке отсутствуют школьные округа с  $STR_i < 12$ ; как можно увидеть на рисунке 4.2, наименьшее значение  $STR$  равно 14. Таким образом,  $NVS_i = 1$  для всех наблюдений. Теперь вспомним, что линейные регрессионные модели со свободным членом могут эквивалентно рассматриваться как модели, в которые включен регрессор  $X_{0i}$ , равный единице для всех  $i$ , как показано в уравнении (6.6). Таким образом, мы можем записать  $NVS_i = 1 \times X_{0i}$  для всех наблюдений в нашей выборке, то есть регрессор  $NVS_i$  может быть записан как точная линейная комбинация регрессоров; в частности, он равен  $X_{0i}$ .

Этот пример иллюстрирует два важных момента, касающихся совершенной мультиколлинеарности. Во-первых, когда регрессия включает в себя свободный член, то один из регрессоров, который может быть причиной совершенной мультиколлинеарности, это постоянный регрессор –  $X_{0i}$ . Во-вторых, совершен-

ная мультиколлинеарность – это некоторое утверждение о выборке, которая находится в вашем распоряжении. И хотя вполне возможно представить себе школьный округ со школами, в которых на одного учителя приходится менее 12 учеников, в нашей выборке таких районов нет, так что мы не можем включать соответствующую переменную в нашу регрессию.

**Пример № 3: Процент школьников, говорящих на английском языке.** Пусть  $PctES$  – процент учеников, «говорящих на английском языке» в  $i$ -ом школьном округе – определяется как процент школьников, которые не изучают английский язык. И снова регрессоры будут совершенно мультиколлинеарны. Как и в предыдущем примере, точное линейное соотношение среди регрессоров включает константу  $X_{0i}$ : для любого округа  $PctEL_i = 100 \times X_{0i} - PctEL_i$ .

Рассмотренный пример является иллюстрацией другой идеи: совершенная мультиколлинеарность – свойство полного набора регрессоров. Если свободный член (который является регрессором  $X_{0i}$ ) или  $PctEL_i$  были бы исключены из этой регрессии, объясняющие переменные не были бы совершенно мультиколлинеарны.

**Ловушка фиктивных переменных.** Еще один возможный источник совершенной мультиколлинеарности появляется, когда несколько бинарных или фиктивных переменных используются в качестве регрессоров. Например, предположим, что вы разделили школьные округа на три категории: сельские, пригородные и городские. Каждый округ попадает в одну (и только одну) категорию. Пусть эти бинарные переменные будут  $Rural_i$ , которая равна единице для сельских районов и нулю в противном случае,  $Suburban_i$  и  $Urban_i$ . Если включить все три бинарные переменные в регрессию наряду с константой, регрессоры будут совершенно мультиколлинеарны, так как каждый район принадлежит к одной и только одной категории,  $Rural_i + Suburban_i + Urban_i = 1 = X_{0i}$ , где  $X_{0i}$  обозначает постоянный регрессор, представленный в выражении (6.6). Таким образом, для оценки регрессии необходимо исключить одну из этих четырех переменных, неважно, фиктивную переменную или свободный член. По соглашению, постоянный член сохраняется, но тогда одна из бинарных переменных исключается. Например, если бы переменная  $Rural_i$  была исключена, то коэффициент при  $Suburban_i$  отражал бы среднюю разность между результатами тестов в пригородных и сельских районах при условии постоянства других переменных в регрессии.

В общем случае, если имеется  $G$  фиктивных переменных, если каждое наблюдение попадает в одну и только одну категорию, если в регрессии есть свободный член и если все  $G$  бинарных переменных включены в качестве регрессоров, то в регрессии будет наблюдаться совершенная мультиколлинеарность. Эта ситуация называется *ловушкой фиктивных переменных*. Обычный способ избежать ловушки фиктивных переменных – это исключить одну из бинарных переменных из множественной регрессии так, чтобы только  $G-1$  переменные из  $G$  фиктивных переменных остались включенными в модель в качестве регрессоров. В этом случае коэффициенты при включенных бинарных переменных представляют собой дополнительный эффект от присутствия наблюдений в рассматриваемой группе по отношению к базовому случаю пропущенной

группы и при постоянстве других регрессоров. В качестве альтернативы все  $G$  фиктивных переменных могут быть включены в регрессию, если свободный член из нее исключен.

**Решение проблемы совершенной мультиколлинеарности.** Совершенная мультиколлинеарность обычно возникает, когда ошибка сделана в спецификации регрессии. Иногда ошибку легко заметить (как в первом примере), но иногда это не так (как во втором примере). В той или иной форме используемый программный пакет даст вам знать, если сделана подобная ошибка, потому что он не сможет вычислить МНК-оценки, если такие ошибки имеют место.

Если такое случится, необходимо изменить спецификацию регрессии для устранения совершенной мультиколлинеарности. Некоторые программные пакеты не являются надежными при наличии совершенной мультиколлинеарности в модели регрессии в том смысле, что сами принимают решение о выборе конкретных переменных, которые остаются в модели при наличии совершенной мультиколлинеарности.

### ***Несовершенная мультиколлинеарность***

Несмотря на похожие названия, несовершенная мультиколлинеарность концептуально отличается от совершенной мультиколлинеарности. *Несовершенная мультиколлинеарность* означает, что два или более регрессора тесно связаны в том смысле, что существует линейная комбинация объясняющих переменных, сильно коррелированная с каким-либо из оставшихся регрессоров. Несовершенная мультиколлинеарность не представляет никакой теоретической проблемы для МНК-оценок, да и цель МНК – разобраться в независимом влиянии различных регрессоров, когда они потенциально коррелированы.

Тем не менее если объясняющие переменные являются несовершенно мультиколлинеарными, то, по крайней мере, при одном из регрессоров будет неточно оценен коэффициент. Например, рассмотрим регрессию *TestScore* на *STR* и *PctEL*. Предположим, что мы должны были добавить третий регрессор, процент жителей района, которые являются иммигрантами в первом поколении. Первое поколение иммигрантов часто говорит на английском языке как на втором языке, так что переменные *PctEL* и процент иммигрантов будут тесно связаны: школьные округа, в которых проживает большое количество недавних иммигрантов, будут иметь тенденцию к наличию в них большого числа школьников, которые все еще изучают английский язык. Так как эти две переменные тесно связаны, было бы трудно использовать их для оценки частного влияния увеличения *PctEL* на результаты тестов, считая постоянной долю иммигрантов. Иными словами, в выборке содержится мало информации о том, что происходит с результатами тестов, если процент изучающих английский низок, но доля иммигрантов высока, или наоборот. Если предположения метода наименьших квадратов выполнены, то МНК-оценка коэффициента при *PctEL* в этой регрессии будет несмещенной; однако она будет иметь большую дисперсию, чем если бы регрессоры *PctEL* и процент иммигрантов были некоррелированы.

Влияние несовершенной мультиколлинеарности на дисперсию МНК-оценок можно вывести математически из уравнения (6.17), что сделано в приложении 6.2, для дисперсии  $\hat{\beta}_1$  во множественной регрессии с двумя регрессорами ( $X_1$  и  $X_2$ ) для специального случая гомоскедастичной ошибки. В этом случае дисперсия  $\hat{\beta}_1$  является обратно пропорциональной  $1 - \rho_{X_1, X_2}^2$ , где  $\rho_{X_1, X_2}^2$  – коэффициент корреляции между  $X_1$  и  $X_2$ . Чем больше корреляция между двумя регрессорами, тем ближе этот член к нулю и тем больше дисперсия  $\hat{\beta}_1$ . В общем случае, когда несколько объясняющих переменных являются несовершенно мультиколлинеарными, коэффициенты при одном или более регрессоре будут оценены неточно, то есть они будут иметь большую выборочную дисперсию.

Совершенная мультиколлинеарность является проблемой, которая часто свидетельствует о наличии логической ошибки. Несовершенная мультиколлинеарность, наоборот, необязательно является ошибкой, а просто особенностью МНК, имеющейся выборки и вопроса, на который вы пытаетесь ответить. Если переменные в регрессии являются теми, которые вы хотите включить – например теми, которые вы выбрали для устранения смещения из-за пропущенных переменных, – то наличие несовершенной мультиколлинеарности среди них предполагает, что будет трудно точно оценить один или несколько частных эффектов с помощью имеющихся данных.

## 6.8. Заключение

Модель парной регрессии очень уязвима из-за проблемы смещения вследствие наличия в ней пропущенной переменной: если пропущенная переменная является определяющей для зависимой переменной и коррелирует с регрессорами, то МНК-оценка коэффициента наклона будет смещенной и будет отражать как влияние включенного регрессора, так и влияние пропущенной переменной. Множественная регрессия позволяет уменьшить смещение из-за пропущенной переменной при помощи ее включения переменной в регрессию. Коэффициент при регрессоре  $X_1$  множественной регрессии характеризует частный эффект, который оказывает изменение объясняющей переменной  $X_1$  на зависимую переменную в предположении постоянства остальных регрессоров. В примере с результатами тестов включение доли изучающих английский язык школьников в качестве регрессора позволило оценить влияние на результаты тестов, которое оказывает изменение соотношения учеников и учителей, считая постоянной долю изучающих английский язык школьников. Это сократило наполовину оцененное влияние на результаты тестов изменений в соотношении учеников и учителей.

С формальной (математической) точки зрения модель множественной линейной регрессии является расширением модели парной линейной регрессии. Предположения метода наименьших квадратов для модели множественной линейной регрессии являются расширениями трех предположений метода наименьших квадратов для модели парной линейной регрессии, и к ним добавляется четвертое предположение, исключающее совершенную мультиколлинеарность. Так как коэффициенты регрессии оцениваются с использованием одной выборки, МНК-

оценки имеют совместное выборочное распределение и, естественно, выборочную неопределенность. Эта выборочная неопределенность должна быть количественно определена в рамках эмпирического исследования. Методы количественной оценки выборочной неопределенности будут рассмотрены в следующей главе.

## **Выводы**

1. Смещение пропущенных переменных возникает, если пропущенная переменная (1) коррелирована с включенным в модель регрессором и (2) влияет на зависимую переменную  $Y$ .
2. Модель множественной регрессии – линейная модель, которая включает несколько объясняющих переменных  $X_1, X_2, \dots, X_k$ . Каждый регрессор связан с соответствующим регрессионным коэффициентом  $\beta_1, \beta_2, \dots, \beta_k$ . Коэффициент  $\beta_1$  характеризует ожидаемое изменение зависимой переменной  $Y$ , связанное с единичным изменением объясняющей переменной  $X_1$  при условии постоянства остальных регрессоров.
3. Коэффициенты множественной регрессии можно оценить с помощью МНК. При выполнении четырех предположений метода наименьших квадратов, перечисленных во вставке «Основные понятия 6.4», МНК-оценки модели множественной линейной регрессии являются несмещенными, состоятельными и асимптотически нормально распределенными.
4. Совершенная мультиколлинеарность, которая возникает в ситуации, когда один регрессор является точной линейной комбинацией других регрессоров, как правило, возникает из-за ошибки в выборе регрессоров, включенных в модель множественной регрессии. Для решения проблемы совершенной мультиколлинеарности необходимо изменить набор включенных в модель регрессоров.
5. Стандартная ошибка регрессии, коэффициент детерминации  $R^2$  и скорректированный коэффициент детерминации  $\bar{R}^2$  являются мерами качества приближения данных моделью множественной линейной регрессии.

## **Основные понятия**

Смещение из-за пропущенной переменной (с. 184).

Модель множественной линейной регрессии (с. 190).

Линия теоретической регрессии (с. 191).

Функция теоретической регрессии (с. 191).

Свободный член (с. 191).

Угловой коэффициент при  $X_{1i}$  (с. 191).

Коэффициент при  $X_{1i}$  (с. 191).

Угловой коэффициент при  $X_{2i}$  (с. 191).

Коэффициент при  $X_{2i}$  (с. 191).

Считая  $X_2$  постоянной (с. 191).

Контролируя  $X_2$  (с. 191).

Частный эффект (с. 192).

- Теоретическая модель множественной регрессии (с. 192).  
Постоянный регрессор (с. 192).  
Постоянный член (константа) (с. 192).  
Гомоскедастичный (с. 193).  
Гетероскедастичный (с. 193).  
Оценка обычного метода наименьших квадратов (МНК)  $\beta_1, \beta_2, \dots, \beta_k$  (с. 194).  
Линия МНК-регрессии (с. 194).  
Предсказанное по модели значение (с. 194).  
МНК-остатки (с. 194).  
Коэффициент детерминации  $R^2$  (с. 197).  
Скорректированный коэффициент детерминации  $R^2$  ( $\bar{R}^2$ ) (с. 198).  
Совершенная мультиколлинеарность (с. 201).  
Ловушка фиктивных переменных (с. 205).  
Несовершенная мультиколлинеарность (с. 206).

### **Вопросы для повторения и закрепления основных понятий**

- 6.1. Исследователя интересует, как использование компьютеров в школах влияет на результаты тестов. Используя данные по школьным округам, похожие на те, что были использованы в данной главе, она оценивает регрессию зависимости среднего балла по тесту в округе от количества компьютеров, приходящихся на одного учащегося. Будет ли  $\hat{\beta}_1$  несмещенной оценкой влияния на результаты тестов увеличения числа компьютеров на одного ученика? Почему да или почему нет? Если вы считаете, что оценка  $\hat{\beta}_1$  смещена, будет ли это смещение переоценкой или недооценкой коэффициента? Почему?
- 6.2. Множественная регрессия включает в себя два регрессора:  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ . Каково ожидаемое изменение  $Y$ , если  $X_1$  увеличивается на 3 единицы и  $X_2$  остается неизменной? Каково ожидаемое изменение  $Y$ , если  $X_2$  уменьшается на 5 единиц и  $X_1$  остается неизменным? Каково ожидаемое изменение  $Y$ , если  $X_1$  увеличивается на 3 единицы и  $X_2$  уменьшается на 5 единиц?
- 6.3. Объясните, почему две совершенно мультиколлинеарные объясняющие переменные не могут быть включены в модель линейной множественной регрессии. Приведите два примера пары совершенно мультиколлинеарных регрессоров.
- 6.4. Объясните, почему трудно точно оценить частный эффект влияния  $X_1$  на зависимую переменную, считая регрессор  $X_2$  постоянным, если  $X_1$  и  $X_2$  тесно взаимосвязаны.

### **Упражнения**

Первые четыре упражнения относятся к таблице оценок регрессии на стр. 211, рассчитанных с использованием данных за 1998 год из текущего обследования населения (CPS). Набор данных состоит из информации по 4000 работникам,

занятым полный рабочий день в течение всего года. В рассматриваемой выборке присутствуют респонденты, имеющие либо аттестат средней школы, либо степень бакалавра. Возраст работника колеблется от 25 до 34 лет. В данных также содержится информация о регионе проживания респондентов, их семейном положении и количестве детей. Для целей этого упражнения обозначим:

*AHE* = средняя почасовая заработка (в долларах 1998 года).

*College* = бинарная переменная (1 – если респондент закончил колледж, 0 – если среднюю школу).

*Female* = бинарная переменная (1 – если респондент является женщиной, 0 – если мужчиной).

*Age* = возраст (в годах).

*Northeast* = бинарная переменная (1 – если респондент проживает на Северо-Востоке, 0 – в противном случае).

*Midwest* = бинарная переменная (1 – если респондент проживает на Среднем Западе, 0 – в противном случае).

*South* = бинарная переменная (1 – если респондент проживает на Юге, 0 – в противном случае).

*West* = бинарная переменная (1 – если регион респондент проживает на Западе, 0 – в противном случае).

- 6.1. Рассчитайте  $\bar{R}^2$  для каждой из регрессий.
- 6.2. Используя результаты оценки регрессии в столбце (1), ответьте на следующие вопросы:
  - а) Зарабатывают ли респонденты, имеющие высшее образование, в среднем больше, чем респонденты только со школьным образованием? Насколько?
  - б) Зарабатывают ли мужчины в среднем больше женщин? Насколько?
- 6.3. Используя результаты регрессии в столбце (2), ответьте на следующие вопросы:
  - а) Является ли возраст фактором, серьезно влияющим на доходы? Объясните.
  - б) Пусть Салли – 29-летняя женщина, выпускница колледжа, а Бетси – 34-летняя женщина, выпускница колледжа. Рассчитайте доходы Салли и Бетси на основе оценок этой регрессии.
- 6.4. Используя результаты регрессии в столбце (3), ответьте на следующие вопросы:
  - а) Имеют ли место региональные различия?
  - б) Почему регрессор *West* исключен из регрессии? Что бы произошло, если бы он был включен?
  - в) Пусть Хуанита – 28-летняя женщина, выпускница колледжа с Юга, а Дженифер – 28-летняя женщина, выпускница колледжа на Среднем Западе. Рассчитайте ожидаемую разницу в доходах между Хуанитой и Дженифер.

**Результаты оценки регрессий среднечасовой зарплаты на бинарные переменные, характеризующие пол, образование и другие характеристики, с использованием данных за 1998 год из базы данных текущего обследования населения США**

Зависимая переменная: средняя почасовая зарплата ( <i>AHE</i> )			
Регрессор	(1)	(2)	(3)
<i>College</i> ( $X_1$ )	5,46	5,48	5,44
<i>Female</i> ( $X_2$ )	-2,64	-2,62	-2,62
<i>Age</i> ( $X_3$ )		0,29	0,29
<i>Northeast</i> ( $X_4$ )			0,69
<i>Midwest</i> ( $X_5$ )			0,60
<i>South</i> ( $X_6$ )			-0,27
Константа	12,69	4,40	3,75
Итоговая статистика			
<i>SER</i>	6,27	6,22	6,21
$R^2$	0,176	0,190	0,194
$\bar{R}^2$			
<i>n</i>	4000	4000	4000

- 6.5. Рассмотрим данные случайной выборки о продажах 220 домов в некотором районе в 2003 году. Пусть *Price* обозначает отпускную цену (в тыс. долл.), *BDR* – количество спален в доме, *Bath* – число ванных комнат, *Hsize* – площадь дома (в квадратных футах), *Lsize* – площадь земельного участка (в квадратных футах), *Age* – возраст дома (в годах) и *Poor* – бинарная переменная, равная единице, если состояние дома считается «плохим». Оцененная регрессия имеет вид:

$$\widehat{\text{Price}} = 119,2 + 0,485 \text{BDR} + 23,4 \text{Bath} + 0,156 \text{Hsize} + 0,002 \text{Lsize} + 0,090 \text{Age} - 48,8 \text{Poor}, \bar{R}^2 = 0,72, \text{SER} = 41,5.$$

- Предположим, что хозяйка дома переделывает часть существующей гостиной в доме в новую ванную. Каково ожидаемое увеличение стоимости дома?
- Предположим, что хозяйка дома добавляет новую ванную, которая увеличивает площадь дома на 100 квадратных футов. Каково ожидаемое увеличение стоимости дома?

- в) Какова оцениваемая потеря стоимости дома, если домовладелец позволит дому состарится настолько, что его состояние оценится как «плохое»?
- г) Вычислите  $R^2$  оцененной регрессии.
- 6.6. Исследователь планирует изучить причинно-следственную связь между числом полицейских и преступностью в округе, используя данные из случайной выборки по округам США. Он планирует построить регрессию уровня преступности в округе в зависимости от числа полицейских (на душу населения) в округе.
- а) Объясните, почему в оценках этой регрессии, скорее всего, будет присутствовать смещение из-за пропущенной переменной. Какие переменные вы бы добавили в регрессию для контроля важных пропущенных переменных?
- б) Используйте результаты пункта (а) и выражение для смещения из-за пропущенной переменной в уравнении (6.1), чтобы определить, будет ли в регрессии пере- или недооценено влияние числа полицейских на уровень преступности. (Иными словами, считаете ли вы, что  $\hat{\beta}_1 > \beta_1$  или  $\hat{\beta}_1 < \beta_1$ ?)
- 6.7. Рассмотрите критически каждый из предлагаемых ниже планов исследования. Опишите любые проблемы, которые могут возникнуть при проведении исследования, а также способы улучшения имеющихся планов исследования. Обсудите, какие дополнительные данные нужно собрать и соответствующие статистические методы, которые можно использовать для анализа данных.
- а) Исследователь хочет выяснить, существует ли дискриминация по половому признаку при установлении заработной платы в крупной аэрокосмической фирме. Чтобы определить потенциальный гендерный разрыв в зарплатах, исследователь собирает данные о зарплате и поле всех инженеров фирмы. Затем исследователь планирует провести тест на равенство средних значений для проверки гипотезы о том, что средняя заработка женщины существенно меньше, чем средняя зарплата мужчин.
- б) Исследователь хочет выяснить, существует ли постоянное влияние наличия у человека срока тюремного заключения на ставку его заработной платы. Он собирает данные из случайной выборки людей, которые вышли из тюрьмы по крайней мере 15 лет назад. Он также собирает аналогичные данные по случайной выборке людей, которые никогда не сидели в тюрьме. Собранные данные включают информацию о текущей заработной плате каждого человека, образовании, возрасте, этнической принадлежности, поле, стаже (время работы в текущей области), роде занятий и статусе в профсоюзе, а также информацию о том, сидел ли этот человек когда-либо в тюрьме или нет. Исследователь планирует оценить влияние факта лишения свободы на заработную плату, оценив регрессию заработной платы на фиктивную переменную, равную единице, если человек сидел в тюрьме,

и нулю – в противном случае, а также на другие факторы, потенциально влияющие на размер заработной платы (образование, стаж, статус в профсоюзе и так далее).

- 6.8. Недавнее исследование показало, что уровень смертности среди людей, которые спят по 6–7 часов в сутки, ниже, чем смертность среди людей, спящих по 8 и более часов. 1,1 млн наблюдений, используемых в данном исследовании, были получены путем случайного опроса американцев в возрасте от 30 до 102 лет. Каждый респондент наблюдался в течение четырех лет. Смертность людей, спящих по 7 часов в сутки, рассчитывалась как отношение числа смертей среди спящих по 7 часов людей, произошедших в течение периода исследования, к общему числу респондентов, спящих по 7 часов в сутки. Аналогичные расчеты были сделаны для людей, спящих по 6 часов в сутки, и так далее. Основываясь на полученном результате, могли бы вы рекомендовать американцам, спящим по 9 часов в сутки, сократить свой сон до 6 или 7 часов, если они хотят увеличить продолжительность своей жизни? Почему да или почему нет? Объясните.
- 6.9. Предположим, что  $(Y_i, X_{1i}, X_{2i})$  удовлетворяют предположениям из вставки «Основные понятия 6.4». Вы заинтересованы в получении оценки коэффициента  $\beta_1$ , отражающего влияние переменной  $X_1$  на  $Y$ . Предположим, что  $X_1$  и  $X_2$  некоррелированы. Вы оцениваете коэффициент  $\beta_1$  в регрессии  $Y$  на  $X_1$  (переменная  $X_2$  не входит в регрессию). Значит ли это, что полученная оценка будет смещена из-за пропущенной переменной? Объясните.
- 6.10. Предположим, что  $(Y_i, X_{1i}, X_{2i})$  удовлетворяют предположениям из вставки «Основные понятия 6.4» и, кроме того,  $\text{var}(u_i | X_{1i}, X_{2i}) = 4$  и  $\text{var}(X_{1i}) = 6$ . Мы рассматриваем случайную выборку из генеральной совокупности, состоящую из  $n=400$  наблюдений.
- Предположим, что  $X_1$  и  $X_2$  независимы. Вычислите дисперсию  $\hat{\beta}_1$ . [Подсказка: воспользуйтесь уравнением (6.17) из приложения 6.2.]
  - Предположим, что  $\text{cor}(X_1, X_2) = 0,5$ . Вычислите дисперсию  $\hat{\beta}_1$ .
  - Прокомментируйте следующие высказывания: «Если  $X_1$  и  $X_2$  коррелированы, дисперсия  $\hat{\beta}_1$  больше, чем она была бы, если бы  $X_1$  и  $X_2$  были некоррелированы. Поэтому  $X_2$  лучше не включать в регрессию, если она коррелирована с  $X_1$ ».
- 6.11. (Требует использования техники математического анализа.) Рассмотрим модель регрессии:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

для  $i=1, \dots, n$ . (Обратите внимание на то, что в регрессии отсутствует свободный член.) Проведите анализ, аналогичный тому, что был сделан в приложении 4.2:

- Запишите функцию суммы квадратов остатков модели (целевую функцию), которая минимизируется с помощью МНК.
- Вычислите частные производные целевой функции из предыдущего пункта по  $b_1$  и  $b_2$ .

- в) Пусть  $\sum_{i=1}^n X_{1i}X_{2i} = 0$ . Покажите, что  $\hat{\beta}_1 = \frac{\sum_{i=1}^n X_{1i}Y_i}{\sum_{i=1}^n X_{1i}^2}$ .
- г) Пусть  $\sum_{i=1}^n X_{1i}X_{2i} \neq 0$ . Выведите выражение для  $\hat{\beta}_1$  в зависимости от  $(Y_i, X_{1i}, X_{2i})$ ,  $i=1, \dots, n$ .
- д) Предположим, что модель включает свободный член:  

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$
. Покажите, что оценка наименьших квадратов удовлетворяет выражению  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ .
- е) Как и в пункте (д), предположим, что модель содержит свободный член. Также предположим, что  $\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0$ . Покажите, что  $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}$ . Как это выражение соотносится с МНК-оценкой коэффициента  $\beta_1$  из регрессии, в которой пропущена переменная  $X_2$ ?

## Компьютерные упражнения

- E6.1. Используя базу данных **TeachingRatings**, описанную в компьютерном упражнении E4.2, выполните следующие задания:
- а) Оцените регрессию *Course\_Eval* от *Beauty*. Чему равен оцененный коэффициент наклона?
  - б) Оцените регрессию *Course\_Eval* от *Beauty*, включив некоторые дополнительные переменные для контроля характеристик курса и профессора. В частности, включите в качестве дополнительных регрессоров переменные *Intro*, *OneCredit*, *Female*, *Minority* и *NNEnglish*. Каково оцененное влияние индекса «красоты» профессора на *Course\_Eval*? Подвержена ли регрессия из пункта (а) смещению из-за пропущенных переменных?
  - в) Рассчитайте коэффициент при индексе «красоты» профессора для модели множественной регрессии в (б), используя трехшаговую процедуру, описанную в приложении 6.3 (теорема Фриша–Во (Frisch–Waugh)). Убедитесь, что эта трехшаговая процедура дает тот же коэффициент при индексе «красоты», как и полученный в пункте (б).
  - г) Профессор Смит – чернокожий мужчина со средним индексом «красоты», являющийся носителем английского языка. Курс, который он преподает, является курсом промежуточного уровня (upper-division) и оценивается тремя кредитными баллами. Рассчитайте предсказанное моделью значение оценки курса профессора Смита.

E6.2. Используя базу данных **CollegeDistance**, описанную в компьютерном упражнении E4.3, выполните следующие задания:

- a) Оцените регрессию числа полных лет обучения (*ED*) от расстояния до ближайшего колледжа (*Dist*). Чему равен оцененный коэффициент наклона?
- б) Оцените регрессию *ED* на *Dist*, включив дополнительные переменные для контроля за характеристиками студентов и их семей, а также местного рынка труда. В частности, включите в качестве дополнительных регрессоров переменные *Bytest*, *Female*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *Cue80*, и *Swmfg80*. Каково оцененное влияние *Dist* на *ED*?
- в) Отличается ли существенно предполагаемое влияние *Dist* на *ED* в регрессии из пункта (б) от предполагаемого влияния *Dist* на *ED* в регрессии из пункта (а)? Основываясь на этом выводе, что вы можете сказать о наличии смещения из-за пропущенной переменной оценок регрессии из пункта (а)?
- г) Сравните качество приближения данных регрессиями из пунктов (а) и (б), используя стандартную ошибку регрессии, коэффициент детерминации  $R^2$  и скорректированный коэффициент детерминации  $\bar{R}^2$ . Почему  $R^2$  и  $\bar{R}^2$  так близки в регрессии (б)?
- д) Значение коэффициента при переменной *DadColl* положительно. Что показывает этот коэффициент?
- е) Объясните, почему переменные *Cue80* и *Swmfg80* появились в регрессии. Считаете ли вы, что знаки оцененных коэффициентов (+ или –) при этих переменных отражают действительность? Дайте интерпретацию величине этих коэффициентов.
- ж) Боб – чернокожий мужчина. Средняя школа, в которой он учился, располагалась в 20 милях от ближайшего колледжа. Его сводная выпускная оценка<sup>1</sup> (*Bytest*) составила 58 баллов. Доход его семьи в 1980 году составил 26 тыс. долл., и семья владела домом. Его мама посещала колледж, а отец – нет. Уровень безработицы в округе составлял 7,5 %, а средняя почасовая заработка плата – 9,75 долл. Рассчитайте предсказанное моделью число полных лет обучения Боба с использованием регрессии из пункта (б).
- з) Джим имеет те же характеристики, что и Боб, за исключением того, что его школа располагалась в 40 милях от ближайшего колледжа. Рассчитайте предсказанное моделью число полных лет обучения Джима с использованием регрессии из пункта (б).

E6.3. Используя базу данных **Growth**, описанную в компьютерном упражнении E4.3, выполните следующие задания, предварительно исключив из нее данные по Мальте.

<sup>1</sup> Base-year composite test score – сводная выпускная оценка в США, рассчитанная на основе результатов выпускных экзаменов из средней школы по нескольким предметам, аналогичным российским ЕГЭ. – Примеч. науч. ред. перевода.

- a) Рассчитайте выборочные статистики показателей *Growth*, *TradeShare*, *YearsSchool*, *Oil*, *Rev\_Coups*, *Assassinations*, *RGDP60*. Постройте таблицу, содержащую их выборочные средние, стандартные отклонения, а также минимальные и максимальные значения (не забудьте записать их единицы измерения).
- б) Оцените регрессию *Growth* на *TradeShare*, *YearsSchool*, *Rev\_Coups*, *Assassinations* и *RGDP60*. Чему равно значение коэффициента при *Rev\_Coups*? Дайте интерпретацию значению этого коэффициента. Является ли он большим или маленьким?
- в) Используйте оцененную регрессию, чтобы предсказать среднегодовые темпы роста для страны, у которой все характеристики равны своему выборочному среднему.
- г) Повторите (в), но теперь предположите, что значение показателя *TradeShare* для этой страны на одно стандартное отклонение больше средневыборочного.
- д) Почему переменная *Oil* исключена из регрессии? Что произошло бы, если бы она была включена в регрессию?

## Приложения

### Приложение 6.1. Вывод уравнения (6.1)

В данном приложении выводится формула для смещения, возникающего из-за пропущенной переменной в уравнении (6.1). Из уравнения (4.30) приложения 4.3 следует:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) u_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (6.16)$$

Тогда из двух последних предположений из вставки «Основные понятия 4.3» получаем, что  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \xrightarrow{p} \sigma_X^2$  и  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 u_i \xrightarrow{p} \text{cov}(u_i, X_i) = = \rho_{Xu} \sigma_u \sigma_X$ . Подставляя эти пределы в выражение (6.16), получаем уравнение (6.1).

### Приложение 6.2. Распределение МНК-оценок в случае двух регрессоров и гомоскедастичных ошибок

Общая формула для дисперсии МНК-оценок во множественной регрессии является довольно громоздкой, но для случая двух регрессоров ( $k=2$ ) и гомо-

скедастичной ошибки регрессии формула упрощается достаточно для того, чтобы дать некоторое представление о распределении МНК-оценок.

Так как ошибки регрессии гомоскедастичны, условная дисперсия  $u_i$  может быть записана в виде  $\text{var}(u_i | X_{1i}, X_{2i}) = \sigma_u^2$ . При наличии двух регрессоров,  $X_{1i}$  и  $X_{2i}$ , и гомоскедастичной ошибки регрессии в больших выборках выборочное распределение  $\hat{\beta}_1$  является нормальным —  $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$ , где дисперсия этого распределения  $\sigma_{\hat{\beta}_1}^2$  равна:

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \left( \frac{1}{1 - \rho_{X_1, X_2}^2} \right) \frac{\sigma_u^2}{\sigma_{X_1}^2}, \quad (6.17)$$

где  $\rho_{X_1, X_2}$  — теоретический коэффициент корреляции между двумя регрессорами  $X_1$  и  $X_2$  и  $\sigma_{X_1}^2$  — дисперсия  $X_{1i}$  в генеральной совокупности.

Дисперсия  $\sigma_{\hat{\beta}_1}^2$  выборочного распределения зависит от квадрата коэффициента корреляции между регрессорами. Если  $X_1$  и  $X_2$  сильно коррелированы, положительно или отрицательно, то  $\rho_{X_1, X_2}^2$  близка к единице и, следовательно, член  $1 - \rho_{X_1, X_2}^2$  в знаменателе выражения (6.17) мал, и тогда дисперсия  $\sigma_{\hat{\beta}_1}^2$  больше, чем она была бы, если бы  $\rho_{X_1, X_2}$  был близок к нулю.

Еще одной особенностью асимптотического совместного нормального распределения с МНК-оценки является то, что  $\hat{\beta}_1$  и  $\hat{\beta}_2$  в общем случае коррелированы. Когда ошибки гомоскедастичны, корреляция между МНК-оценками  $\hat{\beta}_1$  и  $\hat{\beta}_2$  равна коэффициенту корреляции между двумя регрессорами, взятому со знаком минус:

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1, X_2}. \quad (6.18)$$

### Приложение 6.3. Теорема Фриша–Во

МНК-оценки множественной регрессии можно получить, последовательно оценивая некоторый набор регрессий. Рассмотрим модель множественной регрессии, представленную выражением (6.7). МНК-оценка коэффициента  $\beta_1$  может быть вычислена в три этапа:

1. Оцените регрессию  $X_1$  от  $X_2, X_3, \dots, X_k$ . Пусть  $\tilde{X}_1$  обозначает остатки этой регрессии.
2. Оцените регрессию  $Y$  от  $X_2, X_3, \dots, X_k$ , и пусть  $\tilde{Y}$  обозначает остатки этой регрессии.
3. Оцените регрессию  $\tilde{Y}$  от  $\tilde{X}_1$ , где все регрессии включают свободный член. Теорема Фриша–Во (Frisch–Waugh) утверждает, что МНК-оценка коэффициента в пункте 3 равна МНК-оценке коэффициента при  $X_1$  в модели множественной регрессии (6.7).

Данный результат представляет собой математическое утверждение о том, что коэффициент множественной регрессии  $\hat{\beta}_1$  является оценкой влияния переменной  $X_1$  на переменную  $Y$ , контролируя все остальные объясняющие переменные: так как, оценив первые две регрессии (шаги 1 и 2), мы очистили  $Y$  и  $X_1$  от влияния других  $X$ -ов, третья регрессия дает оценку влияния очищенной

переменной  $X_1$  на очищенную переменную  $Y$ . Теорему Фриша–Во предлагаются доказать в упражнении 18.17.

Теорема Фриша–Во показывает, как выражение (6.17) может быть получено из уравнения (5.27). Так как  $\hat{\beta}_1$  является МНК-оценкой коэффициента регрессии  $\tilde{Y}$  от  $\tilde{X}_1$ , уравнение (5.27) предполагает, что дисперсия  $\hat{\beta}_1$  для случая гомоскедастичных ошибок регрессии имеет вид  $\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_u^2}{n\sigma_{\tilde{X}_1}^2}$ , где  $\sigma_{\tilde{X}_1}^2$  – дисперсия  $\tilde{X}_1$ .

Так как  $\tilde{X}_1$  – остатки регрессии  $X_1$  на  $X_2$  (вспомните, что выражение (6.17) относится к модели с  $k=2$  регрессорами), выражение (6.15) подразумевает, что  $s_{\tilde{X}_1}^2 = (1 - \bar{R}_{X_1, X_2}^2)s_{X_1}^2$ , где  $\bar{R}_{X_1, X_2}^2$  – скорректированный коэффициент детерминации  $R^2$  из регрессии  $X_1$  на  $X_2$ . Тогда выражение (6.17) следует из того, что  $s_{\tilde{X}_1}^2 \xrightarrow{p} \sigma_{\tilde{X}_1}^2$ ,  $\bar{R}_{X_1, X_2}^2 \xrightarrow{p} R_{X_1, X_2}^2$  и  $s_{X_1}^2 \xrightarrow{p} \sigma_{X_1}^2$ .

# **Глава 7. Множественная линейная регрессия: проверка гипотез и доверительные интервалы**

Как обсуждалось в главе 6, модель множественной линейной регрессии дает возможность уменьшить смещение оценок, возникающее вследствие пропущенных переменных: включая дополнительные регрессоры, мы, тем самым, учитываем влияние этих дополнительных регрессоров на зависимую переменную. Коэффициенты модели множественной регрессии можно оценить при помощи МНК. Как и все методы оценивания, МНК-оценки обладают свойством неопределенности в конечных выборках, так как значения МНК-оценок отличаются от одной выборки к другой.

В данной главе рассказывается о методах измерения неопределенности МНК-оценок в конечных выборках при помощи стандартных ошибок, проверки статистических гипотез и построения доверительных интервалов. Одна новая возможность, которая появляется во множественных регрессиях, — это проверка совместных гипотез, которые одновременно содержат предположения о двух или более коэффициентах регрессии. Общий подход к проверке таких «совместных» гипотез связан с использованием  $F$ -статистики.

Раздел 7.1 расширяет статистические методы, используемые в парных линейных регрессиях, на случай множественной линейной регрессии. В разделах 7.2 и 7.3 описаны методы проверки гипотез для двух или более коэффициентов регрессии. Раздел 7.4 расширяет понятие доверительных интервалов для одного коэффициента до понятия доверительной области для нескольких коэффициентов. Решение о том, какие переменные включать в регрессию, является важным практическим вопросом, поэтому в разделе 7.5 обсуждаются возможные подходы к решению этой задачи. В разделе 7.6 мы применяем модель множественной линейной регрессии к анализу данных по результатам тестов в Калифорнии, чтобы получить улучшенные оценки влияния соотношения учеников и учителей на результаты тестов.

## **7.1. Проверка гипотез и доверительные интервалы для одного коэффициента**

В этом разделе мы обсудим, как можно вычислить стандартную ошибку, проверить гипотезы и построить доверительные интервалы для одного коэффициента в модели множественной линейной регрессии.

## Стандартные ошибки МНК-оценок

Напомним, что в случае модели парной линейной регрессии дисперсию МНК-оценки коэффициента можно было оценить, подставив выборочные средние вместо математических ожиданий, что приводило к оценке  $\hat{\sigma}_{\hat{\beta}_1}^2$ , представленной в выражении (5.4). В условиях предположений метода наименьших квадратов и из закона больших чисел следует, что эти выборочные средние сходятся к соответствующим генеральным средним, и поэтому, например,  $\hat{\sigma}_{\hat{\beta}_1}^2 / \sigma_{\hat{\beta}_1}^2 \xrightarrow{P} 1$ . Квадратный корень из  $\hat{\sigma}_{\hat{\beta}_1}^2$  – это стандартная ошибка оценки коэффициента  $\hat{\beta}_1$ ,  $SE(\hat{\beta}_1)$ , являющаяся оценкой стандартного отклонения выборочного распределения  $\hat{\beta}_1$ .

Вышесказанное естественным образом расширяется на случай множественной регрессии. У МНК-оценки  $j$ -го коэффициента регрессии ( $\hat{\beta}_j$ ) существует стандартное отклонение, и оценкой этого стандартного отклонения является стандартная ошибка  $SE(\hat{\beta}_j)$ . Формулу для стандартной ошибки легче всего записать в матричных обозначениях (см. раздел 18.2). Важным моментом здесь является то, что в случае стандартных ошибок нет концептуальных различий между случаями парной и множественной регрессий. Ключевые идеи – асимптотическая нормальность оценок и возможность состоятельно оценивать стандартное отклонение их выборочного распределения – остаются неизменными вне зависимости от того, сколько у нас регрессоров: один, два или 12.

## Проверка гипотез относительно одного коэффициента

Предположим, вы хотите проверить гипотезу о том, что изменение соотношения учеников и учителей не влияет на результаты тестов в школьном округе, если процент изучающих английский язык остается неизменным в этом округе. Такая постановка вопроса соответствует гипотезе о том, что истинный коэффициент  $\beta_1$  при соотношении учеников и учителей равен нулю в теоретической регрессии, характеризующей зависимость результатов тестов от  $STR$  и  $PctEL$ . В общем случае нам нужно проверить гипотезу о том, что истинный коэффициент  $\beta_j$  при  $j$ -м регрессоре принимает определенное значение  $\beta_{j,0}$ . Значение  $\beta_{j,0}$  следует либо из экономической теории, либо, как в случае с соотношением учеников и учителей, из контекста эмпирической задачи. Если альтернативная гипотеза является двухсторонней, то обе гипотезы можно формально записать так:

$$H_0 : \beta_j = \beta_{j,0} \text{ против } H_1 : \beta_j \neq \beta_{j,0} \quad (\text{двуихсторонняя альтернатива}). \quad (7.1)$$

Например, если первым регрессором является переменная  $STR$ , тогда нулевая гипотеза о том, что изменение соотношения учеников и учителей не влияет на результаты тестов, соответствует нулевой гипотезе  $\beta_1 = 0$  (т.е.  $\beta_{1,0} = 0$ ). Наша задача заключается в том, чтобы проверить нулевую гипотезу  $H_0$  против альтернативы  $H_1$ , используя выборку данных.

**Проверка нулевой гипотезы  $\beta_j = \beta_{j,0}$   
против альтернативной гипотезы  $\beta_j \neq \beta_{j,0}$**

1. Вычислите стандартную ошибку  $\beta_j$ ,  $SE(\hat{\beta}_j)$ .
2. Вычислите  $t$ -статистику:

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}. \quad (7.2)$$

3. Вычислите  $p$ -значение:

$$p\text{-value} = 2\Phi(-|t^{act}|), \quad (7.3)$$

где  $t^{act}$  равно значению вычисленной в пункте 2  $t$ -статистики. Нулевая гипотеза отвергается на 5%-м уровне значимости, если  $p$ -значение меньше 0,05 или, что эквивалентно,  $|t^{act}| > 1,96$ . Стандартная ошибка и, как правило,  $t$ -статистика и  $p$ -значение вычисляются автоматически в эконометрических программных пакетах.

**ОСНОВНЫЕ ПОНЯТИЯ**

7.1

Во вставке «Основные понятия 5.2» описана процедура проверки этой нулевой гипотезы для случая парной линейной регрессии. На первом шаге этой процедуры необходимо вычислить стандартную ошибку оценки коэффициента при регрессоре. На втором шаге —  $t$ -статистику, используя общую формулу из вставки «Основные понятия 5.1». На третьем шаге необходимо вычислить соответствующее  $p$ -значение, используя распределение нормальной величины из таблицы 1 приложения, или в качестве альтернативы сравнить рассчитанную  $t$ -статистику с критическим значением, соответствующим желаемому уровню статистической значимости теста. Теоретическое обоснование этого метода заключается в том, что МНК-оценка является асимптотически нормальной и в условиях нулевой гипотезы имеет математическое ожидание, равное предполагаемому истинному значению, а дисперсия этого распределения может быть состоятельно оценена.

Такое же теоретическое обоснование остается в силе и для случая множественной регрессии. Как было сформулировано во вставке «Основные понятия 6.5», выборочное распределение  $\hat{\beta}_j$  хорошо приближается нормальным распределением. Если нулевая гипотеза верна, математическое ожидание этого распределения равно  $\beta_{j,0}$ . Дисперсию этого распределения можно оценить состоятельно. Поэтому для того чтобы проверить нулевую гипотезу из выражения (7.1), мы можем просто использовать ту же самую процедуру, что и в случае парной линейной регрессии.

Процедура проверки гипотезы о равенстве одного коэффициенте конкретному значению в модели множественной линейной регрессии описана во вставке «Основные понятия 7.1». В тексте этой вставки фактически вычисленное значение  $t$ -статистики обозначено как  $t^{act}$ . Однако, как правило, это число просто обозначают как  $t$ , и в оставшейся части книги мы будем использовать это краткое обозначение.

## Доверительные интервалы для одного коэффициента

Метод построения доверительных интервалов во множественной линейной регрессии также не меняется по сравнению со случаем парной линейной регрессии. Этот метод изложен во вставке «Основные понятия 7.2».

### ОСНОВНЫЕ ПОНЯТИЯ 7.2

#### Доверительные интервалы для одного коэффициента в модели множественной линейной регрессии

95%-м двухсторонним доверительным интервалом называется интервал, содержащий истинное значение  $\beta_j$  с 95%-й вероятностью; то есть это интервал, содержащий истинное значение  $\beta_j$  для 95% всех возможных случайных выборок. Эквивалентно, двухсторонним доверительным интервалом называется множество всех значений  $\hat{\beta}_j$ , равенство которым не может быть отвергнуто при проверке двухсторонней гипотезы на 5%-м уровне значимости. Для выборки большого размера 95%-й доверительный интервал равен:

$$\beta_j = [\hat{\beta}_j - 1,96SE(\hat{\beta}_j); \hat{\beta}_j + 1,96SE(\hat{\beta}_j)]. \quad (7.4)$$

Чтобы получить 90%-й двухсторонний доверительный интервал, нужно заменить число 1,96 в выражении (7.4) на число 1,64.

Метод проверки гипотез, описанный во вставке «Основные понятия 7.1», и метод построения доверительных интервалов, описанный во вставке «Основные понятия 7.2», основываются на том, что для большой выборки распределение МНК-оценки коэффициента  $\hat{\beta}_j$  хорошо приближается нормальным распределением. Поэтому необходимо помнить, что эти методы измерения неопределенности выборок работают только для случая больших выборок.

### Пример: результаты тестов и соотношение учеников и учителей

Можем ли мы отклонить нулевую гипотезу о том, что изменение соотношения учеников и учителей не влияет на результаты тестов в школьном округе, если мы учитываем влияние показателя процента изучающих английский язык в этом округе на результаты тестов? Чему равен 95 %-й доверительный интервал для оценки того, как на результаты тестов влияет изменение соотношения учеников и учителей, если мы считаем неизменным процент школьников, изучающих английский язык? Теперь мы можем ответить на эти вопросы. МНК-оценки регрессии, характеризующей зависимость результатов тестов от соотношения учеников и учителей (*STR*) и процента изучающих английский язык школьников (*PctEL*), была рассмотрена в уравнении (6.12) и снова приводится здесь (в скобках даны стандартные ошибки коэффициентов регрессии):

$$\widehat{TestScore} = 686,0 - 1,10 \times STR - 0,650 \times PctEL. \quad (7.5)$$

(8.7)	(0.43)	(0.031)
-------	--------	---------

Чтобы проверить гипотезу о том, что истинный коэффициент перед  $STR$  равен нулю, нам сначала необходимо рассчитать значение  $t$ -статистики для соответствующего коэффициента из уравнения (7.2). Так как нулевая гипотеза состоит в том, что истинное значение этого коэффициента равно нулю, рассчитанное нами значение  $t$ -статистики будет равно:  $t = (-1,10 - 0) / 0,43 = -2,54$ . Соответствующее ей  $p$ -значение равно:  $2\Phi(-2,54) = 1,1\%$ . Поскольку  $p$ -значение меньше 5 %, нулевая гипотеза отвергается на 5 %-м уровне значимости (но не на 1 %-м уровне значимости).

Двусторонний 95 %-й доверительный интервал для коэффициента перед  $STR$  равен  $-1,10 \pm 1,96 \times 4,43 = (-1,95; -0,26)$ , то есть мы можем быть уверены на 95 %, что истинное значение этого коэффициента находится между  $-1,95$  и  $-0,26$ . Интерпретируя этот результат в контексте того, что окружной школьный инспектор хотела бы уменьшить соотношение учеников и учителей на 2, мы можем сказать, что 95 %-й доверительный интервал, количественно характеризующий влияние, которое окажет такое уменьшение на результаты тестов, равен  $(-1,95 \times 2; -0,26 \times 2) = (-3,90; -0,52)$ .

**Включение в уравнение расходов на ученика.** Полученные вами оценки множественной линейной регрессии (7.5) убедили окружного школьного инспектора в том, что, судя по накопленному к настоящему времени опыту, уменьшение размеров классов приведет к улучшению результатов тестов в ее школьном округе. Однако теперь перед ней стоит более тонкий вопрос. Если она хочет нанять большее число учителей, ей придется увеличить расходы на зарплату учителей. Это можно сделать либо за счет сокращения других статей бюджета (например, покупки новых компьютеров, расходов на содержание и т.д.), либо за счет увеличения бюджета ее школьного округа, что вряд ли понравится налогоплательщикам. В этой ситуации она задается вопросом: каким будет влияние на результаты тестов от уменьшения соотношения учеников и учителей при неизменных показателях расходов на одного ученика и процента изучающих английский язык школьников?

Чтобы ответить на этот вопрос, оценим регрессию зависимости результатов тестов от соотношения учеников и учителей, суммарные расходы на одного ученика и процент изучающих английский язык школьников. Линия МНК-регрессии имеет вид:

$$\widehat{TestScore} = 649,6 - 0,29 \times STR + 3,87 \times Expn - 0,656 \times PctEL, \quad (7.6)$$

где  $Expn$  – это суммарные годовые расходы на одного ученика в школьном округе, исчисляемые в тысячах долларов.

Результат удивителен. Исходя из полученных оценок, при неизменности показателей расходов на одного ученика и процента изучающих английский язык школьников изменение соотношения учеников и учителей оказывает небольшое влияние на результаты тестов: оцененный коэффициент перед  $STR$  равен  $-1,10$  в регрессии (7.5), но после того как регрессор  $Expn$  добавлен в уравнение (7.6), он равен всего лишь  $-0,29$ . Более того,  $t$ -статистика для проверки того, что истинным значением коэффициента является ноль, теперь принимает

значение  $t = (-0,29 - 0) / 0,48 = -0,60$ . Поэтому гипотеза о том, что этот коэффициент равен нулю, не может быть отвергнута даже на 10 %-ном уровне значимости ( $| -0,60 | < 1,645$ ). Таким образом, регрессия (7.6) не дает основания сделать вывод о том, что найм большего числа учителей улучшит результаты тестов, если суммарные расходы на одного ученика остаются неизменными.

Одним из объяснений результатов, представленных в регрессии (7.6), может быть то, что в используемых данных по Калифорнии школьные инспекторы распределяют бюджеты эффективно. Предположим, что коэффициент перед  $STR$  в регрессии (7.6) был отрицательным и большим. Тогда школьные округа могли бы повысить результаты тестов просто за счет выделения меньшего количества средств на другие цели (учебники, технологическое обеспечение, спорт и т.д.) и перераспределения этих денег для найма большего числа учителей. Размер классов уменьшился бы, в то время как расходы остались бы неизменными. Однако маленький и статистически незначимый коэффициент перед  $STR$  в регрессии (7.6) говорит о том, что такое перераспределение средств окажет незначительное влияние на результаты тестов. Другими словами, школьные округа и так уже распределяют свои ресурсы эффективно.

Обратите внимание на то, что стандартная ошибка для  $STR$  увеличилась с 0,43 в регрессии (7.5) до 0,48 в регрессии (7.6), когда мы добавили переменную  $Expt$ . Этот момент является иллюстрацией более общей ситуации, которую мы обсуждали в разделе 6.7 в контексте проблемы несовершенной мультиколлинеарности, говорящей о том, что наличие корреляции между регрессорами может сделать МНК-оценки менее точными (коэффициент корреляции между  $STR$  и  $Expt$  равен  $-0,62$ ).

И наконец, что мы можем сказать о нашем сердитом налогоплательщике? Он утверждает, что в генеральной совокупности оба коэффициента — и коэффициент перед соотношением учеников и учителей ( $\beta_1$ ), и коэффициент перед показателем расходов на одного ученика ( $\beta_2$ ) — равны нулю; то есть он выдвигает гипотезу о том, что  $\beta_1 = 0$  и  $\beta_2 = 0$ . Может показаться, что мы отвергнем эту гипотезу, поскольку значение соответствующей  $t$ -статистики для проверки гипотезы  $\beta_2 = 0$  в регрессии (7.6) равно  $t = 3,87 / 1,59 = 2,43$ , но мы должны понимать, что это рассуждение является ошибочным. Гипотеза налогоплательщика является совместной гипотезой, и для того чтобы ее проверить, нам необходим новый инструмент —  $F$ -статистика.

## 7.2. Проверка совместных гипотез

Данный раздел описывает, как нужно формулировать совместные гипотезы для нескольких коэффициентов множественной линейной регрессии и как проверять эти гипотезы при помощи  $F$ -статистики.

### Проверка гипотез для двух или более коэффициентов

**Совместная нулевая гипотеза.** Рассмотрим оценки (7.6) регрессии зависимости результатов тестов от соотношения учеников и учителей, расходов на од-

ного ученика и процента изучающих английский язык школьников. Наш сердитый налогоплательщик выдвигает гипотезу о том, что ни соотношение учеников и учителей, ни расходы на одного ученика не оказывают влияния на результаты тестов, если мы контролируем процент изучающих английский язык. Поскольку  $STR$  – первый регрессор в регрессии (7.6), а  $Expn$  – второй, мы можем формально записать эту гипотезу так:

$$H_0 : \beta_1 = 0 \text{ и } \beta_2 = 0 \text{ против } H_1 : \beta_1 \neq 0 \text{ и/или } \beta_2 \neq 0. \quad (7.7)$$

Гипотеза о том, что *как* коэффициент перед соотношением числа учеников и учителей ( $\beta_1$ ), *так и* коэффициент перед показателем расходов на одного ученика ( $\beta_2$ ) равны нулю – это пример совместной гипотезы о коэффициентах в модели множественной регрессии. В данном случае нулевая гипотеза ограничивает значения двух коэффициентов, и если использовать правильную терминологию, мы можем сказать, что сформулированная в выражении (7.7) нулевая гипотеза налагает два *ограничения* на коэффициенты модели множественной линейной регрессии:  $\beta_1 = 0$  и  $\beta_2 = 0$ .

В общем случае *совместной гипотезой* называется гипотеза, которая налагивает два или более ограничений на коэффициенты регрессии. Мы рассматриваем совместную нулевую и альтернативную гипотезы в виде:

$$\begin{aligned} H_0 : \beta_j &= \beta_{j,0} \quad \beta_m = \beta_{m,0} \dots, \text{ для всех ограничений} \\ \text{против } H_1 &: \text{одно или более из } q \text{ ограничений} \\ &\text{в условии гипотезы } H_0 \text{ не выполнены,} \end{aligned} \quad (7.8)$$

где  $\beta_j, \beta_m, \dots$  – это различные коэффициенты регрессии, а  $\beta_{j,0}, \beta_{m,0}, \dots$  – значения этих коэффициентов в условиях нулевой гипотезы. Нулевая гипотеза в выражении (7.7) является примером более общей нулевой гипотезы из выражения (7.8). Другим примером будет совместная нулевая гипотеза для регрессии с  $k=6$  регрессорами, состоящая в том, что коэффициенты перед 2, 4 и 5-м регрессорами равны нулю; то есть  $\beta_2 = 0, \beta_4 = 0$  и  $\beta_5 = 0$  – всего  $q=3$  ограничений. В общем случае нулевая гипотеза  $H_0$  содержит  $q$  таких ограничений.

Если хотя бы одно (или больше чем одно) из равенств в условии нулевой гипотезы  $H_0$  из (7.8) неверно, то и сама совместная нулевая гипотеза неверна. Поэтому альтернативная гипотеза и состоит в том, что хотя бы одно из равенств в нулевой гипотезе  $H_0$  не выполняется.

**Почему нельзя проверить каждое ограничение отдельно?** Складывается впечатление, что можно было бы проверить совместную гипотезу, используя обычные  $t$ -статистики для проверки каждого ограничения в отдельности. Покажем, что такой подход неверен. Допустим, вы хотите проверить совместную нулевую гипотезу из (7.6) о том, что  $\beta_1 = 0$  и  $\beta_2 = 0$ . Пусть значение  $t$ -статистики для проверки нулевой гипотезы  $\beta_1 = 0$  равно  $t_1$ , а значение  $t$ -статистики для проверки нулевой гипотезы  $\beta_2 = 0$  равно  $t_2$ . Что произойдет, если вы используете процедуру проверки ограничений по отдельности, т.е. отвергнете совместную нулевую гипотезу, если абсолютное значение либо  $t_1$ , либо  $t_2$  превысит 1,96?

Поскольку этот вопрос касается двух случайных величин  $t_1$  и  $t_2$ , то чтобы ответить на него, нужно охарактеризовать совместное выборочное

распределение  $t_1$  и  $t_2$ . Как уже упоминалось в разделе 6.6, в больших выборках  $\beta_1 = 0$  и  $\beta_2 = 0$  имеют совместное нормальное распределение, поэтому в условиях совместной нулевой гипотезы  $t$ -статистики  $t_1$  и  $t_2$  распределены по двумерному нормальному распределению, в котором среднее каждой  $t$ -статистики равно нулю, а дисперсия равна единице.

Рассмотрим сначала частный случай, в котором  $t$ -статистики некоррелированы и поэтому независимы. Чему равен размер теста, когда мы проверяем каждое ограничение отдельно; то есть какова вероятность отвержения нулевой гипотезы, когда она верна? Больше 5%! В этом частном случае мы можем точно посчитать вероятность отклонения нулевой гипотезы. Нулевая гипотеза не отвергается только в том случае, если выполнены оба неравенства  $|t_1| \leq 1,96$  и  $|t_2| \leq 1,96$ . Поскольку  $t$ -статистики независимы,  $\Pr(|t_1| \leq 1,96 \text{ и } |t_2| \leq 1,96) = \Pr(|t_1| \leq 1,96) \times \Pr(|t_2| \leq 1,96) = 0,95^2 = 0,9025 = 90,25\%$ . Таким образом, вероятность отвержения нулевой гипотезы, когда она справедлива, равна  $1 - 0,9025 = 9,75\%$ . Рассмотренный метод тестирования каждого ограничения отдельно отклоняет нулевую гипотезу слишком часто, потому что увеличивает шанс отвергнуть гипотезу: если вы не смогли отвергнуть гипотезу, используя первую  $t$ -статистику, вы получаете возможность попробовать еще раз, используя вторую.

Если же регрессоры коррелированы, ситуация еще более сложная. Размер теста для проверки каждого ограничения отдельно зависит от величины коэффициента корреляции между регрессорами. Поскольку тестирование каждого ограничения отдельно приводит к неправильному размеру теста – то есть вероятность отклонения нулевой гипотезы не равна желаемому уровню значимости, – нужно использовать новый подход.

Одним из возможных подходов является так называемый метод Бонферрони, описанный в приложении 7.1. Его суть заключается в использовании скорректированных критических значений при проверке каждого ограничения отдельно, чтобы размер критерия совпадал с его уровнем значимости. Преимуществом метода Бонферрони является то, что он применим в самом общем виде, недостатком – возможная низкая мощность процедуры: она часто не в состоянии отклонить нулевую гипотезу, когда на самом деле верна альтернативная гипотеза.

К счастью, существует еще один поход к проверке совместных гипотез, который является более мощным, особенно в ситуации сильной коррелированности регрессоров. Этот подход основан на  $F$ -статистике.

## ***F-стatisстика***

***F-стatisстика*** используется для проверки совместной гипотезы о коэффициентах модели множественной линейной регрессии. Формулы для расчета  $F$ -статистики включены в современные программные приложения, используемые для регрессионного анализа. Сначала мы обсудим случай с двумя ограничениями, а затем перейдем к общему случаю с  $q$  ограничениями.

***F-стatisстика для случая  $q=2$  ограничений.*** Когда совместная нулевая гипотеза содержит два ограничения  $\beta_1 = 0$  и  $\beta_2 = 0$ ,  $F$ -статистика объединяет две  $t$ -статистики  $t_1$  и  $t_2$  и рассчитывается по формуле:

$$F = \frac{1}{2} \left( \frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2}{1 - \hat{\rho}_{t_1, t_2}^2} \right), \quad (7.9)$$

где  $\hat{\rho}_{t_1, t_2}$  – оценка корреляции между двумя  $t$ -статистиками.

Для понимания формулы  $F$ -статистики из выражения (7.9), предположим сначала, что мы знаем, что  $t$ -статистики некоррелированы, и поэтому мы можем опустить члены, содержащие  $\hat{\rho}_{t_1, t_2}$ . Тогда выражение (7.9) упрощается и  $F = \frac{1}{2}(t_1^2 + t_2^2)$ ; то есть  $F$ -статистика является средним значение квадратов  $t$ -статистик. В условиях нулевой гипотезы  $t_1$  и  $t_2$  являются независимыми стандартными нормальными случайными величинами (так как  $t$ -статистики не коррелированы по предположению), поэтому в условиях нулевой гипотезы  $F$  распределена согласно распределению Фишера  $F_{2, \infty}$  (раздел 2.4). В соответствии с альтернативной гипотезой либо  $\beta_1$ , либо  $\beta_2$ , либо оба коэффициента отличны от нуля, тогда либо  $t_1$ , либо  $t_2$ , либо обе статистики будут большими, что приведет к отверждению нулевой гипотезы.

В общем случае  $t$ -статистики коррелированы, и тогда формула для расчета  $F$ -статистики из выражения (7.9) скорректирована на эту корреляцию. Коррекция осуществлена таким образом, что в условиях нулевой гипотезы  $F$ -статистика асимптотически распределена по распределению Фишера  $F_{2, \infty}$  вне зависимости от того, коррелированы ли  $t$ -статистики или нет.

**$F$ -статистика для случая  $q$  ограничений.** Формула для расчета  $F$ -статистики, устойчивой к гетероскедастичности и проверяющей выполнение  $q$  ограничений из совместной нулевой гипотезы из выражения (7.8), приведена в разделе 18.3. Эта формула включена в программные пакеты, используемые для регрессионного анализа, что делает  $F$ -статистику легко вычисляемой на практике.

В условиях нулевой гипотезы  $F$ -статистика имеет выборочное распределение, которое в больших выборках аппроксимируется распределением Фишера  $F_{q, \infty}$ . То есть в больших выборках в условиях нулевой гипотезы имеем:

$$F\text{-статистика распределена как } F_{q, \infty}. \quad (7.10)$$

Поэтому критические значения для  $F$ -статистики можно получить из таблиц для распределения Фишера  $F_{q, \infty}$ , приведенных в таблице 4 приложения, для соответствующего значения  $q$  и желаемого уровня значимости.

**Вычисление устойчивых к гетероскедастичности  $F$ -статистик в статистических программных приложениях.** Если  $F$ -статистика вычисляется по общей формуле, использующей устойчивые к гетероскедастичности стандартные ошибки, то в условиях нулевой гипотезы ее распределением в большой выборке является  $F_{q, \infty}$ , вне зависимости от того, являются ли ошибки регрессии гомоскедастичными или гетероскедастичными. Как отмечалось в разделе 5.4, в силу исторических причин большая часть статистических программных пакетов по умолчанию вычисляет стандартные ошибки оценок регрессии только для случая гомоскедастичных ошибок регрессии. Поэтому в некоторых программных пакетах необходимо специально выбирать среди пункта меню программы «устойчивый» (robust) вариант, чтобы  $F$ -статистика вычислялась с использованием устойчивых к гетероскедастичности стандартных ошибок (или, другими словами, устойчивой к гетероскедастичности оценки «ковариационной матрицы»).

Версия  $F$ -статистики для случая гомоскедастичных ошибок регрессии обсуждается в конце этого раздела.

**Вычисление  $p$ -значения для  $F$ -статистики.** Для вычисления  $p$ -значения для  $F$ -статистики используется распределение Фишера  $F_{q,\infty}$ , которое является приближением истинного распределения в больших выборках. Пусть  $F^{act}$  обозначает фактически вычисленное значение  $F$ -статистики. Поскольку  $F_{q,\infty}$  является распределением  $F$ -статистики в больших выборках,  $p$ -значение равно:

$$p\text{-значение} = \Pr[F_{q,\infty} > F^{act}]. \quad (7.11)$$

Для вычисления  $p$ -значения из (7.11) можно использовать таблицу распределения Фишера  $F_{q,\infty}$  (или, в качестве альтернативы, таблицу распределения  $\chi_q^2$ , поскольку случайная величина с распределением  $\chi_q^2$  в  $q$  раз больше случайной величины с распределением  $F_{q,\infty}$ ). Кроме того,  $p$ -значение может быть вычислено с помощью компьютера, потому что формулы для распределений Фишера и хи-квадрат включены в большинство современных статистических программных приложений.

**$F$ -статистика для проверки значимости регрессии.**  $F$ -статистика для проверки значимости регрессии тестирует гипотезу о том, что все коэффициенты наклона в регрессии равны нулю. В этом случае нулевая и альтернативная гипотезы формально записываются так:

$$\begin{aligned} H_0 : \beta_1 = 0, \beta_2 = 0, \dots, \beta_k = 0 \text{ против } H_1 : \beta_j \neq 0, \\ \text{по меньшей мере для одного } j, j=1, \dots, k. \end{aligned} \quad (7.12)$$

Если эта нулевая гипотеза справедлива, то ни один из регрессоров не объясняет никакую часть изменений зависимой переменной  $Y_i$ , хотя свободный член в регрессии (который в условиях нулевой гипотезы является математическим ожиданием  $Y_i$ ) может быть отличен от нуля. Нулевая гипотеза из (7.12) является специальным случаем нулевой гипотезы из (7.8), и  $F$ -статистика для проверки значимости регрессии является  $F$ -статистикой, вычисленной для нулевой гипотезы из (7.12). Если нулевая гипотеза верна, то  $F$ -статистика для проверки значимости регрессии асимптотически распределена согласно распределению Фишера  $F_{k,\infty}$ .

**$F$ -статистика для случая  $q=1$ .** Если  $q=1$ , то  $F$ -статистика проверяет одно единственное ограничение. В этом случае нулевая совместная гипотеза сводится к нулевой гипотезе об одном коэффициенте регрессии, и тогда  $F$ -статистика равна квадрату  $t$ -статистики.

### Пример: результаты тестов и соотношение учеников и учителей

Теперь мы готовы проверить нулевую гипотезу о том, что и коэффициент при соотношении учеников и учителей, и коэффициент при расходах на одного ученика равны нулю, против альтернативной гипотезы о том, что, по крайней мере, один из коэффициентов отличен от нуля, считая неизменным процент изучающих английский язык детей в школьном округе.

Для проверки этой гипотезы нам необходимо вычислить  $F$ -статистику, используя устойчивые к гетероскедастичности стандартные ошибки, для ограни-

чений  $\beta_1$  и  $\beta_2$ , используя оценки регрессии  $TestScore$  на  $STR$ ,  $Expn$  и  $PctEL$ , приведенные в уравнении (7.6). Эта  $F$ -статистика равна 5,43. В условиях нулевой гипотезы распределением этой статистики в больших выборках является распределение Фишера  $F_{2,\infty}$ . Критическое значение распределения  $F_{2,\infty}$  на 5%-м уровне значимости равно 3,00, а на 1%-м уровне значимости равно 4,61. Расчитанное на основе имеющихся данных, значение  $F$ -статистики, равное 5,43, больше чем 4,61, поэтому нулевая гипотеза отклоняется на 1%-м уровне значимости. Очень маловероятно, что случайная выборка могла бы дать такое большое значение  $F$ -статистики, если нулевая гипотеза была бы действительно справедлива (поскольку  $p$ -значение равно 0,005). Следовательно, исходя из результатов оценок, приведенных в уравнении (7.6), и на основе рассчитанной  $F$ -статистики мы можем отклонить гипотезу налогоплательщика о том, что ни соотношение учеников и учителей, ни расходы на одного ученика не влияют на результаты тестов (при неизменном проценте изучающих английский язык).

### ***F-стatisстика для случая гомоскедастичных ошибок регрессии***

Вопрос, на который отвечает  $F$ -стatisстика, можно переформулировать следующим образом. Предположим, что мы ослабим наши  $q$  ограничений, формирующих нулевую гипотезу. Улучшится ли от этого качество приближения данных моделью настолько, чтобы было маловероятно, что это улучшение является всего лишь результатом случайной ошибки при формировании случайной выборки при условии, что нулевая гипотеза верна? Данная формулировка наводит на мысль о том, что существует связь между  $F$ -стatisстикой и коэффициентом детерминации  $R^2$ : большая  $F$ -стatisтика должна быть, по всей видимости, связана с существенным увеличением коэффициента детерминации  $R^2$ . Оказывается, что если ошибки регрессии  $u_i$  гомоскедастичны, то наша интуиция имеет точное математическое обоснование. А именно если ошибки регрессии являются гомоскедастичными,  $F$ -стatisтика может быть записана в терминах улучшения качества приближения данных моделью регрессии, которое выражается либо в уменьшении остаточной суммы квадратов, либо в увеличении коэффициента детерминации  $R^2$ . Посчитанную таким образом  $F$ -стatisтику называют  $F$ -стatisстикой, рассчитанной для случая гомоскедастичных ошибок регрессии (или, кратко,  $F$ -стatisстикой при условии гомоскедастичности), поскольку она верна только при выполнении условия гомоскедастичности ошибок регрессии. В противоположность этому устойчивая к гетероскедастичности ошибок регрессии  $F$ -стatisтика, посчитанная по приведенной в разделе 18.3 формуле, применима вне зависимости от того, являются ли ошибки регрессии гомоскедастичными или гетероскедастичными. Несмотря на это существенное ограничение, в условиях гомоскедастичности ошибок регрессии формула для расчета  $F$ -стatisстике довольно проста и дает представление о том, как  $F$ -стatisтика работает. К тому же по этой формуле можно вычислить  $F$ -стatisтику с помощью известных характеристик качества подгонки данных моделью, которые обычно приводятся в оценках регрессий – таблица с результатами оценки регрессии может содержать коэффициент детерминации  $R^2$ , но не  $F$ -стatisтику.

При условии гомоскедастичности ошибок регрессии  $F$ -статистика вычисляется по простой формуле, основанной на сумме квадратов остатков от двух регрессий. Первая модель регрессии, называемая *регрессия с ограничениями*, формулируется так, как если бы нулевая гипотеза была верна. Если, например, нулевая гипотеза имеет вид, как записано в выражении (7.8), в котором все гипотетические значения равны нулю, моделью с ограничениями является модель регрессии, в которой эти коэффициенты устанавливаются равными нулю; другими словами, соответствующие регрессоры исключаются из модели. Вторая регрессия, которая называется *регрессия без ограничений*, формулируется так, как если бы альтернативная гипотеза была верна. Если сумма квадратов остатков в регрессии без ограничений существенно меньше, чем в регрессии с ограничениями, то нулевая гипотеза отвергается.

***F-статистика, рассчитанная при условии гомоскедастичности ошибок регрессии***, вычисляется по формуле:

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}}) / q}{SSR_{\text{unrestricted}} (n - k_{\text{unrestricted}} - 1)}, \quad (7.13)$$

где  $SSR_{\text{restricted}}$  является суммой квадратов остатков в регрессии с ограничениями,  $SSR_{\text{unrestricted}}$  является суммой квадратов остатков в регрессии без ограничений,  $q$  равно числу ограничений в нулевой гипотезе и  $k_{\text{unrestricted}}$  равно числу регрессоров в регрессии без ограничений. Альтернативная (эквивалентная) формула  $F$ -статистики при условии гомоскедастичности ошибок основана на коэффициентах детерминации  $R^2$  двух регрессий:

$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}}) / q}{(1 - R^2_{\text{unrestricted}}) / (n - k_{\text{unrestricted}} - 1)}. \quad (7.14)$$

Если ошибки гомоскедастичны, то разница между  $F$ -статистикой, рассчитанной при условии гомоскедастичности ошибок регрессии с помощью уравнения (7.13) или (7.14), и  $F$ -статистикой, устойчивой к гетероскедастичности ошибок регрессии, исчезает по мере роста размера выборки  $n$ . Таким образом, если ошибки гомоскедастичны и верна нулевая гипотеза, то выборочным распределением  $F$ -статистики, рассчитанной при условии гомоскедастичности ошибок регрессии, для больших выборок является  $F_{q,\infty}$ .

Эти формулы легко вычислять, и у них есть интуитивная интерпретация в терминах того, как хорошо модели регрессий с ограничениями и без них приближают данные. К сожалению, эти формулы применимы, только если ошибки регрессии гомоскедастичны. Поскольку гомоскедастичность является частным случаем, на который мы не можем полагаться в приложениях с экономическими данными или, более широко, с данными, с которыми обычно приходится работать в общественных науках, на практике  $F$ -статистика, рассчитанная при предположении гомоскедастичности ошибок регрессии, не является хорошей заменой для  $F$ -статистики, которая рассчитана для предположения гетероскедастичности ошибок регрессии.

**Использование  $F$ -статистики, рассчитанной для случая гомоскедастичных ошибок регрессии, для небольших  $n$ .** Если ошибки регрессии гомоскедастичны и являются независимо и одинаково распределенными нормальными случайными величинами, тогда  $F$ -статистика, рассчитанная в предположении гомоскедастичности ошибок регрессии и определенная в уравнениях (7.13) и (7.14), имеет распределение Фишера  $F_{q, n-k_{unrestricted}-1}$ , если нулевая гипотеза верна. Критические значения этого распределения зависят от числа степеней свободы  $q$  и  $n-k_{unrestricted}-1$ ; они приведены в таблице 5 приложения. Как уже говорилось в разделе 2.4, распределение  $F_{q, n-k_{unrestricted}-1}$  сходится к распределению  $F_{q,\infty}$  с ростом  $n$ ; и для выборок большого размера различия между этими двумя распределениями являются незначительными. Однако для маленьких выборок критические значения этих двух распределений различаются.

**Пример: результаты тестов и соотношение учеников и учителей.** Чтобы проверить нулевую гипотезу о том, что теоретические коэффициенты при  $STR$  и  $Expn$  равны нулю при постоянстве  $PctEL$ , нам нужно рассчитать  $SSR$  (или  $R^2$ ) для моделей с ограничениями и без ограничений. Модель регрессии без ограничений содержит объясняющие переменные  $STR$ ,  $Expn$  и  $PctEL$ , и ее оценки приведены в уравнении (7.6); ее  $R^2$  равен 0,4366; то есть  $R^2_{unrestricted}=0,4366$ . Регрессия с ограничениями оценена с учетом того, что в условиях нулевой гипотезы коэффициенты при  $STR$  и  $Expn$  равны нулю; то есть в условиях нулевой гипотезы  $STR$  и  $Expn$  не входят в модель теоретической регрессии, в то время как  $PctEL$  входит (нулевая гипотеза не налагает ограничений на коэффициент перед  $PctEL$ ). В результате получаем следующие МНК-оценки регрессии с ограничениями:

$$\widehat{TestScore} = 664,7 - 0,671 \times PctEL, R^2 = 0,4149. \quad (7.15)$$

Таким образом,  $R^2=0,4149$ . Число ограничений равно  $q=2$ , число наблюдений равно  $n=420$ , число регрессоров в регрессии без ограничений равно  $k=3$ .  $F$ -статистика, рассчитанная при предположении гомоскедастичности ошибок регрессии с использованием уравнения (7.14), равна:

$$F = \frac{(0,4366 - 0,4149)/2}{(1 - 0,4366)/(42 - 3 - 1)} = 8,01.$$

Поскольку 8,01 превышает 4,61 (1 %-е критическое значение), то в соответствии с предположением гомоскедастичности ошибок регрессии нулевая гипотеза отвергается на 1 %-м уровне значимости.

Этот пример показывает преимущества и недостатки  $F$ -статистики, рассчитанной в предположении гомоскедастичности ошибок регрессии. Ее преимущество заключается в том, что она может быть посчитана с помощью калькулятора. А недостатком – то, что значения рассчитанной в предположении гомоскедастичности ошибок  $F$ -статистики и  $F$ -статистики, являющейся устойчивой к наличию гетероскедастичности ошибок регрессии, могут сильно различаться: устойчивая к гетероскедастичности  $F$ -статистика для проверки этой совместной

гипотезы равна 5,43, что довольно сильно отличается от значения  $F$ -статистики для случая гомоскедастичности, которое равно 8,43.

### 7.3. Тестирование одного ограничения, включающего несколько коэффициентов модели

Иногда из экономической теории следует одно ограничение, содержащее два или более коэффициентов регрессии. Например, из теории может следовать нулевая гипотеза о равенстве коэффициентов модели ( $\beta_1 = \beta_2$ ); то есть из теории следует, что первый и второй регрессоры оказывают одинаковое влияние на зависимую переменную. В этом случае наша задача состоит в том, чтобы проверить нулевую гипотезу о равенстве коэффициентов против альтернативы о том, что эти коэффициенты различны:

$$H_0 : \beta_1 = \beta_2 \text{ против } H_1 : \beta_1 \neq \beta_2. \quad (7.16)$$

Эта нулевая гипотеза состоит только из одного ограничения, то есть  $q=1$ , однако это ограничение содержит несколько коэффициентов ( $\beta_1$  и  $\beta_2$ ). Мы должны изменить методы, рассмотренные ранее, чтобы иметь возможность проверить эту гипотезу. Существует два подхода; и в зависимости от имеющегося в вашем распоряжении программного обеспечения проще может оказаться один из них.

**Подход № 1: Непосредственная проверка ограничения.** Некоторые статистические программные пакеты содержат специальную команду для проверки ограничений вида (7.16). Результатом является  $F$ -статистика, которая имеет распределение  $F_{1,\infty}$ , поскольку  $q=1$ , если нулевая гипотеза верна. (Напомним из раздела 2.4, что квадратный корень из стандартной нормальной величины имеет распределение  $F_{1,\infty}$ ; поэтому 95 %-й процентиль распределения  $F_{1,\infty}$  равен  $1,96^2=3,84$ .)

**Подход № 2: Преобразование регрессии.** Если используемый вами статистический программный пакет не может проверить это ограничение непосредственно, то нулевую гипотезу из (7.16) можно протестировать с помощью следующего приема: нужно переписать исходную модель регрессии таким образом, чтобы ограничение из нулевой гипотезы (7.16) превратилось в ограничение на один коэффициент регрессии. Рассмотрим пример. Предположим для определенности, что регрессия содержит только два регрессора,  $X_1$  и  $X_2$ , то есть теоретическая модель регрессии имеет вид:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i. \quad (7.17)$$

Описываемый прием заключается в том, что если мы вычтем и прибавим  $\beta_2 X_{1i}$ , то получим, что  $\beta_1 X_{1i} + \beta_2 X_{2i} = \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} = (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) = \gamma_1 X_{1i} + \beta_2 W_i$ , где  $\gamma_1 = \beta_1 - \beta_2$  и  $W_i = X_{1i} + X_{2i}$ . Таким образом, теоретическую регрессию из уравнения (7.17) можно представить в таком виде:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i. \quad (7.18)$$

Поскольку коэффициент  $\gamma_1$  в этом уравнении равен  $\gamma_1 = \beta_1 - \beta_2$ , то нулевая гипотеза переписывается в виде  $\gamma_1 = 0$ , а альтернативная — как  $\gamma_1 \neq 0$ . Таким образом, представив уравнение (7.17) в виде (7.18), мы превратили ограничение на два коэффициента регрессии в ограничение на один коэффициент.

Поскольку полученное ограничение содержит один-единственный коэффициент  $\gamma_1$ , нулевую гипотезу из (7.16) можно проверить с помощью  $t$ -статистики из раздела 7.1. На практике для этого сначала строят новый регрессор  $W_i$  как сумму двух исходных регрессоров, а затем оценивают регрессию  $Y_i$  на  $X_{1i}$  и  $W_i$ . В этом случае 95 %-й доверительный интервал для разности коэффициентов  $\beta_1 - \beta_2$  можно построить как  $\hat{\gamma}_1 \pm 1,96SE(\hat{\gamma}_1)$ .

Используя этот прием, данный метод можно расширить и на другие ограничения в моделях регрессий (см. упражнение 7.9).

Два описанных выше метода (подходы № 1 и 2) эквивалентны друг другу в том смысле, что  $F$ -статистика, используемая для проверки нулевой гипотезы в первом методе, равна квадрату  $t$ -статистики, которая используется для проверки нулевой гипотезы во втором методе.

**Расширение на случай  $q > 1$ .** В общем случае нулевая гипотеза может состоять из  $q$  ограничений, все или некоторые из которых могут содержать несколько коэффициентов.  $F$ -статистика из раздела 7.2 распространяется и на этот тип совместных гипотез. Для вычисления этой  $F$ -статистики можно использовать любой из методов, описанных выше для случая  $q=1$ . Как это лучше всего делать на практике — зависит от того, каким программным пакетом для регрессионного анализа вы пользуетесь.

## 7.4. Доверительные области для нескольких коэффициентов

В этом разделе объясняется, как построить доверительные области<sup>1</sup> для двух или более коэффициентов регрессии. Представленный здесь метод концептуально похож на метод построения доверительного интервала для одного коэффициента с помощью  $t$ -статистики из раздела 7.1, за исключением того, что доверительная область для нескольких коэффициентов строится на основе  $F$ -статистики.

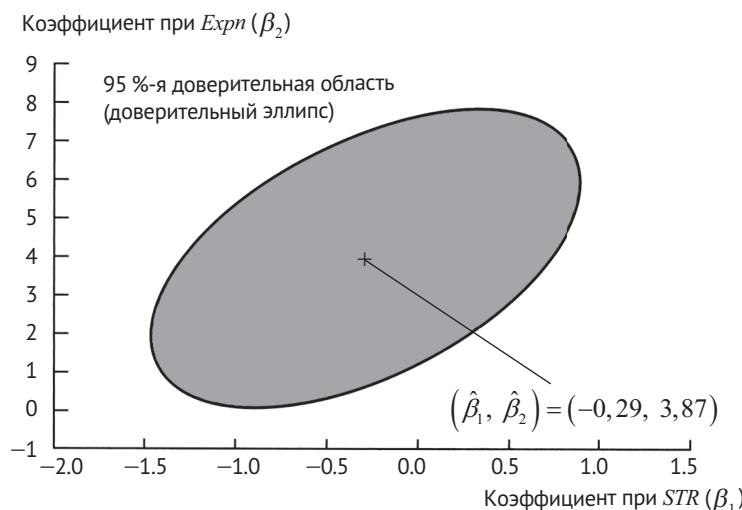
Множество, содержащее истинные значения двух или нескольких коэффициентов в 95 % случайных выборок, называют 95 %-я доверительная область для этих коэффициентов. Таким образом, доверительная область обобщает понятие доверительного интервала для одного коэффициента на случай с двумя или более коэффициентами.

Напомним, что 95 %-й доверительный интервал вычисляют как множество значений коэффициента, гипотеза о равенстве которым не отклоняется  $t$ -статистикой на 5 %-м уровне значимости. Данный подход можно расширить на случай с несколькими коэффициентами. Для определенности предположим, что вы хотите построить доверительную область для двух коэффициентов,  $\beta_1$

<sup>1</sup> В русскоязычных учебниках по эконометрике можно встретить название доверительные эллипсоиды (для случая двух переменных — доверительные эллипсы). — Примеч. науч. ред. перевода.

и  $\beta_2$ . В разделе 7.2 описывалось, как применять  $F$ -статистику для проверки нулевой гипотезы о том, что  $\beta_1 = \beta_{1,0}$  и  $\beta_2 = \beta_{2,0}$ . Допустим, вы проверите каждое возможное значение  $\beta_{1,0}$  и  $\beta_{2,0}$  на 5 %-м уровне значимости. Для каждой пары кандидатов  $(\beta_{1,0}, \beta_{2,0})$  вы посчитаете  $F$ -статистику и отвергнете соответствующую нулевую гипотезу, если  $F$ -статистика превышает 3-5 %-е критическое значение. Поскольку уровень значимости критерия равен 5 %, истинные значения  $\beta_1$  и  $\beta_2$  не будут отвергнуты в 95 % всех выборок. Таким образом, множество значений, гипотеза о равенстве которым не отвергается этой  $F$ -статистикой на 5 %-м уровне значимости, будет 95 %-й доверительной областью для  $\beta_1$  и  $\beta_2$ .

Несмотря на то что такой метод проверки всех возможных значений  $\beta_{1,0}$  и  $\beta_{2,0}$  работает в теории, на практике гораздо проще использовать явную формулу для доверительной области. Такая формула для произвольного числа коэффициентов основана на формуле для  $F$ -статистики. В случае двух коэффициентов полученные доверительные области являются эллипсами.



**Рисунок 7.1. 95 %-е доверительное множество для коэффициентов при  $STR$  и  $Exprn$  из регрессии (7.6)**

95 %-я доверительная область для коэффициентов при  $STR(\beta_1)$  и  $Exprn(\beta_2)$  является эллипсом. Этот эллипс содержит пары значений  $\beta_1$  и  $\beta_2$ , гипотеза о равенстве которых не может быть отвергнута на 5 %-м уровне значимости с использованием  $F$ -статистики.

В качестве иллюстрации рассмотрим рисунок 7.1, на котором изображена 95 %-я доверительная область (доверительный эллипс) для коэффициентов при соотношении учеников и учителей и расходов на одного ученика, при неизменном проценте изучающих английский язык школьников, построенная на основе оценок регрессии (7.6). Этот эллипс не содержит точку  $(0; 0)$ . Это означает, что нулевая гипотеза о том, что эти два коэффициента одновременно равны нулю, отвергается с помощью  $F$ -статистики на 5 %-м уровне значимости — что мы уже выяснили в разделе 7.2. Изображенный доверительный эл-

липс похож на толстую сосиску, длинная часть которой направлена из левого нижнего угла к правому верхнему. Причиной такого расположения является наличие положительной корреляции между  $\hat{\beta}_1$  и  $\hat{\beta}_2$ , которая, в свою очередь, вызвана отрицательной корреляцией между регрессорами *STR* и *Expn* (в школах с большими расходами на одного ребенка, как правило, приходится меньше учеников на одного учителя).

## 7.5. Выбор спецификации модели множественной регрессии

Проблема выбора переменных, включаемых в модель множественной регрессии, то есть проблема выбора спецификации регрессии может быть довольно сложной, и нет единого для всех ситуаций правила. Но отчаиваться не стоит, поскольку существуют некоторые полезные рекомендации. В качестве отправной точки для выбора спецификации регрессии можно проанализировать возможные источники смещения оценок из-за пропущенных переменных. Здесь важно полагаться на ваше экспертное мнение относительно эмпирической проблемы и со средоточиться на том, чтобы получить несмещенную оценку интересующих вас эффектов; поэтому не следует полностью полагаться на чисто статистические показатели качества приближения данных моделью регрессии, такие как коэффициент детерминации  $R^2$  или скорректированный коэффициент детерминации  $\bar{R}^2$ .

### *Смещение из-за пропущенной переменной в модели множественной регрессии*

Если в регрессию не включен фактор, влияющий на  $Y_i$  и коррелированный хотя бы с одним из регрессоров, то МНК-оценка коэффициентов регрессии будет смещена. Например, ученики из богатых семей часто имеют больше возможностей для обучения за пределами школы (они имеют больше возможностей читать дома, путешествовать, посещать музеи и т.д.), чем их менее обеспеченные сверстники, что может приводить к лучшим результатам учебы. Более того, у школ в обеспеченных районах, как правило, более крупные бюджеты и меньше учеников в классах. Если это так, то показатель, характеризующий наличие дополнительных возможностей для обучения, и соотношение учеников и учителей будут отрицательно коррелированы, и МНК-оценка коэффициента перед соотношением учеников и учителей будет частично отражать эффект от влияния этих дополнительных возможностей для обучения, даже если мы будем контролировать процент изучающих английский язык. Таким образом, если в регрессию не включена характеристика дополнительных возможностей для обучения (и другие переменные, связанные с финансовыми возможностями ученика), то мы можем получить смещенную оценку регрессии результатов тестов на показатели соотношения учеников и учителей и процента изучающих английский язык школьников.

Общие условия, приводящие к смещению оценки из-за пропущенной переменной в множественной регрессии, похожи на условия для случая парной

регрессии: если пропущенная переменная влияет на  $Y_i$  и коррелирует, по крайней мере, с одним регрессором, то МНК-оценка по крайней мере одного коэффициента будет смещена. Эти два условия смещения оценки из-за пропущенной переменной во множественной регрессии сформулированы во вставке «Основные понятия 7.3».

С математической точки зрения если оба этих условия смещения оценки выполнены, то как минимум один из регрессоров коррелирован с ошибкой. Это означает, что условное математическое ожидание  $u_i$  относительно  $X_{1i}, \dots, X_{ki}$  отлично от нуля, и, таким образом, первое условие метода наименьших квадратов нарушено. В результате смещение оценки сохраняется даже при большом размере выборки, то есть смещение из-за пропущенной переменной приводит к несостоятельности МНК-оценки.

**ОСНОВНЫЕ  
ПОНЯТИЯ**

**7.3**

**Смещение из-за пропущенных переменных в модели множественной регрессии**

Смещение МНК-оценки из-за пропущенной переменной возникает, если один или несколько включенных регрессоров коррелированы с пропущенной переменной. Двумя условиями для появления смещения оценки из-за пропущенной переменной являются:

1. Хотя бы один из включенных регрессоров должен быть коррелирован с пропущенной переменной.
2. Пропущенная переменная должна оказывать влияние на зависимую переменную  $Y$ .

**Роль контрольных переменных в множественной регрессии**

До сих пор мы неявно различали регрессор, эффект от влияния которого мы хотим оценить, то есть интересующую нас переменную и контрольные переменные. Теперь мы обсудим это различие более подробно.

Контрольными переменными называются показатели, которые не являются непосредственным предметом анализа; это, скорее, регрессоры, включенные для того, чтобы избежать смещения оценки влияния изучаемых факторов, которое может возникнуть, если пренебречь этими регрессорами. В предположениях метода наименьших квадратов для множественной регрессии (раздел 6.5) все регрессоры считаются равноправными. В этом подразделе мы представляем альтернативу первому из предположений метода наименьших квадратов, в котором мы проводим явное различие между изучаемой и контрольной переменными. Если это альтернативное предположение выполняется, то МНК-оценка эффекта влияния изучаемой переменной является несмешенной, в то время как МНК-оценки коэффициентов при контрольных переменных, в общем случае,

смещены и не могут быть интерпретированы как величина влияния соответствующих факторов.

Рассмотрим, например, смещение из-за пропущенной переменной, потенциально возникающее, если регрессия результатов тестов не содержит характеристики дополнительных возможностей обучения. Хотя понятие «дополнительные возможности обучения» является довольно общим и трудно измеримым, эти возможности коррелированы с финансовыми возможностями ученика (точнее, его семьи), которые могут быть измерены. Поэтому мера финансовых возможностей ученика может быть включена в регрессию результатов тестов в качестве контрольной переменной факторов, влияющих на результаты тестов, например таких как дополнительные возможности обучения. Для этого мы скорректируем регрессию результатов тестов на  $STR$  и  $PctEL$  и добавим в нее процент учеников, получающих бесплатные или льготные школьные обеды ( $LchPct$ ). Поскольку ученики имеют право на эту программу, если доход их семьи меньше определенного порога (примерно 150 % от черты бедности),  $LchPct$  измеряет долю экономически неблагополучных детей в школьном округе. Оценка этой регрессии дает:

$$\widehat{TestScore} = 700,2 - 1,00 \times STR - 0,122 \times PctEL - 0,547 \times LchPct. \quad (7.19)$$

После включения контрольной переменной  $LchPct$  выводы о влиянии размера класса на результаты тестов существенно не меняются: коэффициент перед  $STR$  изменяется совсем немного – с  $-1,10$  в уравнении (7.5) на  $-1,00$  в уравнении (7.19) – и остается статистически значимым на 1 %-м уровне значимости.

Что же означает коэффициент перед  $LchPct$  в уравнении (7.19)? Этот коэффициент является довольно большим: различия в результатах тестов между школьными округами с  $LchPct=0\%$  и  $LchPct=50\%$ , по нашей оценке, составят 27,4 процентных пункта  $[= 0,547 \times (50 - 0)]$  – это примерно разница между 75 и 25-м процентилями результатов тестов в таблице 4.1. Можно ли интерпретировать этот коэффициент с точки зрения причинно-следственных связей? Представим себе, что школьный инспектор, увидев уравнение (7.19), предложила отменить программу льготных обедов, чтобы  $LchPct$  был равен нулю в ее школьном округе. Позволит ли отмена этой программы улучшить результаты тестов в ее школьном округе? Здравый смысл подсказывает, что ответ будет отрицательным; более того, устранение программы льготных обедов может привести к противоположному эффекту, так как некоторые ученики останутся голодными. Однако имеет ли смысл то, что коэффициент перед изучаемым фактором,  $STR$ , мы можем интерпретировать с точки зрения причинно-следственных связей, тогда как коэффициент перед контрольной переменной  $LchPct$  – не можем?

Различие между изучаемыми факторами и контрольными переменными можно формализовать путем замены первого из предположений метода наименьших квадратов, приведенных во вставке «Основные понятия 6.4», – то есть предположения о нулевом условном среднем, – на предположение о независимости условного среднего. Рассмотрим регрессию с двумя переменными,

в которой  $X_{1i}$  является изучаемым фактором, а  $X_{2i}$  – контрольной переменной. Независимость условного среднего имеет место, если математическое ожидание ошибки регрессии  $u_i$  относительно  $X_{1i}$  и  $X_{2i}$  не зависит от  $X_{1i}$ , но может зависеть от  $X_{2i}$ . То есть:

$$E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i}) \text{ (независимость условного среднего).} \quad (7.20)$$

Как показано в приложении 7.2, если предположение (7.20) о независимости условного среднего выполняется, то коэффициент при  $X_{1i}$  можно интерпретировать с точки зрения причинно-следственных взаимосвязей, в то время как коэффициент при  $X_{2i}$  – нельзя.

Идея независимости условного среднего заключается в том, что если контролировать на  $X_{2i}$ , то переменную  $X_{1i}$  можно рассматривать так, как если бы ее величина устанавливалась случайным образом, в том смысле что условное среднее ошибки не зависит от значения  $X_{1i}$ . После включения  $X_{2i}$  в качестве контрольной переменной  $X_{1i}$  становится не коррелированной с ошибкой, и поэтому МНК может оценить влияние изменений  $X_{1i}$  на  $Y_i$ . Контрольная переменная, однако, остается коррелирована с ошибкой, и поэтому оценка коэффициента перед контрольной переменной смещена из-за пропущенных переменных и не имеет причинно-следственной интерпретации.

Терминология, связанная с контрольными переменными, может ввести в заблуждение. Контрольная переменная  $X_{2i}$  включена потому, что она контролирует пропущенные факторы, влияющие на  $Y_i$  и коррелированные с  $X_{1i}$ , и, возможно (но не обязательно), сама влияет на  $Y_i$ . Поэтому коэффициент при  $X_{1i}$  измеряет влияние  $X_{1i}$  на  $Y_i$ , в то время как переменная  $X_{2i}$  используется одновременно для того, чтобы держать неизменным непосредственное влияние  $X_{2i}$  на  $Y_i$ , а также контролировать факторы, коррелированные с  $X_{2i}$ . Поскольку использовать такую терминологию трудно, принято просто говорить, что коэффициент при  $X_{1i}$  измеряет влияние  $X_{1i}$  на  $Y_i$ , если контролировать на  $X_{2i}$ . Если контрольная переменная используется, то это делается для того, чтобы контролировать как ее непосредственное влияние (если такое имеется), так и влияние коррелированных с ней пропущенных факторов с целью обеспечить выполнение условия независимости условного среднего.

В примере с размером класса переменная  $LchPct$  может быть коррелирована с такими факторами, как возможности обучения вне школы, которые входят в ошибку; более того, именно из-за этой корреляции  $LchPct$  и является полезной контрольной переменной. Наличие корреляции между  $LchPct$  и ошибкой означает, что оценка коэффициента перед  $LchPct$  не может быть интерпретирована с точки зрения причинно-следственных связей. В предположении о независимости условного среднего требуется, чтобы при фиксированных значениях контрольных переменных ( $PctEL$  и  $LchPct$ ) математическое ожидание ошибки не зависело от соотношения учеников и учителей. Другими словами, независимость условного среднего ошибки регрессии означает, что среди школ с одинаковыми значениями  $PctEL$  и  $LchPct$  размер класса как будто бы определяется случайным образом: включение  $PctEL$  и  $LchPct$  в регрессию позволяет

контролировать пропущенные факторы, так что переменная  $STR$  не коррелирована с ошибкой. Если это так, то у коэффициента перед соотношением учеников и учителей есть причинно-следственная интерпретация, а у коэффициента перед  $LchPct$  такой интерпретации нет: окружной школьный инспектор, желающий повысить результаты тестов, не может этого сделать за счет сокращения бесплатных обедов.

### ***Спецификация модели в теории и на практике***

В теории, если имеются данные по пропущенной переменной, то проблема смещения из-за пропущенной переменной решается путем ее включения в регрессию. На практике же решение о том, включать определенную переменную или нет, может быть непростым и требовать рассудительности.

Мы предлагаем двухшаговый подход к решению проблемы возможного смещения из-за пропущенных переменных. Сначала следует выбрать основной или базовый набор регрессоров, используя комбинацию экспертной оценки, выводов экономической теории и знания метода сбора данных; регрессию с использованием этого базового набора регрессоров иногда называют *базовой спецификацией*. В этой базовой спецификации должны присутствовать непосредственно интересующие нас переменные, а также контрольные переменные, подсказанные экспертной оценкой и экономической теорией. Однако экспертная оценка и экономическая теория нередко бывают неопределенными в том смысле, что зачастую у нас нет данных по тем переменным, необходимость использования которых вытекает, например, из экономической теории. Поэтому на следующем шаге необходимо составить список *альтернативных спецификаций*, то есть альтернативных наборов регрессоров. Если оценки интересующих нас коэффициентов близки по значению для разных альтернативных спецификаций, то это свидетельствует о том, что оценки для базовой спецификации являются надежными. Если же, с другой стороны, оценки интересующих нас коэффициентов существенно отличаются для разных спецификаций, то это часто свидетельствует о том, что исходная спецификация дает смещенную из-за пропущенных переменных оценку. Мы остановимся подробно на этом подходе к спецификации модели в разделе 9.2 после изучения некоторых инструментов для выбора спецификации регрессий.

### ***Интерпретация $R^2$ и скорректированного $R^2$ на практике***

Если  $R^2$  или  $\bar{R}^2$  близки к единице, то это означает, что регрессоры хорошо приближают значения зависимой переменной в выборке, если же  $R^2$  или  $\bar{R}^2$  близки к нулю, то значения зависимой переменной приближаются моделью плохо. Поэтому эти статистики полезны для описания возможностей приближения данных моделью регрессии. Однако в них легко увидеть больше, чем они на самом деле заслуживают.

Существуют четыре логические ошибки, которые потенциально можно сделать при использовании  $R^2$  или  $\bar{R}^2$  и о которых необходимо помнить:

1. **Увеличение  $R^2$  или  $\bar{R}^2$  не обязательно означает, что добавленная переменная является статистически значимой.** Коэффициент детерминации  $R^2$  всегда увеличивается при добавлении регрессора вне зависимости от того, является ли регрессор статистически значимым или нет. Скорректированный коэффициент детерминации  $\bar{R}^2$  увеличивается не всегда, но если он все же увеличивается, то это не обязательно означает, что коэффициент при добавленном регрессором является статистически значимым. Чтобы проверить, является ли добавленная переменная статистически значимой, необходимо проверить гипотезу с использованием  $t$ -статистики.
2. **Высокие значения  $R^2$  или  $\bar{R}^2$  не означают, что регрессоры влияют на зависимую переменную.** Представим себе регрессию зависимости результатов тестов от показателя парковочного пространства, приходящегося на одного ученика (площадь места парковки для учителя, деленная на число учеников). Этот показатель коррелирован с соотношением учеников и учителей, с показателем, характеризующим месторасположение школы (в пригороде или большом городе), и, возможно, с доходами школьного округа — с теми факторами, которые коррелированы с результатами тестов. Поэтому регрессия результатов тестов на показатель парковочного пространства, приходящегося на одного ученика, может иметь высокие  $R^2$  и  $\bar{R}^2$ , но его нельзя интерпретировать с точки зрения причинно-следственных связей (попробуйте сказать школьному инспектору, что для улучшения результатов тестов нужно увеличить количество парковочных мест!).
3. **Высокие значения  $R^2$  или  $\bar{R}^2$  не означают, что в оценке отсутствует смещение из-за пропущенных переменных.** Вспомните обсуждение в разделе 6.1, которое касалось смещения оценки из-за пропущенных переменных в регрессии результатов тестов на соотношение учеников и учителей. Коэффициент детерминации  $R^2$  ни разу не упоминался в этом обсуждении, так как он не играл в нем никакой логической роли. Смещение оценки из-за пропущенных переменных может произойти в регрессиях с низким  $R^2$ , средним  $R^2$  или высоким  $R^2$ . И наоборот, низкий  $R^2$  не обязательно означает, что оценка смещена из-за пропущенных переменных.
4. **Высокие значения  $R^2$  или  $\bar{R}^2$  не обязательно означают, что регрессоры подобраны наиболее подходящим образом, равно как и низкие значения  $R^2$  или  $\bar{R}^2$  не обязательно означают, что регрессоры подобраны плохо.** Вопрос о том, что представляет собой правильный набор регрессоров в модели множественной регрессии, является трудным, и мы будем к нему возвращаться на протяжении всего учебника. Выбирая набор регрессоров, необходимо принимать во внимание смещение оценки из-за пропущенных переменных, наличие данных, качество данных и, самое главное, выводы экономической теории и характер основных решаемых вопросов. Ни один из этих вопросов нельзя ответить, просто имея высокое (или низкое) значение  $R^2$  или  $\bar{R}^2$  для регрессии.

Эти выводы сформулированы во вставке «Основные понятия 7.4».

### **$R^2$ и скорректированный $R^2$ : о чем они говорят**

#### **и о чем они не говорят**

$R^2$  и  $\bar{R}^2$  показывают, хорошо ли регрессоры приближают или «объясняют» значения зависимой переменной в имеющейся у вас выборке данных. Если значение  $R^2$  (или  $\bar{R}^2$ ) близко к единице, то регрессоры хорошо приближают (предсказывают) зависимую переменную в этой выборке в том смысле, что дисперсия МНК-остатков мала по сравнению с дисперсией зависимой переменной. Если значение  $R^2$  (или  $\bar{R}^2$ ) близко к нулю, то верно обратное.

$R^2$  и  $\bar{R}^2$  не показывают:

- является ли добавленная переменная статистически значимой;
- являются ли регрессоры истинной причиной изменений зависимой переменной;
- смешена ли оценка из-за пропущенных переменных;
- подобраны ли регрессоры наиболее подходящим образом.

## **ОСНОВНЫЕ ПОНЯТИЯ**

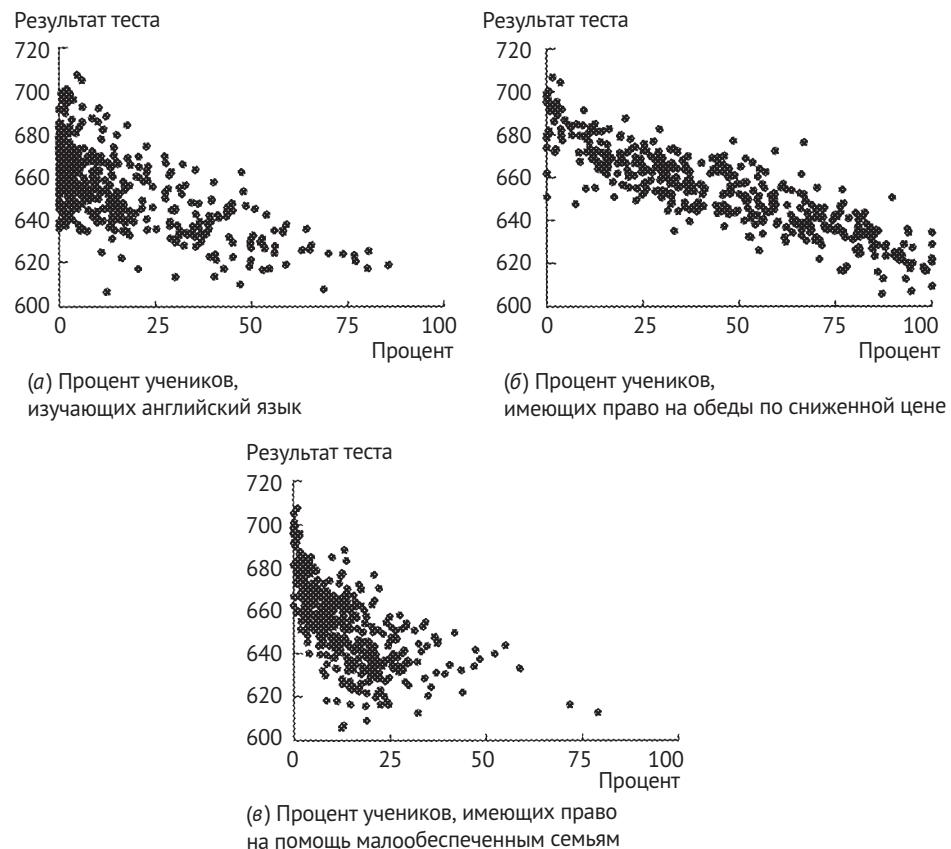
### **7.4**

## **7.6. Анализ данных по результатам тестов**

В данном разделе представлены результаты анализа влияния на результаты тестов соотношения учеников и учителей на примере данных по Калифорнии. Основная цель: привести пример использования модели множественной линейной регрессии для снижения смещения оценок из-за пропущенных переменных. В качестве другой цели мы рассматриваем необходимость демонстрации использования таблиц для представления результатов оценки регрессии.

**Обсуждение базовой и альтернативных спецификаций.** Как и ранее, мы оцениваем влияние соотношения учеников и учителей на результаты тестов при неизменных характеристиках учеников, которые окружной школьный инспектор не может контролировать. На средний балл за тест в школьных округах потенциально влияют многие факторы. Некоторые из этих факторов коррелированы с соотношением учеников и учителей, и поэтому оценка будет смещенной, если они не включены в регрессию. Поскольку эти факторы (такие как дополнительные возможности обучения) не могут быть непосредственно измерены, мы включаем контрольные переменные, которые коррелированы с этими факторами. Если контрольные переменные являются адекватными в том смысле, что выполняется предположение о независимости условного среднего, то коэффициент перед соотношением учеников и учителей измеряет эффект от изменения данного отношения при неизменных других факторах.

Здесь мы рассмотрим три переменные, контролирующие те характеристики учеников, которые могут влиять на результаты тестов: долю учеников, которые до сих пор учат английский язык; процент учеников которые имеют право получать субсидированный или бесплатный обед в школе, и новую переменную – процент учеников в школьном округе, чьи семьи могут претендовать на программу помощи малообеспеченным семьям. Право на участие в этой программе отчасти зависит от дохода семьи, но имеет более низкий (строгий) порог, чем программа субсидированных обедов. Таким образом, две последние переменные являются различными мерами доли экономически неблагополучных детей в школьных округах (их коэффициент корреляции равен 0,74). Ни теория, ни экспертное мнение не говорят о том, какую из этих двух переменных нам необходимо использовать, чтобы контролировать оценки на факторы, влияющие на результаты тестов и связанные с экономическими характеристиками учеников. Для нашей базовой спецификации мы используем процент школьников, имеющих право получать субсидированные обеды, но мы также рассматриваем альтернативную спецификацию с долей семей, которые могут претендовать на программу помощи малообеспеченным семьям.



**Рисунок 7.2. Диаграммы рассеяния результатов тестов от трех характеристик школьников**

Диаграммы рассеяния показывают отрицательную связь между результатами тестов и (a) процентом изучающих английский язык школьников (корреляция = -0,64); (б) процентом учеников, которые имеют право на обеды по сниженной цене (корреляция = -0,87); (в) процентом детей, имеющих право на помочь малообеспеченным семьям (корреляция = -0,63).

Диаграммы рассеяния результатов тестов от этих трех переменных представлены на рисунке 7.2. Каждая из этих переменных показывает отрицательную корреляцию с результатами тестов. Корреляция между результатами тестов и процентом изучающих английский язык учеников равна  $-0,64$ ; между результатами тестов и процентом детей, имеющих право на субсидированные обеды, она равна  $-0,87$ ; между результатами тестов и процентом школьников, имеющих право на помочь малообеспеченным семьям,  $-0,63$ .

**Какой масштаб (единицы измерения) следует использовать для регрессоров?** Одним из практических вопросов, возникающих в регрессионном анализе, является вопрос о том, какой масштаб следует использовать для объясняющих переменных. На рисунке 7.2 переменные измерены в процентах, поэтому максимально возможный диапазон данных составляет от 0 до 100. Альтернативно, мы могли бы определить эти переменные как доли, а не проценты; например, переменная  $PctEL$  может быть заменена на долю изучающих английский язык,  $FracEL (=PctEL/100)$ , которая принимает значения между 0 и 1 вместо 0 и 100. В общем случае в регрессионном анализе обычно необходимо принять какое-то решение о масштабе как зависимой, так и объясняющих переменных. По какому принципу следует тогда выбирать масштаб или единицы измерения переменных?

Общий ответ на этот вопрос заключается в том, что масштаб переменных нужно выбирать таким образом, чтобы результаты регрессии было легко читать и интерпретировать. В нашем примере с результатами тестов естественной единицей измерения зависимой переменной является количество баллов за тест. В регрессии  $TestScore$  на  $STR$  и  $PctEL$ , представленной в уравнении (7.5), коэффициент перед  $PctEL$  равен  $-0,650$ . Если бы мы вместо этого использовали регрессор  $FracEL$ , то  $R^2$  и SER для этой регрессии не изменились бы, однако коэффициент перед  $FracEL$  был бы равен  $-65,0$ . В спецификации с  $PctEL$  оценка коэффициента перед этой переменной интерпретируется как предсказание изменения результатов тестов при увеличении процента изучающих английский язык на 1 процентный пункт при неизменном  $STR$ . В спецификации с  $FracEL$  оценка коэффициента перед этой переменной интерпретируется как предсказание изменения результатов тестов при увеличении доли изучающих английский язык на 1, то есть на 100%, при неизменном  $STR$ . Несмотря на то что эти две спецификации математически эквивалентны, для целей интерпретации спецификация с  $PctEL$  нам кажется более естественной.

Другим соображением при принятии решения о масштабе является выбор единиц измерения регрессоров таким образом, чтобы оценки коэффициентов было легко читать. Например, если регрессор измеряется в долларах США и имеет коэффициент 0,000 003 56, результат регрессии будет легче читать, если перевести этот регрессор в миллионы долларов и привести коэффициент 3,56.

**Представление результатов в таблице.** Теперь перед нами стоит проблема представления результатов. Как наилучшим способом показать результаты нескольких многомерных регрессий, которые содержат разные подмножества возможных регрессоров? До сих пор мы представляли результаты регрессий,

записывая их в виде уравнений как в (7.6) и (7.19). Этот способ хорош, когда мы имеем дело всего лишь с несколькими регрессорами и несколькими уравнениями, но с большим числом регрессоров и уравнений такой способ представления результатов может сбивать с толку. Более подходящим способом представления результатов нескольких регрессий является таблица.

Таблица 7.1

**Регрессии результатов тестов от соотношения учеников  
и учителей и характеристики школьников с использованием данных  
по школьным округам в Калифорнии**

<b>Зависимая переменная: средний балл за тест в школьном округе</b>					
Объясняющая переменная	(1)	(2)	(3)	(4)	(5)
Соотношение учеников и учителей ( $X_1$ )	-2,28** (0,52)	-1,10* (0,43)	-1,00** (0,27)	-1,31** (0,34)	-1,01** (0,27)
Процент изучающих английский язык ( $X_2$ )		-0,650** (0,031)	-0,122** (0,033)	-0,488** (0,030)	-0,130** (0,036)
Процент имеющих право на субсидированные обеды ( $X_3$ )			-0,547** (0,024)		-0,529** (0,038)
Процент имеющих право на помочь малообеспеченным семьям ( $X_4$ )				-0,790** (0,068)	0,048 (0,059)
Свободный член	698,9** (10,4)	686,0** (8,7)	700,2** (5,6)	698,0** (6,9)	700,4** (5,5)
Статистики качества приближения данных моделью					
SER	18,58	14,46	9,08	11,65	9,08
$\bar{R}^2$	0,049	0,424	0,773	0,626	0,773
<i>n</i>	420	420	420	420	420

*Примечание.* Регрессии оценены с использованием данных по школьным округам (К-8 — начальные школы) в Калифорнии, описанным в приложении 4.1. В скобках под коэффициентами приведены устойчивые к гетероскедастичности стандартные ошибки. Коэффициент значим на 5%-м уровне\* или 1%-м уровне\*\* значимости в соответствии с двусторонним критерием.

В таблице 7.1 представлено несколько регрессий зависимости результатов тестов от различных наборов регрессоров. В каждой колонке приведена отдельная регрессия. Зависимая переменная в каждой регрессии одна и та же — результаты тестов. Первые пять строк содержат оценки коэффициентов регрессии и стандартные ошибки, приведенные в скобках под коэффициентами. Звездочки означают, что  $t$ -статистика, проверяющая гипотезу о том, что соответствующий коэффициент равен нулю, значима на 5%-м уровне (одна звездочка) или на 1%-м уровне (две звездочки). Последние три строчки содержат статистики, характеризующие качество приближения данных моделью, для каждой регрессии (стандартную ошибку регрессии, SER, и скорректированный  $R^2$  и размер выборки (он один и тот же для всех регрессий — 420 наблюдений).

Вся информация, которую мы до сих пор представляли в виде уравнения, представлена одним столбцом в этой таблице. Например, рассмотрим регрессию

результатов тестов на соотношение учеников и учителей без контрольных переменных. В виде уравнения эта регрессия записывается следующим образом:

$$\widehat{TestScore} = 698,9 - 2,28 \times STR, \bar{R}^2 = 0,049, SER = 18,58, n = 420. \quad (7.21)$$

Эта информация содержится в столбце (1) таблицы 7.1. Оценка коэффициента перед отношением числа учеников и учителей (-2,28) содержится в первой строке чисел, а ее стандартная ошибка (0,52) записана в скобках прямо под этой оценкой. Свободный член (698,9) и его стандартная ошибка (10,4) приведены в строке с названием «Свободный член». (Иногда эту строку также называют «константой», поскольку, как говорилось в разделе 6.2, свободный член можно рассматривать как коэффициент при регрессоре, который всегда равен 1.) Наконец,  $\bar{R}^2$  (0,049),  $SER$  (18,58) и размер выборки  $n$  (420) приведены в последних строчках. Пустые записи в строчках для остальных регрессоров означают, что эти объясняющие переменные не включены в данную регрессию.

Немногая на то что в этой таблице не приводятся  $t$ -статистики, их можно посчитать, исходя из имеющейся в таблице информации; например,  $t$ -статистика для проверки гипотезы о том, что коэффициент перед отношением числа учеников к учителям в столбце (1) равен нулю, равна  $-2,28/0,52 = -4,38$ . Эта гипотеза отклоняется на 1%-м уровне значимости, что обозначено двумя звездочками рядом с оценкой коэффициента в таблице.

Регрессии, в которые включены контрольные переменные, измеряющие характеристики учеников, приведены в колонках (2) – (5). В колонке (2) приведены оценки регрессии зависимости результатов тестов от соотношения учеников и учителей и от процента изучающих английский язык школьников. Ранее эту информацию мы записывали в виде уравнения (7.5).

В столбце (3) представлена наша базовая спецификация, в которой регрессорами являются соотношение учеников и учителей и две контрольные переменные: процент изучающих английский язык детей и процент школьников, имеющих право на бесплатный обед.

В столбцах (4) и (5) представлены альтернативные спецификации, с помощью которых проверяется, как разные способы измерения финансовых возможностей учеников влияют на оценки. В регрессию из столбца (4) включен процент учеников, пользующихся программой помощи малообеспеченным семьям, а в столбец (5) включены обе характеристики финансовых возможностей учеников.

**Обсуждение эмпирических результатов.** Из представленных выше результатов следуют три вывода:

- Если контролировать характеристики школьников, то оценка влияния соотношения учеников и учителей на результаты тестов уменьшается вдвое. Эта оценка не слишком чувствительна к тому, какие конкретно контрольные переменные включаются в регрессию. Во всех случаях коэффициент перед соотношением учеников и учителей остается статистически значимым на 5%-м уровне значимости. В четырех спецификациях с контрольными переменными – регрессии (2) – (5) – уменьшение соотношения учеников

- и учителей на единицу (т.е. на одного ученика, приходящегося на одного учителя) увеличивает средний балл за тесты приблизительно на 1 балл, если другие характеристики учеников не изменяются.
2. Характеристики школьников являются важными показателями, позволяющими предсказывать результаты тестов. Соотношение учеников и учителей само по себе объясняет лишь маленькую долю дисперсии результатов тестов:  $\bar{R}^2$  в столбце (1) равен 0,049. В то же время  $\bar{R}^2$  резко возрастает, если мы добавляем характеристики школьников. Например,  $\bar{R}^2$  для базовой спецификации – регрессия (3) – равен 0,773. Знаки коэффициентов перед демографическими характеристиками учеников соответствуют закономерностям, наблюдаемым в данных (см. рис. 7.2): результаты тестов хуже в школьных округах с большим числом изучающих английский язык детей и в школьных округах с большим числом малообеспеченных семей.
3. Контрольные переменные не всегда статистически значимы по отдельности: в спецификации (5) мы не можем отклонить на 5 %-м уровне значимости гипотезу о том, что коэффициент перед процентом детей, имеющих право на помочь малообеспеченным семьям, равен нулю ( $t$ -статистика равна -0,82). Поскольку после добавления этой контрольной переменной в базовую спецификацию (3) оценка коэффициента перед соотношением учеников и учителей и ее стандартная ошибка меняются незначительно и поскольку коэффициент перед этой контрольной переменной не является статистически значимым в спецификации (5), эта дополнительная контрольная переменная является лишней, по крайней мере, для целей данного анализа.

## 7.7. Заключение

Глава 6 началась с описания проблемы: в регрессии зависимости результатов тестов от соотношения учеников и учителей пропущенные характеристики учеников, возможно, коррелированы с соотношением учеников и учителей в школьном округе, и если это так, то оценка коэффициента перед соотношением учеников и учителей в школьном округе может отражать часть влияния на результаты тестов пропущенных характеристик учеников. Вследствие этого МНК-оценка может быть смещена. Чтобы уменьшить это возможное смещение, мы расширили регрессию и добавили в нее переменные, которые контролируют различные характеристики учеников (процент изучающих английский язык и две меры финансовых возможностей учеников). Это вдвое сократило оцененный эффект от изменения соотношения учеников и учителей, при этом нулевая гипотеза о том, что в генеральной совокупности соотношение учеников и учителей не влияет на результаты тестов, если считать контрольные переменные неизменными, по-прежнему не может быть отклонена на 5 %-м уровне значимости. Поскольку множественная регрессия позволяет устраниТЬ смещение оценки из-за пропущенных характеристик учеников, оценка множественных регрессий, проверка гипотез и доверительные интервалы гораздо более по-

лезны с точки зрения окружного школьного инспектора, чем оценка парных регрессий из глав 4 и 5.

Анализ, проведенный в этой и предыдущей главах, предполагал, что функция теоретической регрессии линейно зависит от регрессоров, то есть условное математическое ожидание  $Y_i$  относительно объясняющих переменных является прямой линией. Однако нет особых оснований полагать, что это так. В самом деле, эффект от снижения соотношения учеников и учителей может сильно отличаться в районах с большими классами от районов, в которых классы уже и так небольшие. Если это действительно так, то теоретическая функция регрессии не является линейной функцией от  $X$ -ов, а, скорее, нелинейно зависит от  $X$ -ов. Однако для того чтобы расширить возможности анализа на случай нелинейной по  $X$ -ам регрессии, нам необходимы инструменты, о которых рассказывается в следующей главе.

## **Выходы**

1. Проверка гипотез и построение доверительных интервалов для одного коэффициента регрессии осуществляются, по сути, при помощи тех же самых процедур, что мы использовали в случае модели парной линейной регрессии из главы 5. Например, 95 %-й доверительный интервал для  $\beta_1$  строится как  $\hat{\beta}_1 \pm 1,96SE(\hat{\beta}_1)$ .
2. Гипотезы, содержащие более одного ограничения на коэффициенты, называются совместными гипотезами. Совместные гипотезы можно проверять при помощи  $F$ -статистики.
3. При выборе спецификации регрессии сначала определяют базовую спецификацию, которая подбирается так, чтобы решить проблему смещения оценки из-за пропущенных переменных. Затем базовую спецификацию можно изменить и добавить дополнительные регрессоры, которые устроят другие потенциальные источники смещения из-за пропущенных переменных. Если просто выбирать спецификацию с наибольшим  $R^2$ , то можно получить регрессионную модель, которая не оценивает влияние интересующих нас факторов.

## **Основные понятия**

Ограничения (с. 225).

Совместная гипотеза (с. 225).

$F$ -статистика (с. 226).

Регрессия с ограничениями (с. 230).

Регрессия без ограничений (с. 230).

$F$ -статистика, рассчитанная при условии гомоскедастичности ошибок регрессии (с. 230).

95 %-я доверительная область (с. 233).

Контрольная переменная (с. 236).

Независимость условного среднего (с. 238).

Базовая спецификация (с. 239).

Альтернативные спецификации (с. 239).

Критерий Бонферрони (с. 253).

## **Вопросы для повторения и закрепления основных понятий**

7.1. Объясните, как вы будете проверять нулевую гипотезу о том, что  $\beta_1 = 0$  в модели множественной линейной регрессии  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ .

Объясните, как вы будете проверять нулевую гипотезу о том, что  $\beta_2 = 0$ .

Объясните, как вы будете проверять совместную нулевую гипотезу о том, что  $\beta_1 = 0$  и  $\beta_2 = 0$ . Объясните, почему результат совместной гипотезы не следует из результатов проверки первых двух гипотез.

7.2. Приведите пример регрессии, которая, возможно, будет иметь высокое значение  $R^2$ , но которая приведет к смещенным и несостоительным оценкам коэффициента/коэффициентов регрессии. Объясните, почему  $R^2$ , скорее всего, будет высоким. Объясните, почему МНК-оценки будут смещенными и несостоительными.

## **Упражнения**

Первые шесть упражнений относятся к таблице с результатами оценки регрессий на стр. 249, которые были оценены с использованием данных из текущего обследования населения (CPS) за 1998 год. Эта база данных состоит из информации о 4000 респондентах (работниках), работавших в течение года полный рабочий день. Все респонденты выборки имеют либо аттестат о среднем образовании, либо диплом бакалавра. Их возраст варьируется от 25 до 34 лет. В данных также содержится информация о регионе проживания респондента, его семейном положении и количестве детей. Для этого упражнения определим следующие переменные:

*AHE* = средняя почасовая заработка (в долларах 1998 года).

*College* = бинарная переменная (1 – если респондент закончил колледж, 0 – если среднюю школу).

*Female* = бинарная переменная (1 – если респондент является женщиной, 0 – если мужчиной).

*Age* = возраст (в годах).

*Northeast* = бинарная переменная (1 – если респондент проживает на Северо-Востоке, 0 – в противном случае).

*Midwest* = бинарная переменная (1 – если респондент проживает на Среднем Западе, 0 – в противном случае).

*South* = бинарная переменная (1 – если респондент проживает на Юге, 0 – в противном случае).

*West* = бинарная переменная (1 – если регион респондент проживает на Западе, 0 – в противном случае).

**Результаты оценки регрессий среднечасовой зарплаты на бинарные переменные, характеризующие пол, образование и другие характеристики, с использованием данных за 1998 год из базы данных текущего обследования населения США**

**Зависимая переменная: средняя почасовая зарплата (*AHE*)**

Регрессор	(1)	(2)	(3)
<i>College</i> ( $X_1$ )	5,46 (0,21)	5,48 (0,21)	5,44 (0,21)
<i>Female</i> ( $X_2$ )	-2,64 (0,20)	-2,62 (0,20)	-2,62 (0,20)
<i>Age</i> ( $X_3$ )		0,29 (0,04)	0,29 (0,04)
<i>Northeast</i> ( $X_4$ )			0,69 (0,30)
<i>Midwest</i> ( $X_5$ )			0,60 (0,28)
<i>South</i> ( $X_6$ )			-0,27 (0,26)
Константа	12,69 (0,14)	4,40 (1,05)	3,75 (1,06)

**Итоговая статистика и совместные тесты**

<i>F</i> -статистика для региональных эффектов=0			6,10
SER	6,27	6,22	6,21
<i>R</i> <sup>2</sup>	0,176	0,190	0,194
<i>n</i>	4000	4000	4000

- 7.1. Обозначьте в таблице уровень статистической значимости коэффициентов с помощью символов «\*» (5 %) и «\*\*» (1 %).
- 7.2. Для результатов регрессии из колонки (1):
  - a) Является ли статистически значимыми на 5 %-м уровне значимости различия в доходах работников с дипломом бакалавра и школьным дипломом? Постройте 95 %-й доверительный интервал для этих различий.
  - б) Являются ли статистически значимыми на 5 %-м уровне значимости различия в доходах женщин и мужчин? Постройте 95 %-й доверительный интервал для этих различий.
- 7.3. Для результатов регрессии из колонки (2):
  - a) Является ли возраст важным фактором, определяющим доход? Аргументируйте свой ответ с помощью соответствующего статистического теста и/или доверительного интервала.
  - б) Салли – 29-летняя женщина с дипломом бакалавра. Бетси – 34-летняя женщина с дипломом бакалавра. Постройте 95 %-й доверительный интервал для ожидаемых различий в их доходах.
- 7.4. Для результатов регрессии из колонки (3):
  - а) Видите ли вы важные региональные различия в доходах? Аргументируйте свой ответ при помощи проверки соответствующей гипотезы.
  - б) Хуанита – 28-летняя женщина с Юга с дипломом бакалавра. Молли – 28-летняя женщина с Запада с дипломом бакалавра. Дженифер – 28-летняя женщина со Среднего Запада с дипломом бакалавра.

- (i) Постройте 95 %-й доверительный интервал для различий в доходах Хуаниты и Молли.
- (ii) Объясните, как вы могли бы построить 95 %-й доверительный интервал для различий в ожидаемых доходах Хуаниты и Дженнифер. (Подсказка: что произойдет, если вы включите в регрессию *West* и исключите *Midwest*?)
- 7.5. Регрессия из столбца (2) была снова оценена, но на этот раз с использованием данных за 1992 год (4000 наблюдений, выбранных случайным образом из CPS, составленном в марте 1993 года, и переведенных в доллары 1998 года, используя индекс потребительских цен). Результат оценки:
- $$\widehat{AHE} = 0,77 + 5,29 \text{College} - 2,59 \text{Female} + 4,40 \text{Age}, \quad SER = 5,85, \quad \bar{R}^2 = 0,21.$$
- (0,98)      (0,20)      (0,18)      (0,03)
- Если сравнить эту регрессию с регрессией для 1998 года, приведенной в столбце (2), можно ли говорить о наличии статистически значимого изменения коэффициента при объясняющей переменной *College*?
- 7.6. Прокомментируйте следующее утверждение: «Во всех приведенных регрессиях коэффициент перед *Female* является отрицательным, большим и статистически значимым. Это служит убедительным доказательством гендерной дискриминации на рынке труда США».
- 7.7. В упражнении 6.5 была оценена следующая регрессия (здесь мы добавили к ней стандартные ошибки):
- $$\begin{aligned} Price = & 119,2 + 0,485 BDR + 23,4 Bath + 0,156 Hsize + 0,002 Lsize + . \\ & (23,9) \quad (2,61) \quad (8,94) \quad (0,011) \quad (0,00048) \\ & + 0,090 Age - 48,8 Poor, \quad \bar{R}^2 = 0,72, \quad SER = 41,5. \\ & (0,311) \quad (10,5) \end{aligned}$$
- а) Можно ли утверждать, что коэффициент при *BDR* статистически значимо отличается от нуля?
- б) Как правило, дома с пятью комнатами стоят гораздо дороже, чем дома с двумя комнатами. Согласуется ли это с вашими выводами из пункта (а) и с регрессией в целом?
- в) Домовладелец покупает соседний (прилегающий) участок земли площадью 2000 квадратных футов. Постройте 99 %-й доверительный интервал для изменения в стоимости ее дома.
- г) Размер земельного участка измеряется в квадратных футах. Как вы думаете, может быть, другие единицы измерения были бы более подходящими? Поясните.
- д) *F*-статистика, используемая для проверки нулевой гипотезы о равенстве нулю коэффициентов при *BDR* и *Age*, равна  $F=0,08$ . Отличаются ли коэффициенты при *BDR* и *Age* значимо от нуля на 10 %-м уровне значимости?
- 7.8. Для таблицы 7.1 из текста главы:
- а) Посчитайте  $R^2$  для каждой регрессии.
- б) Для регрессии из столбца (5) вычислите *F*-статистику для проверки нулевой гипотезы  $\beta_3 = \beta_4 = 0$ , предполагая гомоскедастичность ошибок регрессии. Можно ли отвергнуть эту гипотезу на 5 %-м уровне значимости?

- в) Используя критерий Бонферрони, описанный в приложении 7.1, проверьте гипотезу  $\beta_3 = \beta_4 = 0$  для регрессии из столбца (5).
- г) Для регрессии из столбца (5) постройте 99 %-й доверительный интервал для коэффициента  $\beta_1$ .
- 7.9. Рассмотрим модель регрессии  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ . Используя подход № 2 из раздела 7.3, преобразуйте эту регрессию таким образом, чтобы вы смогли при помощи  $t$ -статистики проверить следующие гипотезы:
- $\beta_1 = \beta_2$ ;
  - $\beta_1 + a\beta_2 = 0$ , где  $a$  является константой;
  - $\beta_1 + \beta_2 = 1$ . (Подсказка: переопределите зависимую переменную в регрессии.)
- 7.10. В уравнениях (7.13) и (7.14) приводятся две формулы  $F$ -статистики, вычисленные в предположении гомоскедастичности ошибок регрессии. Покажите, что эти две формулы эквивалентны.
- 7.11. Чтобы оценить влияние размера класса на результаты тестов во вторых классах, школьный округ проводит эксперимент. Из учеников, которые в прошлом году учились в первом классе в этом округе, 50 % распределяется в маленькие вторые классы (18 учеников в классе) и 50 % – в классы обычного размера (21 ученик в классе). Новые же ученики в этом округе распределяются иначе: 20 % случайным образом распределяется в маленькие классы и 80 % – в классы обычного размера. В конце учебного года каждому ученику второго класса дается стандартизованный тест. Пусть  $Y_i$  обозначает количество баллов за тест  $i$ -го ученика;  $X_{1i}$  обозначает бинарную переменную, которая равна единице, если ученик был распределен в маленький класс, и  $X_{2i}$  обозначает бинарную переменную, которая равна единице для новых учеников в школьном округе. Пусть  $\beta_1$  обозначает эффект влияния на результаты тестов, возникающий вследствие уменьшения размера класса с обычного до маленького.
- Рассмотрим регрессию  $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ . Как вы считаете, будет ли выполнено условие  $E(u_i | X_{1i}) = 0$ ? Является ли МНК-оценка  $\beta_1$  несмешанной и состоятельной? Объясните.
  - Рассмотрим регрессию  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ . Как вы считаете, зависит ли  $E(u_i | X_{1i}, X_{2i})$  от  $X_1$ ? Является ли МНК-оценка  $\beta_1$  несмешанной и состоятельной? Объясните. Как вы считаете, зависит ли  $E(u_i | X_{1i}, X_{2i})$  от  $X_2$ ? Является ли МНК-оценка  $\beta_2$  несмешанной и состоятельной оценкой эффекта перехода в новую школу (т.е. эффекта для новых учеников в округе)? Объясните.

### Компьютерные упражнения

E7.1. Используя базу данных **CPS08**, описанную в Е4.1, ответьте на следующие вопросы:

- a) Оцените регрессию средней зарплаты в час (*AHE*) на возраст (*Age*). Чему равна оценка свободного члена? Чему равна оценка коэффициента наклона?
- б) Оцените регрессию *AHE* на *Age*, пол (*Female*) и образование (*Bachelor*). Чему равна оценка влияния переменной *Age* на доход? Постройте 95 %-й доверительный интервал для коэффициента при *Age* в регрессии.
- в) Существенны ли различия в оценках влияния переменной *Age* на *AHE* в регрессиях из пунктов (a) и (б)? Можно ли сказать, что оценка соответствующего коэффициента в регрессии из пункта (a) смещена из-за пропущенных переменных?
- г) Боб – 26-летний мужчина, имеющий школьный аттестат. Рассчитайте доход Боба, используя оцененную в пункте (б) регрессию. Алексис – 30-летняя женщина с дипломом бакалавра. Рассчитайте доход Алексис, используя ту же самую регрессию.
- д) Используя стандартные ошибки регрессии,  $R^2$  и  $\bar{R}^2$ , сравните, насколько хорошо регрессии из пунктов (a) и (б) объясняют данные. Почему  $R^2$  и  $\bar{R}^2$  так близки между собой в регрессии из пункта (б)?
- е) Являются ли пол и образование факторами, определяющими доход? Проверьте нулевую гипотезу о том, что переменная *Female* может быть исключена из регрессии. Проверьте нулевую гипотезу о том, что переменная *Bachelor* может быть исключена из регрессии. Проверьте нулевую гипотезу о том, что обе переменные, *Female* и *Bachelor*, могут быть исключены из регрессии.
- ж) Оценка регрессии будет смещена из-за пропущенных переменных, если выполняются два условия. Какие это условия? Можно ли сказать, что эти условия выполняются здесь?

E7.2. Используя базу данных **TeachingRatings**, описанную в E4.2, выполните следующие упражнения:

- а) Оцените регрессию *Course\_Eval* от *Beauty*. Постройте 95 %-й доверительный интервал оценки влияния переменной *Beauty* на *Course\_Eval*?
- б) Рассмотрите различные контрольные переменные, имеющиеся в данных. Какие из них, на ваш взгляд, должны быть включены в регрессию? Используя таблицу, подобную таблице 7.1, проанализируйте чувствительность построенного вами в (a) доверительного интервала. Что является разумным значением для 95 %-го доверительного интервала для эффекта влияния переменной *Beauty* на *Course\_Eval*?

E7.3. Используя базу данных **CollegeDistance**, описанную E4.3, ответьте на следующие вопросы:

- а) Одна организация, выступающая с пропагандой образования, утверждает, что в среднем уровень образования человека повысится примерно на 0,15 года, если расстояние до ближайшего высшего учебного заведения снизится на 20 миль. Оцените регрессию числа полных лет обучения (*ED*) на расстояние до ближайшего колледжа (*Dist*). Со-

гласуются ли результаты оценки регрессии с тем, что утверждает эта организация? Поясните.

- б) Есть и другие факторы, которые влияют на число полных лет обучения. Изменится ли оценка влияния расстояния до ближайшего колледжа на число полных лет обучения, если контролировать другие факторы? Для ответа на данный вопрос постройте таблицу, подобную таблице 7.1. Включите в нее простую спецификацию [построенную в пункте (а)], базовую спецификацию (которая включает набор важных контрольных переменных) и несколько модификаций базовой спецификации. Обсудите различия в оценках коэффициента при *Dist* на *ED* между этими спецификациями.
- в) Некоторые исследователи считают, что если контролировать другие факторы, то число полных лет обучения чернокожих и испаноязычных граждан будет больше, чем белых. Согласуется ли этот результат с регрессиями, которые вы построили в пункте (б)?

E7.4. Используя базу данных **Growth**, описанную в Е 4.4, но не включая данные по Мальте (Malta), выполните следующие упражнения:

- а) Оцените регрессию *Growth* от переменных *TradeShare*, *YearsSchool*, *Rev\_Coups*, *Assassinations* и *RGDP60*. Постройте 95 %-й доверительный интервал для коэффициента при *TradeShare*. Является ли этот коэффициент статистически значимым на 5 %-м уровне значимости?
- б) Проверьте гипотезу о том, что *YearsSchool*, *Rev\_Coups*, *Assassinations* и *RGDP60* можно одновременно исключить из регрессии. Чему равно *p*-значение *F*-статистики?

## Приложения

### Приложение 7.1. Критерий Бонферрони для проверки совместных гипотез

Совместные гипотезы для множественных регрессий лучше всего тестировать, используя метод, описанный в разделе 7.2. Однако если автор некоторого исследования приводит результаты регрессии, но не проверяет интересующее вас совместное ограничение, и если у вас нет исходных данных, тогда вы не сможете посчитать *F*-статистику из раздела 7.2. В этом приложении мы описываем способ проверки совместных гипотез, который можно использовать, когда у вас есть только таблица с результатами оценки регрессии. Этот метод является приложением весьма общего подхода, основанного на неравенстве Бонферрони.

Критерий Бонферрони используется для проверки совместных гипотез при помощи индивидуальных *t*-статистик, вычисленных для отдельных гипотез; то есть критерий Бонферрони – это обсуждавшийся в разделе 7.2 метод проверки совместной значимости коэффициентов множественной регрессии, использующий *t*-статистики для проверки значимости каждого коэффициента в отдельности, выполненный корректным образом. *Критерий Бонферрони*, для

проверки совместной нулевой гипотезы  $\beta_1 = \beta_{1,0}$  и  $\beta_2 = \beta_{2,0}$  использующий критическое значение  $c > 0$ , использует следующее правило:

*Не отвергаем, если  $|t_1| \leq c$  и  $|t_2| \leq c$ ; в противном случае отвергаем (критерий Бонферрони для индивидуальных t-статистик),* (7.22)

где  $t_1$  и  $t_2$  –  $t$ -статистики для проверки ограничений на коэффициенты  $\beta_1$  и  $\beta_2$ .

Используемый прием заключается в том, чтобы выбрать критическое значение таким образом, чтобы вероятность отвержения нулевой гипотезы в случае ее справедливости при использовании критерия, учитывающего индивидуальные  $t$ -статистики, была бы не больше, чем желаемый уровень значимости, скажем, 5 %. Это можно сделать с помощью неравенства Бонферрони –  $c$  выбирается таким образом, чтобы одновременно учесть и то, что проверяются два ограничения, и то, что  $t_1$  и  $t_2$  могут быть коррелированы.

### **Неравенство Бонферрони**

Неравенство Бонферрони является одним из основных результатов теории вероятностей. Допустим,  $A$  и  $B$  – два события. Назовем  $A \cap B$  пересечением событий (т.е. событием, заключающимся в том, что  $A$  и  $B$  происходят одновременно), и пусть  $A \cup B$  будет объединением событий (т.е. в данный момент времени происходит или  $A$ , или  $B$ , или оба). Тогда  $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ . Поскольку  $\Pr(A \cap B) \geq 0$ , следовательно,  $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ . Из этого неравенства, в свою очередь, следует, что  $1 - \Pr(A \cup B) \geq 1 - [\Pr(A) + \Pr(B)]$ . Обозначим через  $A^c$  и  $B^c$  дополнения к событиям  $A$  и  $B$ , то есть события «не  $A$ » и «не  $B$ ». Поскольку дополнением к событию  $A \cup B$  является событие  $A^c \cap B^c$ , то  $1 - \Pr(A \cup B) = \Pr(A^c \cap B^c)$ , откуда следует неравенство Бонферрони:  $\Pr(A^c \cap B^c) \geq 1 - [\Pr(A) + \Pr(B)]$ .

Пусть теперь событие  $A$  – это  $|t_1| > c$ , а событие  $B$  –  $|t_2| > c$ . Тогда неравенство  $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$  переписывается следующим образом:

$$\Pr(|t_1| > c \text{ или } |t_2| > c \text{ или оба}) \leq \Pr(|t_1| > c) + \Pr(|t_2| > c). \quad (7.23)$$

### **Критерий Бонферрони**

Поскольку событие « $|t_1| > c$  или  $|t_2| > c$  или оба» описывает область отверждения нулевой гипотезы, выражение (7.23) определяет корректное критическое значение для критерия с использованием отдельных  $t$ -статистик. Если нулевая гипотеза справедлива, то  $\Pr(|t_1| > c) = \Pr(|t_2| > c) = \Pr(|Z| > c)$ . Поэтому из выражения (7.23) следует, что если нулевая гипотеза справедлива, то в больших выборках вероятность того, что критерий с использованием отдельных  $t$ -статистик отклоняет нулевую гипотезу, равна:

$$\Pr_{H_0} (\text{вероятность отверждения индивидуальной гипотезы}) \leq 2 \Pr(|Z| > c). \quad (7.24)$$

Неравенство в выражении (7.24) дает возможность выбрать критическое значение таким образом, что вероятность отклонения нулевой гипотезы, если она справедлива, равна желаемому уровню значимости. Подход Бонферрони может быть расширен на случай более чем двух коэффициентов; если нулевая гипотеза состоит из  $q$  ограничений, то умножение на 2 в правой части выражения (7.24) заменяется умножением на  $q$ .

Таблица 7.2

**Критические значения  $c$  для критерия Бонферрони,  
использующего индивидуальные  $t$ -статистики для проверки совместной гипотезы**

Число ограничений ( $q$ )	Уровень значимости		
	10 %	5 %	1 %
2	1,960	2,241	2,807
3	2,128	2,394	2,935
4	2,241	2,498	3,023

В таблице (7.2) приведены критические значения критерия Бонферрони для разных уровней значимости и  $q=2,3$  и  $4$ . Например, предположим, что желаемый уровень значимости равен 5 % и  $q=2$ . В соответствии с таблицей 7.2, критическое значение  $c$  равно 2,241. Данное критическое значение является 1,25-м процентилем стандартного нормального распределения, так что  $\Pr(|Z|=2,241)=2,5\%$ . Таким образом, выражение (7.24) показывает, что если нулевая гипотеза справедлива, то в больших выборках критерий из выражения (7.22) будет отвергать ее не более, чем в 5 % случаях.

Критические значения, представленные в таблице 7.2, больше, чем критические значения для проверки одного ограничения. Например, для  $q=2$  критерий Бонферрони отвергает нулевую гипотезу, если хотя бы одна  $t$ -статистика превышает 2,241 по абсолютной величине. Это критическое значение больше, чем 1,96, потому что оно учитывает тот факт, что, проверяя две  $t$ -статистики, вы получаете второй шанс отклонить совместную нулевую гипотезу (о чем уже говорилось в разделе 7.2).

Если отдельные  $t$ -статистики основаны на устойчивых к гетероскедастичности стандартных ошибках, то критерий Бонферрони корректен вне зависимости от того, есть ли гетероскедастичность или нет; однако если  $t$ -статистики рассчитываются исходя из предположения гомоскедастичности ошибок регрессии, то критерий Бонферрони корректен только при условии наличия гомоскедастичности ошибок регрессии.

### **Применение к результатам тестов**

Используя оценки регрессии (7.6), мы можем посчитать  $t$ -статистики для проверки совместной нулевой гипотезы о том, что истинные коэффициенты при соотношении учеников и учителей и расходов на одного ученика равны нулю. Эти  $t$ -статистики равны, соответственно,  $t_1=-0,60$  и  $t_2=2,43$ . Хотя и  $|t_1|<2,241$ ,

но поскольку  $|t_2| > 2,241$ , то в соответствии с критерием Бонферрони мы можем отклонить эту совместную нулевую гипотезу на 5%-м уровне значимости. Однако обе  $t$ -статистики,  $t_1$  и  $t_2$ , меньше по абсолютному значению, чем 2,807, поэтому, используя критерий Бонферрони, мы не можем отклонить данную нулевую гипотезу на 1 %-м уровне значимости. В противоположность этому в разделе 7.2 мы отклонили эту гипотезу на 1 %-м уровне значимости, используя  $F$ -статистику.

### **Приложение 7.2. Независимость условного среднего**

В данном приложении показано, что при выполнении условия независимости условного среднего, введенного в разделе 7.5 [уравнение (7.20)], МНК-оценки коэффициентов при изучаемых переменных являются несмещенными, но не перед контрольными переменными.

Рассмотрим регрессию с двумя регрессорами  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ . Если  $E(u_i | X_{1i}, X_{2i}) = 0$  – что будет выполнено, если значения  $X_{1i}$  и  $X_{2i}$  получены случайным образом в рамках эксперимента, – то МНК-оценки  $\hat{\beta}_1$  и  $\hat{\beta}_2$  являются несмещенными оценками коэффициентов  $\beta_1$  и  $\beta_2$ .

Теперь предположим, что  $X_{1i}$  является переменной, чье влияние изучается, а  $X_{2i}$  – контрольная переменная, которая коррелирована с пропущенными факторами (с ошибкой). Хотя предположение о равенстве нулю условного среднего тогда не выполняется, предположим, однако, что выполнено предположение о независимости условного среднего, то есть что  $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ . Также для удобства предположим, что  $E(u_i | X_{2i})$  линейно зависит от  $X_{2i}$ , так что  $E(u_i | X_{2i}) = \gamma_0 + \gamma_2 X_{2i}$ , где  $\gamma_0$  и  $\gamma_2$  являются константами (это предположение о линейности обсуждается ниже). Определим  $v_i$  как разность между  $u_i$  и условным математическим ожиданием  $u_i$  относительно  $X_{1i}$  и  $X_{2i}$ , то есть  $v_i = u_i - E(u_i | X_{1i}, X_{2i})$ . Тогда условное среднее  $v_i$  относительно  $X_{1i}$  и  $X_{2i}$  равно нулю:  $E(v_i | X_{1i}, X_{2i}) = E[(u_i - E(u_i | X_{1i}, X_{2i})) | X_{1i}, X_{2i}] = E(u_i | X_{1i}, X_{2i}) - E(u_i | X_{1i}, X_{2i}) = 0$ . Отсюда:

$$\begin{aligned}
 Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i = \\
 &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E(u_i | X_{1i}, X_{2i}) + v_i = (\text{используя определение } v_i) \\
 &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E(u_i | X_{2i}) + v_i = (\text{используя независимость условного среднего}) \\
 &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + (\gamma_0 + \gamma_2 X_{2i}) + v_i = [\text{используя линейность } E(u_i | X_{2i})] \\
 &= (\beta_0 + \gamma_0) + \beta_1 X_{1i} + (\beta_2 + \gamma_2) X_{2i} + v_i = (\text{общие факторы}) \\
 &= \delta_0 + \beta_1 X_{1i} + \delta_2 X_{2i} + v_i,
 \end{aligned} \tag{7.25}$$

где  $\delta_0 = \beta_0 + \gamma_0$  и  $\delta_2 = \beta_2 + \gamma_2$ .

Условное среднее ошибки  $v_i$  в выражении (7.25) равно нулю; то есть  $E(v_i | X_{1i}, X_{2i}) = 0$ . Следовательно, первое предположение МНК для множественной регрессии выполняется для последней строки выражения (7.25),

и если выполнены и другие три предположения МНК для множественной регрессии, то МНК-регрессия  $Y_i$  на константу,  $X_{1i}$  и  $X_{2i}$  даст несмешенную и состоятельную оценку коэффициентов  $\delta_0$ ,  $\beta_1$  и  $\beta_2$ . Поэтому МНК-оценка коэффициента при  $X_{1i}$  является несмешенной оценкой влияния этой переменной на независимую переменную. Однако МНК-оценка коэффициента перед  $X_{2i}$  является смещенной оценкой  $\beta_2$  и, на самом деле, оценивает сумму коэффициентов  $\beta_2$  и  $\gamma_2$ , последний из которых возникает из-за корреляции между контрольной переменной  $X_{2i}$  с исходной ошибкой  $u_i$ .

Вывод выражения (7.25) верен для любого значения  $\beta_2$ , включая ноль. Переменная  $X_{2i}$  является полезной контрольной переменной, если выполняется условие независимости условного среднего; она не должна непосредственно влиять на  $Y_i$ .

В четвертой строке выражения (7.25) используется предположение о том, что  $E(u_i | X_{2i})$  линейно зависит от  $X_{2i}$ . Как уже говорилось в разделе (2.4), это условие выполняется, если  $u_i$  и  $X_{2i}$  совместно нормально распределены. Предположение о линейности может быть ослаблено путем использования методов, которые будут обсуждаться в главе 8. В упражнении 18.9 повторяется вывод выражения (7.25) для нелинейных условных математических ожиданий, нескольких изучаемых переменных и нескольких контрольных переменных.

В терминах примера из раздела 7.5 [регрессия (7.19)], если  $X_{2i}$  – это *LchPct*, то  $\beta_2$  измеряет причинно-следственное влияние программы субсидированных обедов (коэффициент  $\beta_2$  положителен, если эта программа улучшает результаты тестов);  $\gamma_2$  является отрицательным, так как переменная *LchPct* отрицательно коррелирована с пропущенными характеристиками дополнительных возможностей обучения, которые улучшают результаты тестов, и  $\delta_2 = \beta_2 + \gamma_2$  будет отрицательным, если вклад смещения из-за пропущенных переменных (т.е.  $\gamma_2$ ) не компенсируется положительным эффектом от субсидированных обедов (т.е.  $\beta_2$ ).

Чтобы лучше понять предположение о независимости условного среднего, обратимся снова к случайному контролируемому эксперименту. Как уже говорилось в разделе 4.4, если  $X_{1i}$  выбирается случайным образом, то в регрессии  $Y_i$  на  $X_{1i}$  выполняется предположение о равенстве нулю условного среднего. Однако если значение  $X_{1i}$  выбирается случайным образом условно относительно другой переменной  $X_{2i}$ , тогда выполняется предположение о независимости условного среднего; при этом если переменная  $X_{2i}$  коррелирована с  $u_i$ , то предположение о равенстве нулю условного среднего не выполняется. Например, рассмотрим эксперимент по изучению того, как обязательные домашние задания влияют на оценки по эконометрике по сравнению с необязательными домашними заданиями. Из числа студентов-экономистов ( $X_{2i} = 1$ ) 75 % помещены в исследуемую группу (обязательные домашние задания:  $X_{1i} = 1$ ), в то время как из числа студентов, специализирующихся не в экономике, только 25 % помещаются в контрольную группу. Поскольку как студенты-экономисты, так и студенты других специальностей отбираются в исследуемую группу случайнным образом, то  $u_i$  не зависит от  $X_{1i}$  при данном  $X_{2i}$  и, следовательно,  $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ . Если выбор специальности коррелирует с другими характеристиками (такими как, например, способности к математике),

которые определяют, насколько хорошо студент справляется с курсом эконометрики, тогда  $E(u_i | X_{2i}) \neq 0$ . Кроме того, в этом случае оценки регрессии результатов итогового экзамена ( $Y_i$ ) только на  $X_{1i}$  будут смещены ( $X_{1i}$  коррелирует с выбором специальности и, поэтому, с другими пропущенными факторами, влияющими на итоговый результат). Включение специальности ( $X_{2i}$ ) в регрессию устраняет это смещение из-за пропущенных переменных (попадание в ис следуемую группу случайно для экономистов/неэкономистов), и поэтому МНК-оценка коэффициента при  $X_{1i}$  будет несмешенной оценкой того, как обязательные домашние задания влияют на оценки по эконометрике. Однако МНК-оценка коэффициента при  $X_{2i}$  не является несмешенной оценкой того, как смена специальности на экономику влияет на оценки по эконометрике. Это происходит потому, что специальность выбирается неслучайным образом и коррелирует с другими пропущенными факторами, которые не могут измениться (как, например, способности к математике), если студент сменит специальность.

# Глава 8. Нелинейные регрессионные модели

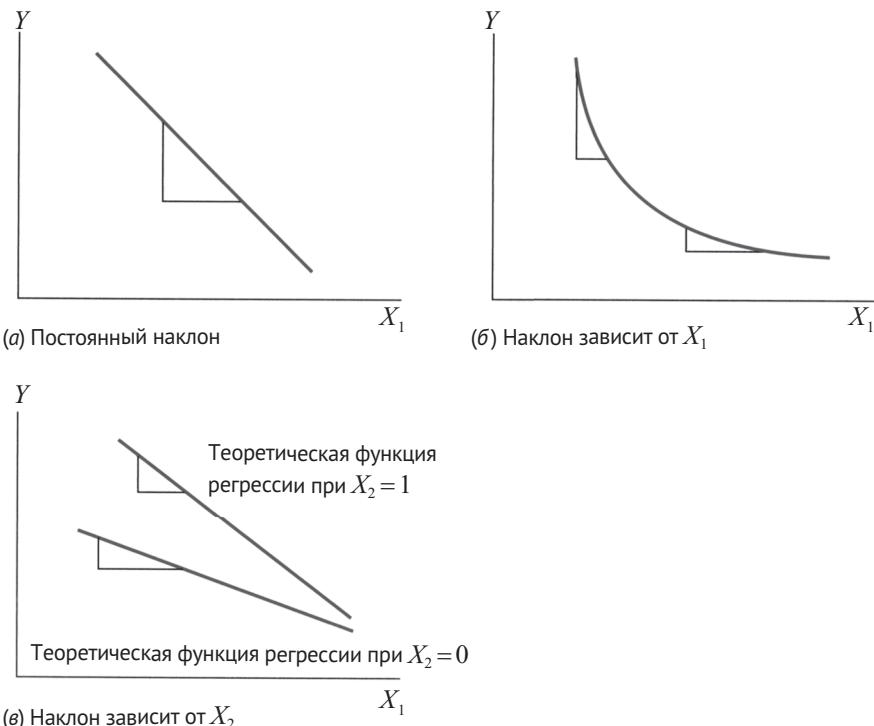
В главах 4–7 предполагалось, что теоретическая функция регрессии является линейной. Другими словами, угловой коэффициент теоретической функции регрессии предполагался равным константе, то есть коэффициент, характеризующий влияние единичного изменения  $X$  на  $Y$ , не зависел от значения  $X$ . Но что случится, если изменение  $Y$  в зависимости от изменения  $X$  будет зависеть от значения одной или более объясняющих переменных? Если это так, теоретическая функция регрессии не будет линейной.

В данной главе рассматриваются две группы методов, используемых для моделирования теоретической функции нелинейной регрессии. Методы из первой группы применяются, когда эффект влияния на  $Y$  от изменения в одной независимой переменной,  $X_1$ , зависит от значения этой переменной ( $X_1$ ). Например, уменьшение соотношения учеников и учителей на единицу может иметь больший эффект, если размеры класса малы, чем в ситуации, когда классы настолько велики, что учитель в основном занят тем, что держит класс под контролем. Если это так, результаты тестов ( $Y$ ) являются нелинейной по соотношению учеников и учителей ( $X_1$ ) функцией, и эта функция круче, когда  $X_1$  мал. Пример нелинейной функции регрессии с такой особенностью приведен на рисунке 8.1. В то время как линейная теоретическая функция регрессии на рисунке 8.1а имеет постоянный угловой коэффициент, угловой коэффициент нелинейной теоретической функции регрессии на рисунке 8.1б меняется в зависимости от  $X_1$ : он больше, когда  $X_1$  мал. Первая группа методов рассматривается в разделе 8.2.

Вторая группа методов рассматривает ситуации, когда эффект влияния на  $Y$  от изменения  $X_1$  зависит от значения другой независимой переменной, скажем,  $X_2$ . Например, ученики, все еще изучающие английский язык, могут начать лучше учиться, если учителя будут уделять им больше внимания; если это так, эффект влияния на результаты тестов от уменьшения соотношения учеников и учителей будет выше в округах с большим количеством учеников, все еще изучающих английский язык, чем в округах с малым количеством изучающих английский язык школьников. В этом примере эффект влияния на результаты тестов ( $Y$ ) от уменьшения соотношения учеников и учителей ( $X_1$ ) зависит от процента изучающих английский язык в округе ( $X_2$ ). Как показано на рисунке 8.1в, угловой коэффициент такого типа теоретических функций регрессии зависит от значения  $X_2$ . Эта вторая группа методов представлена в разделе 8.3.

В моделях в разделах 8.2 и 8.3 теоретическая функция регрессии является нелинейной функцией от независимых переменных; то есть условное

математическое ожидание  $E(Y_i|X_{1i}, \dots, X_{ki})$  является нелинейной функцией от одного или более  $X$ -ов. Несмотря на то что функция нелинейна по  $X$ -ам, такие модели являются линейными функциями от неизвестных коэффициентов (или параметров) теоретической модели регрессии и, таким образом, являются версиями модели множественной регрессии, рассмотренной в главах 6 и 7. Следовательно, неизвестные параметры этих нелинейных функций регрессии могут быть оценены и тестиированы, используя МНК и методы из глав 6 и 7.



**Рисунок 8.1. Теоретические функции регрессии с различными угловыми коэффициентами**

На рисунке 8.1а теоретическая функция регрессии имеет постоянный угловой коэффициент. На рисунке 8.1б угловой коэффициент теоретической функции регрессии зависит от значения  $X_1$ . На рисунке 8.1в угловой коэффициент теоретической функции регрессии зависит от значения  $X_2$ .

В разделах 8.1 и 8.2 вводится в рассмотрение нелинейная функция парной регрессии, а в разделе 8.3 эта модель расширяется до случая двух независимых переменных. Чтобы не усложнять изложение, дополнительные контрольные переменные опускаются в эмпирических примерах в разделах 8.1–8.3, однако важно помнить о необходимости включения в нелинейную функцию регрессии переменных, которые учитывают пропущенные факторы, включая, в том числе, и контрольные переменные. В разделе 8.4 мы применяем модели нелинейных регрессий к примеру о влиянии соотношения учеников и учителей на результаты тестов, учитывая дополнительные контрольные переменные. В некоторых

эмпирических приложениях функция регрессии нелинейна по  $X$ -ам и по параметрам. Если это так, параметры регрессии не могут быть оценены МНК, но могут быть оценены нелинейным методом наименьших квадратов. В приложении 8.1 предоставлены примеры таких функций и описана оценка нелинейного метода наименьших квадратов.

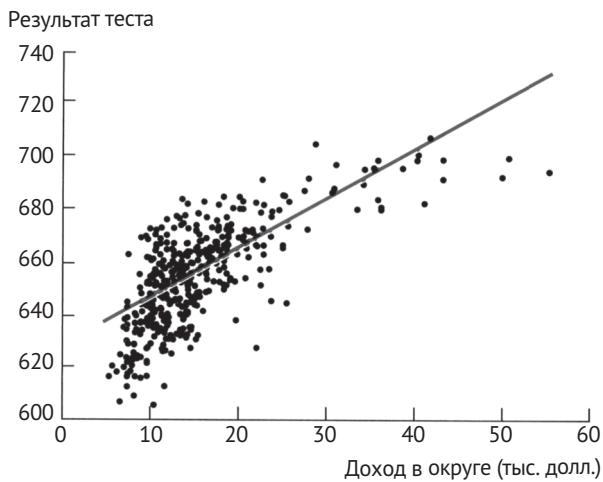
## 8.1. Общая стратегия моделирования функции нелинейной регрессии

В разделе излагается общая стратегия моделирования нелинейной теоретической функции регрессии. С этой точки зрения нелинейные модели являются обобщениями линейной модели множественной регрессии и, следовательно, могут быть оценены и тестиированы с использованием инструментария глав 6 и 7. Сначала, однако, мы возвращаемся к данным по результатам тестов в Калифорнии и рассматриваем соотношение между результатами тестов и доходом округа.

### *Результаты тестов и доход округа*

В главе 7 мы выяснили, что характеристики экономического положения учеников являются важными факторами в объяснении результатов стандартизованных тестов. Наш анализ включал две характеристики экономического положения (процент учеников, получающих субсидированные обеды, и процент семей в округе, получающих помощь в качестве малообеспеченных), чтобы измерить долю учеников в округе, выходцев из малообеспеченных семей. Другая, более широкая характеристика экономического положения – это среднегодовой доход на душу населения в школьном округе. База данных по школьным округам Калифорнии включает доход в округе, измеренный в тысячах долларов 1998 года. Разброс подушевых доходов в базе данных довольно высок: для 420 округов в нашей выборке медианный округ составляет 13,7 (т.е. 13 700 долл. на человека), и он варьируется с 5,3 (5300 долл. на человека) до 55,3 (т.е. 55 300 долл. на человека).

На рисунке 8.2 приведена диаграмма рассеяния результатов тестов пятиклассников и подушевых доходов в школьных округах Калифорнии вместе с линией МНК-регрессии, включающей эти две объясняющие переменные. Результаты тестов и средний доход сильно положительно коррелированы с коэффициентом корреляции, равным 0,71; ученики из богатых округов имеют более высокие результаты тестов, чем ученики из бедных округов. Но диаграмма рассеяния имеет одну особенность: мы видим много точек, расположенных ниже линии МНК, когда доход очень низкий (ниже 10 тыс. долл.) или очень высокий (выше 40 тыс.) долл., но есть много точек, расположенных выше нее, если доход расположен между 15 тыс. и 30 тыс. долл. Кажется, что линия регрессии несколько искривлена, и эта особенность не учитывается в модели линейной регрессии.



**Рисунок 8.2. Диаграмма рассеяния результатов тестов и подушевых доходов в школьных округах Калифорнии и линия линейной МНК-регрессии**

Существует положительная корреляция между результатами тестов и доходами в округе (корреляция = 0,71), но линия линейной МНК-регрессии неадекватно описывает соотношение между этими переменными.

Говоря кратко, из диаграммы рассеяния видно, что соотношение между доходами в округе и результатами тестов не является прямой линией. Скорее, оно является нелинейным. Нелинейная функция – это функция с коэффициентом наклона, не являющимся константой: функция  $f(X)$  линейна, если угловой коэффициент  $f'(X)$  одинаков для всех значений  $X$ , но если угловой коэффициент зависит от значения  $X$ , тогда  $f(X)$  нелинейна.

Что мы можем сделать, если прямая линия неадекватно описывает соотношение между доходами в округе и результатами тестов? Представьте, что мы пытаемся начертить кривую, которая соответствует точкам на рисунке 8.2. Эта кривая будет иметь более крутой наклон для округов с низкими значениями доходов, а затем будет выравниваться с увеличением доходов в округе. Один из способов аппроксимировать такую кривую – это моделировать соотношение как квадратичную функцию. То есть мы могли бы моделировать результаты тестов как функцию от дохода и квадрата дохода.

Модель квадратичной теоретической регрессии, описывающая влияние подушевых доходов на результаты тестов, формулируется следующим образом:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i, \quad (8.1)$$

где  $\beta_0$ ,  $\beta_1$  и  $\beta_2$  являются коэффициентами,  $Income_i$  – доход в  $i$ -ом округе,  $Income_i^2$  – квадрат дохода в  $i$ -ом округе и  $u_i$  является компонентой ошибок, которая, как обычно, представляет все другие неучтенные факторы, определяющие результаты тестов. Уравнение (8.1) называется моделью квадратичной ре-

грессии, поскольку теоретическая функция регрессии,  $E(\text{TestScore}_i | \text{Income}_i) = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Income}_i^2$ , является квадратичной функцией от независимой переменной  $\text{Income}$ .

Если бы вы знали теоретические коэффициенты  $\beta_0$ ,  $\beta_1$  и  $\beta_2$  в модели (8.1), вы могли бы предсказать средний результат теста в округе на основе информации о среднем доходе в нем. Но эти коэффициенты теоретической модели неизвестны и, следовательно, должны быть оценены, используя выборку данных.

Сначала может показаться, что сложно найти коэффициенты квадратичной функции, которые наилучшим образом описывают данные, изображенные на рисунке 8.2. Если вы сравните модель (8.1) с моделью множественной регрессии из вставки «Основные понятия 6.2», вы увидите, однако, что модель (8.1) фактически представляет собой версию модели множественной регрессии с двумя объясняющими переменными: первая – это  $\text{Income}$ , а вторая –  $\text{Income}^2$ . Вы можете создать второй регрессор механически, генерируя новую переменную, которая равна квадрату  $\text{Income}$ , например, как дополнительный столбец в таблице. Таким образом, после определения регрессоров как  $\text{Income}$  и  $\text{Income}^2$  нелинейная модель (8.1) представляет собой модель множественной регрессии с двумя регрессорами!

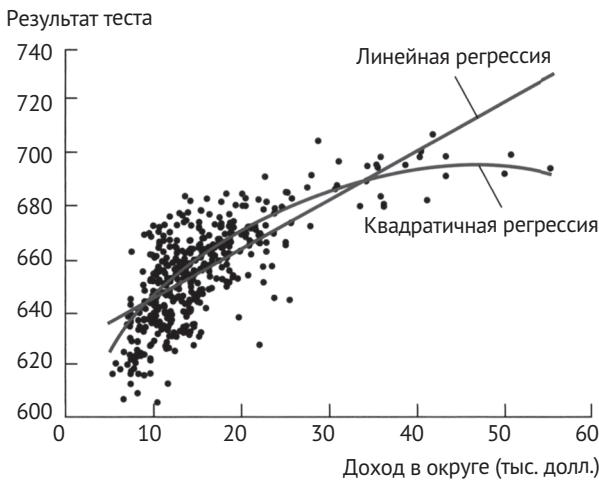
Поскольку модель квадратичной регрессии является вариантом множественной регрессии, ее неизвестные теоретические коэффициенты могут быть оценены и проверены, используя методы МНК, описанные в главах 6 и 7. МНК-оценки модели регрессии (8.1) для 420 наблюдений, изображенных на рисунке 8.2, выглядят следующим образом:

$$\widehat{\text{TestScore}} = 607,3 + 3,85 \text{Income} - 0,0423 \text{Income}^2, \bar{R}^2 = 0,554, \quad (8.2)$$

(2,9)                    (0,27)                    (0,0048)

где, как и обычно, стандартные ошибки оцененных коэффициентов приведены в скобках. Оцененная функция регрессии уравнения (8.2) изображена на рисунке 8.3, наложенная на диаграмму рассеяния данных. Квадратичная функция покрывает кривизну в диаграмме рассеяния: она крутая для низких значений дохода в округе, но выравнивается, когда доход в округе высокий. Короче говоря, квадратичная регрессионная функция лучше соответствует данным, чем линейная.

Мы можем пойти на один шаг дальше этого визуального сравнения и формально проверить гипотезу, что соотношение между доходом и результатами теста является линейным против альтернативы, что она нелинейна. Если соотношение линейно, тогда регрессионная функция корректно специфицирована, за исключением того, что регрессор  $\text{Income}^2$  отсутствует, то есть если соотношение линейно, тогда уравнение (8.1) выполняется с  $\beta_2 = 0$ . Таким образом, мы можем проверить нулевую гипотезу о том, что регрессионная функция генеральной совокупности линейна против альтернативы, которая квадратична, на основе проверки нулевой гипотезы о том, что  $\beta_2 = 0$  против альтернативы, что  $\beta_2 \neq 0$ .



**Рисунок 8.3. Диаграмма рассеяния результатов тестов и подушевых доходов в школьных округах Калифорнии с линиями линейной и квадратичной МНК-регрессий**

Квадратичная МНК-регрессии лучше соответствует данным, чем линейная МНК-регрессия.

Поскольку модель (8.1) является вариантом модели множественной регрессии, нулевая гипотеза о том, что  $\beta_2 = 0$ , может быть проверена при помощи стандартной  $t$ -статистики, то есть  $t = (\hat{\beta}_2 - 0) / SE(\hat{\beta}_2)$ . Для МНК-оценок из уравнения (8.2) эта  $t$ -статистика равна  $t = -0,0423 / 0,0048 = -8,81$ . Рассчитанное значение тестовой статистики по абсолютному значению превышает 5 %-е критическое значение (которое равно 1,96). Действительно,  $p$ -значение для  $t$ -статистики меньше, чем 0,01%, поэтому мы отвергаем нулевую гипотезу о том, что  $\beta_2 = 0$  на всех разумных уровнях значимости. Таким образом, формальное тестирующее гипотезы не противоречит нашей неформальной проверке на рисунках 8.2 и 8.3: квадратичная модель соответствует данным лучше, чем линейная модель.

### **Влияние на $Y$ единичного изменения переменной $X$ в нелинейной модели**

Отложим в сторону пример с результатами тестов и рассмотрим общие проблемы. Вы хотите знать, как изменится зависимая переменная  $Y$ , если независимая переменная  $X_1$  изменится на величину  $\Delta X_1$ , а остальные объясняющие переменные  $X_2, \dots, X_k$  не изменятся. Если теоретическая функция регрессии линейна, этот эффект вычислить легко: как показано в выражении (6.4), ожидаемое изменение  $Y$  есть  $\Delta Y = \beta_1 \Delta X_1$ , где  $\beta_1$  – теоретический коэффициент соответствующей регрессии, умноженный на  $\Delta X_1$ . Однако если функция регрессии нелинейна, ожидаемое изменение  $Y$  вычислить сложнее, поскольку оно может зависеть от значений независимых переменных.

**Общий вид функции нелинейной теоретической регрессии<sup>1</sup>.** Модели нелинейных регрессий, рассматриваемые в этой главе, имеют вид:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i, \quad i = 1, \dots, n, \quad (8.3)$$

где  $f(X_{1i}, X_{2i}, \dots, X_{ki})$  – теоретическая функция *нелинейной регрессии*, то есть (возможно) нелинейная функция от независимых переменных  $X_{1i}, X_{2i}, \dots, X_{ki}$ , а  $u_i$  является компонентой ошибок. Например, в модели квадратичной регрессии (8.1) присутствует только одна независимая переменная, поэтому  $X_1 = Income$ , и теоретическая функция регрессии имеет вид:  $f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$ .

Поскольку теоретическая функция регрессии представляет собой условное математическое ожидание  $Y_i$  относительно  $X_{1i}, X_{2i}, \dots, X_{ki}$ , то в выражении (8.3) мы допускаем возможность, что это условное математическое ожидание является нелинейной функцией от  $X_{1i}, X_{2i}, \dots, X_{ki}$ ; то есть  $E(Y_i | X_{1i}, X_{2i}, \dots, X_{ki}) = f(X_{1i}, X_{2i}, \dots, X_{ki})$ , где  $f$  может быть нелинейной функцией. Если теоретическая функция регрессии является линейной, то  $f(X_{1i}, X_{2i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ , и выражение (8.3) становится моделью линейной регрессии, которая описана во вставке «Основные понятия 6.2». Однако выражение (8.3) допускает также нелинейную функцию.

**Величина изменения  $Y$  в зависимости от изменения  $X_1$ .** Как обсуждалось в разделе 6.2, величина изменения  $Y$  при изменении  $X_1$  на  $\Delta X_1$  и постоянстве  $X_2, \dots, X_k$  представляет собой разность между ожидаемым значением  $Y$  в предположении, что независимые переменные принимают значения  $X_1 + \Delta X_1, X_2, \dots, X_k$ , и ожидаемым значением  $Y$  в предположении, что независимые переменные принимают значения  $X_1, X_2, \dots, X_k$ . Разность между этими двумя ожидаемыми значениями, назовем ее  $\Delta Y$ , характеризует изменение  $Y$  в среднем в генеральной совокупности, если  $X_1$  изменится на значение  $\Delta X_1$ , а остальные переменные  $X_2, \dots, X_k$  останутся постоянными. В модели нелинейной регрессии (8.3) величина изменения  $Y$  будет равна  $\Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) - f(X_1, X_2, \dots, X_k)$ .

Поскольку функция регрессии  $f$  неизвестна, изменение  $Y$  в генеральной совокупности в зависимости от изменения  $X_1$  также неизвестно. Чтобы получить оценку такого изменения, сначала нужно получить оценку теоретической функции регрессии. В общем случае, обозначим эту оцененную функцию через  $\hat{f}$ ; пример такой оцененной функции есть оцененная в выражении (8.2) квадратичная регрессия. Оцененное изменение  $Y$  (обозначенное  $\hat{\Delta}Y$ ) в зависимости от изменения  $X_1$  – разность между предсказанным значением  $Y$ , когда неза-

<sup>1</sup> Термин «нелинейная регрессия» применяется для двух концептуально различных классов моделей. В первом классе функция теоретической регрессии является нелинейной функцией по объясняющим переменным (по  $X$ -ам), но является линейной функцией по неизвестным параметрам ( $\beta$ -ам). Во втором – функция теоретической регрессии является нелинейной функцией по неизвестным параметрам и может быть или может не быть нелинейной функцией по  $X$ -ам. В этой главе мы рассматриваем модели из первого класса. В приложении 8.1 рассматриваются модели из второго класса.

висимая переменная принимает значения  $X_1 + \Delta X_1, X_2, \dots, X_k$ , и предсказанным значением  $Y$ , когда независимая переменная принимает значения  $X_1, X_2, \dots, X_k$ .

## ОСНОВНЫЕ ПОНЯТИЯ

### 8.1

#### Ожидаемое изменение $Y$ в зависимости от изменения $X_1$ в модели нелинейной регрессии

Ожидаемое изменение  $Y, \Delta Y$ , являющееся следствием изменения  $X_1, \Delta X_1$ , при постоянстве  $X_2, \dots, X_k$ , представляет собой разность между значениями теоретической функции регрессии до и после изменения  $X_1$  при неизменности переменных  $X_2, \dots, X_k$ . То есть ожидаемое изменение  $Y$  равно:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) - f(X_1, X_2, \dots, X_k). \quad (8.4)$$

Оценка этой неизвестной теоретической разности представляет собой разность между предсказанными значениями для этих двух случаев. Пусть  $\hat{f}(X_1, X_2, \dots, X_k)$  будет предсказанным значением  $Y$ , полученным на основе оценки  $\hat{f}$  теоретической функции регрессии. Тогда предсказанное изменение  $Y$  равно

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \dots, X_k) - \hat{f}(X_1, X_2, \dots, X_k). \quad (8.5)$$

Метод вычисления ожидаемого изменения  $Y$  в зависимости от изменения  $X_1$  представлен во вставке «Основные понятия 8.1». Описанный метод работает вне зависимости от того, мало или велико  $\Delta X_1$ , и вне зависимости от того, непрерывны ли регрессоры или дискретны. В приложении 8.2 показывается, как, используя дифференциальное исчисление, можно оценить угловой коэффициент в специальном случае единственного непрерывного регрессора с малым изменением  $\Delta X_1$ .

**Пример: результаты тестов и доход.** Чему будет равно предсказанное изменение результатов тестов, связанное с изменением дохода в округе на 1000 долл., рассчитанное на основе оценки функции квадратичной регрессии (8.2)? Так как функция регрессии является квадратичной, этот эффект зависит от начального дохода в округе. Поэтому мы рассмотрим случай единственного непрерывного регрессора с малым  $\Delta X_1$ .

Чтобы вычислить изменение  $\Delta \hat{Y}$ , связанное с изменением в доходе с 10 до 11 (с 10 тыс. до 11 тыс. долл. на душу населения в округе), мы можем применить общую формулу из выражения (8.5) к модели квадратичной регрессии. Получим:

$$\Delta \hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2), \quad (8.6)$$

где  $\hat{\beta}_0, \hat{\beta}_1$  и  $\hat{\beta}_2$  являются МНК-оценками нелинейной регрессии.

Компонента в первой скобке в равенстве (8.6) – предсказанное значение  $Y$ , когда  $Income = 11$ , а компонента во второй скобке – предсказанное значение  $Y$ , когда  $Income = 10$ . Эти предсказанные значения вычислены с использованием МНК-оценок коэффициентов в регрессии (8.2). Соответственно, когда  $Income = 10$ ,

предсказанное значение результатов теста составляет  $607,3 + 3,85 \times 10 - 0,0423 \times 10^2 = 641,57$ . Когда  $Income = 14$ , предсказанное значение составляет  $607,3 + 3,85 \times 11 - 0,0423 \times 11^2 = 644,53$ . Разность между двумя предсказанными значениями составляет  $\Delta\hat{Y} = 644,53 - 641,57 = 2,96$  пункта, то есть предсказанная разность результатов тестов между округом со средним доходом в 11 тыс. долл. и округом со средним доходом в 10 тыс. долл. составляет 2,96 пункта.

Во втором случае, когда доход изменяется с 40 тыс. до 41 тыс. долл., разность в предсказанных значениях в равенстве (8.6) равна:  $\Delta\hat{Y} = (607,3 + 3,85 \times 41 - 0,0423 \times 41^2) - (607,3 + 3,85 \times 40 - 0,0423 \times 40^2) = 694,04 - 693,62 = 0,42$  пункта. Таким образом, изменение в доходе на 1000 долл. вызывает более высокое изменение предсказанных результатов тестов, если начальный доход составляет 10 тыс. долл., чем если бы он составлял 40 тыс. долл. (предсказанное изменение 2,96 пункта против 0,42 пункта). Другими словами, угловой коэффициент оцененной функции квадратичной регрессии на рисунке 8.3 больше при низких значениях дохода (таких как 10 тыс. долл.), чем при высоких значениях дохода (как 40 тыс. долл.).

**Стандартная ошибка ожидаемого изменения.** Оценка изменения  $Y$  в зависимости от изменения  $X_1$ , связана с оценками теоретической функции регрессии,  $\hat{f}$ , которая, вообще говоря, изменяется в зависимости от выборки. Следовательно, оцененный эффект содержит ошибку, связанную с неопределенностью выборки. Одним из способов количественно измерить неопределенность выборки является вычисление доверительного интервала для истинного изменения  $Y$  в генеральной совокупности. Для этого мы должны вычислить стандартную ошибку  $\Delta\hat{Y}$  в выражении (8.5).

Если функция регрессии линейна, то стандартную ошибку  $\Delta\hat{Y}$  вычислить легко. В этом случае предсказанное изменение  $Y$  в зависимости от изменения  $X_1$  равно  $\hat{\beta}_1\Delta X_1$ , и поэтому 95 %-й доверительный интервал для этого предсказанного изменения есть  $\hat{\beta}_1\Delta X_1 \pm 1,96SE(\hat{\beta}_1)\Delta X_1$ .

В моделях нелинейных регрессий, рассматриваемых в этой главе, стандартная ошибка  $\Delta\hat{Y}$  может быть вычислена с использованием инструментария, введенного в разделе 7.3 для тестирования единственного ограничения, включающего несколько коэффициентов. Чтобы проиллюстрировать этот метод, рассмотрим оцененное изменение результатов тестов, связанное с изменением в доходах с 10 до 11 в регрессии (8.6). Это изменение равно:  $\Delta\hat{Y} = \hat{\beta}_1 \times (11 - 10) + \hat{\beta}_2 \times (11^2 - 10^2) = \hat{\beta}_1 + 21\hat{\beta}_2$ . Поэтому стандартная ошибка предсказанного изменения равна:

$$SE(\Delta\hat{Y}) = SE(\hat{\beta}_1 + 21\hat{\beta}_2). \quad (8.7)$$

Таким образом, если мы можем вычислить стандартную ошибку  $\hat{\beta}_1 + 21\hat{\beta}_2$ , то мы можем вычислить и стандартную ошибку  $\Delta\hat{Y}$ . Существуют два способа сделать это с помощью стандартных программных пакетов, которые соответствуют двум подходам из раздела 7.3, используемым для тестирования единственного ограничения на несколько коэффициентов модели.

Первый метод состоит в использовании «Подхода № 1» из раздела 7.3, который заключается в вычислении  $F$ -статистики для тестирования нулевой гипотезы о том, что  $\beta_1 + 21\beta_2 = 0$ . Стандартная ошибка  $\Delta\hat{Y}$  в этом случае задается так<sup>1</sup>:

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}. \quad (8.8)$$

Применяя все это для случая квадратичной регрессии (8.2), получаем, что  $F$ -статистика для тестирования нулевой гипотезы о том, что  $\beta_1 + 21\beta_2 = 0$ , равна  $F = 299,94$ . Поскольку  $\Delta\hat{Y} = 2,96$ , использование равенства (8.8) приводит к  $SE(\Delta\hat{Y}) = 2,96 / \sqrt{299,94} = 0,17$ . Таким образом, 95 %-й доверительный интервал для ожидаемого изменения  $Y$  равен  $2,96 \pm 1,96 \times 0,17$  или  $(2,63; 3,29)$ .

Второй метод заключается в использовании «Подхода № 2» из раздела 7.3, согласно которому нужно преобразовать регрессоры таким образом, чтобы в новой регрессии один из коэффициентов был  $\beta_1 + 21\beta_2$ . Выполнение этого преобразования мы оставляем в качестве упражнения (упражнение 8.9).

**Комментарий к интерпретации коэффициентов в нелинейной спецификации.** В модели множественной регрессии из глав 6 и 7 коэффициенты регрессии имели естественную интерпретацию. Например,  $\beta_1$  представляло собой ожидаемое изменение  $Y$ , связанное с изменением  $X_1$ , фиксируя остальные регрессоры постоянными. Но, как мы уже видели, такая интерпретация не всегда логична для случая нелинейной модели. Например, не имеет смысла думать о  $\beta_1$  в модели (8.1) как об эффекте от изменения дохода в округе, считая при этом квадрат дохода в округе постоянным. В нелинейных моделях функцию регрессии лучше всего интерпретировать, изображая ее графически и вычисляя предсказанное изменение  $Y$  в зависимости от изменения одной или более независимых переменных.

### Общий подход к моделированию нелинейности с использованием множественной регрессии

Общий подход к моделированию нелинейных регрессионных функций, принятый в данной главе, состоит из пяти шагов:

1. *Идентифицируйте возможное нелинейное соотношение.* Лучший способ сделать это – использовать свои знания из экономической теории и эмпирики, чтобы предложить возможное нелинейное соотношение. Еще до того как вы посмотрите на данные, спросите себя, может ли угловой коэффициент функции регрессии  $Y$  от  $X$  зависеть каким-либо разумным образом от значения  $X$  или любой другой объясняющей переменной? Почему такая нелинейная зависимость может существовать? Какую нелинейную форму она предполагает? Например, думая о динамике успехов класса 11-летних школьников, можно ли предположить, что сокращение размера класса

<sup>1</sup> Уравнение (8.8) можно вывести, заметив, что  $F$ -статистика равна квадрату  $t$ -статистики для проверки этой гипотезы, то есть  $F = t^2 = [(\hat{\beta}_1 + 21\hat{\beta}_2) / SE(\hat{\beta}_1 + 21\hat{\beta}_2)]^2 = [\Delta\hat{Y} / SE(\Delta\hat{Y})]^2$ , и, преобразовав последнее равенство так, получаем выражение для  $SE(\Delta\hat{Y})$ .

- с 18 учеников до 17 будет иметь большее влияние на зависимую переменную, чем его сокращение с 30 до 29 учеников.
2. *Специфицируйте нелинейную функцию и оцените ее параметры при помощи МНК.* В разделах 8.2 и 8.3 содержатся примеры разнообразных моделей нелинейных регрессий, которые могут быть оценены при помощи МНК. После изучения этих разделов вы поймете специфические особенности каждой из этих функций.
  3. *Определите, улучшает ли нелинейная модель линейную.* Просто тот факт, что вы думаете, что функция регрессии является нелинейной, не означает, что это так на самом деле! Вы должны определить эмпирически, соответствует ли ваша нелинейная модель данным. В большинстве случаев вы можете использовать  $t$ -статистики и  $F$ -статистики, чтобы тестировать нулевую гипотезу о том, что теоретическая функция регрессии линейна против альтернативы о том, что она нелинейна.
  4. *Изобразите графически оцененную функцию нелинейной регрессии.* Хорошо ли описывает данные оцененная модель регрессии? Посмотрите на рисунки 8.2 и 8.3, предполагающие, что квадратичная модель соответствует данным лучше, чем линейная модель.
  5. *Оцените изменение  $Y$  в зависимости от изменения  $X$ .* И наконец, последний шаг — используйте оцененную регрессию, чтобы вычислить изменение  $Y$  в зависимости от изменения одного или более регрессоров  $X$ , используя метод, описанный во вставке «Основные понятия 8.1».

## 8.2. Функции парных нелинейных регрессий

В данном разделе представлены два метода, используемые для моделирования нелинейных регрессий. Чтобы не усложнять изложение, мы рассматриваем эти методы для случая парных нелинейных регрессий, то есть для функций, которые включают только одну независимую переменную  $X$ . Тем не менее, как мы увидим в разделе 8.5, эти модели могут быть модифицированы для случая нескольких независимых переменных.

Первый метод, обсуждаемый в этом разделе, — это полиномиальная регрессия, обобщение квадратичной регрессии, использованной в предыдущем разделе в модели взаимосвязи между результатами тестов и доходами. Второй метод использует логарифмы  $X$ ,  $Y$  или  $X$  и  $Y$ . Несмотря на то что эти методы представлены раздельно, они могут использоваться в комбинации.

В приложении 8.2 модели из данного раздела рассматриваются с использованием дифференциального исчисления.

### Полиномы

Одной из простейших спецификаций функции нелинейной регрессии является полиномиальная функция от  $X$ . В общем случае, пусть  $r$  обозначает наибольшую степень  $X$ , которая включена в регрессию. Тогда модель полиномиальной регрессии степени  $r$  — это:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_r X_i^r + u_i . \quad (8.9)$$

Если  $r = 2$ , уравнение (8.9) является моделью квадратичной регрессии, обсуждаемой в разделе 8.1. Если  $r = 3$ , так, что наибольшая включенная степень  $X$  есть  $X^3$ , уравнение (8.9) называется *моделью кубической регрессии*.

Модель полиномиальной регрессии аналогична модели множественной регрессии из главы 6, за исключением того, что в главе 6 регрессоры были различными независимыми переменными, тогда как здесь регрессоры есть степени одной и той же независимой переменной  $X$ , то есть в данном случае регрессорами являются  $X, X^2, X^3$  и так далее. Таким образом, техника для оценивания и статистической проверки, разработанная для множественной регрессии, может применяться и здесь. В частности, неизвестные коэффициенты  $\beta_0, \beta_1, \dots, \beta_r$  в уравнении (8.9) могут быть оценены при помощи МНК-регрессии  $Y_i$  от  $X_i, X_i^2, \dots, X_i^r$ .

**Тестирование нулевой гипотезы о том, что теоретическая функция регрессии является линейной.** Если теоретическая функция регрессии линейна, то регрессоры, являющиеся степенями объясняющей переменной, начиная со второй и выше, не входят в нее. Следовательно, нулевая гипотеза ( $H_0$ ) о том, что функция регрессии линейна и соответствующая ей альтернативная гипотеза ( $H_1$ ) о том, что она является полиномом степени  $r$ , имеют вид:

$$\begin{aligned} H_0 : \beta_2 &= 0, \beta_3 = 0, \dots, \beta_r = 0 \text{ против} \\ H_1 : \text{по крайней мере один } \beta_j &\neq 0, j = 2, \dots, r . \end{aligned} \quad (8.10)$$

Нулевая гипотеза о том, что теоретическая функция регрессии линейна, может быть проверена против альтернативы, что она полиномиальная степени  $r$ , если проверять  $H_0$  против  $H_1$  из (8.10). Поскольку  $H_0$  является совместной нулевой гипотезой с  $q = r - 1$  ограничениями на коэффициенты теоретической модели полиномиальной регрессии, она может быть проверена с использованием  $F$ -статистики, как описано в разделе 7.2.

**Полином какой степени следует использовать?** То есть как много степеней объясняющей переменной  $X$  нужно включать в полиномиальную регрессию? Ответ представляет собой некий компромисс между гибкостью модели и ее статистической точностью. Увеличение степени  $r$  дает больше гибкости при выборе функции регрессии и позволяет ей лучшим образом отображать форму данных; полином степени  $r$  может иметь до  $r - 1$  изгиба (т.е. точек перегиба) на своем графике. Но увеличение  $r$  означает добавление большего числа регрессоров, которые могут уменьшить точность оцененных коэффициентов.

Таким образом, ответ на вопрос о количестве включаемых в модель регрессоров заключается в том, что их нужно включать в количестве, достаточном для адекватного моделирования функции нелинейной регрессии, но не более того. К сожалению, этот ответ является не очень полезным с практической точки зрения!

Практический способ определения степени полинома заключается в том, чтобы выяснить, являются ли в регрессии (8.9) нулевыми коэффициенты при

высоких степенях объясняющей переменной. Если это так, то эти регрессоры могут быть исключены из модели. Соответствующая процедура называется процедурой последовательного тестирования гипотез, так как индивидуальные гипотезы тестируются друг за другом, и заключается в следующем:

1. Выберите максимальное значение степени полинома  $r$  и оцените полиномиальную регрессию для этого  $r$ .
2. Используйте  $t$ -статистику для тестирования нулевой гипотезы о том, что коэффициент при  $X^r$  [ $\beta_r$  в регрессии (8.9)] равен нулю. Если вы отвергаете эту гипотезу, тогда  $X^r$  необходимо оставить в регрессии и использовать полином степени  $r$  для моделирования зависимости.
3. Если вы не отвергаете  $\beta_r = 0$  на шаге 2, исключите  $X^r$  из регрессии и оцените полиномиальную регрессию степени  $r-1$ . Проверьте, является ли коэффициент при  $X^{r-1}$  нулевым. Если гипотеза отвергается, используйте полином степени  $r-1$ .
4. Если вы не отвергаете  $\beta_{r-1} = 0$  на шаге 3, продолжайте эту процедуру до тех пор, пока коэффициент при наибольшей степени в вашем полиноме не будет статистически значимым.

В этой процедуре есть один недостающий элемент: информация о начальной степени полинома  $r$ . Во многих приложениях, включающих экономические данные, используются гладкие нелинейные функции, то есть функции, у которых нет резких скачков или пиков. Если это так, то соответствующим выбором будет небольшая максимальная степень для полинома – 2, 3 или 4, то есть необходимо начать со степени  $r = 2$  или 3 или 4 на шаге 1.

**Пример: Влияние средних доходов в округе на результаты тестов.** Рассмотрим оценку кубической регрессии, характеризующей влияние средних доходов в округе на результаты тестов:

$$\begin{aligned}\widehat{\text{TestScore}} &= 600,1 + 5,02 \frac{\text{Income}}{(5,1)} - 0,096 \frac{\text{Income}^2}{(0,029)} + 0,00069 \frac{\text{Income}^3}{(0,00035)}, \\ R^2 &= 0,555,\end{aligned}\tag{8.11}$$

$t$ -статистика для коэффициента при  $\text{Income}^3$  равна 1,97, поэтому нулевая гипотеза о том, что функция регрессии является квадратичной, отвергается против альтернативы о том, что она кубическая на 5 %-м уровне значимости. Кроме того,  $F$ -статистика для тестирования совместной нулевой гипотезы о том, что коэффициенты при  $\text{Income}^2$  и  $\text{Income}^3$  оба равны нулю, составляет 37,5, с  $p$ -значением менее чем 0,01 %, поэтому нулевая гипотеза о линейности функции регрессии отвергается против альтернативы о том, что она или квадратичная, или кубическая.

**Интерпретация коэффициентов в модели полиномиальной регрессии.** У коэффициентов полиномиальных регрессий нет простой интерпретации. Наилучший способ интерпретировать коэффициенты в полиномиальных регрессиях – нарисовать оцененную функцию регрессии и вычислить оцененное изменение  $Y$ , связанное с изменением  $X$  для одного или более значений  $X$ .

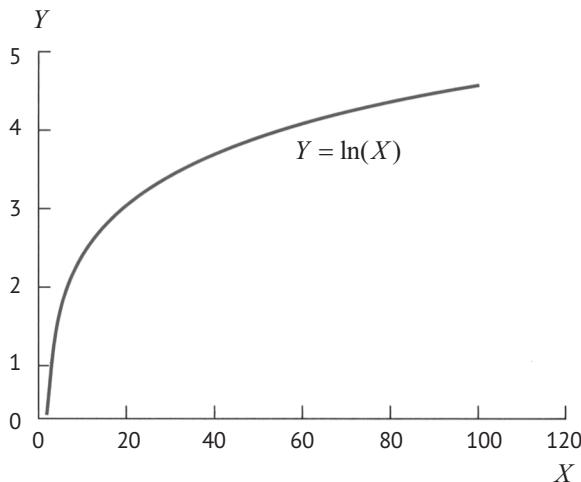
## Логарифмы

Существует еще один распространенный способ специфицировать функцию нелинейной регрессии: можно использовать натуральный логарифм  $Y$  или  $X$ . Логарифмы очень удобны с той точки зрения, что они конвертируют изменения в переменных в их процентные изменения, и многие соотношения естественным образом выражаются в терминах процентов. Рассмотрим несколько примеров:

- Во вставке из главы 3 «Гендерный разрыв в заработных платах выпускников колледжей в Соединенных Штатах» был рассмотрен пример о разрыве в зарплатах мужчин и женщин, являющихся выпускниками колледжей. В этом обсуждении разрыв в зарплате измерялся в терминах долларов. Однако было бы легче сравнивать этот разрыв в разрезе профессий и времени, если бы зарплаты выражались в процентах.
- В разделе 8.1 мы обнаружили, что зависимость между средним подушевым доходом в округе и результатами тестов нелинейна. Будет ли это соотношение линейным, если использовать процентные изменения переменных? То есть может ли быть, что изменение в доходе округа на 1 % – а не на 1 000 долл. – меняет результаты тестов практически одинаково для различных значений дохода?
- В экономическом анализе потребительского спроса часто предполагается, что увеличение цены на 1 % приводит к некоторому процентному снижению объема спроса. Процентное уменьшение потребительского спроса в результате 1 %-го роста цены называется ценовая эластичность.

Спецификация регрессии, использующая натуральный логарифм, позволяет оценивать процентные соотношения, аналогичные рассмотренным выше. Прежде чем ввести эти спецификации, рассмотрим экспоненциальную функцию и функцию натурального логарифма.

**Экспоненциальная функция и натуральный логарифм.** Экспоненциальная функция и обратная к ней – натуральный логарифм – играют важную роль в моделировании нелинейных функций. Экспоненциальная функция от  $x$  равна  $e^x$  (т.е. число  $e$ , возведенное в степень  $x$ ), где  $e$  является константой, равной 2,718 28...; экспоненциальная функция также часто записывается как  $\exp(x)$ . Натуральный логарифм является функцией, обратной к экспоненциальной, то есть натуральный логарифм – это функция, для которой  $x = \ln(e^x)$  или, что эквивалентно,  $x = \ln[\exp(x)]$ . Основание натурального логарифма –  $e$ . Хотя существуют логарифмы с другими основаниями, такими как 10 (десятичный логарифм), в этой книге мы рассмотрим только логарифмы по основанию  $e$ , то есть натуральные логарифмы, поэтому, когда мы используем термин «логарифм», всегда подразумеваем «натуральный логарифм».

**Рисунок 8.4. Логарифмическая функция  $Y = \ln(X)$** 

Логарифмическая функция  $Y = \ln(X)$  имеет более крутой наклон для маленьких значений  $X$  и меньший – для больших, определена только для  $X > 0$ , и ее наклон равен  $1/X$ .

Логарифмическая функция  $y = \ln(x)$  изображена на рисунке 8.4. Заметим, что она определена только для положительных значений  $x$ . Логарифмическая функция имеет крутой наклон в начале своей области определения, а затем выравнивается (хотя и продолжает возрастать). Наклон логарифмической функции  $\ln(x)$  равен  $1/X$ .

Отметим полезные свойства логарифмической функции:

$$\ln(1/x) = -\ln(x); \quad (8.12)$$

$$\ln(ax) = \ln(a) + \ln(x); \quad (8.13)$$

$$\ln(x/a) = \ln(x) - \ln(a) \text{ и} \quad (8.14)$$

$$\ln(x^a) = a \ln(x). \quad (8.15)$$

**Логарифм и проценты.** Связь между логарифмом и процентами следует из того, что когда  $\Delta x$  мало, разность между логарифмом  $x + \Delta x$  и логарифмом  $x$  приблизительно равна  $\Delta x/x$ , что является процентным изменением  $x$ , деленным на 100. То есть

$$\ln(x + \Delta x) - \ln(x) \cong \frac{\Delta x}{x} \quad (\text{когда } \frac{\Delta x}{x} \text{ мало}), \quad (8.16)$$

где « $\cong$ » означает «приблизительно равно». Вывести это приближенное равенство можно с использованием дифференциального исчисления, но его легко проиллюстрировать, перебирая различные значения  $x$  и  $\Delta x$ . Например, если  $x = 100$

и  $\Delta x = 1$ , тогда  $\Delta x / x = 1/100 = 0,01$  (или 1 %), в то время как  $\ln(x + \Delta x) - \ln(x) = \ln(101) - \ln(100) = 0,00995$  (или 0,995 %). Таким образом,  $\Delta x / x$  (равное 0,01) очень близко к  $\ln(x + \Delta x) - \ln(x)$  (которое равно 0,00995). Если  $\Delta x = 5$ ,  $\Delta x / x = 5/100 = 0,05$ , при этом  $\ln(x + \Delta x) - \ln(x) = \ln(105) - \ln(100) = 0,04879$ .

**Три модели логарифмических регрессий.** Существует три различных модели логарифмических регрессий: зависимость  $Y$  от логарифма  $X$ ; зависимость логарифма  $Y$  от  $X$ ; и зависимость логарифма  $Y$  от логарифма  $X$ . Интерпретация коэффициентов регрессии различна в каждой из рассматриваемых моделей. Обсудим эти три случая.

**Случай I: Зависимость  $Y$  от логарифма  $X$ .** В этом случае модель регрессии имеет вид:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i, \quad i = 1, \dots, n. \quad (8.17)$$

Поскольку здесь  $Y$  берется не в логарифмах, а  $X$  – в логарифмах, эту модель иногда называют линейно-логарифмической.

В линейно-логарифмической модели 1 %-е изменение  $X$  влечет за собой изменение  $Y$  на  $0,01\beta_1$ . Чтобы увидеть это, рассмотрим разность между теоретической функцией регрессии в точках  $X + \Delta X$  и  $X$ :

$$[\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1 [\ln(X + \Delta X) - \ln(X)] \cong \beta_1 (\Delta X / X),$$

где последнее приближенное равенство использует аппроксимацию из выражения (8.16). Если  $X$  изменяется на 1 %, тогда  $\Delta X / X = 0,01$ ; таким образом, в этой модели 1 % изменение в  $X$  влечет за собой изменение  $Y$  на  $0,01\beta_1$ .

Единственное различие между моделью регрессии (8.17) и моделью парной регрессии из главы 4 заключается в том, что переменная в правой части теперь логарифм  $X$ , а не сама переменная  $X$ . Чтобы оценить коэффициенты  $\beta_0$  и  $\beta_1$  в регрессии (8.17), нужно вычислить новую переменную  $\ln(X)$ , что легко сделать, имея компьютер. Тогда  $\beta_0$  и  $\beta_1$  могут быть оценены МНК-регрессией  $Y_i$  на  $\ln(X_i)$ , гипотезы о  $\beta_1$  могут проверяться при помощи обычных  $t$ -статистик, а 95 %-й доверительный интервал для  $\beta_1$  может быть построен как  $\hat{\beta}_1 \pm 1,96SE(\hat{\beta}_1)$ .

В качестве примера вернемся к соотношению между доходом и результатами тестов в школьных округах. Вместо квадратичной спецификации мы можем использовать линейно-логарифмическую спецификацию (8.17). Оценивание этой регрессии при помощи МНК приводит к:

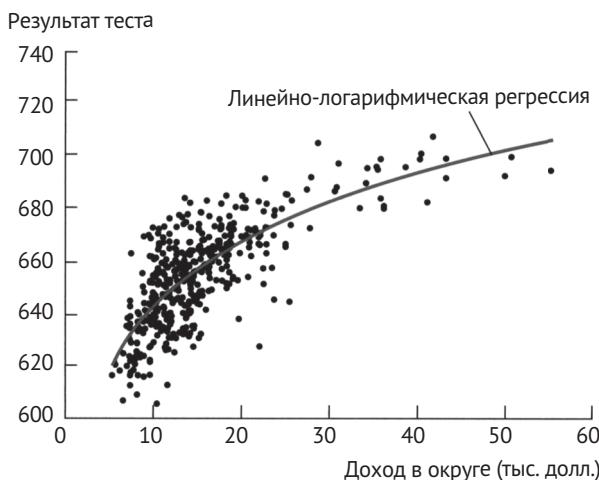
$$\widehat{TestScore} = 557,8 + 36,42 \ln(Income), \quad \bar{R}^2 = 0,561. \quad (8.18)$$

Как следует из уравнения (8.18), 1 %-й рост средних доходов влечет улучшение результатов тестов на  $0,01 \times 36,42 = 0,36$  пунктов.

Чтобы оценить, как влияет на  $Y$  изменение  $X$  в его исходных единицах измерения, то есть в тысячах долларов (а не в логарифмах), мы можем использовать метод из вставки «Основные понятия 8.1». Например, каким будет предсказанное различие в результатах тестов для округов со средними доходами в 10 тыс. и 11 тыс. долл.? Оцененное значение  $\Delta Y$  представляет собой

разность между предсказанными значениями для двух различных значений  $X$ :  $\Delta\hat{Y} = [557,8 + 36,42 \ln(11)] - [557,8 + 36,42 \ln(10)] = 36,42 \times [\ln(11) - \ln(10)] = 3,47$ . Аналогично предсказанное различие между округом со средним доходом в 40 тыс. долл. и округом со средним доходом в 41 тыс. долл. составляет  $36,42 \times [\ln(41) - \ln(40)] = 0,90$ . Таким образом, как и в случае квадратичной регрессии, линейно-логарифмическая регрессия предсказывает, что возрастание среднего дохода на 1000 долл. окажет больший эффект на результаты тестов в бедном округе по сравнению богатым округом.

График оцененной в (8.18) линейно-логарифмической регрессии изображен на рисунке 8.5. Поскольку в уравнении (8.18) регрессором является натуральный логарифм дохода, а не доход, оцененная функция регрессии не является прямой линией. Как и функция квадратичной регрессии на рисунке 8.3, она имеет более крутой наклон в начале своей области определения, но затем выравнивается для высоких уровней дохода.



**Рисунок 8.5. Функция линейно-логарифмической регрессии**

Оцененная линейно-логарифмическая регрессия  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$  хорошо описывает нелинейную связь между результатами тестов и доходом в округе.

**Случай II: Зависимость логарифма  $Y$  от  $X$ .** В этом случае модель регрессии имеет вид:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i. \quad (8.19)$$

Поскольку  $Y$  берется в логарифмах, а  $X$  – нет, такую модель иногда называют логарифмически-линейной моделью.

В логарифмически-линейной модели изменение  $X$  на единицу (т.е.  $\Delta X = 1$ ) влечет изменение  $Y$  в размере  $100 \times \beta_1 \%$ . Для того чтобы увидеть это, сравним ожидаемые значения  $\ln(Y)$  в точках, которые отличаются на  $\Delta X$ . Ожидаемое значение  $\ln(Y)$  при заданном  $X$  есть  $\ln(Y) = \beta_0 + \beta_1 X$ . Если  $X$  меняется до  $X + \Delta X$ , ожидаемое значение равно  $\ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X)$ . Таким образом, разность между этими ожидаемыми значениями составляет  $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1 (X + \Delta X)] - [\beta_0 + \beta_1 X] = \beta_1 \Delta X$ . Из приближения в уравнении (8.16), следует, однако, что если  $\beta_1 \Delta X$  мало, то  $\ln(Y + \Delta Y) - \ln(Y) \cong \Delta Y / Y$ . Значит,

$\Delta Y / Y \cong \beta_1 \Delta X$ . Если  $\Delta X = 1$ , то есть  $X$  изменяется на одну единицу, то  $\Delta Y / Y$  изменяется на  $\beta_1$ . Переводя в проценты, изменение  $X$  на единицу влечет за собой  $100 \times \beta_1\%$ -е изменение  $Y$ .

В качестве иллюстрации мы возвращаемся к эмпирическому примеру из раздела 3.7 о соотношении между возрастом и доходами выпускников колледжей. По условиям трудовых контрактов заработная плата многих работников увеличивается на определенный процент по окончании каждого дополнительного года работы. Такое процентное соотношение предполагает оценку логарифмически-линейной регрессии (8.19), связывающую каждый дополнительный год возраста ( $X$ ) с некоторым постоянным процентом увеличения зарплаты ( $Y$ ). Для оценки такой регрессии сначала необходимо вычислить новую зависимую переменную  $\ln(Earnings_i)$ , после чего неизвестные коэффициенты  $\beta_0$  и  $\beta_1$  могут быть оценены при помощи МНК-регрессии  $\ln(Earnings_i)$  от  $Age_i$ . Оценка этой модели с использованием 14 407 наблюдений о выпускниках колледжей из текущего обследования населения за март 2009 года (данные приведены в Приложении 3.1) дала следующие результаты:

$$\ln(Earnings) = 2,805 + 0,0087 Age, \bar{R}^2 = 0,027. \quad (8.20)$$

Согласно этой регрессии, зарплата предсказывается возрастающей на 0,87% [ $(100 \times 0,087)\%$ ] для каждого дополнительного года возраста.

**Случай III: Зависимость логарифма  $Y$  от логарифма  $X$ .** В этом случае модель регрессии имеет вид:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i. \quad (8.21)$$

Поскольку оба  $Y$  и  $X$  берутся в логарифмах, эта модель иногда называется *линейной в логарифмах моделью*<sup>1</sup>.

В линейной в логарифмах модели на 1%-е изменение  $X$  приходится  $\beta_1\%$ -е изменение  $Y$ . Таким образом, в этой спецификации  $\beta_1$  является эластичностью  $Y$  относительно  $X$ . Для того чтобы понять это, снова вспомним информацию, представленную во вставке «Основные понятия 8.1»; откуда, таким образом, следует  $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1 [\ln(X + \Delta X) - \ln(X)]$ . Применение равенства (8.16) к обеим частям этого уравнения приводит к такому виду:

$$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

или  $\beta_1 = \frac{\Delta Y / Y}{\Delta X / X} = \frac{100 \times (\Delta Y / Y)}{100 \times (\Delta X / X)} = \frac{\text{процентное изменение } Y}{\text{процентное изменение } X}.$  (8.22)

Таким образом, в линейной в логарифмах спецификации  $\beta_1$  является отношением процентного изменения  $Y$ , связанного с процентным изменением  $X$ . Если процентное изменение  $X$  составляет 1% (т.е. если  $\Delta X = 0,01X$ ), тогда  $\beta_1$  есть процентное изменение  $Y$ , связанное с 1%-м изменением  $X$ . То есть  $\beta_1$  является эластичностью  $Y$  относительно  $X$ .

<sup>1</sup> Иногда такую модель называют двойной логарифмической моделью. – Примеч. науч. ред. перевода.

В качестве иллюстрации вернемся к соотношению между доходами и результатами тестов. Оценим регрессию логарифма результатов тестов на логарифм среднего подушевого дохода в школьном округе. Результирующие оценки имеют вид:

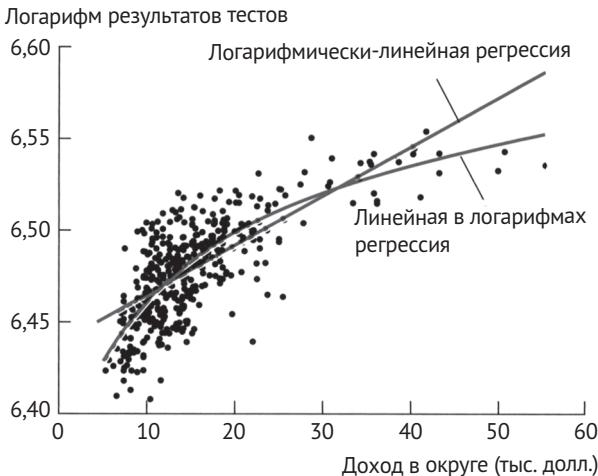
$$\widehat{\ln(\text{TestScore})} = 6,336 + 0,0554 \ln(\text{Income}), \bar{R}^2 = 0,557. \quad (8.23)$$

Согласно полученным оценкам, увеличение дохода на 1 % повлечет за собой соответствующее улучшение результатов тестов на 0,055 4 %.

Линия линейной в логарифмах регрессии, оценки которой представлены в уравнении (8.23), изображена на рисунке 8.6. Поскольку зависимая переменная берется в логарифмах, вертикальная ось на рисунке 8.6 – это логарифм результатов тестов, а построенная диаграмма рассеяния представляет собой диаграмму рассеяния логарифма результатов тестов и средних подушевых доходов в округе. Для целей сравнения на рисунке 8.6 также приведена линия оцененной логарифмически-линейной регрессии:

$$\widehat{\ln(\text{TestScore})} = 6,439 + 0,00284 \ln(\text{Income}), \bar{R}^2 = 0,497, \quad (8.24)$$

Поскольку вертикальная ось – это логарифм результатов тестов, линия регрессии (8.24) является прямой линией на рисунке 8.6.



**Рисунок 8.6. Функции логарифмически-линейной и линейной в логарифмах регрессий**

В функции логарифмически-линейной регрессии  $\ln(Y)$  является линейной функцией от  $X$ . В функции линейной в логарифмах регрессии  $\ln(Y)$  является линейной функцией от  $\ln(X)$ .

На рисунке 8.6 можно видеть, что линейная в логарифмах спецификация несколько лучше, чем логарифмически-линейная спецификация. Это отражено в более высоком  $\bar{R}^2$  для линейной в логарифмах регрессии (0,557), чем для логарифмически-линейной регрессии (0,497). Несмотря на это, линейная в логарифмах спецификация не очень хорошо приближает данные: для низких значений дохода большинство наблюдений оказываются ниже логарифмической

кривой, в то время как в среднем диапазоне дохода большинство наблюдений попадают выше предполагаемой функции регрессии.

Основные моменты, касающиеся всех трех моделей логарифмических регрессий, представлены во вставке «Основные понятия 8.2».

## ОСНОВНЫЕ ПОНЯТИЯ 8.2

### Логарифмы в регрессии: три случая

Логарифмы могут использоваться для преобразования зависимой переменной  $Y$ , независимой переменной  $X$  или обеих (но переменные, к которым применяется логарифмическое преобразование, должны быть положительными). В следующей таблице собраны три возможных случая и приведена интерпретация коэффициента  $\beta_1$  регрессии. В каждом случае коэффициент  $\beta_1$  может быть оценен при помощи МНК после взятия логарифма зависимой и/или независимой переменной.

Случай	Спецификация регрессии	Интерпретация $\beta_1$
I	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	Изменение $X$ на 1 % вызывает изменение $Y$ на $0,01\beta_1$
II	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	Изменение $X$ на единицу ( $\Delta X = 1$ ) вызывает изменение $Y$ на $100\beta_1\%$
III	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	Изменение $X$ на 1 % вызывает изменение $Y$ на $\beta_1\%$ , поэтому $\beta_1$ является эластичностью $Y$ по $X$

### Проблема, возникающая при сравнении логарифмических спецификаций.

Какая из моделей логарифмических регрессий лучше соответствует данным? Как мы уже видели при обсуждении регрессий (8.23) и (8.24),  $\bar{R}^2$  может использоваться для сравнения логарифмически-линейной и линейной в логарифмах моделей; в нашем случае линейная в логарифмах модель имела более высокий  $\bar{R}^2$ . Аналогично  $\bar{R}^2$  может быть использован для сравнения модели линейно-логарифмической регрессии (8.18) и модели линейной регрессии  $Y$  от  $X$ . В регрессии результатов тестов от средних доходов линейно-логарифмическая регрессия имеет  $\bar{R}^2$ , равный 0,561, в то время как линейная регрессия имеет  $\bar{R}^2$ , равный 0,508, поэтому линейно-логарифмическая модель лучше соответствует данным.

Но как сравнить линейно-логарифмическую и линейную в логарифмах модели? К сожалению,  $\bar{R}^2$  не может использоваться для сравнения этих двух регрессий, поскольку зависимые переменные в них различны [в первом случае это  $Y_i$ , а во втором —  $\ln(Y_i)$ ]. Вспомним, что  $\bar{R}^2$  измеряет долю дисперсии зависимости переменной, объясненную регрессорами. Так как зависимые перемен-

ные в линейной в логарифмах и линейно-логарифмической моделях различны, нет смысла сравнивать их  $\bar{R}^2$ -ы.

Поэтому самое лучшее, что можно сделать при решении определенной задачи, – это решить, используя экономическую теорию либо ваше или другое экспертное мнение о проблеме, есть ли смысл специфицировать  $Y$  в логарифмах. Например, в экономике труда обычно моделируют доходы, используя логарифмы, поскольку сравнение зарплат, увеличение заработной платы и т.д. часто более естественно обсуждать в процентах. В моделировании результатов тестов кажется (для нас, во всяком случае) естественным обсуждать результаты тестов в терминах баллов за тест, а не в процентном возрастании результатов теста, поэтому мы акцентируем внимание на моделях, в которых зависимой переменной является результат теста, а не ее логарифм.

**Вычисление предсказанных значений зависимой переменной, когда она взята в логарифмах.**<sup>1</sup> Если в качестве зависимой переменной мы берем логарифм  $Y$ , оцененная регрессия может быть использована непосредственно для вычисления предсказанного значения  $\ln(Y)$ . Однако немного сложнее вычислить предсказанное значение самого  $Y$ .

Для иллюстрации рассмотрим модель логарифмически-линейной регрессии в уравнении (8.19) и перепишем ее так, чтобы зависимой переменной была  $Y$ , а не  $\ln(Y)$ . Чтобы сделать это, применим экспоненциальное преобразование к обеим частям уравнения (8.19) и получим:

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i) = e^{\beta_0 + \beta_1 X_i} e^{u_i}. \quad (8.25)$$

Тогда ожидаемое значение  $Y_i$  при заданном  $X_i$  равно  $E(Y_i | X_i) = E(e^{\beta_0 + \beta_1 X_i} e^{u_i} | X_i) = e^{\beta_0 + \beta_1 X_i} E(e^{u_i} | X_i)$ . Проблема заключается в том, что даже если  $E(u_i | X_i) = 0$ ,  $E(e^{u_i} | X_i) \neq 1$ . Таким образом, соответствующее предсказанное значение  $Y_i$  не может быть получено простым возведением экспоненты в степень  $\hat{\beta}_0 + \hat{\beta}_1 X_i$ , то есть считая, что  $\hat{Y}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}$ : это предсказанное значение смещено из-за пропущенного фактора  $E(e^{u_i} | X_i)$ .

В качестве одного из возможных решений этой проблемы можно оценить фактор  $E(e^{u_i} | X_i)$  и использовать эту оценку, когда вычисляется прогнозное значение  $Y$ . В упражнении 17.12 рассматривается несколько способов оценки  $E(e^{u_i} | X_i)$ , но такая оценка становится сложной, в частности, если  $u_i$  гетероскедастична, и тогда мы не можем ее найти.

Другим решением, используемым в этой книге, может быть вычисление предсказанных значений логарифма  $Y$  без вычисления предсказанных значений самой переменной  $Y$ . Этот подход часто используется на практике, поскольку, если зависимая переменная берется в логарифмах, чаще всего более естественно просто использовать эту логарифмическую спецификацию (и связанные интерпретации в процентах) на протяжении всего анализа.

---

<sup>1</sup> Этот материал является материалом более продвинутого уровня и может быть опущен без потери общности.

## Полиномиальные и логарифмические модели зависимости результатов тестов от доходов в округе

На практике экономическая теория или экспертное суждение могут предложить функциональную форму, которую целесообразно использовать в каждом конкретном случае, но в конце концов истинный вид теоретической функции регрессии неизвестен. Таким образом, на практике соответствие данных нелинейной функции влечет за собой решение, какой из методов или их комбинация работает лучше всего. В качестве иллюстрации мы сравниваем логарифмическую и полиномиальную модели зависимости результатов тестов и доходов в школьных округах.

**Полиномиальные спецификации.** Рассмотрим две полиномиальные спецификации, заданные степенями *Income*: квадратичную [уравнение (8.2)] и кубическую [уравнение (8.11)]. Поскольку коэффициент при *Income*<sup>3</sup> в уравнении (8.11) был значимым на 5 %-м уровне, кубическая спецификация улучшает квадратичную, поэтому мы выбираем кубическую модель как предпочтительную полиномиальную спецификацию.

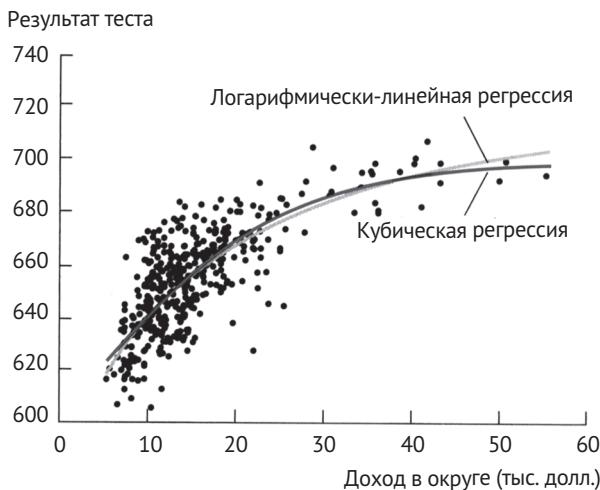
**Логарифмические спецификации.** Логарифмическая спецификация из уравнения (8.18), казалось бы, хорошо согласуется с рассматриваемыми данными, но мы не проверяли это формально. Один из способов сделать такую проверку – расширить модель, включив более высокие степени логарифма дохода. Если эти дополнительные компоненты статистически не отличаются от нуля, тогда мы можем заключить, что спецификация из уравнения (8.18) адекватна в том смысле, что не может быть отвергнута против полиномиальной функции логарифма. Соответственно, оцененная кубическая регрессия (специфицированная в степенях логарифма дохода) имеет вид:

$$\begin{aligned} \widehat{\text{TestScore}} = & 486,1 + 113,4 \ln(\text{Income}) - 26,9 \left[ \ln(\text{Income}) \right]^2 + \\ & + 3,06 \left[ \ln(\text{Income}) \right]^3, \bar{R}^2 = 0,560. \end{aligned} \quad (8.26)$$

*t*-статистика коэффициента при третьей степени объясняющей переменной составляет 0,818, поэтому нулевая гипотеза о том, что истинный коэффициент равен нулю, не отвергается на 10 %-м уровне значимости. *F*-статистика для тестирования совместной гипотезы о том, что истинные коэффициенты при второй и третьей степенях объясняющей переменной равны нулю, составляет 0,44 с *p*-значением 0,64, поэтому эта совместная нулевая гипотеза не отвергается на 10 %-м уровне значимости. Таким образом, кубическая логарифмическая модель (8.26) не дает статистически значимого улучшения линейно-логарифмической модели (8.18).

**Сравнение кубической и линейно-логарифмической спецификаций.** На рисунке 8.7 представлены оцененные функции кубической регрессии (8.11) и линейно-логарифмической регрессии (8.18). Эти оцененные функции регрессии довольно похожи. Одним из статистических инструментов для сравнения этих

спецификаций является  $\bar{R}^2$ .  $\bar{R}^2$  линейно-логарифмической регрессии составляет 0,561, а для кубической регрессии – 0,555. Поскольку линейно-логарифмическая спецификация все-таки имеет небольшое преимущество в терминах  $\bar{R}^2$  и потому, что эта спецификация не требует включения высоких степеней логарифма дохода, чтобы соответствовать моделируемым данным, мы выбираем линейно-логарифмическую спецификацию (8.18).



**Рисунок 8.7. Функции линейно-логарифмической и кубической регрессий**

Оцененная функция кубической регрессии [уравнение (8.11)] и оцененная функция линейно-логарифмической регрессии [уравнение (8.18)] почти идентичны в рассматриваемой выборке.

### 8.3. Взаимодействия между независимыми переменными

Во введении к данной главе мы задавались вопросом, может ли иметь большое влияние факт снижения соотношения учеников и учителей на результаты тестов в школьных округах, в которых многие ученики изучают английский язык, по сравнению с округами, в которых немногие школьники изучают английский язык. Например, такая ситуация может возникнуть, если ученики, которые все еще изучают английский язык, получают различные преимущества от индивидуальных занятий и от занятий в маленьких группах. Если это так, наличие большого количества школьников, изучающих английский язык в округе, будет взаимодействовать с соотношением учеников и учителей таким образом, что влияние на результаты тестов, оказываемое при изменении соотношения учеников и учителей, будет зависеть от доли изучающих английский язык.

В данном разделе рассматриваются методы, при помощи которых можно включить взаимодействие между двумя независимыми переменными в модель множественной регрессии. Возможная взаимосвязь между соотношением учеников и учителей и долей изучающих английский язык является примером более стандартной ситуации, в которой эффект влияния на  $Y$  от изменения независимой переменной на единицу зависит от значения другой независимой

переменной. Мы рассмотрим три случая: обе независимые переменные являются бинарными, одна – бинарная, а вторая – непрерывная и обе – непрерывные.

### **Взаимодействие между двумя бинарными переменными**

Рассмотрим теоретическую регрессию логарифма доходов [ $Y_i$ , где  $Y_i = \ln(Earnings_i)$ ] от двух бинарных переменных, характеризующих наличие у индивида высшего образования ( $D_{1i}$ , где  $D_{1i} = 1$ , если  $i$ -й индивид закончил колледж) и пол индивида ( $D_{2i}$ , где  $D_{2i} = 1$ , если  $i$ -й индивид – женщина). Линейная теоретическая регрессия  $Y_i$  от этих двух переменных имеет вид:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i. \quad (8.27)$$

В этой модели регрессии  $\beta_1$  характеризует эффект от наличия высшего образования при фиксированном поле индивида, а  $\beta_2$  является характеристикой эффекта влияния на логарифм доходов от принадлежности к женскому полу при постоянном образовании.

У спецификации регрессии (8.27) есть одно важное ограничение: эффект наличия высшего образования в этой спецификации при фиксированном поле индивида является одинаковым и для мужчин и для женщин. Однако нет основания полагать, что это так на самом деле. Выражаясь математически, эффект влияния  $D_{1i}$  на  $Y_i$  при постоянном  $D_{2i}$  может зависеть от  $D_{2i}$ . Другими словами, может присутствовать взаимосвязь между наличием высшего образования и полом индивида, такая что наличие у индивида высшего образования зависит от того, мужчина он или женщина.

Несмотря на то что спецификация регрессии (8.27) не позволяет учесть такую связь между наличием высшего образования и полом, ее легко модифицировать так, чтобы эта связь учитывалась, введением еще одного регрессора: произведения двух бинарных переменных  $D_{1i} \times D_{2i}$ . Результатирующая регрессия тогда имеет вид:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i. \quad (8.28)$$

Новый регрессор – произведение  $D_{1i} \times D_{2i}$  – называется *компонентой взаимодействия*, или *регрессором взаимодействия*, а теоретическая модель регрессии (8.28) называется *моделью регрессии с бинарной компонентой взаимодействия*.

Компонента взаимодействия в уравнении (8.28) позволяет смоделировать наличие в генеральной совокупности эффекта влияния наличия высшего образования (изменяя  $D_{1i}$  с  $D_{1i} = 0$  до  $D_{1i} = 1$ ) на логарифм доходов ( $Y_i$ ) и зависимости этого эффекта от пола индивида ( $D_{2i}$ ). Покажем это математически, вычисляя теоретический эффект от изменения  $D_{1i}$ , используя общий метод, изложенный во вставке «Основные понятия 8.1». На первом шаге вычислим условное математическое ожидание  $Y_i$  при  $D_{1i} = 0$  для заданного  $D_{2i}$ , предполагая равенство нулю условного среднего ошибки,  $E(u_i | D_{1i}, D_{2i}) = 0$ ; то есть  $E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_1 \times 0 + \beta_2 \times d_2 + \beta_3 \times (0 \times d_2) = \beta_0 + \beta_2 d_2$ . Затем вычислим условное математическое ожидание  $Y_i$  после изменения, то есть для

$D_{1i} = 1$  – при том же значении  $D_{2i}$  и в тех же предположениях об условном среднем, то есть  $E(Y_i | D_{1i} = 1, D_{2i} = d_2) = \beta_0 + \beta_1 \times 1 + \beta_2 \times d_2 + \beta_3 \times (1 \times d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$ . Эффект от такого изменения есть разность ожидаемых значений [т.е. разность из уравнения (8.4)]:

$$E(Y_i | D_{1i} = 1, D_{2i} = d_2) - E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_1 + \beta_3 d_2. \quad (8.29)$$

Таким образом, в модели регрессии с бинарной компонентой взаимодействия (8.28) эффект влияния наличия высшего образования (единичное изменение в  $D_{1i}$ ) на доход зависит от пола индивида [значение  $D_{2i}$ , которое представлено как  $d_2$  в уравнении (8.29)]. Если это мужчина ( $d_2 = 0$ ), эффект влияния от наличия высшего образования составляет  $\beta_1$ , но если это женщина ( $d_2 = 1$ ), то этот эффект равен  $\beta_1 + \beta_3$ . Коэффициент  $\beta_3$  при компоненте взаимодействия – это разность эффектов влияния факта наличия высшего образования у женщин и мужчин на доход.

### Интерпретация коэффициентов в регрессиях с бинарными объясняющими переменными

На первом шаге вычислите математическое ожидание  $Y$  для всех возможных комбинаций значений бинарных переменных. Далее сравните полученные математические ожидания между собой. Тогда каждый коэффициент может быть выражен либо как математическое ожидание, либо как разность между двумя или более математическими ожиданиями.

### ОСНОВНЫЕ ПОНЯТИЯ

8.3

Несмотря на то что мы рассмотрели проблему взаимодействия объясняющих переменных на примере, используя зависимость логарифма доходов от наличия и высшего образования и пола индивида, такой подход верен и в общем случае. Модель регрессии с бинарной компонентой взаимодействия допускает наличие зависимости влияния, которое оказывает на объясняемую переменную изменение одной из бинарных объясняющих переменных на единицу, от значений других бинарных переменных.

Метод, который мы используем здесь для интерпретации коэффициентов, по сути, может быть применен в любой возможной комбинации бинарных переменных. Рассмотренный нами метод, применяемый ко всем регрессиям с бинарными объясняющими переменными, описан во вставке «Основные понятия 8.3».

**Пример: соотношение учеников и учителей и процент изучающих английский язык.** Пусть  $HiSTR_i$  – бинарная переменная, которая равна 1, если соотношение учеников и учителей равно 20 или более, и равна 0 – в противном случае, и пусть  $HiEL_i$  – бинарная переменная, равная 1, если процент изучающих английский язык равен 10 % или более, и равна 0 – в противном случае. Регрессия с бинарной компонентой взаимодействия результатов тестов от переменных  $HiSTR_i$  и  $HiEL_i$  имеет вид:

$$\widehat{TestScore} = 664,1 - \underset{(1,4)}{1,9} HiSTR - \underset{(1,9)}{18,2} HiEL - \underset{(2,3)}{3,5} (HiSTR \times HiEL), \\ \bar{R}^2 = 0,290. \quad (8.30)$$

Предсказанный эффект влияния на результаты тестов от перехода из округа с низким соотношением учеников и учителей к округу с высоким соотношением учеников и учителей, считая постоянной долю изучающих английский язык школьников (либо большую, либо маленькую), задается уравнением (8.29) при замене теоретических коэффициентов на их оценки. Таким образом, согласно оценкам из выражения (8.30), такой эффект равен  $-1,9 - 3,5 HiEL$ . То есть если доля изучающих английский язык низка ( $HiEL = 0$ ), то эффект влияния на результаты тестов от перехода с  $HiSTR = 0$  к  $HiSTR = 1$  уменьшит результаты на 1,9 пункта. Если доля изучающих английский язык высока, тогда по оценке результаты тестов уменьшатся на  $1,9 + 3,5 = 5,4$  пункта.

Оцененная в уравнении (8.30) регрессия также может быть использована для оценивания среднего результата тестов для каждой из четырех возможных комбинаций бинарных переменных. Сделаем это, используя процедуру из вставки «Основные понятия 8.3». Получаем, что средний по выборке результат теста для округа с низким соотношением учеников и учителей ( $HiSTR_i = 0$ ) и низкой долей изучающих английский ( $HiEL_i = 0$ ) составляет 664,1. Для округов с  $HiSTR_i = 1$  (высокое соотношение учеников и учителей) и  $HiEL_i = 0$  (низкая доля изучающих английский язык), среднее по выборке значение равно 662,2 ( $=664,1 - 1,9$ ). Когда  $HiSTR_i = 0$  и  $HiEL_i = 1$ , средневыборочное значение равно 645,9 ( $=664,1 - 18,2$ ), а когда  $HiSTR_i = 1$  и  $HiEL_i = 1$ , среднее по выборке равно 640,5 ( $=664,1 - 1,9 - 18,2 - 3,5$ ).

### **Взаимодействие между непрерывной и бинарной переменными**

Рассмотрим далее теоретическую регрессию логарифма дохода [ $Y_i = \ln(Earnings_i)$ ] от одной непрерывной переменной, характеризующей трудовой стаж индивида ( $X_i$ ), и бинарной переменной  $D_i$ , характеризующей наличие высшего образования у  $i$ -го индивида ( $D_i = 1$ ) или его отсутствие ( $D_i = 0$ ). Как показано на рисунке 8.8, линия регрессии генеральной совокупности, связывающая  $Y$  и непрерывную переменную  $X$ , может зависеть от бинарной переменной  $D$  тремя различными способами.

На рисунке 8.8а две линии регрессии различаются только своими константами. Соответствующая теоретическая модель регрессии имеет вид:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i. \quad (8.31)$$

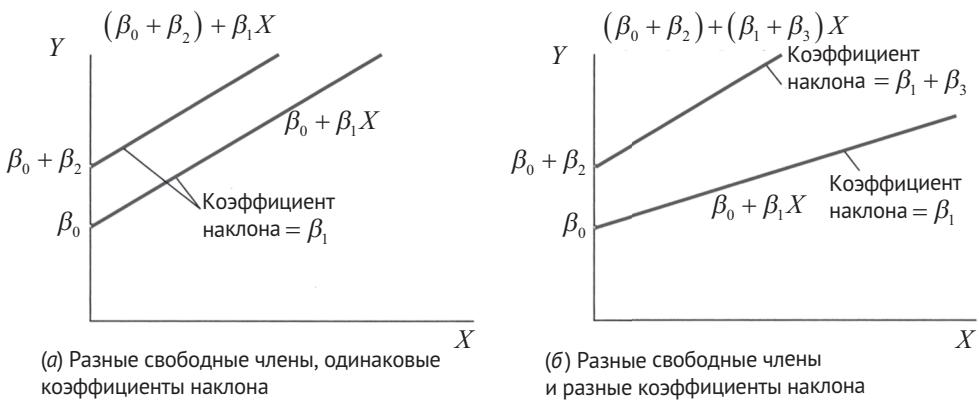
Это знакомая модель множественной регрессии с теоретической функцией регрессии, линейной по  $X_i$  и  $D_i$ . Если  $D_i = 0$ , теоретическая функция регрессии имеет вид:  $\beta_0 + \beta_1 X_i$ , поэтому свободный член равен  $\beta_0$ , а угловой коэффициент —  $\beta_1$ . Если  $D_i = 1$ , то теоретическая функция регрессии имеет вид:  $\beta_0 + \beta_1 X_i + \beta_2$ , поэтому угловой коэффициент остается равным  $\beta_1$ , но констан-

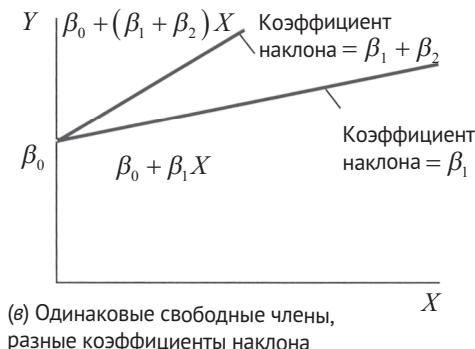
та теперь равна  $\beta_0 + \beta_2$ . Таким образом,  $\beta_2$  представляет собой разность между свободными членами двух линий регрессии, как показано на рисунке 8.8а. Говоря в терминах примера с доходами,  $\beta_1$  характеризует эффект влияния дополнительного года трудового стажа на логарифм доходов при неизменности характеристики наличия/отсутствия высшего образования, а  $\beta_2$  отражает эффект влияния факта наличия высшего образования на логарифм зарплаты при постоянном трудовом стаже. В этой спецификации эффект влияния дополнительного года трудового стажа одинаков для индивидов, имеющих и не имеющих высшее образование, то есть две линии на рисунке 8.8а имеют одинаковый угловой коэффициент.

На рисунке 8.8б две линии имеют различные угловые коэффициенты и константы. Различные угловые коэффициенты предполагают, что для лиц, окончивших колледж, эффект влияния дополнительного года трудового стажа на логарифм дохода не такой, как для лиц, не окончивших колледж. Чтобы учесть наличие различных угловых коэффициентов, добавим компоненту взаимодействия в уравнение (8.31):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i, \quad (8.32)$$

где  $X_i \times D_i$  является новой переменной, произведением  $X_i$  и  $D_i$ . Чтобы интерпретировать коэффициенты этой регрессии, воспользуемся схемой из вставки «Основные понятия 8.3». Получим, что если  $D_i = 0$ , теоретическая функция регрессии равна  $\beta_0 + \beta_1 X_i$ , в то время как если  $D_i = 1$ , теоретическая функция регрессии равна  $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$ . Таким образом, эта спецификация допускает две различные теоретические функции регрессии, связывающие  $Y_i$  и  $X_i$ , и зависящие от значения  $D_i$ , как показано на рисунке 8.8б. Разность между двумя константами –  $\beta_2$ , а разность между двумя угловыми коэффициентами –  $\beta_3$ . В примере с доходами  $\beta_1$  отражает эффект влияния дополнительного года трудового стажа на логарифм доходов тех, кто не имеет высшего образования ( $D_i = 0$ ), а  $\beta_1 + \beta_3$  – соответствующий эффект для имеющих высшее образование, поэтому  $\beta_3$  есть *разность* эффектов наличия дополнительного года трудового стажа у выпускников колледжа и тех, у кого нет высшего образования.





**Рисунок 8.8. Регрессионные функции с использованием бинарной и непрерывной переменных**

Взаимодействие бинарной переменной и непрерывной переменной может происходить тремя различными путями, что реализуется тремя различными функциями регрессии: (а)  $\beta_0 + \beta_1 X + \beta_2 D$  допускает различные константы, но имеет одинаковые угловые коэффициенты, (б)  $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$  допускает различные константы и различные угловые коэффициенты и (в)  $\beta_0 + \beta_1 X + \beta_3 (X \times D)$  имеет одинаковые константы, но допускает различные угловые коэффициенты.

## ОСНОВНЫЕ ПОНЯТИЯ

### 8.4

#### Взаимодействие между бинарной и непрерывной переменными

Благодаря использованию компоненты взаимодействия  $X_i \times D_i$ , теоретическая линия регрессии, связывающая  $Y_i$  и непрерывную переменную  $X_i$ , может иметь коэффициент наклона, зависящий от бинарной переменной  $D_i$ . Существует три возможности:

1. Различная константа, одинаковый угловой коэффициент (рисунок 8.8а):  

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i.$$
2. Различны и константа, и угловой коэффициент (рисунок 8.8б):  

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i.$$
3. Однаковая константа, различный угловой коэффициент (рисунок 8.8в):  

$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 (X_i \times D_i) + u_i.$$

И, наконец, третья ситуация, изложенная на рисунке 8.8в, заключается в том, что две линии имеют различный угловой коэффициент, но одинаковый свободный член. Модель регрессии с компонентой взаимодействия для этого случая имеет вид:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i. \quad (8.33)$$

Коэффициент этой спецификации также может быть интерпретирован с использованием схемы из вставки «Основные понятия 8.3». В терминах примера с доходами эта спецификация допускает различные эффекты влияния

длительности трудового стажа на логарифм доходов для индивидов, имеющих и не имеющих высшего образования, но требует, чтобы ожидаемый логарифм доходов был одинаковым для обеих групп, если индивиды не имеют никакого предварительного опыта работы. Другими словами, эта спецификация предполагает наличие в генеральной совокупности некоторого среднего начального уровня зарплаты, который является одинаковым для тех, кто окончил колледж, и для тех, кто его не окончил. Это предположение не имеет смысла в данном примере, и на практике такая спецификация используется реже, чем спецификация (8.32), которая допускает различные константы и угловые коэффициенты.

Все три спецификации – уравнения (8.31), (8.32) и (8.33) – представляют собой версии модели множественной регрессии из главы 6, и на основе новой переменной  $X_i \times D_i$  коэффициенты всех трех моделей могут оцениваться при помощи МНК.

Три типа моделей регрессии с бинарной и непрерывной независимыми переменными описываются во вставке «Основные понятия 8.4».

**Пример: Соотношение учеников и учителей и процент изучающих английский язык.** Присутствует ли эффект влияния от сокращения соотношения учеников и учителей на результаты тестов в зависимости от того, мал или велик процент изучающих английский язык школьников в округе? Один из способов ответить на этот вопрос – использовать спецификацию, которая допускает две различные линии регрессии, зависящие от того, высок или низок процент изучающих английский язык школьников в округе. Это достигается с использованием спецификации, допускающей наличие различных констант и различных угловых коэффициентов:

$$\widehat{TestScore} = 682,2 - 0,97 STR_{(11,9)} + 5,6 HiEL_{(0,59)} - 1,28 (STR \times HiEL_{(19,5)}) - 0,97,$$

$$\bar{R}^2 = 0,305, \quad (8.34)$$

где бинарная переменная  $HiEL_i$  равна 1, если процент школьников, изучающих английский язык, больше 10 %, и равна 0 – в противном случае.

Для округов с низкой долей изучающих английский ( $HiEL_i = 0$ ) оцененная линия регрессии имеет вид  $682,2 - 0,97 STR_i$ . Для школьных округов с высокой долей изучающих английский язык школьников ( $HiEL_i = 1$ ) оцененная линия регрессии равна  $682,2 + 5,6 - 0,97 STR_i - 1,28 STR_i = 687,8 - 2,25 STR_i$ . Согласно этим оценкам, уменьшение соотношения учеников и учителей на единицу повлечет увеличение результатов тестов на 0,97 пунктов в округах с низкой долей изучающих английский язык учеников, но на 2,25 пункта – в округах с высокой долей изучающих английский язык. Разность между этими двумя эффектами, 1,28 пунктов, равна коэффициенту при компоненте взаимодействия в уравнении (8.34).

Модель регрессии с компонентой взаимодействия (8.34) позволяет нам оценить эффект более тонких мер, чем уменьшение размеров классов во всех школах, рассмотренное до сих пор. Например, предположим, что государство рассмотрело политику уменьшения соотношения учеников и учителей на 2 в округе с высокой долей изучающих английский язык ( $HiEL_i = 1$ ), но при этом

оставляет неизменным размеры классов в других округах. Применение способа, описанного во вставке «Основные понятия 8.1», к уравнениям (8.32) и (8.34) показывает, что оцененный эффект влияния на результаты тестов от такого уменьшения для округов с высокой долей изучающих английский язык школьников ( $HiEL_i = 1$ ) составляет  $-2(\hat{\beta}_1 + \hat{\beta}_3) = 4,50$ . Стандартная ошибка этого оцененного эффекта равна  $SE(-2\hat{\beta}_1 - 2\hat{\beta}_3) = 1,53$ , и она может быть вычислена с использованием уравнения (8.8) и методов из раздела 7.3.

МНК-регрессия (8.34) может быть использована для тестирования различных гипотез о теоретической линии регрессии. Во-первых, гипотеза о том, что две линии фактически одинаковы, может быть проверена путем вычисления  $F$ -статистики для проверки совместной гипотезы о том, что коэффициенты при  $HiEL_i$  и при компоненте взаимодействия  $STR_i \times HiEL_i$  одновременно равны нулю. Соответствующая  $F$ -статистика составляет 89,9, из чего следует, что нулевая гипотеза отклоняется на уровне значимости 1 %.

Во-вторых, гипотеза о том, что две линии имеют одинаковые угловые коэффициенты, может быть протестирована путем проверки того, является ли коэффициент при компоненте взаимодействия нулевым.  $t$ -статистика,  $-1,28 / 0,97 = -1,32$ , меньше, чем 1,645 по абсолютному значению, поэтому нулевая гипотеза о том, что две линии имеют одинаковый наклон, не может быть отвергнута с использованием двухстороннего теста на 10 %-м уровне значимости.

В-третьих, гипотеза о том, что линии регрессии имеют одинаковые свободные члены, соответствует ограничению о том, что теоретический коэффициент при  $HiEL$  равен нулю.  $t$ -статистика для проверки этого ограничения равна:  $t = 5,6 / 19,5 = 0,29$ , поэтому гипотеза о том, что линии имеют одинаковую константу, не может быть отвергнута на 5 %-м уровне значимости.

Результаты этих трех тестов дают кажущиеся противоречивыми результаты:  $F$ -статистика отвергает совместную гипотезу о том, что угловые коэффициенты и константы в двух регрессиях совпадают, но тесты для проверки индивидуальных гипотез, использующие  $t$ -статистику, не в состоянии отвергнуть эти гипотезы. Причина этого парадокса заключается в том, что регрессоры,  $HiEL$  и  $STR \times HiEL$ , сильно коррелированы. Это приводит к большим стандартным ошибкам индивидуальных коэффициентов. Несмотря на то что невозможно сказать, какой из коэффициентов ненулевой, есть серьезное свидетельство против гипотезы о том, что оба коэффициента равны нулю.

Наконец, гипотеза о том, что соотношение учеников и учителей не входит в эту спецификацию, может быть проверена при помощи  $F$ -статистики для совместной гипотезы о том, что коэффициенты при  $STR$  и компоненте взаимодействия нулевые. Эта  $F$ -статистика равна 5,64 и имеет  $p$ -значение, равное 0,004. Таким образом, коэффициенты при соотношении учеников и учителей статистически значимы на 1 %-м уровне значимости.

### **Взаимодействие между двумя непрерывными переменными**

Предположим теперь, что обе независимые переменные ( $X_{1i}$  и  $X_{2i}$ ) непрерывны. Например,  $Y_i$  – логарифм доходов  $i$ -го индивида,  $X_{1i}$  – его или ее количе-

ство лет трудового стажа, а  $X_{2i}$  – число лет обучения в школе. Если теоретическая функция регрессии линейна, эффект влияния дополнительного года трудового стажа на зарплату не зависит от числа лет образования, или, эквивалентно, эффект от наличия дополнительного года образования не зависит от трудового стажа. В реальности, однако, может присутствовать взаимодействие между этими двумя переменными, поэтому эффект влияния дополнительного года трудового стажа на зарплату зависит от числа лет образования. Это взаимодействие может быть смоделировано путем расширения модели до модели линейной регрессии с компонентой взаимодействия, которая является произведением  $X_{1i}$  и  $X_{2i}$ :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i. \quad (8.35)$$

Компонента взаимодействия допускает наличие эффекта от единичного изменения в  $X_1$ , зависящего от  $X_2$ . Чтобы увидеть это, применим общий метод для вычисления эффектов в модели нелинейной регрессии из вставки «Основные понятия 8.1». Разность в выражении (8.4), вычисленная для функции регрессии с компонентой взаимодействия (8.35), равна  $\Delta Y = (\beta_1 + \beta_3 X_2) \Delta X_1$  [упражнение 8.10 (а)]. Таким образом, эффект влияния изменения  $X_1$  на  $Y$  при постоянном  $X_2$  равен:

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2, \quad (8.36)$$

который зависит от  $X_2$ . В примере с доходами, если  $\beta_3$  положителен, эффект влияния дополнительного года трудового стажа на логарифм доходов выше на величину  $\beta_3$  для каждого дополнительного года образования работника.

Аналогичные вычисления показывают, что эффект влияния изменения  $X_2$  (на  $\Delta Y$ ) на  $Y$  при постоянном  $X_1$  равен  $\Delta Y / \Delta X_2 = (\beta_2 + \beta_3 X_1)$ .

Введение этих двух эффектов одновременно показывает, что коэффициент  $\beta_3$  при компоненте взаимодействия характеризует эффект влияния единичного роста  $X_1$  и  $X_2$ , сверх суммы отдельных эффектов влияния увеличения на единицу только  $X_1$  и увеличения на единицу только  $X_2$ . То есть если  $X_1$  изменяется на  $\Delta X_1$  и  $X_2$  изменяется на  $\Delta X_2$ , тогда ожидаемое изменение  $Y$  равно  $\Delta Y = (\beta_1 + \beta_3 X_2) \Delta X_1 + (\beta_2 + \beta_3 X_1) \Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$  [упражнение 8.10 (в)]. Первая компонента равна эффекту влияния изменения  $X_1$  при постоянном  $X_2$ ; вторая компонента равна эффекту влияния изменения  $X_2$  при постоянном  $X_1$ ; последняя компонента,  $\beta_3 \Delta X_1 \Delta X_2$ , есть дополнительный эффект влияния изменения  $X_1$  и  $X_2$ .

Описанные выше механизмы взаимодействия между двумя переменными кратко изложены во вставке «Основные понятия 8.5».

Если компонента взаимодействия объединяется с логарифмическими преобразованиями, то такие модели могут быть использованы для получения оценок эластичностей по цене, когда ценовая эластичность зависит от характеристик товара (см. вставку «Спрос на экономические журналы» на странице 293 для примера).



### ***Отдача от образования и гендерный разрыв***

Помимо интеллектуального удовлетворения образование дает и экономические выгоды. Как показывают вставки в главах 3 и 5, работники с более высоким уровнем образования, как правило, зарабатывают больше, чем их коллеги с более низким уровнем образования. Однако анализ, проведенный в этих вставках, является неполным, по крайней мере, по трем причинам. Во-первых, регрессии в нем не контролируются относительно других факторов, определяющих доходы, которые могут быть коррелированы с уровнем образования, поэтому МНК-оценка коэффициента при образовании могут иметь смещение из-за пропущенных переменных. Во-вторых, функциональная форма, использованная в главе 5 – простое линейное соотношение, – предполагает, что доходы изменяются в долларовом выражении на постоянную величину за каждый дополнительный год образования, в то время как можно было бы подозревать, что такое изменение доходов на самом деле будет большим для более высоких уровней образования. В-третьих, вставка из главы 5 игнорирует гендерные различия в доходах, подчеркнутые во вставке в главе 3.

Все эти ограничения могут быть решены с помощью модели множественной регрессии, в которой есть возможность контролировать на факторы, определяющие зарплатные платы, пропуск которых может привести к смещению из-за пропущенных переменных, и в которой используется нелинейная функциональная форма, связывающая образование и доходы. В таблице 8.1 приведены результаты оценок регрессий, использующих данные об индивидах в возрасте 30–64 лет, работающих полный рабочий день, из текущего обследования населения (данные CPS описаны в приложении 3.1). Зависимая переменная – логарифм почасовой зарплаты, то есть еще один год обучения связан с постоянным увеличением доли (а не увеличением зарплаты в долларах) в доходах.

Из таблицы 8.1 можно сделать четыре основных вывода. Во-первых, пропуск переменной, характеризующей пол индивида, в регрессии (1) не приводит к существенному смещению из-за пропущенной переменной: несмотря на то что пол индивида входит в регрессию (2) значимо и с большим коэффициентом, пол индивида и число лет образования не коррелированы, то есть в среднем мужчины и женщины имеют почти одинаковый уровень образования. Во-вторых, отдача от образования экономически и статистически различна для мужчин и женщин: в регрессии (3) *t*-статистика для проверки гипотезы о том, что они одинаковы, составляет 7,02 ( $=0,0121 / 1,0017$ ). В-третьих, регрессия (4), оценки которой контролируются относительно региона проживания, позволяет решить потенциальную проблему смещения из-за пропущенной переменной, которая может возникнуть, если число лет образования индивидов систематически различается по регионам. Учет региона проживания в регрессии приводит к небольшому различию в оценках коэффициентов при переменной, характеризующей уровень образования, по сравнению с оценками, полученными в регрессии (3). В-четвертых, регрессия (4) учитывает потенциальный опыт работы индивидов, измения его числом лет после завершения ими обучения. Оцененные коэффициенты предполагают уменьшение предельного значения для каждого следующего года работы.

Оцененная экономическая отдача от образования в регрессии (4) составляет 10,32% для каждого года образования для мужчин и 11,66% ( $=0,1032 + 0,0134$ , в процентах) для женщин. Поскольку функции регрессий для мужчин и женщин имеют различные угловые коэффициенты, гендерный разрыв зависит от числа лет образова-

ния. Для 12 лет образования гендерный разрыв оценен равным 29,0% ( $= 0,0134 \times 12 - 0,451$ , в процентах); для 16 лет образования гендерный разрыв меньше в процентном выражении – 23,7%.

Такие оценки отдачи от образования и размера гендерного разрыва все еще имеют ограничения, включающие, в том числе, возможность наличия других пропущенных переменных – в частности, переменных, характеризующих национальные особенности индивида, а также потенциальные проблемы, связанные с тем, каким образом измеряются данные в CPS. Тем не менее оценки в таблице 8.1 согласуются с результатами, полученными экономистами, которые аккуратно учитывали эти ограничения. Из анализа десятков эмпирических исследований, проведенных эконометристом Дэвидом Кардом (David Card, 1999), делается вывод о том, что лучшие оценки отдачи от образования, полученные экономистами, специализирующимися в экономике труда, в целом находятся между 8 и 11 %, и отдача зависит от качества образования. Если вам интересна дополнительная информация об экономической отдаче от образования, см. работу Карда (1999).

Таблица 8.1

**Отдача от образования и гендерный разрыв:  
результаты оценки регрессий для США в 2008 году**

Зависимая переменная: логарифм Hourly Earnings				
Регрессор	(1)	(2)	(3)	(4)
<i>Years of Education</i>	0,103 5** (0,000 9)	0,105 0** (0,000 9)	0,100 1** (0,001 1)	0,103 2** (0,001 2)
<i>Female</i>		-0,263** (0,004)	-0,432** (0,024)	-0,451** (0,024)
<i>Female Years of Education</i>			-0,012 1** (0,001 7)	0,013 4** (0,001 7)
<i>Potential Experience</i>				0,014 3** (0,001 2)
<i>Potential Experience</i> <sup>2</sup>				-0,000 211** (0,000 023)
<i>Midwest</i>				-0,095** (0,006)
<i>South</i>				-0,092** (0,006)
<i>West</i>				-0,023** (0,007)
Константа	1,533** (0,012)	1,629** (0,012)	1,697** (0,016)	1,503** (0,023)
$\bar{R}^2$	0,208	0,258	0,258	0,267

*Примечание:* Данные взяты из текущего обследования населения, проведенного в марте 2009 года (см. приложение 3.1). Размер выборки равен  $n=52\,790$  наблюдений для каждой регрессии. *Years of Education* – число лет образования индивида; *Female* является фиктивной переменной, равной 1 для женщин и 0 для мужчин; *Potential Experience* – трудовой стаж после окончания обучения. *Midwest*, *South* и *West* являются фиктивными переменными, обозначающими регион Соединенных Штатов, в котором живут индивиды: например, *Midwest* равен 1, если работник живет на Среднем Западе, и равен 0 — в противном случае (опущенный регион – *Northeast*). Стандартные ошибки приведены в скобках под оценками коэффициентов. Индивидуальные коэффициенты статистически значимы на 5 %-м\* или 1 %-м уровне\*\* значимости.



## ОСНОВНЫЕ ПОНЯТИЯ

### 8.5

#### Взаимодействие во множественной регрессии

Компонента взаимодействия между двумя независимыми переменными  $X_1$  и  $X_2$  равна их произведению  $X_1 \times X_2$ . Включение этой компоненты взаимодействия позволяет моделировать эффект влияния изменения  $X_1$  на  $Y$ , зависящий от значения  $X_2$ , и наоборот, допускает, чтобы эффект влияния изменения  $X_2$  зависел от значения  $X_1$ .

Коэффициент при  $X_1 \times X_2$  характеризует эффект влияния увеличения на единицу  $X_1$  и  $X_2$ , сверх суммы отдельных эффектов влияния увеличения на единицу только  $X_1$  и увеличения на единицу только  $X_2$ . Это верно независимо от того, непрерывны или бинарны  $X_1$  и/или  $X_2$ .

**Пример: Соотношение учеников и учителей и процент изучающих английский язык.** В предыдущем примере мы рассмотрели взаимодействие между соотношением учеников и учителей и бинарной переменной, указывающей на то, велик или мал процент изучающих английский язык школьников в округе. Изучить это взаимодействие можно еще одним способом, если включить в регрессию переменную, характеризующую взаимодействие между соотношением учеников и учителей и непрерывной переменной – процентом изучающих английский ( $PctEL$ ). Оцененная регрессия взаимодействия имеет вид:

$$\widehat{TestScore} = \underset{(11,8)}{686,3} - \underset{(0,59)}{1,12} STR - \underset{(0,37)}{0,67} PctEl + \underset{(0,019)}{0,0012}(STR \times PctEL), \\ \bar{R}^2 = 0,422. \quad (8.37)$$

Если процент изучающих английский язык находится на медиане ( $PctEL = 8,85$ ), угловой коэффициент линии регрессии результатов теста от соотношения учеников и учителей оценивается равным  $-1,11$  ( $= -1,12 + 0,0012 \times 8,85$ ). Если процент изучающих английский язык совпадает с 75 %-м процентилем ( $PctEL = 23,0$ ), то эта линия оценивается более плоской с угловым коэффициентом, равным  $-1,09$  ( $= -1,12 + 0,0012 \times 23,0$ ). То есть для округа с 8,85 % изучающих английский язык оцененный эффект влияния сокращения (на единицу) соотношения учеников и учителей на результаты тестов равен увеличению на 1,11 пункта, но для округа с 23,0 % изучающих английский язык уменьшение соотношения учеников и учителей на единицу предсказывает увеличение результатов тестов на 1,09 пункта. Однако разность между этими оцененными эффектами статистически незначима:  $t$ -статистика для проверки гипотезы о том, является ли коэффициент при компоненте взаимодействия нулевым, равна  $t = 0,0012 / 0,019 = 0,06$ , что незначимо на 10 %-м уровне значимости.

Для того чтобы сосредоточить внимание на обсуждении нелинейных моделей, в спецификациях в разделах 8.1–8.3 исключены дополнительные контрольные переменные, такие как экономическое положение семей студентов. Следовательно, эти результаты, вполне вероятно, могут иметь смещение из-за пропущенных переменных. Чтобы сделать содержательные выводы о влиянии снижения соотношения учеников и учителей на результаты тестов, эти нели-

нейные спецификации должны быть дополнены контрольными переменными, и далее мы сделаем именно это.

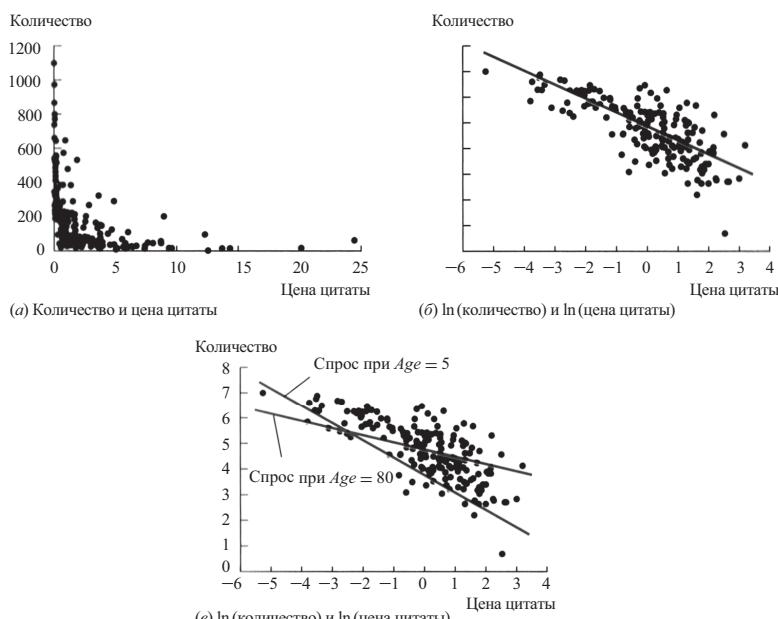
## 8.4. Нелинейные эффекты влияния изменения соотношения учеников и учителей на результаты тестов

В данном разделе мы рассматриваем три конкретных вопроса, касающихся влияния соотношения учеников и учителей на результаты тестов. Во-первых, будет ли зависеть эффект влияния соотношения учеников и учителей на результаты тестов после учета различий в экономических характеристиках различных округов? Во-вторых, зависит ли этот эффект от показателя соотношения учеников и учителей? В-третьих, и это самое главное, после принятия во внимание экономических факторов и нелинейности каков будет оцененный эффект влияния снижения соотношения учеников и учителей (на двух учеников на учителя, как предлагает в главе 4 наш окружной школьный инспектор) на результаты тестов?



### Спрос на экономические журналы

Профессиональные экономисты следят за самыми последними исследованиями в своих областях, результаты которых публикуются в экономических журналах, поэтому экономисты – или библиотеки – подписываются на экономические журналы.



**Рисунок 8.9. Библиотечная подписка и цены на экономические журналы**

Существует обратная нелинейная зависимость между числом подписок в библиотеках США (количество) и ценой за цитату (цена), как показано на рисунке 8.9а для 180 экономических журналов в 2000 году. Но, как видно на рисунке 8.9б, соотношение между логарифмом количества и логарифмом цены представляется близким к линейной. Рисунок 8.9в показывает, что спрос более эластичен для новых журналов ( $Age=5$ ), чем для старых журналов ( $Age=80$ ).

Насколько эластичным является спрос библиотек на экономические журналы?

Чтобы выяснить это, мы проанализировали взаимосвязь между количеством подписок на журналы в библиотеках США ( $y_i$ ) и ценой подписки, используя данные за 2000 год для 180 экономических журналов. Поскольку продуктом журнала является не бумага, на которой он напечатан, а идеи, которые он содержит, его цену логично измерять не в долларах за годовую подписку или в долларах за страницу, а в долларах за идею. Несмотря на то что мы не можем измерить стоимость «идеи» непосредственно, хорошим косвенным показателем является количество последующих цитирований журнальных статей в работах других исследователей. Соответственно, мы измеряем цену как «цену цитаты» в журнале. Диапазон таких цен огромен, с  $\frac{1}{2} \text{¢}$  за цитату *American Economic Review* до 20¢ за цитату и более. Некоторые журналы являются очень дорогими с точки зрения стоимости цитаты, поскольку их редко цитируют, другие – поскольку годовая стоимость библиотечной подписки на них очень высока. В 2010 году библиотечная подписка на бумажную версию *Journal of Econometrics* стоила 3264 долл. по сравнению с 455 долл. за подписку всех семи журналов, издаваемых Американской экономической ассоциацией, включая *American Economic Review*!

Поскольку мы интересовались оценкой эластичностей, мы используем линейную в логарифмах спецификацию (вставка «Основные понятия 8.2»). Диаграммы рассеяния на рисунках 8.9а и 8.9б предлагают эмпирическое обоснование использования этого преобразования. Поскольку некоторые из старейших и самых престижных журналов являются самыми дешевыми с точки зрения стоимости цитаты, регрессия логарифма количества подписок от логарифма цены ( $\ln(\text{Price per citation})$ ) может иметь смещение из-за пропущенных переменных. Наши регрессии, следовательно, включают две контрольные переменные: логарифм возраста журнала (сколько лет он издается –  $\ln(\text{Age})$ ) и логарифм числа символов в статьях в журнале в год ( $\ln(\text{Characters} \div 1\,000\,000)$ ).

Результаты оценок регрессий представлены в таблице 8.2. Эти результаты приводят к следующим выводам (смотрите, сможете ли вы найти основания этих выводов в таблице!):

1. Спрос на более старые журналы менее эластичен по сравнению с журналами более молодыми.
2. Линейная зависимость от логарифма цены предпочтительнее кубической.
3. Спрос на журналы с большим числом символов (при постоянных ценах и возрасте) выше.

Таблица 8.2

**Оценки спроса на экономические журналы**

Зависимая переменная: логарифм числа подписок в библиотеках США в 2000 году, 180 наблюдений				
Регрессор	(1)	(2)	(3)	(4)
$\ln(\text{Price per citation})$	-0,533** (0,034)	-0,408** (0,044)	-0,961** (0,160)	-0,899** (0,145)
$[\ln(\text{Price per citation})]^2$			0,017 (0,025)	

Окончание таблицы 8.2

<b>Зависимая переменная: логарифм числа подписок в библиотеках США в 2000 году, 180 наблюдений</b>				
<b>Регрессор</b>	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>
$[\ln(Price\ per\ citation)]^3$			0,0037 (0,005 5)	
$\ln(Age)$		0,424** (0,119)	0,373** (0,118)	0,374** (0,118)
$\ln(Age) \times \ln(Price\ per\ citation)$			0,156** (0,052)	0,141** (0,040)
$\ln(Characters \div 1\ 000\ 000)$		0,206* (0,098)	0,235* (0,098)	0,229* (0,096)
Константа	4,77** (0,055)	3,21** (0,38)	3,41** (0,38)	3,43** (0,38)
<i>F</i> -статистики и статистики качества регрессий				
<i>F</i> -статистика для коэффициентов при квадратичном и кубическом регрессорах ( <i>p</i> -значение)			0,25 (0,779)	
SER	0,750	0,705	0,691	0,688
$\bar{R}^2$	0,555	0,607	0,622	0,626

Примечание. *F*-статистика проверяет нулевую гипотезу о том, что коэффициенты при переменных  $[\ln(Price\ per\ citation)]^2$  и  $[\ln(Price\ per\ citation)]^3$  равны нулю. Стандартные ошибки оценок коэффициентов приведены в скобках под оценками коэффициентов, *p*-значение приведено в скобках для *F*-статистики. Коэффициенты значимы на уровнях значимости \*5% и \*\*1%.

Так какова же эластичность спроса на экономические журналы? Она зависит от возраста журнала. Кривые спроса на 80-летние и пятилетние журналы накладываются на диаграммы рассеяния на рисунке 8.9; эластичность спроса для более старых журналов составляет  $-0,28$  ( $SE = 0,06$ ), в то время как для более молодых равна  $-0,67$  ( $SE = 0,08$ ).

Этот спрос очень неэластичен: он очень нечувствителен к цене, особенно для старых журналов. Для библиотек наличие самых последних исследований в руках является необходимостью, а не роскошью. Для сравнения: по оценкам экспертов, эластичность спроса на сигареты находится в диапазоне от  $-0,3$  до  $-0,5$ . Необходимость читать экономические журналы похожа на привыкание к сигаретам, но это намного лучше для вашего здоровья!\*

\* Данные были любезно предоставлены профессором Теодором Бергстромом с экономического факультета Калифорнийского университета в Санта-Барбара (Theodore Bergstrom of the Department of Economics at the University of California, Santa Barbara). Если вы заинтересованы в получении дополнительной информации об экономике экономических журналов, см. Bergstrom (2001).



Мы отвечаем на эти вопросы, рассматривая спецификации нелинейных регрессий, аналогичные рассмотренным нами в разделах 8.2 и 8.3 и расширенные включением двух характеристик экономического положения учеников: процента школьников, имеющих право на субсидированные обеды, и логарифма среднего дохода в округе. Логарифм доходов использовался, поскольку из результатов эмпирического анализа, проведенного в разделе 8.2, следует, что эта

спецификация отражает нелинейную связь между результатами тестов и доходами. Как и в разделе 7.6, мы не включаем в регрессию показатель расходов на одного ученика в качестве регрессора, и при этом рассматриваем эффект влияния снижения соотношения учеников и учителей, позволяя расходам на одного ученика возрастать (т.е. расходы на одного ученика у нас непостоянны).

### **Обсуждение результатов оценок регрессий**

Результаты МНК-оценок регрессий представлены в таблице 8.3. Каждый из столбцов, обозначенных с (1) по (7), представляет отдельную регрессию. В таблице приведены оценки коэффициентов, их стандартные ошибки, определенные  $F$ -статистики и их  $p$ -значения и статистики качества регрессий, как указано в описании в каждой строке.

Таблица 8.3

#### **Нелинейные регрессионные модели зависимости результатов тестов от различных характеристик**

<b>Зависимая переменная: средняя оценка за тест в школьном округе; 420 наблюдений</b>							
<b>Объясняющая переменная</b>	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Соотношение учеников и учителей ( $STR$ )	-1,00** (0,27)	-0,73** (0,26)	-0,97 (0,59)	-0,53 (0,34)	64,33** (24,86)	83,70** (28,50)	65,29** (25,26)
$STR^2$					-3,42** (1,25)	-4,38** (1,44)	-3,47** (1,27)
$STR^3$					0,059** (0,021)	0,075** (0,024)	0,060** (0,021)
% школьников, изучающих английский язык	-0,122** (0,033)	-0,176** (0,034)					-0,166 (0,034)
Бинарная переменная, равная единице, если процент школьников, изучающих английский язык, не меньше 10 % ( $HiEL$ )			5,64 (19,51)	5,50 (9,80)	-5,47** (1,03)	816,1* (327,7)	
$HiEL \times STR$			-1,28 (0,97)	-0,58 (0,50)		-123,3* (50,2)	
$HiEL \times STR^2$						6,12* (2,54)	
$HiEL \times STR^3$						-0,101* (0,043)	
% школьников, имеющих право на субсидированный обед	-0,547** (0,024)	-0,398** (0,033)		-0,411** (0,029)	-0,420** (0,029)	-0,418** (0,029)	-0,402** (0,033)
Логарифм среднего подушевого дохода в округе		11,57** (1,81)		12,12** (1,80)	11,75** (1,78)	11,80** (1,78)	11,51** (1,81)
Константа	700,2** (5,6)	658,6** (8,6)	682,2** (11,9)	653,6** (9,9)	252,0 (163,6)	122,3 (185,5)	244,8 (165,7)
<i>F</i> -статистики и $p$ -значения для проверки совместных гипотез							
(a) Все переменные $STR$ и компоненты взаимодействия равны нулю			5,64 (0,004)	5,92 (0,003)	6,31 (<0,001)	4,96 (<0,001)	5,91 (0,001)

Окончание таблицы 8.3

<b>Зависимая переменная: средняя оценка за тест в школьном округе; 420 наблюдений</b>							
Объясняющая переменная	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(6) $STR^2, STR^3 = 0$					6,17 (<0,001)	5,81 (0,003)	5,96 (0,003)
(в) $HiEL \times STR$ , $HiEL \times STR^2$ , $HiEL \times STR^3 = 0$						2,69 (0,046)	
$STR$	9,08	8,64	15,88	8,63	8,56	8,55	8,57
$\bar{R}^2$	0,773	0,794	0,305	0,795	0,798	0,799	0,798

Примечание. Регрессии оценены на данных по школьным округам Калифорнии, описанным в приложении 4.1. В скобках под коэффициентами приведены стандартные ошибки; в скобках под F-статистиками приведены соответствующие им p-значения. Коэффициенты значимы на \*5 %-м или \*\*1 %-м уровнях значимости.

В первом столбце таблицы, обозначенном (1), представлены оценки регрессии, которые совпадают с оценками регрессии (3) в таблице 7.1 и повторяются здесь для удобства. Данная регрессия не учитывает доход, поэтому первое, что мы делаем, это проверяем, существенно ли изменяются результаты оценки при включении логарифма доходов в качестве дополнительной экономической контрольной переменной. Результаты приведены в регрессии (2) в таблице 8.3. Логарифм дохода является статистически значимым на 1 %-м уровне значимости, а коэффициент при соотношении учеников и учителей становится более близким к нулю, изменяясь с -1,00 до -0,73, хотя и остается статистически значимым на 1 %-м уровне значимости. Изменение коэффициента при  $STR$  при переходе от регрессии (1) к регрессии (2) достаточно велико для обоснования необходимости включения логарифма дохода и в другие регрессии в качестве фактора, корректирующего возможное смещение из-за пропущенных переменных.

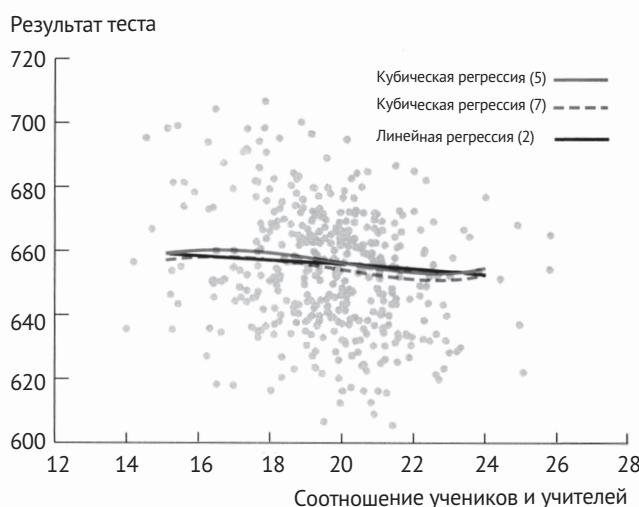
Регрессия (3) в таблице 8.3 – это регрессия с компонентой взаимодействия из уравнения (8.34) с бинарной переменной, характеризующей округа с высоким процентом изучающих английский язык школьников, но при отсутствии экономических контрольных переменных. Если добавить экономические контрольные переменные (процент имеющих право на субсидированные обеды школьников и логарифм доходов) [регрессия (4) в таблице], коэффициенты изменятся, но ни в одном из рассмотренных случаев коэффициент при компоненте взаимодействия не является значимым на 5 %-м уровне значимости. Основываясь на полученных в регрессии (4) оценках, мы не можем отвергнуть гипотезу о том, что эффект влияния  $STR$  одинаков для округов с низким и высоким процентом изучающих английский язык школьников: гипотеза не может быть отвергнута на 5 %-м уровне значимости ( $t$ -статистика равна  $t = -0,58 / 0,50 = -1,16$ ).

Регрессия (5) проверяет, зависит ли эффект влияния изменения соотношения учеников и учителей от величины соотношения учеников и учителей, включая кубическую спецификацию для  $STR$  в дополнение к другим контрольным переменным в регрессии (4) [компоненты взаимодействия,  $HiEL \times STR$ , была отброшена, поскольку она не была значима в регрессии (4) на 10 %-м уровне значимости]. Оценки в регрессии (5) согласуются с предположением о том, что соотношение учеников и учителей влияет на результаты тестов нелинейно. Нулевая гипотеза

о том, что это влияние линейно, отвергается на 1%-м уровне значимости против альтернативы, что оно кубическое ( $F$ -статистика для проверки гипотезы о том, что истинные коэффициенты при  $STR^2$  и  $STR^3$  нулевые, составляет 6,17 с  $p$ -значением  $<0,001$ ).

В рассмотренной далее регрессии (6) анализируется, зависит ли эффект влияния соотношения учеников и учителей в школе только от самого соотношения учеников и учителей, но также и от доли изучающих английский язык школьников. Включая взаимодействие между  $HIEL$  и  $STR$ ,  $STR^2$  и  $STR^3$ , мы можем проверить, различны ли (возможно кубические) теоретические функции регрессии, связывающие результаты тестов и  $STR$  для школ с низким и высоким процентом изучающих английский язык учеников. Для проверки этого мы тестируем гипотезу о том, что коэффициенты при всех трех компонентах взаимодействия нулевые. Результатирующая  $F$ -статистика равна 2,69 с соответствующим  $p$ -значением, равным 0,046, и, таким образом, значима на 5%-м, но не на 1%-м уровне значимости. Все это некоторым образом свидетельствует о том, что функции регрессий различны для округов с высоким и низким процентом изучающих английский язык школьников, однако сравнение регрессий (6) и (4) говорит, скорее, о том, что различия между этими регрессиями связаны с включением в регрессию (6) квадратичной и кубической компонент.

Регрессия (7) является модификацией регрессии (5), в которой непрерывная переменная  $PctEL$  использована вместо бинарной переменной  $HIEL$  для учета процента изучающих английский язык в округе. В результате этой модификации коэффициенты при остальных регрессорах не изменяются существенно, указывая на то, что результаты оценки регрессии (5) не чувствительны к тому, какая мера процента изучающих английский язык фактически используется в регрессии.



**Рисунок 8.10. Три функции регрессии зависимости результатов тестов от соотношения учеников и учителей**

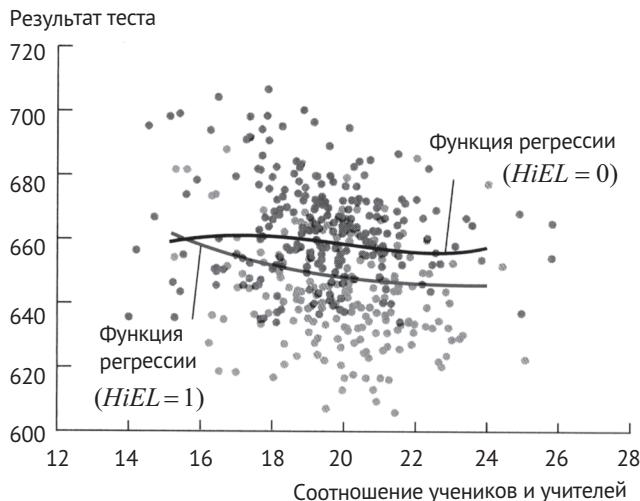
Кубические регрессии из столбцов (5) и (7) в таблице 8.3 почти идентичны. Они указывают на некоторую нелинейность в зависимости результатов тестов от соотношения учеников и учителей.

Во всех спецификациях гипотеза о том, что соотношение учеников и учителей не входит в регрессию, отвергается на 1 %-м уровне значимости.

Нелинейную спецификацию из таблицы 8.3 легче всего изобразить графически. На рисунке 8.10 изображены оцененные функции регрессии, характеризующей влияние соотношения учеников и учителей на результаты тестов для линейной спецификации (2) и кубических спецификаций (5) и (7), а также диаграмма рассеяния данных<sup>1</sup>. Оцененные регрессии показывают зависимость предсказанных значений результатов тестов от соотношения учеников и учителей при фиксированных остальных переменных в регрессии. Все оцененные функции регрессии близки друг к другу, хотя кубические регрессии являются более плоскими для больших значений соотношения учеников и учителей.

Регрессия (6) показывает наличие статистически значимых различий в оценках кубических регрессий, характеризующих зависимость результатов тестов от  $STR$  и учитывающих размер доли изучающих английский язык школьников в округе. На рисунке 8.11 изображены эти две оцененные функции регрессии так, чтобы мы могли видеть, имеет ли это различие, в дополнение к статистической значимости, практическую важность. Как показано на рисунке 8.11, если значение соотношения учеников и учителей находится в диапазоне между 17 и 23, то есть в диапазоне, который включает 88 % наблюдений, две функции отделены приблизительно десятью пунктами, но в остальном очень похожи. Таким образом, для значений  $STR$  между 17 и 23 округа с низким процентом изучающих английский язык школьников сдают экзамены лучше при постоянном соотношении учеников и учителей, но эффект от изменения соотношения учеников и учителей по сути одинаков для этих двух групп. Две функции регрессии различны для значений отношения учеников и учителей ниже 16,5, но нам следует с осторожностью интерпретировать результаты оценок для таких значений соотношения учеников и учителей. Округа со значениями  $STR < 16,5$  составляют всего 6 % наблюдений, поэтому различие между нелинейными функциями регрессий отражает различие в этих немногих округах с очень низким соотношением учеников и учителей. Таким образом, на основе рисунка 8.11 мы заключаем, что эффект влияния изменений в соотношении учеников и учителей на результаты тестов не зависит от процента изучающих английский язык для того диапазона значений соотношения учеников и учителей, для которого мы имеем большинство наблюдений, в том смысле что при фиксированном значении числа изучающих английский язык школьников изменение соотношения учеников и учителей практически не влияет на результаты тестов в округе.

<sup>1</sup> Для каждой кривой предсказанное значение результата теста было вычислено путем замены каждой независимой переменной, кроме  $STR$ , его выборочным средним значением и вычисления предсказанного значения путем умножения этих фиксированных значений независимых переменных на их соответствующие оцененные коэффициенты из таблицы 8.3. Это было сделано для различных значений  $STR$ , и график таким образом предсказанных значений представляет собой оцененную функцию регрессии, характеризующую зависимость результатов тестов от  $STR$  при равенстве других независимых переменных их выборочным средним.



**Рисунок 8.11. Функции регрессии для округов с высоким и низким процентом изучающих английский язык школьников**

Округа с низким процентом изучающих английский язык ( $HiEL = 0$ ) отмечены темными точками, а округа с  $HiEL = 1$  — светлыми точками. График функции кубической регрессии для  $HiEL = 1$  в регрессии (6) в таблице 8.3 располагается приблизительно на 10 пунктов ниже графика функции кубической регрессии для  $HiEL = 0$  для значений  $17 \leq STR \leq 23$ , но в остальном эти две функции имеют аналогичную форму и углы наклона в этом диапазоне. Угловые коэффициенты функций регрессии отличаются больше всего для очень больших и очень малых значений  $STR$ , число наблюдений которых в выборке мало.

### *Краткое изложение результатов*

Полученные результаты позволяют ответить на три вопроса, поднятые в начале данного раздела.

Во-первых, после учета экономического положения семей в школьном округе или числа изучающих английский язык школьников можно говорить о том, что изменение соотношения учеников и учителей не оказывает существенного влияния на результаты тестов в округе. В линейных спецификациях не обнаружено статистически значимого свидетельства такого различия. Кубическая спецификация в регрессии (6) предоставляет статистически значимое свидетельство (на 5 %-м уровне значимости) того, что функции регрессии различны для округов с высоким и низким процентом изучающих английский язык. Однако, как показано на рисунке 8.11, оцененные функции регрессии имеют похожие угловые коэффициенты в диапазоне значений соотношения учеников и учителей, в котором содержится большая часть наблюдений.

Во-вторых, после учета экономического положения семей в округе обнаружено свидетельство существования нелинейности эффекта влияния соотношения учеников и учителей на результаты тестов. Этот эффект статистически значим на 1 %-м уровне значимости (коэффициенты при  $STR^2$  и  $STR^3$  всегда значимы на 1 %-м уровне значимости).

В-третьих, теперь мы можем вернуться к проблеме, связанной с окружным школьным инспектором, которая была поставлена в главе 4. Окружной школь-

ный инспектор хочет знать, какой эффект на результаты тестов окажет уменьшение соотношения учеников и учителей на 2. В линейной спецификации (2) этот эффект не зависит от самого соотношения учеников и учителей, и оцененный эффект такого уменьшения улучшает результаты тестов на 1,46 ( $= -0,73 \times -2$ ) пункта. В нелинейных спецификациях этот эффект зависит от значения соотношения учеников и учителей. Если в школьном округе в настоящее время соотношение учеников и учителей равно 20 и окружной школьный инспектор рассматривает возможность его уменьшения до 18, тогда на основании регрессии (5) оцененный эффект от этого уменьшения улучшает результаты тестов на 3,00 пункта, в то время как на основании регрессии (7) эта оценка составляет 2,93. Если в ее округе в настоящее время соотношение учеников и учителей составляет 22 и она рассматривает его сокращение до 20, тогда на основании регрессии (5) оцененный эффект от этого уменьшения улучшает результаты теста на 1,93 пункта, в то время как на основании регрессии (7) эта оценка составляет 1,90. Оценки нелинейных спецификаций предполагают, что сокращение соотношения учеников и учителей оказывает больший эффект, если это отношение уже мало.

## 8.5. Заключение

В данной главе рассмотрено несколько способов моделирования нелинейных функций регрессии. Поскольку рассмотренные модели являются различными вариантами модели множественной регрессии, неизвестные коэффициенты могут быть оценены при помощи МНК, а гипотезы о значениях этих коэффициентов могут быть проверены, используя  $t$ - и  $F$ -статистики, как описано в главе 7. В этих моделях ожидаемый эффект влияния на  $Y$  от изменения независимой переменной  $X_1$  на единицу при постоянстве других независимых переменных  $X_2, \dots, X_k$  в общем случае зависит от значений  $X_1, X_2, \dots, X_k$ .

Мы рассмотрели большое количество различных моделей, поэтому может возникнуть некоторое недоумение по поводу того, какую модель использовать в данном примере. Как следует анализировать возможные виды нелинейности на практике? В разделе 8.1 изложен общий подход к такому анализу, но этот подход требует от вас принятия и осуществления некого набора шагов при его реализации. Было бы удобно, если бы существовал единственный рецепт, которому вы могли бы следовать и который работал бы всегда и во всех приложениях, но на практике редко бывает все так просто.

Самым важным шагом при спецификации нелинейных функций регрессии является «использование своей собственной головы». Прежде чем посмотреть на данные, могли ли вы поразмышлять о причине того, почему угловой коэффициент теоретической функции регрессии может зависеть от значения той или иной независимой переменной и как это следует из экономической теории? Если да, то какую зависимость вы могли бы ожидать? И самое главное, нелинейности какого типа (если таковые имеются) могут оказывать серьезное влияние при ответе на основные вопросы, рассматриваемые

в вашем исследовании? Ответы на эти вопросы позволят вам провести анализ аккуратно. Например, в приложении к результатам тестов такие рассуждения привели нас к исследованию того, дает ли найм дополнительных учителей больший эффект в районах с большим процентом учеников, все еще изучающих английский язык; это происходит, возможно, вследствие того, что в данном случае школьники будут извлекать личную пользу от наличия большего персонального внимания со стороны учителя. Задавая этот конкретный вопрос, мы смогли найти точный ответ: после учета экономического положения студентов мы не обнаружили статистически значимых доказательств такого взаимодействия.

## **Выводы**

1. В нелинейной регрессии угловой коэффициент теоретической функции регрессии зависит от значения одной или более независимых переменных.
2. Влияние на  $Y$  изменения независимой(ых) переменной(ых) может быть вычислено как разность предсказанных значений зависимой переменной при двух различных значениях независимой(ых) переменной(ых). Данная процедура описана во вставке «Основные понятия 8.1».
3. Полиномиальная регрессия включает степени  $X$  в качестве регрессоров. Квадратичная регрессия включает  $X$  и  $X^2$ , а кубическая регрессия включает  $X$ ,  $X^2$  и  $X^3$ .
4. Небольшие изменения в логарифмах могут быть интерпретированы как пропорциональные или процентные изменения в переменной. Регрессии, включающие логарифмы, используются для оценивания пропорциональных изменений и эластичностей.
5. Произведение двух переменных называется компонентой взаимодействия. Если компоненты взаимодействия включены в модель в качестве регрессоров, то угловой коэффициент регрессии одной переменной зависит от другой переменной.

## **Основные понятия**

- Модель квадратичной регрессии (с. 262).  
Функция нелинейной регрессии (с. 265).  
Модель полиномиальной регрессии (с. 269).  
Модель кубической регрессии (с. 270).  
Эластичность (с. 272).  
Экспоненциальная функция (с. 272).  
Натуральный логарифм (с. 272).  
Линейно-логарифмическая модель (с. 274).  
Логарифмически-линейная модель (с. 275).  
Линейная в логарифмах модель (с. 276).  
Компонента взаимодействия (с. 282).

Регрессор взаимодействия (с. 282).

Модель регрессии с компонентой взаимодействия (с. 282).

Нелинейный метод наименьших квадратов (с. 314).

Оценки нелинейного метода наименьших квадратов (с. 315).

## **Вопросы для повторения и закрепления основных понятий**

- 8.1. Нарисуйте функцию регрессии, которая будет возрастать (иметь положительный наклон) и будет более крутой для малых значений  $X$ , но менее крутой для больших значений  $X$ . Объясните, как вы должны специфицировать нелинейную регрессию для моделирования такой формы. Можете ли вы привести экономический пример такого соотношения?
- 8.2. Производственная функция Кобба–Дугласа устанавливает связь между выпуском ( $Q$ ) и факторами производства: капиталом ( $K$ ), трудом ( $L$ ) и сырьем ( $M$ ) и компонентой ошибок  $u$ , используя соотношение  $Q = \lambda K^{\beta_1} L^{\beta_2} M^{\beta_3} e^u$ , где  $\lambda$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  являются параметрами производственной функции. Предположим, что у вас есть данные по выпуску и факторам производства из случайной выборки фирм с одинаковой производственной функцией Кобба–Дугласа. Как бы вы использовали регрессионный анализ для оценки параметров этой производственной функции?
- 8.3. Стандартная функция «спроса на деньги», которая используется макроэкономистами, имеет форму  $\ln(m) = \beta_0 + \beta_1 \ln(GDP) + \beta_2 R$ , где  $m$  – (реальная) денежная масса,  $GDP$  – значение (реального) валового внутреннего продукта и  $R$  – значение номинальной процентной ставки, измеренной в процентах за год. Предположим, что  $\beta_1 = 1,0$  и  $\beta_2 = -0,02$ . Что случится со значением  $m$ , если  $GDP$  увеличится на 2 %? Что случится с  $m$ , если процентная ставка увеличится с 4 до 5 %?
- 8.4. Вы оценили модель линейной регрессии  $Y$  от  $X$ . Ваш профессор сказал: «Я думаю, что соотношение между  $Y$  и  $X$  нелинейное». Объясните, как можно проверить адекватность вашей линейной регрессии.
- 8.5. Предположим, что в упражнении 8.2 вы думали, что значение  $\beta_2$  не является постоянным, а возрастает при росте  $K$ . Как вы могли бы использовать компоненту взаимодействия, чтобы учесть этот эффект?

## **Упражнения**

- 8.1. В 2009 году продажи в компании составили 196 млн долл. и возросли до 198 млн долл. в 2010 году.
  - a) Вычислите процентное изменение продаж, используя обычную формулу  $100 \times \frac{Sales_{2010} - Sales_{2009}}{Sales_{2009}}$ . Сравните это значение с аппроксимацией  $100 \times [\ln(Sales_{2010}) - \ln(Sales_{2009})]$ .

- б) Повторите (a), предполагая  $Sales_{2010} = 205$ ,  $Sales_{2010} = 250$  и  $Sales_{2010} = 500$ .  
 в) Насколько хороша аппроксимация, когда изменение мало? Ухудшается ли качество аппроксимации при увеличении процентного изменения?
- 8.2. Предположим, что исследователь собирает данные о домах, которые проходятся в определенном районе за прошедший год, и получает оценки регрессий, представленные в таблице ниже.
- а) Используя результаты из столбца (1), определите ожидаемое изменение цены дома, если его площадь увеличится на 500 квадратных футов. Постройте 95 %-й доверительный интервал для процентного изменения стоимости дома.

### Результаты оценки регрессий для упражнения 8.2

<b>Зависимая переменная: <math>\ln(Price)</math></b>					
<b>Объясняющая переменная</b>	(1)	(2)	(3)	(4)	(5)
<i>Size</i>	0,000 42 (0,000 038)				
$\ln(Size)$		0,69 (0,054)	0,068 (0,087)	0,57 (2,03)	0,69 (0,055)
$\ln(Size)^2$				0,0078 (0,14)	
<i>Bedrooms</i>			0,0036 (0,037)		
<i>Pool</i>	0,082 (0,032)	0,071 (0,034)	0,071 (0,034)	0,071 (0,036)	0,071 (0,035)
<i>View</i>	0,037 (0,029)	0,027 (0,028)	0,026 (0,026)	0,027 (0,029)	0,027 (0,030)
<i>Pool</i> $\times$ <i>View</i>					0,002 2 (0,10)
<i>Condition</i>	0,13 (0,045)	0,12 (0,035)	0,12 (0,035)	0,12 (0,036)	0,12 (0,035)
Константа	10,97 (0,069)	6,60 (0,39)	6,63 (0,53)	7,02 (7,50)	6,60 (0,40)
<b>Статистики качества регрессий</b>					
<i>SER</i>	0,102	0,098	0,099	0,099	0,099
$\bar{R}^2$	0,72	0,74	0,73	0,73	0,73

Примечание. Определения переменных: *Price* = цена продажи (долл.); *Size* = площадь дома (в квадратных футах); *Bedrooms* = число спален; *Pool* = бинарная переменная (1 – если дом имеет бассейн, 0 – в противном случае); *View* = бинарная переменная (1 – если у дома прекрасный вид, 0 – в противном случае); *Condition* = бинарная переменная (1 – если агент по недвижимости сообщает об отличном состоянии дома, 0 – в противном случае).

- б) Сравнивая столбцы (1) и (2), скажите, какую переменную лучше использовать для объяснения стоимости домов: *Size* или  $\ln(\text{Size})$ ?  
 в) Используя столбец (2), выясните, чему равен оцененный эффект влияния наличия бассейна в доме на его цену? (Убедитесь, что единицы измерения корректны.) Постройте 95 %-й доверительный интервал для этого эффекта.

- г) В регрессию из столбца (3) добавлена переменная, характеризующая число спален в доме. Насколько велик оцененный эффект от наличия дополнительной спальни? Является ли этот эффект статистически значимым? Почему, как вы думаете, оцененный эффект так мал? (Подсказка: какие другие переменные остаются постоянными?)
- д) Важна ли квадратичная компонента  $\ln(\text{Size})^2$ ?
- е) Используйте регрессию из столбца (5) для вычисления ожидаемого изменения в цене при наличии бассейна в доме без красивого вида. Является ли различие большим? Является ли различие статистически значимым?
- 8.3. После прочтения раздела об анализе зависимости результатов тестов от размера класса учитель говорит: «По моему опыту, успеваемость учащихся зависит от размера класса, но не так, как показывают ваши регрессии. Скорее всего, ученики пишут тест хорошо, когда размер класса составляет менее 20 учеников, и выполняют его очень плохо, когда его размер больше 25. Не существует никаких выгод от снижения размера класса ниже 20 учеников, влияние размера класса постоянно в промежуточной области между 20 и 25 учениками, и нет потери от увеличения размера класса, когда он уже больше, чем 25». Учитель описывает «пороговый эффект», в котором производительность является постоянной для класса размером менее 20, а затем перескакивает и является постоянной для класса размером от 20 до 25, а затем снова перескакивает для класса размером более 25. Для моделирования этих пороговых эффектов определим бинарные переменные:
- $STRsmall = 1$ , если  $STR < 20$  и  $STRsmall = 0$  – в противном случае;
- $STRmoderate = 1$ , если  $20 \leq STR \leq 25$  и  $STRmoderate = 0$  – в противном случае; и
- $STRlarge = 1$ , если  $STR > 25$  и  $STRlarge = 0$  – в противном случае.
- а) Рассмотрите регрессию  $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$ . Начертите график функции регрессии, описывающей зависимость  $TestScore$  от  $STR$  для гипотетических значений коэффициентов регрессии, которые соответствуют заявлению учителя.
- б) Исследователь пытается оценить регрессию  $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge_i + u_i$  и обнаруживает, что в его компьютере происходит сбой. Почему?
- 8.4. Прочитайте вставку «Отдача от образования и гендерный разрыв» из раздела 8.3.
- а) Рассмотрите мужчину, проживающего на Западе, число лет образования которого равно 16 годам, а трудовой стаж – двум. Используйте результаты столбца (4) таблицы 8.1 и метод, описанный во вставке «Основные понятия 8.1», для оценки ожидаемого изменения логарифма средней почасовой зарплаты ( $AHE$ ) от дополнительного года трудового стажа.

- б) Повторите (a), предполагая, что трудовой стаж равен 10 годам.
- в) Объясните, почему ответы в пунктах (a) и (б) различны?
- г) Значимо ли статистически различие между пунктами (a) и (б) на 5 %-м уровне значимости? Объясните.
- д) Как изменились бы ответы на вопросы с (a) по (г), если индивид был бы женщиной? Если бы индивид жил на Юге? Объясните.
- е) Как вы изменили бы регрессию, если подозревали, что эффект влияния трудового стажа на заработную плату был бы различным для мужчин и женщин?

8.5. Прочитайте вставку «Спрос на экономические журналы» из раздела 8.3.

- а) Во вставке получено три результата. Скажите, на основе чего сделан каждый из этих выводов?
- б) Из оценок регрессии (4), следует, что эластичность спроса для 80-летних журналов составляет  $-0,28$ .
  - (i) Как это значение было определено из оцененной регрессии?
  - (ii) Во вставке говорится, что стандартная ошибка для оцененной эластичности составляет 0,06. Как вы могли бы вычислить эту стандартную ошибку? (Подсказка: Смотрите обсуждение «Стандартная ошибка ожидаемого изменения» на с. 267.)
- в) Предположим, что переменная Characters была разделена на 1 000 вместо 1 000 000. Как изменятся результаты, представленные в столбце (4)?

8.6. Вернемся к оценкам, представленным в таблице 8.3.

- а) Исследователь подозревает, что переменная, характеризующая процент школьников, имеющих право на субсидированный обед, оказывает нелинейный эффект на результаты тестов. В частности, он предполагает, что возрастание этой переменной с 10 до 20 % оказывает небольшой эффект на результаты тестов, но ее изменение с 50 до 60 % оказывает гораздо больший эффект.
  - (i) Опишите нелинейную спецификацию, которая может быть использована для моделирования этой формы нелинейности.
  - (ii) Как вы могли бы проверить, является ли предположение исследователя лучшим, чем линейная спецификация в столбце (7) таблицы 8.3?
- б) Исследователь подозревает, что влияние доходов семей на результаты тестов различны для округов с небольшими классами и для округов с большими классами.
  - (i) Опишите нелинейную спецификацию, которая может быть использована для моделирования этой формы нелинейности.
  - (ii) Как вы могли бы проверить, является ли предположение исследователя лучшим, чем линейная спецификация в столбце (7) таблицы 8.3?

8.7. Это упражнение придумано под влиянием исследования «гендерного разрыва» в доходах топ-менеджеров открытых акционерных обществ, проведенного Берtrandом и Хэллоком (Bertrand and Hallock, 2001). В исследовании сравниваются суммы полных компенсаций топ-менеджеров в большой выборке ОАО в США в 1990 году. (Каждый год эти ОАО обязаны публиковать информацию об общем уровне компенсаций для первой пятерки своих топ-менеджеров.)

- а) Пусть  $Female$  будет фиктивной переменной, которая равна 1 для женщин и 0 для мужчин. Регрессия логарифма доходов на  $Female$  дает оценки:

$$\widehat{\ln(Earnings)} = 6,48 - 0,44 \widehat{Female}, SER = 2,65.$$

(0,01)      (0,05)

- (i) Оцененный коэффициент при  $Female$  равен – 0,44. Объясните, что означает это значение?  
(ii)  $SER$  равна 2,65. Объясните, что означает это значение?  
(iii) Предполагает ли эта регрессия, что женщины топ-менеджеры зарабатывают меньше, чем топ-менеджеры мужчины? Объясните.  
(iv) Предполагает ли эта регрессия, что существует половая дискриминация? Объясните.
- б) Две новых переменных, рыночная стоимость фирмы (показатель размера фирмы в млн долл.,  $MarketValue$ ) и доходность акций (показатель деятельности фирмы в процентных пунктах,  $Return$ ), добавлены в регрессию:

$$\widehat{\ln(Earnings)} = 3,86 - 0,28 \widehat{Female} + 0,37 \ln(MarketValue) + 0,004 \widehat{Return},$$

(0,03)      (0,04)      (0,004)      (0,003)

$$n = 46670, R^2 = 0,345.$$

- (i) Коэффициент при  $\ln(MarketValue)$  равен 0,37. Объясните, что означает это значение.  
(ii) Коэффициент при  $Female$  равен –0,28. Объясните, почему он изменился по сравнению с регрессией в пункте (a).  
в) Можно ли сказать, что в больших фирмах женщины топ-менеджеры встречаются чаще, чем в небольших? Объясните.
- 8.8. Пусть  $X$  является непрерывной переменной, которая принимает значения между 5 и 100.  $Z$  является бинарной переменной. Начертите схематически графики следующих функций регрессии (со значениями  $X$  между 5 и 100 на горизонтальной оси и значениями  $\hat{Y}$  на вертикальной оси):
- а)  $\hat{Y} = 2,0 + 3,0 \times \ln(X)$ .  
б)  $\hat{Y} = 2,0 - 3,0 \times \ln(X)$ .  
в) (i)  $\hat{Y} = 2,0 + 3,0 \times \ln(X) + 4,0Z$  с  $Z = 1$ .  
(ii) Аналогично (i), но с  $Z = 0$ .

- г) (i)  $\hat{Y} = 2,0 + 3,0 \times \ln(X) + 4,0Z - 1,0 \times Z \times \ln(X)$  с  $Z = 1$ .  
 (ii) Аналогично (i), но с  $Z = 0$ .  
 д)  $\hat{Y} = 1,0 + 125,0X - 0,01X^2$ .
- 8.9. Объясните, как вы будете использовать «Подход № 2» из раздела 7.3 для расчета доверительного интервала обсуждаемого ниже уравнения (8.8). [Подсказка: оцените новую регрессию с использованием различных определений регрессоров и зависимой переменной. См. упражнение (7.9).]
- 8.10. Рассмотрим модель регрессии  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$ . Используя вставку «Основные понятия 8.1», покажите:
- а)  $\Delta Y / \Delta X_1 = \beta_1 + \beta_3 X_2$  (эффект от изменения  $X_1$  при постоянной  $X_2$ ).
  - б)  $\Delta Y / \Delta X_2 = \beta_2 + \beta_3 X_1$  (эффект от изменения  $X_2$  при постоянной  $X_1$ ).
  - в) Если  $X_1$  изменяется на  $\Delta X_1$ , а  $X_2$  изменяется на  $\Delta X_2$ , тогда  $\Delta Y = (\beta_1 + \beta_3 X_2) \Delta X_1 + (\beta_2 + \beta_3 X_1) \Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$ .
- 8.11. Получите выражение для эластичностей, приведенных в приложении 8.2, для линейной и линейной в логарифмах моделей. (Подсказка: для линейной в логарифмах модели предположите, что  $u$  и  $X$  независимы, как сделано в приложении 8.2 для логарифмически-линейной модели.)
- 8.12. Следующее за выражением (8.28) обсуждение дает интерпретацию коэффициента при бинарной компоненте взаимодействия, используя предположение о равенстве нулю условного среднего. Данное упражнение показывает, что такая интерпретация также может быть использована при независимости условного среднего. Рассмотрим гипотетический эксперимент, описанный в упражнении 7.11.
- а) Предположим, что вы оцениваете регрессии  $Y_i = \gamma_0 + \gamma_1 X_{1i} + u_i$ , используя только старых учеников (т.е. учившихся в школе и раньше). Покажите, что  $\gamma_1$  – это эффект влияния размера класса для старых учеников, то есть что  $\gamma_1 = E(Y_i | X_{1i} = 1, X_{2i} = 0) - E(Y_i | X_{1i} = 0, X_{2i} = 0)$ . Объясните, почему  $\hat{\gamma}_1$  является несмещенной оценкой  $\gamma_1$ ?
  - б) Предположим, что вы оцениваете регрессию  $Y_i = \delta_0 + \delta_1 X_{1i} + u_i$ , используя только данные о новых учениках. Покажите, что  $\delta_1$  есть эффект влияния размера класса для новых учеников, то есть что  $\delta_1 = E(Y_i | X_{1i} = 1, X_{2i} = 1) - E(Y_i | X_{1i} = 0, X_{2i} = 1)$ . Объясните, почему  $\hat{\delta}_1$  является несмещенной оценкой  $\delta_1$ ?
  - в) Рассмотрим регрессию для старых и новых учеников  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$ . Используйте предположение независимости условного среднего  $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ , чтобы показать, что  $\beta_1 = \gamma_1$ ,  $\beta_1 + \beta_3 = \delta_1$  и  $\beta_3 = \delta_1 - \gamma_1$  (разность в эффектах влияния размера класса).
  - г) Предположим, что вы оцениваете регрессию взаимодействия в (с), используя комбинированные данные, и что  $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ .

Покажите, что  $\hat{\beta}_1$  и  $\hat{\beta}_3$  являются несмешенными, но  $\hat{\beta}_2$ , в общем случае, смещена.

## Компьютерные упражнения

E8.1. Используя базу данных *CPS08*, описанную в E4.1, ответьте на следующие вопросы:

- a) Оцените регрессию средней почасовой зарплаты (*AHE*) на возраст (*Age*), пол (*Female*) и образование (*Bachelor*). Каково ожидаемое изменение зарплаты, если *Age* увеличивается с 25 до 26? Каково ожидаемое изменение зарплаты, если *Age* возрастет с 33 до 34?
- б) Оцените регрессию логарифма средней почасовой зарплаты  $\ln(AHE)$  на *Age*, *Female* и *Bachelor*. Каково ожидаемое изменение зарплаты, если *Age* увеличивается с 25 до 26? Каково ожидаемое изменение зарплаты, если *Age* возрастет с 33 до 34?
- в) Оцените регрессию логарифма средней почасовой зарплаты  $\ln(AHE)$  на  $\ln(\text{Age})$ , *Female* и *Bachelor*. Каково ожидаемое изменение зарплаты, если *Age* увеличивается с 25 до 26? Каково ожидаемое изменение зарплаты, если *Age* возрастет с 33 до 34?
- г) Оцените регрессию логарифма средней почасовой зарплаты  $\ln(AHE)$  на *Age*,  $\text{Age}^2$ , *Female* и *Bachelor*. Каково ожидаемое изменение зарплаты, если *Age* увеличивается с 25 до 26? Каково ожидаемое изменение зарплаты, если *Age* возрастет с 33 до 34?
- д) Предпочитаете ли вы регрессию из пункта (в) регрессии из пункта (б)? Объясните.
- е) Предпочитаете ли вы регрессию из пункта (г) регрессии из пункта (б)? Объясните.
- ж) Предпочитаете ли вы регрессию из пункта (г) регрессии из пункта (в)? Объясните.
- з) Начертите графики функций регрессии между *Age* и  $\ln(AHE)$  из пунктов (б), (в) и (г) для мужчин, имеющих высшее образование. Объясните сходства и различия между оцененными функциями регрессий. Изменится ли ваш ответ, если вы начертите графики функций регрессии для женщин с высшим образованием?
- и) Оцените регрессию  $\ln(AHE)$  на *Age*,  $\text{Age}^2$ , *Female* и *Bachelor* и компоненту взаимодействия  $\text{Female} \times \text{Bachelor}$ . Что измеряет коэффициент при компоненте взаимодействия? Алексис является 30-летней женщиной с дипломом бакалавра. Чему равно предсказанное значение  $\ln(AHE)$  для Алексис? Джейн является 30-летней женщиной со средним образованием. Чему равно предсказанное значение  $\ln(AHE)$  для Джейн? Какова предсказанная разность между зарплатами Алексис и Джейн? Боб является 30-летним мужчиной

с дипломом бакалавра. Чему равно предсказанное значение  $\ln(AHE)$  для него? Джим является 30-летним мужчиной со средним образованием. Чему равно предсказанное значение  $\ln(AHE)$  для Джима? Какова предсказанная разность между зарплатами Боба и Джима?

- к) Различается ли эффект влияния переменной  $Age$  на доходы мужчин и женщин? Специфицируйте и оцените регрессию, которую вы можете использовать для ответа на этот вопрос.
- л) Различается ли эффект влияния переменной  $Age$  на доходы людей, имеющих среднее и высшее образование? Специфицируйте и оцените регрессию, которую вы можете использовать для ответа на этот вопрос.
- м) Оценив все эти регрессии (и любые другие, которые вы хотите оценить), сделайте выводы об эффекте влияния возраста на доходы молодых работников.

E8.2. Используя базу данных *TeachingRating*, описанную в E4.2, выполните следующие упражнения:

- а) Оцените регрессию переменной *Course\_Eval* от *Beauty*, *Intro*, *OneCredit*, *Female*, *Minority* и *NNEnglish*.
- б) Добавьте переменные *Age* и  $Age^2$  в регрессию. Существует ли свидетельство того, что эффект влияния на *Course\_Eval* переменной *Age* не является линейным? Можно ли сказать, что переменная *Age* никак не влияет на *Course\_Eval*?
- в) Модифицируйте регрессию из пункта (а) так, чтобы эффект влияния индекса «красоты» *Beauty* на *Course\_Eval* был различен для мужчин и женщин. Является ли значимой разность эффектов влияния *Beauty* для мужчин и женщин?
- г) Профессор Смит – мужчина. Он сделал косметическую операцию, которая увеличила его индекс «красоты» с одного стандартного отклонения ниже среднего до одного стандартного отклонения выше среднего. Чему было равно его значение *Beauty* до операции? После операции? Используя регрессию из пункта (в) постройте 95 %-й доверительный интервал для роста оценки его курса.
- д) Повторите пункт (г) для случая, когда профессор Джонс является женщиной.

E8.3. Используйте базу данных *CollegeDistance*, описанную в E4.3, для ответа на следующие вопросы:

- а) Оцените регрессию переменной *ED* от *Dist*, *Female*, *Bytest*, *Tuition*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *MomColl*, *Cue80* и *Stwmfg80*. Чему равно ожидаемое изменение числа полных лет образования, если *Dist* увеличивается с 2 до 3 (т.е. с 20 до 30 миль)? Чему равно ожидаемое изменение числа полных лет образования, если *Dist* увеличивается с 6 до 7 (т.е. с 60 до 70 миль)?
- б) Оцените регрессию  $\ln(ED)$  от *Dist*, *Female*, *Bytest*, *Tuition*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *MomColl*, *Cue80* и *Stwmfg80*. Чему равно ожидаемое изменение числа полных лет образования,

если  $Dist$  увеличивается с 2 до 3 (т.е. с 20 до 30 миль)? Чему равно ожидаемое изменение числа полных лет образования, если  $Dist$  увеличивается с 6 до 7 (т.е. с 60 до 70 миль)?

- в) Оцените регрессию  $ED$  от  $Dist$ ,  $Dist^2$ ,  $Female$ ,  $Bytest$ ,  $Tuition$ ,  $Black$ ,  $Hispanic$ ,  $Incomehi$ ,  $Ownhome$ ,  $DadColl$ ,  $MomColl$ ,  $Cue80$  и  $Stwmfg80$ . Чему равно ожидаемое изменение числа полных лет образования, если  $Dist$  увеличивается с 2 до 3 (т.е. с 20 до 30 миль)? Чему равно ожидаемое изменение числа полных лет образования, если  $Dist$  увеличивается с 6 до 7 (т.е. с 60 до 70 миль)?
- г) Предпочитаете ли вы регрессию из пункта (в) регрессии из пункта (а)? Объясните.
- д) Рассмотрим латиноамериканок, у которых  $Tuition = \$950$ ,  $Bytest = 58$ ,  $Incomehi = 0$ ,  $Ownhome = 0$ ,  $Dadcoll = 1$ ,  $MomColl = 1$ ,  $Cue80 = 7,1$  и  $Stwmfg = \$10,06$ .
  - (i) Начертите графики функций регрессии между  $Dist$  и  $ED$  из пунктов (а) и (в) для  $Dist$ , находящейся в диапазоне от 0 до 10 (от 0 до 100 миль). Опишите сходства и различия между оцененными функциями регрессии. Изменится ли ваш ответ, если вы построите функцию регрессии для белых мужчин с теми же самыми характеристиками?
  - (ii) Как ведет себя функция регрессии из пункта (в) для  $Dist > 10$ ? На сколько много наблюдений имеется в выборке при  $Dist > 10$ ?
- е) Добавьте компоненту взаимодействия  $DadColl \times MomCall$  в регрессию из пункта (в). Что измеряет коэффициент при компоненте взаимодействия?
- ж) У Мэри, Джейн, Алексис и Бонни одинаковые значения переменных  $Dist$ ,  $Bytest$ ,  $Tuition$ ,  $Female$ ,  $Black$ ,  $Hispanic$ ,  $Fincome$ ,  $Ownhome$ ,  $Cue80$  и  $Stwmfg80$ . Ни один из родителей Мэри не учился в колледже. Отец Джейн учился в колледже, но мать не училась. Мать Алексис училась в колледже, но ее отец не учился. Оба родителя Бонни учились в колледже. Используя регрессии из пункта (е), ответьте на вопросы:
  - (i) Чему равна предсказанная по этой регрессии разность между числом лет полного образования Джейн и Мэри?
  - (ii) Чему равна предсказанная по этой регрессии разность между числом лет полного образования Алексис и Мэри?
  - (iii) Чему равна предсказанная по этой регрессии разность между числом лет полного образования Бонни и Мэри?
- з) Существует ли какое-нибудь свидетельство того, что эффект влияния расстояния  $Dist$  на  $ED$  зависит от дохода семьи?
- и) Оценив все эти регрессии (и любые другие, которые вы хотите оценить), сделайте выводы об эффекте влияния  $Dist$  на число лет полного образования.

E8.4. Используя базу данных *Growth*, описанную в Е4.4, за исключением данных для Мальты, оцените следующие пять регрессий переменной *Growth* на: (1) *TradeShare* и *YearsSchool*; (2) *TradeShare* и *ln(YearsSchool)*; (3) *TradeShare*, *ln(YearsSchool)*, *Rev\_Coups*, *Assassinations* и *ln(RGDP60)*;

- (4)  $\text{TradeShare}$ ,  $\ln(\text{YearsSchool})$ ,  $\text{Rev_Coups}$ ,  $\text{Assassinations}$ ,  $\ln(\text{RGDP60})$  и  $\text{TradeShare} \times \ln(\text{YearsSchool})$  и (5)  $\text{TradeShare}$ ,  $\text{TradeShare}^2$ ,  $\text{TradeShare}^3$ ,  $\ln(\text{YearsSchool})$ ,  $\text{Rev_Coups}$ ,  $\text{Assassinations}$  и  $\ln(\text{RGDP60})$ .
- a) Постройте диаграмму рассеяния переменных  $\text{Growth}$  и  $\text{YearsSchool}$ . Кажется ли полученное соотношение линейным или нет? Объясните. Используйте график для объяснения, почему регрессия (2) подходит для описания данных лучше, чем регрессия (1).
  - b) В 1960 году страна намеревалась провести образовательную политику, которая увеличит среднюю продолжительность обучения с 4 лет до 6 лет. Используйте регрессию (1), чтобы рассчитать предсказанный рост  $\text{Growth}$ . Используйте регрессию (2), чтобы рассчитать предсказанный рост  $\text{Growth}$ .
  - v) Проверьте, равны ли коэффициенты при  $\text{Assassinations}$  и  $\text{Rev_Coups}$  нулю, используя регрессию (3).
  - g) Используя регрессию (4), скажите, существует ли свидетельство того, что эффект влияния  $\text{TradeShare}$  на  $\text{Growth}$  зависит от уровня образования в стране?
  - d) Используя регрессию (5), скажите, существует ли свидетельство нелинейности эффекта влияния  $\text{TradeShare}$  на  $\text{Growth}$ ?
  - e) В 1960 году страна намеревалась провести торговую политику, которая увеличит среднее значение  $\text{TradeShare}$  с 0,5 до 1. Используйте регрессию (3), чтобы рассчитать предсказанный рост  $\text{Growth}$ . Используйте регрессию (5), чтобы рассчитать предсказанный рост  $\text{Growth}$ .

## Приложения

### Приложение 8.1. Нелинейные по параметрам функции регрессии

Нелинейные функции регрессии, рассмотренные в разделах 8.2 и 8.3, являются нелинейными функциями от  $X$ -ов, но линейными функциями от неизвестных параметров. Поскольку они линейны по неизвестным параметрам, эти параметры могут быть оценены при помощи МНК после определения новых регрессоров, которые являются нелинейными преобразованиями от исходных  $X$ -ов. Такое семейство нелинейных функций регрессии и многочленно, и удобно для использования. Однако в некоторых приложениях по экономическим причинам приходится использовать функции регрессии, являющиеся нелинейными по параметрам. Несмотря на то что такие функции регрессии не могут быть оценены при помощи МНК, они могут быть оценены с использованием обобщения МНК, называемого нелинейным методом наименьших квадратов.

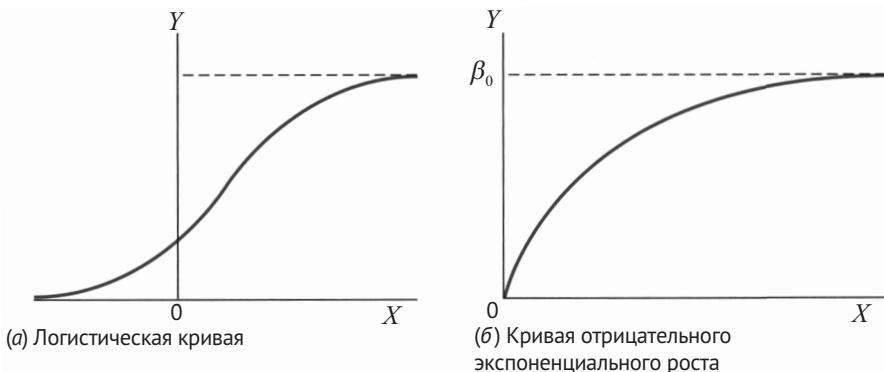
#### Нелинейные по параметрам функции

Мы начнем данный раздел с рассмотрения примеров двух функций, которые являются нелинейными по параметрам. Затем мы предложим более общую формулировку.

**Логистическая кривая.** Предположим, что вы изучаете объемы проникновения на рынок технологий, таких как программное обеспечение, используемое для управления компаниями в различных отраслях промышленности. Пусть зависимой переменной является доля фирм в отрасли, которые используют программное обеспечение, а единственная независимая переменная  $X$  описывает характеристики отрасли, и у вас есть данные о  $n$  отраслях промышленности. Значения зависимой переменной находятся в интервале между от 0 (не используется) и 1 (100 % использование). Поскольку модель линейной регрессии может давать прогнозные значения, меньшие 0 или большие 1, то имеет смысл использовать вместо нее функцию, которая будет принимать значения между нулем и единицей.

Логистическая функция гладко растет с минимума, равного нулю, до максимума, равного единице. Модель логистической регрессии с единственным regressором  $X$  имеет вид:

$$Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} + u_i. \quad (8.38)$$



**Рисунок 8.12. Две функции, нелинейные по своим параметрам**

На рисунке (а) изображен график логистической функции, заданной функцией (8.38), которая принимает значения, лежащие между нулем и единицей. На рисунке (б) изображен график функции отрицательного экспоненциального роста, заданной функцией (8.39), которая всегда имеет положительный наклон, уменьшающийся при увеличении  $X$ , и стремится к  $\beta_0$  при стремлении  $X$  к бесконечности.

Логистическая функция с единственным  $X$  изображена на рисунке 8.12а. Как можно видеть на графике, логистическая функция имеет вытянутую S-образную форму. Для небольших значений  $X$  значения этой функции близки к нулю, а наклон практически отсутствует (она плоская); она имеет более крутой наклон для умеренных значений  $X$  и для больших значений  $X$  функция приближается к единице и снова становится практически плоской.

**Отрицательный экспоненциальный рост.** Функции, использованные в разделе 8.2 для моделирования связи между результатами тестов и доходами, имеют некоторые недостатки. Например, у полиномиальных моделей может быть отрицательный наклон для некоторых значений дохода, что является невероятным. Логарифмическая кривая имеет положительный наклон для всех значений дохода; однако при очень большом доходе предсказанные значения увеличиваются без

ограничения, поэтому для некоторых значений доходов предсказанное значение для округа будет превышать максимально возможное значение результата теста.

Модель отрицательного экспоненциального роста предоставляет нелинейную спецификацию, которая имеет положительный наклон для всех значений дохода, и этот наклон является наибольшим при низких значениях доходов и уменьшается по мере их увеличения, а также имеет верхнюю границу (т.е. асимптоту при стремлении доходов к бесконечности). Регрессионная модель отрицательного экспоненциального роста имеет вид:

$$Y_i = \beta_0 [1 - e^{-\beta_1(X_i - \beta_2)}] + u_i . \quad (8.39)$$

Функция отрицательного экспоненциального роста изображена на рисунке 8.12б. Ее наклон является более крутым для низких значений  $X$ , но при увеличении  $X$  он достигает асимптоты  $\beta_0$ .

**Общий вид функции, нелинейной по параметрам.** Модель логистической регрессии и модель отрицательного экспоненциального роста являются специальными случаями более общей нелинейной регрессионной модели:

$$Y_i = f(X_{1i}, \dots, X_{ki}; \beta_0, \dots, \beta_m) + u_i , \quad (8.40)$$

в которой есть  $k$  независимых переменных и  $m+1$  параметров,  $\beta_0, \dots, \beta_m$ . В моделях из разделов 8.2 и 8.3  $X$ -ы входили в эту функцию нелинейно, но параметры – линейно. В примерах, рассмотренных в данном приложении, параметры входят в функцию также нелинейно. Если параметры известны, тогда предсказанные эффекты могут вычисляться с использованием метода, описанного в разделе 8.1. В приложениях, однако, параметры неизвестны и должны быть оценены по имеющимся данным. Параметры, которые входят нелинейно, не могут быть оценены при помощи МНК, а могут быть оценены нелинейным методом наименьших квадратов.

## **Оценка нелинейным методом наименьших квадратов**

Нелинейный метод наименьших квадратов является общим методом, используемым для оценки неизвестных параметров функции регрессии, в случае их нелинейного вхождения в теоретическую функцию регрессии.

Вспомним обсуждение из раздела 5.3 МНК-оценки коэффициентов линейной модели множественной регрессии. МНК-оценка минимизирует сумму квадратов остатков в уравнении (5.8),  $\sum_{i=1}^n [Y_i - (b_0 + b_1 X_{1i} + \dots + b_k X_{ki})]^2$ . В принципе, МНК-оценка может быть вычислена путем проверки множества пробных значений  $b_0, \dots, b_k$  и выборки значений, которые минимизируют сумму квадратов остатков.

Тот же самый подход может быть использован для получения оценок параметров общей модели нелинейной регрессии (8.40). Поскольку функция регрессии не является линейной по коэффициентам, этот метод называется **нелинейным методом наименьших квадратов**. Для множества пробных значений параметров  $b_0, b_1, \dots, b_k$  построим сумму квадратов прогнозных ошибок:

$$\sum_{i=1}^n [Y_i - f(X_{1i}, \dots, X_{ki}; b_1, \dots, b_m)]^2. \quad (8.41)$$

Оценки нелинейного метода наименьших квадратов  $\beta_0, \beta_1, \dots, \beta_m$  представляют собой значения  $b_0, b_1, \dots, b_k$ , которые минимизируют сумму квадратов остатков (8.41).

В линейной регрессии МНК-оценки выражаются относительно простой формулой как функции от данных. К сожалению, не существует такой общей формулы для нелинейного метода наименьших квадратов, поэтому оценка нелинейного метода наименьших квадратов должна основываться на численных методах и требует использования компьютера. Программное обеспечение, используемое в эконометрике, включает алгоритмы для решения задачи минимизации нелинейного метода наименьших квадратов, что упрощает задачу вычисления оценок нелинейного метода наименьших квадратов на практике.

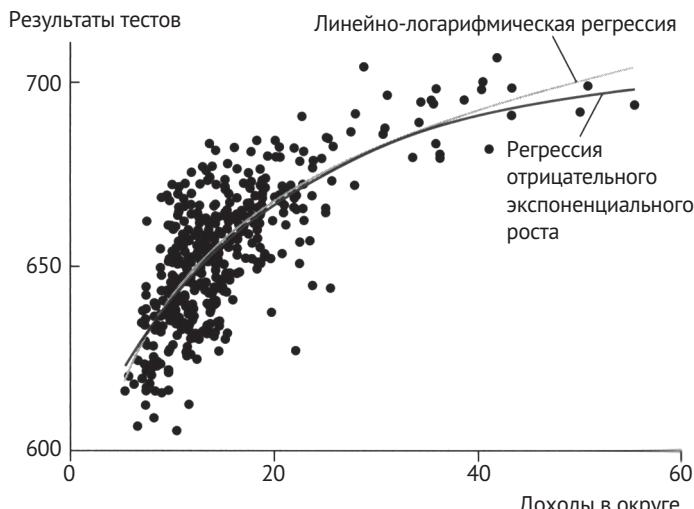
При общих условиях на функцию  $f$  и  $X$ -ы оценка нелинейного метода наименьших квадратов обладает двумя основными свойствами МНК-оценка в модели линейной регрессии: она состоятельна и асимптотически нормально распределена. В эконометрических программных пакетах, включающих нелинейный метод наименьших квадратов, в качестве итоговых результатов обычно сообщаются стандартные ошибки оцененных параметров. Как следствие, статистическая проверка параметров может производиться обычным способом; в частности,  $t$ -статистики могут быть построены с использованием общего подхода, описанного во вставке «Основные понятия 5.1», а 95 %-й доверительный интервал может быть построен как оцененный коэффициент  $\pm 1,96$  стандартной ошибки. Как и в линейной регрессии, остаточный член в нелинейной регрессионной модели может быть гетероскедастичен, поэтому должны быть использованы стандартные ошибки, устойчивые по отношению к гетероскедастичности.

### **Применение к зависимости результатов тестов от доходов**

Модель отрицательного экспоненциального роста связывает доходы в округе ( $X$ ) и результаты тестов ( $Y$ ) и имеет желательные особенности наклона, который всегда положителен [если параметр  $\beta_1$  в уравнении (8.39) положителен] и стремится к  $\beta_0$  при стремлении доходов к бесконечности. В результате оценивания  $\beta_0$ ,  $\beta_1$  и  $\beta_2$  в уравнении (8.39), используя данные результатов тестов в Калифорнии, мы получаем, что  $\hat{\beta}_0 = 703,2$  (устойчивые к гетероскедастичности стандартные ошибки = 4,44),  $\hat{\beta}_1 = 0,0552$  ( $SE = 0,0068$ ) и  $\hat{\beta}_2 = -34,0$  ( $SE = 4,48$ ). Таким образом, оцененная функция нелинейной регрессии (со стандартными ошибками, приведенными ниже оценок параметров) имеет вид:

$$\widehat{TestScore} = 703,2 \left[ 1 - e^{-0,0552(Income + 34,0)} \right] \quad (8.42)$$

Оцененная функция регрессии изображена на рисунке 8.13 вместе с функцией логарифмической регрессии и диаграммой рассеяния данных. Эти две спецификации довольно похожи. Одно из отличий заключается в том, что кривая отрицательного экспоненциального роста выравнивается при самых высоких значениях доходов, что соответствует наличию асимптоты.



**Рисунок 8.13. Графики функций регрессии отрицательного экспоненциального роста и линейно-логарифмической регрессии**

Обе функции регрессии отрицательного экспоненциального роста [уравнение (8.42)] и линейно-логарифмической регрессии [уравнение (8.18)] описывают нелинейную связь между результатами тестов и доходами в округе. Одно из различий между ними заключается в том, что модель отрицательного экспоненциального роста имеет асимптоту при стремлении  $Income$  к бесконечности, а функция линейно-логарифмической регрессии ее не имеет.

## Приложение 8.2. Угловые коэффициенты и эластичности для функции нелинейной регрессии

В данном приложении мы используем методы дифференциального исчисления для получения оценок угловых коэффициентов и эластичностей функции нелинейной регрессии с непрерывными регрессорами. Мы ориентируемся на случай из раздела 8.2 для единственного  $X$ . Рассматриваемый подход может быть обобщен на несколько  $X$ -ов, используя частные производные.

Рассмотрим модель нелинейной регрессии  $Y_i = f(X_i) + u_i$  с  $E(u_i | X_i) = 0$ . Угловой коэффициент теоретической функции регрессии  $f(X)$  в точке  $X = x$  — это производная функции  $f$ , то есть  $df(X) / dX|_{X=x}$ . Для функции полиномиальной регрессии (8.9) получаем, что  $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_r X^r$  и  $dX^a / dX = aX^{a-1}$  для любой постоянной  $a$ , поэтому  $df(X) / dX|_{X=x} = \beta_1 + 2\beta_2 x + \dots + r\beta_r x^{r-1}$ . Оцененный угловой коэффициент при  $x$  равен тогда  $\hat{df}(X) / dX|_{X=x} = \hat{\beta}_1 + 2\hat{\beta}_2 x + \dots + r\hat{\beta}_r x^{r-1}$ . Стандартная ошибка оцененного углового коэффициента равна  $SE(\hat{\beta}_1 + 2\hat{\beta}_2 x + \dots + r\hat{\beta}_r x^{r-1})$  и для заданного значения  $x$  представляет собой стан-

дартную ошибку взвешенной суммы коэффициентов регрессии, которая может быть вычислена с использованием методов из раздела 7.3 и уравнения (8.8).

Эластичность  $Y$  по  $X$  представляет собой процентное изменение  $Y$  для заданного процентного изменения  $X$ . Формально мы рассматриваем предел при стремлении процентного изменения  $X$  к нулю, поэтому угловой коэффициент, присутствующий в определении в уравнении (8.22), заменяется на производную, и эластичность  $Y$  по  $X$  равна:

$$\frac{dY}{dX} \times \frac{X}{Y} = \frac{d \ln Y}{d \ln X}.$$

В модели регрессии  $Y$  зависит и от  $X$  и от компоненты ошибок  $u$ . Поскольку  $u$  является случайной, удобно оценить эластичность не как процентное изменение  $Y$ , а как процентное изменение предсказанного значения  $Y$ , то есть процентное изменение  $E(Y | X)$ . Соответственно, эластичность  $E(Y | X)$  по  $X$  равна:

$$\frac{dE(Y | X)}{dX} \times \frac{X}{E(Y | X)} = \frac{d \ln E(Y | X)}{d \ln X}.$$

Формулы для расчета эластичностей для линейной модели и для трех видов логарифмических моделей приведены во вставке «Основные понятия 8.2» и повторяются в таблице.

Модель	Теоретическая модель регрессии	Эластичность $E(Y   X)$ по $X$
Линейная	$Y = \beta_0 + \beta_1 X + u$	$\frac{\beta_1 X}{\beta_0 + \beta_1 X}$
Линейно-логарифмическая	$Y = \beta_0 + \beta_1 \ln(X) + u$	$\frac{\beta_1}{\beta_0 + \beta_1 \ln(X)}$
Логарифмически-линейная	$\ln(Y) = \beta_0 + \beta_1 X + u$	$\beta_1 X$
Линейная в логарифмах	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$	$\beta_1$

Линейная в логарифмах модель имеет постоянную эластичность, но в других трех спецификациях эластичность зависит от  $X$ .

Выведем выражения для линейно-логарифмической и логарифмически-линейной моделей. Для линейно-логарифмической модели  $E(Y | X) = \beta_0 + \beta_1 \ln(X)$ . Поскольку  $d \ln(X) / dX = 1/X$ , получаем соотношение  $dE(Y | X) / dX = \beta_1 / X$ . Таким образом, в случае линейно-логарифмической функции эластичность составляет  $dE(Y | X) / dX \times X / E(Y | X) = (\beta_1 / X) \times X / [\beta_0 + \beta_1 \ln(X)] = \beta_1 / [\beta_0 + \beta_1 \ln(X)]$ , как и дано в таблице. Для логарифмически-линейной модели удобно сделать

дополнительное предположение о том, что  $u$  и  $X$  независимо распределены, поэтому выражение для  $E(Y|X)$ , заданное уравнением (8.25), принимает вид:  $E(Y|X) = ce^{\beta_0 + \beta_1 X}$ , где  $c = E(e^u)$  является константой, которая не зависит от  $X$  из-за дополнительного предположения о независимости  $u$  и  $X$ . Таким образом,  $dE(Y|X)/dX = ce^{\beta_0 + \beta_1 X} \beta_1$  и эластичность равна  $dE(Y|X)/dX \times X / E(Y|X) = ce^{\beta_0 + \beta_1 X} \beta_1 \times X / (ce^{\beta_0 + \beta_1 X}) = \beta_1 X$ . Вывод формул для эластичностей линейной и линейно-логарифмической моделей остается в качестве упражнения 8.11.

# **Глава 9. Оценка исследований, основанных на множественной регрессии**

В предыдущих пяти главах мы пытались объяснить, как можно использовать множественную регрессию для анализа взаимосвязей между реальными данными. В данной главе мы немного отступим назад и попытаемся понять, из-за чего исследование, использующее множественный регрессионный анализ, может считаться надежным или ненадежным. Мы сосредоточимся на исследованиях, целью которых является оценка причинного эффекта от изменения некоторой независимой переменной (например размера класса) на зависимую переменную (например на результаты тестов). Мы попытаемся понять, когда множественная регрессия является полезной для такой оценки, а когда (и это не менее важно) она не в состоянии быть полезной.

Для ответа на эти вопросы в данной главе предложена схема оценки статистических исследований независимо от того, используется в них регрессионный анализ или нет. В основе предложенной схемы лежат понятия внутренней и внешней обоснованности (или аргументированности). Мы полагаем, что исследование является внутренне обоснованным, если статистические выводы о причинных эффектах, полученные на его основе, верны для исследованной генеральной совокупности и в рамках заданных предположений, и исследование является внешне обоснованным, если статистические выводы из него могут быть обобщены на другие генеральные совокупности и при других предпосылках. В разделах 9.1 и 9.2 мы обсуждаем понятия внутренней и внешней обоснованности, перечисляем различные возможные угрозы для них и обсуждаем, как определить эти угрозы на практике. Наше внимание в разделах 9.1 и 9.2 сосредоточено на оценке причинных эффектов, существующих у наблюдаемых данных. В разделе 9.3 обсуждается другое приложение моделей регрессии – прогнозирование и перечисляются различные причины, которые могут повлечь за собой возникновение сомнений в обоснованности прогнозов, полученных с использованием моделей регрессии.

В качестве иллюстрации понятий внутренней и внешней обоснованности в разделе 9.4 мы оцениваем внутреннюю и внешнюю обоснованность исследования эффекта влияния сокращения соотношения учеников и учителей на результаты тестов, представленные в главах 4–8.

## 9.1. Внутренняя и внешняя обоснованность

Из понятий внутренней и внешней обоснованности, определенных во вставке «Основные понятия 9.1», вытекает схема оценки того, отвечает ли статистическое или экономическое исследование на конкретные интересующие нас вопросы.

Для введения понятий внутренней и внешней обоснованности (аргументированности) необходимо определить такие понятия, как «изучаемая генеральная совокупность» и «заданные условия» (предпосылки) и «целевая генеральная совокупность» и «условия, для которых обобщаются результаты». *Изучаемой<sup>1</sup> генеральной совокупностью* называется генеральная совокупность объектов – людей, компаний, школьных округов и так далее – из которой извлечена выборка. Генеральная совокупность, на которую обобщаются результаты, или *целевая<sup>2</sup> генеральная совокупность*, – это генеральная совокупность объектов, на которую распространяются результаты статистического исследования, руководствуясь причинно-следственными связями. Например, директор средней школы (с 9 по 12 классы) может обобщить ваши выводы о влиянии размера классов и результатов тестов в округах начальной школы (изучаемая генеральная совокупность) на генеральную совокупность средних школ (целевая генеральная совокупность).

### ОСНОВНЫЕ ПОНЯТИЯ 9.1

#### Внутренняя и внешняя обоснованность (аргументированность)

Считается, что статистический анализ является внутренне обоснованным, если статистическая проверка причинных эффектов обоснована для изучаемой генеральной совокупности. Также считается, что анализ является внешне обоснованным, если его статистическая проверка и выводы могут быть обобщены с исследованной генеральной совокупности и заданных условий (предпосылок) на другие генеральные совокупности и желаемые условия.

Под «условиями» (предпосылками) мы подразумеваем институциональные, правовые и экономические условия. Например, было бы важно знать, могут ли результаты лабораторного эксперимента оценивания методов выращивания органических томатов быть обобщены на поле, то есть работают ли органические методы одинаково в условиях лаборатории и в условиях реального мира. Мы будем приводить и другие примеры отличий генеральных совокупностей и условий (предпосылок) по мере изложения материала.

<sup>1</sup> В английском варианте учебника используется термин «the population studied». Мы в качестве русскоязычных аналогов будем использовать слова «изучаемая», «изученная», «используемая», «исходная» генеральная совокупность. – Примеч. научн. ред. перевода.

<sup>2</sup> В английском варианте учебника используется термин «the population of interest». В данном случае мы будем переводить этот термин как «целевая» или «интересующая нас» генеральная совокупность. – Примеч. научн. ред. перевода.

## **Угрозы для внутренней обоснованности**

Для того чтобы исследование было внутренне обоснованным, необходимо выполнение двух условий. Во-первых, оценка причинного эффекта должна быть несмещенной и состоятельной. Например, если  $\hat{\beta}_{STR}$  – МНК-оценка эффекта влияния единичного изменения соотношения учеников и учителей на результаты тестов в определенной регрессии, то  $\hat{\beta}_{STR}$  должна быть несмещенной и состоятельной оценкой истинного причинного эффекта от изменения соотношения учеников и учителей в генеральной совокупности, то есть  $\beta_{STR}$ .

Во-вторых, тестирование гипотез должно иметь желаемый уровень значимости (фактическая частота отвержения нулевой гипотезы при заданной нулевой гипотезе должна быть равна желаемому уровню значимости), и доверительные интервалы должны иметь желаемый уровень доверия (доверительную вероятность). Например, если доверительный интервал строится как  $\hat{\beta}_{STR} \pm 1,96SE(\hat{\beta}_{STR})$ , этот доверительный интервал должен содержать истинные причинные эффекты, присутствующие в генеральной совокупности ( $\beta_{STR}$ ), с вероятностью 95 % по повторяющимся выборкам.

В регрессионном анализе причинные эффекты оцениваются с использованием оценки функции регрессии, а тестирование гипотез проводится с использованием оцененных коэффициентов регрессии и их стандартных ошибок. Соответственно, в исследовании, основанном на МНК-регрессии, требования для внутренней обоснованности заключаются в том, чтобы МНК-оценка была несмещенной и состоятельной, а стандартные ошибки вычислены так, чтобы построенные с их использованием доверительные интервалы имели бы желаемую доверительную вероятность. По различным причинам эти требования могут не выполняться, и тогда эти причины представляют собой угрозу для внутренней обоснованности. Рассматриваемые угрозы (проблемы) приводят к нарушению одного или более предположений метода наименьших квадратов, сформулированных во вставке «Основные понятия 6.4». Например, одной из проблем, которую мы подробно обсудили, является смещение из-за пропущенной переменной; это приводит к корреляции между одним или более регрессорами и компонентой ошибок, что нарушает первое предположение метода наименьших квадратов. Если имеются данные о пропущенной переменной или адекватной контрольной переменной, тогда данную проблему можно избежать, включая такую переменную в качестве дополнительного регрессора.

В разделе 9.2 содержится детальное обсуждение различных угроз для внутренней обоснованности во множественном регрессионном анализе и предлагаются методы их устранения.

## **Угрозы для внешней обоснованности**

Потенциальные угрозы для внешней обоснованности появляются из-за различий между изучаемой генеральной совокупности и имеющихся условий и целевой генеральной совокупностью и желаемыми условиями.

**Различия в генеральных совокупностях.** Различия между изучаемой и целевой генеральными совокупностями могут представлять угрозу для внешней обоснованности статистических методов. Например, в лабораторных исследованиях токсического воздействия химических веществ используются генеральные совокупности животных, таких как мыши (исследуемая генеральная совокупность), но результаты такого исследования используются для того, чтобы сформулировать набор требований по охране здоровья для генеральной совокупности людей (целевая генеральная совокупность). Являются ли различия между мышами и людьми статистически значимыми, чтобы представлять угрозу для внешней обоснованности таких исследований – в этом и заключается предмет дискуссий.

В более общем случае истинный причинный эффект может не быть одинаковым для имеющейся и целевой генеральных совокупностей. Это может случиться из-за того, что генеральная совокупность, на примере которой мы изучаем некоторое явление, была выбрана способом, делающим ее отличной от интересующей нас генеральной совокупности из-за различий в характеристиках генеральных совокупностей, из-за географических различий или потому, что исследование устарело.

**Различия в условиях.** Даже если изучаемая и целевая генеральные совокупности являются идентичными, может быть невозможно обобщить результаты исследования, если различны изучаемые и желаемые условия. Например, исследование эффекта влияния антиалкогольной компании на пьянство в колледже не может быть обобщено на другую такую же группу студентов, если правовые санкции за распитие спиртных напитков в двух колледжах различны. В этом случае правовые условия, при которых проводилось исследование, отличаются от правовых условий, на которые распространяются его результаты.

В более общем случае примеры различий между имеющимися и желаемыми условиями включают различия в институциональных условиях (государственные университеты и религиозные вузы), различия в законах (различия в правовых санкциях) или различия природных условий (например, мы наблюдаем существенно разные условия для проведения пикников в южной Калифорнии и в городе Фэрбанкс на Аляске).

**Применение к анализу зависимости результатов тестов от соотношения учеников и учителей.** В главах 7 и 8 обсуждаются статистически значимые, но небольшие улучшения в оценках зависимости результатов тестов при уменьшении соотношения учеников и учителей. Наш анализ был основан на данных по результатам тестов для школьных округов штата Калифорния. Предположим, что эти результаты внутренне обоснованы. На какие еще генеральные совокупности и условия эти выводы могут быть обобщены?

Чем ближе изучаемая генеральная совокупность и имеющиеся условия исследования к интересующим нас, тем больше наше исследование является внешне обоснованным. Например, обучение в колледже и его студенты сильно отличаются от обучения в начальной школе и ее учеников, поэтому представля-

ется неправдоподобным, что эффект от уменьшения размеров класса, оцененный при использовании данных по школьным округам Калифорнии, может быть обобщен на колледжи. С другой стороны, учащиеся начальной школы, учебные программы и организации в целом похожи на всей территории Соединенных Штатов, так что вполне вероятно, что выводы из анализа зависимости результатов тестов от различных показателей в Калифорнии могли бы быть обобщены и на стандартизованные тесты в других округах начальных школ США.

**Как оценить внешнюю обоснованность исследования.** Внешняя обоснованность должна оцениваться с использованием специфических знаний об изучаемой и целевой генеральных совокупностях и имеющихся и желаемых условиях. Наличие серьезных различий между этими генеральными совокупностями и условиями поставит под сомнение внешнюю обоснованность исследования.

Иногда существует два или более различных исследования связанных генеральных совокупностей. В этом случае внешняя обоснованность обоих исследований может быть проверена путем сравнения их результатов. Например, в разделе 9.4 мы анализируем данные по результатам тестов и размерам классов для округов начальных школ в штате Массачусетс и сравниваем результаты для Массачусетса и Калифорнии. В общем случае похожие результаты двух или более исследований укрепят требования внешней обоснованности, в то время как различия в их выводах, которые нелегко объяснить, поставят под сомнение их внешнюю обоснованность<sup>1</sup>.

**Как создать исследование с внешней обоснованностью.** Поскольку угрозы внешней обоснованности связаны с отсутствием сопоставимости генеральных совокупностей и условий, эти угрозы наилучшим образом надо сводить к минимуму на ранних стадиях исследования, до сбора данных. Структура исследования выходит за рамки этого учебника, поэтому заинтересованный читатель может обратиться к следующей литературе: Shadish, Cook, Campbell (2002).

## 9.2. Угрозы для внутренней обоснованности множественного регрессионного анализа

Исследования, основанные на регрессионном анализе, являются внутренне обоснованными, если оцененные коэффициенты регрессии являются несмещанными и состоятельными и если их стандартные ошибки приводят к доверительным интервалам с желаемым уровнем значимости. В данном разделе мы описываем пять причин, из-за которых МНК-оценка коэффициентов

<sup>1</sup> Сравнение многих связанных исследований на одну и ту же тему называется мета-анализом. Например, обсуждение, которое было проведено во вставке «Эффект Моцарта: смещение из-за пропущенных переменных?» в главе 6, основано на мета-анализе. Применение мета-анализа для многих исследований имеет свои собственные проблемы. По каким критериям исследование можно отнести к хорошему или плохому? Как сравнить исследования, в которых зависимые переменные различны? Нужно ли придавать большее внимание (и больше доверять) исследованиям на больших выборках? Обсуждение мета-анализа и его проблемы выходит за рамки данного учебника. Заинтересованного читателя мы отсылаем к книгам: Hedges, Olkin (1985) и Cooper, Hedges (1994).

множественной регрессии может быть смещенной даже в больших выборках: пропущенные переменные, неправильная спецификация функциональной формы регрессии, неточное измерение независимых переменных («ошибки в переменных»), отбор наблюдений и одновременная причинность. Все пять источников ошибок возникают из-за регрессора, коррелированного с компонентой ошибок в теоретической регрессии, нарушая первое из предположений метода наименьших квадратов, перечисленных во вставке «Основные понятия 6.4». Для каждого источника смещения мы обсуждаем, что можно сделать для уменьшения получаемого смещения. В конце раздела обсуждаются причины, приводящие к несостоятельности стандартных ошибок, и методы борьбы с ними.

### ***Смещение из-за пропущенных переменных***

Вспомним, что смещение из-за пропущенных переменных возникает в ситуации, когда переменная, которая одновременно и влияет на  $Y$ , и коррелирована с одной или более включенными в регрессию независимыми переменными, пропущена в регрессии. Смещение сохраняется даже в больших выборках, поэтому МНК-оценка является несостоятельной. Как наилучшим образом минимизировать смещение из-за пропущенной переменной, зависит от того, доступны или нет переменные, которые адекватно описывают потенциальную пропущенную переменную.

***Методы решения проблемы смещения из-за пропущенной переменной при наличии самой переменной или адекватной контрольной переменной.*** Если пропущенная переменная наблюдаема (т.е. у вас есть соответствующие данные), тогда включение этой переменной во множественную регрессию является простейшим способом решения проблемы пропущенной переменной. Кроме того, если у вас есть данные по одной или более контрольным переменным и если эти контрольные переменные адекватны в том смысле, что они приводят к независимости условного среднего [уравнение (7.20)], тогда включение этих контрольных переменных устраниет потенциальное смещение в коэффициенте при интересующей вас переменной.

Добавление переменной в регрессию имеет свои плюсы и минусы. С одной стороны, пропуск переменной приводит к смещению из-за пропущенной переменной, с другой – включение переменной, не оказывающей влияния на объясняемую переменную (т.е. когда ее коэффициент в теоретической регрессии равен нулю), уменьшает точность оценок других коэффициентов регрессии. Иными словами, принятие решения о том, включать ли переменную в регрессию или нет, связано с выбором между потенциальным смещением и возможным ростом дисперсии интересующего вас коэффициента. На практике существуют четыре шага, которые помогут вам принять решение о необходимости включения переменной (или множества переменных) в регрессию.

Первый шаг заключается в идентификации ключевого коэффициента или коэффициентов в вашей регрессии. В регрессии результатов тестов таким коэффициентом является коэффициент при соотношении учеников и учителей, поскольку изначально поставленный вопрос касается эффекта влияния уменьшения соотношения учеников и учителей на результаты тестов.

На втором шаге спросите себя: каковы наиболее вероятные источники смещения из-за пропущенных переменных в этой регрессии? Ответ на этот вопрос требует знания экономической теории и экспертных оценок, и вы должны знать его до того, как оцените любые регрессии; поскольку этот шаг должен быть сделан до непосредственного анализа данных, его называют априорным (*a priori* – «перед фактом») суждением. В примере с результатами тестов этот шаг влечет за собой идентификацию тех детерминант результатов тестов, которые могут смещать нашу оценку влияния размеров классов, если их игнорировать. Результатами этого шага являются базовая спецификация регрессии, отправная точка для эмпирического анализа и список дополнительных «сомнительных» переменных, которые могут помочь смягчить возможное смещение из-за пропущенных переменных.

Третий шаг заключается в расширении вашей базовой спецификации до регрессий, включающих дополнительные сомнительные контрольные переменные, выявленные на втором шаге. Если коэффициенты при дополнительных контрольных переменных статистически значимы или оценки интересующих вас коэффициентов заметно меняются при включении дополнительной переменной, тогда они должны остаться в спецификации, а вы – модифицировать вашу базовую спецификацию. Если нет, тогда эти дополнительные переменные могут быть исключены из регрессии.

На четвертом шаге приводится сводная таблица всех результатов ваших оценок. Эта таблица обеспечивает «полное раскрытие» для потенциальных скептиков, которые могут использовать ее или сделать свои собственные выводы. Таблицы 7.1 и 8.3 являются примерами реализации этой стратегии. Например, в таблице 8.3 мы могли бы представить только регрессию из столбца (7), потому что эта регрессия отражает соответствующие эффекты и нелинейности, присутствующие в других регрессиях из этой таблицы. Представление других регрессий, однако, позволяет скептически настроенному читателю сделать свои выводы.

Рассмотренные шаги перечислены во вставке «Основные понятия 9.2».

**Методы решения проблемы смещения из-за пропущенной переменной при отсутствии адекватной контрольной переменной.** Включение пропущенной переменной в регрессию может не стать решением проблемы пропущенных переменных, если данных по этой переменной или данных по адекватной контрольной переменной не существует. Тем не менее можно предложить еще три способа решения проблемы смещения из-за пропущенной переменной. Каждый из этих трех способов решает проблему пропущенных переменных за счет использования различных типов данных.

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**9.2**

**Смещение из-за пропущенных переменных: должен ли я включать большее число переменных в мою регрессию?**

Включение дополнительной переменной во множественную регрессию может устраниТЬ потенциальное смещение из-за пропущенных переменных, возникающее вследствие того, что эта переменная исключена, но при этом дисперсия оценки интересующего вас коэффициента может возрасти. Ниже приведены некоторые рекомендации, которые помогут вам решить, следует ли включать в регрессию дополнительные переменные:

1. Определите интересующие вас коэффициенты (или коэффициент).
2. Используйте ваши априорные представления для выявления самых важных потенциальных источников смещения из-за пропущенных переменных, приводящих к базовой спецификации и некоторым «сомнительным» переменным.
3. Проверьте, имеют ли дополнительные «сомнительные» переменные ненулевые коэффициенты.
4. Представьте ваши результаты максимально подробно («полное раскрытие») для того, чтобы другие исследователи могли видеть, как влияет включение сомнительных переменных на оценки коэффициента(ов) при интересующей(их) переменной(ых). Меняются ли результаты ваших оценок при включении контрольных переменных?

Первое решение заключается в использовании данных, которые наблюдаются в различные моменты времени. Например, результаты тестов и связанные с ними данные могут быть собраны для одного и того же округа в 1995 году и снова в 2000. Такие данные называются панельными. Как объясняется в главе 10, панельные данные дают возможность учитывать ненаблюдаемые пропущенные переменные в предположении, что эти пропущенные переменные не меняются с течением времени.

Второе решение – использовать регрессию с инструментальными переменными. Этот метод основывается на использовании новой переменной, называемой инструментальной. Регрессия с инструментальными переменными обсуждается в главе 12.

Третье решение заключается в использовании специального исследования, в котором интересующий нас эффект (например эффект влияния сокращения размеров классов на успеваемость учащихся) изучается с использованием случайного управляемого (контролируемого) эксперимента. Случайные контролируемые эксперименты обсуждаются в главе 13.

***Неправильная спецификация функциональной формы регрессии***

Если истинная теоретическая функция регрессии нелинейна, а оцененная регрессия линейна, то эта *неправильно специфицированная функциональная*

форма приводит к смещению МНК-оценки. Такое смещение представляет собой один из вариантов смещения из-за пропущенных переменных, при котором пропущенные переменные являются компонентами, отражающими отсутствие нелинейных характеристик в функции регрессии. Например, если теоретическая функция регрессии является квадратичной, то регрессия с пропущенным квадратом независимой переменной будет иметь смещение из-за пропущенной переменной. Основные моменты, связанные со смещением, возникающим из-за неправильной спецификации функциональной формы, описаны во вставке «Основные понятия 9.3».

### **Неправильная спецификация функциональной формы**

Неправильная спецификация функциональной формы возникает, когда функциональная форма оцененной функции регрессии отличается от функциональной формы теоретической функции регрессии. Если функциональная форма неправильно специфицирована, то оценка частного эффекта от изменения одной или более переменных будет, в общем случае, смещенной. Неправильная спецификация функциональной формы часто может быть обнаружена, если изобразить графически данные и оцененную функцию регрессии, и скорректирована с использованием другой функциональной формы.

## **ОСНОВНЫЕ ПОНЯТИЯ**

**9.3**

**Методы решения проблемы неправильной спецификации функциональной формы.** Если зависимая переменная непрерывна (как, например, результаты тестов в нашем случае), то проблема потенциальной нелинейности может быть решена с использованием методов из главы 8. Если, однако, зависимая переменная является дискретной или бинарной (например,  $Y_i$  равна 1, если  $i$ -й человек учился в колледже, и равна 0 – в противном случае), то все усложняется. Регрессия с дискретной зависимой переменной обсуждается в главе 11.

### **Смещение из-за ошибок измерения объясняющих переменных**

Предположим, что в нашей регрессии зависимости результатов тестов от отношения учеников и учителей мы случайно перепутали данные таким образом, что в конечном счете мы оценили регрессию зависимости результатов тестов пятиклассников от соотношения учеников и учителей в десятых классах школьного округа. Хотя соотношения учеников и учителей для учащихся начальной школы и пятиклассников могут быть коррелированными, они неодинаковы, поэтому эта путаница могла бы привести к смещению в оцененном коэффициенте. Данная ситуация является примером *смещения из-за ошибок в переменных*, поскольку его источником является ошибка в измерении (в нашем случае – в выборе) независимой переменной. Это смещение сохраняется даже в очень больших выборках, поэтому если имеются ошибки измерения, МНК-оценка не является состоятельной.

Существует много возможных источников погрешности измерений. Если данные собраны при помощи опроса, респондент может дать неправильный ответ. Например, один из вопросов в Текущем обследовании населения США включает в себя доходы за прошлый год. Респондент может не знать свой точный доход или может умышленно назвать неправильную сумму дохода по какой-то другой причине. Если же данные получены из официальных административных баз данных, то в них могут присутствовать типографские ошибки, возникшие при введении данных в память компьютера.

Для того чтобы увидеть, каким образом ошибки в переменных могут привести к корреляции между регрессором и компонентой ошибок, предположим, что есть единственный регрессор  $X_i$  (скажем, фактические доходы), но  $X_i$  измеряется неточно как  $\tilde{X}_i$  (заявленные доходы респондента). Поскольку мы наблюдаем переменную  $\tilde{X}_i$ , а не  $X_i$ , уравнение регрессии фактически оценивается на основе  $\tilde{X}_i$ . Следовательно, если переписать уравнение теоретической регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$  при помощи неточно измеренной (наблюданной) переменной  $\tilde{X}_i$ , то оно принимает вид:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] = \beta_0 + \beta_1 \tilde{X}_i + v_i, \quad (9.1)$$

где  $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$ . Таким образом, уравнение теоретической регрессии, записанное в терминах  $\tilde{X}_i$ , имеет компоненту ошибок, которая содержит ошибки измерения, представленные разностью между  $\tilde{X}_i$  и  $X_i$ . Если эта разность коррелирована с наблюдаемой переменной  $\tilde{X}_i$ , то регрессор  $\tilde{X}_i$  будет коррелирован с компонентой ошибок и  $\hat{\beta}_1$  будет смещенной и несостоительной.

Точный размер и направление смещения  $\hat{\beta}_1$  зависит от корреляции между  $\tilde{X}_i$  и ошибкой измерения  $\tilde{X}_i - X_i$ . В свою очередь эта корреляция зависит от природы ошибки измерения.

Предположим, например, что наблюдаемая переменная  $\tilde{X}_i$  равна сумме фактического неизмеренного (ненаблюданного) значения  $X_i$  и случайной компоненты (ошибки)  $w_i$ , которая имеет нулевое среднее и дисперсию  $\sigma_w^2$ . Поскольку ошибка случайная, мы можем предположить, что  $w_i$  некоррелирована с  $X_i$  и с компонентой ошибок  $u_i$ . На этом предположении основана классическая модель с ошибками измерения в объясняющих переменных, в которой  $\tilde{X}_i = X_i + w_i$ , где  $\text{corr}(w_i, X_i) = 0$  и  $\text{corr}(w_i, u_i) = 0$ . При помощи небольших алгебраических преобразований можно показать, что в классической модели с ошибками измерений<sup>1</sup>  $\hat{\beta}_1$  сходится по вероятности к такому виду:

$$\hat{\beta}_1 \xrightarrow{P} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1. \quad (9.2)$$

То есть если ошибка измерения выражается в простом добавлении случайного элемента к фактическому значению независимой переменной, то  $\hat{\beta}_1$  является

<sup>1</sup> При этом предположении получаем, что в модели с ошибками в измерении  $v_i = \beta_1(X_i - \tilde{X}_i) + u_i = -\beta_1 w_i + u_i$ ,  $\text{cov}(X_i, u_i) = 0$  и  $\text{cov}(\tilde{X}_i, w_i) = \text{cov}(X_i + w_i, w_i) = \sigma_w^2$ , поэтому  $\text{cov}(\tilde{X}_i, v_i) = -\beta_1 \text{cov}(\tilde{X}_i, w_i) + \text{cov}(\tilde{X}_i, u_i) = -\beta_1 \sigma_w^2$ . Таким образом, из уравнения (6.1) имеем:  $\hat{\beta}_1 \xrightarrow{P} \beta_1 - \beta_1 \sigma_w^2 / \sigma_{\tilde{X}}^2$ . И поскольку из предположения модели с ошибками в измерении следует, что  $\sigma_{\tilde{X}}^2 = \sigma_X^2 + \sigma_w^2$ , получаем:  $\hat{\beta}_1 \xrightarrow{P} \beta_1 - \beta_1 \sigma_w^2 / (\sigma_X^2 + \sigma_w^2) = [\sigma_X^2 / (\sigma_X^2 + \sigma_w^2)] \beta_1$ .

ется несостоительной. Поскольку отношение  $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2}$  меньше 1,  $\hat{\beta}_1$  будет смещена к нулю даже в больших выборках. В крайнем случае, когда ошибка измерения настолько велика, что по существу не остается информации об  $X_i$ , то есть отношение дисперсий последнего выражения в уравнении (9.2) стремится к нулю,  $\hat{\beta}_1$  также сходится по вероятности к нулю. В другом крайнем случае, когда нет ошибок измерения, то есть  $\sigma_w^2 = 0$ , получаем  $\hat{\beta}_1 \xrightarrow{p} \beta_1$ .

### Смещение из-за ошибок в независимых переменных

Смещение из-за ошибок в независимых переменных возникает при оценке МНК, когда независимая переменная измерена неточно. Величина смещения зависит от природы ошибок измерения и сохраняется даже в выборках большого размера. Если измеренная переменная равна сумме фактического значения и независимо распределенной случайной ошибки с нулевым средним, то МНК-оценка регрессии с единственной независимой переменной смещена в сторону нуля и ее предел по вероятности задан выражением (9.2).

## ОСНОВНЫЕ ПОНЯТИЯ

9.4

Различные модели с ошибками измерения предполагают, что респондент дает свою лучшую оценку истинного значения переменной. В такой модели «лучшего предположения» предполагается, что ответ  $\tilde{X}_i$  представляет собой условное среднее  $X_i$  при заданной информации, доступной респонденту. Поскольку  $\tilde{X}_i$  является лучшим предположением, ошибка измерения  $\tilde{X}_i - X_i$  некоррелирована с ответом  $\tilde{X}_i$  (если бы ошибка измерения была коррелирована с  $\tilde{X}_i$ , то тогда информация об этой корреляции была бы полезной для прогнозирования  $X_i$ , и в этом случае ответ  $\tilde{X}_i$  не был бы лучшим предположением относительно  $X_i$ ). Таким образом,  $E[(\tilde{X}_i - X_i)\tilde{X}_i] = 0$ , и если информация респондента не коррелирована с  $u_i$ , то  $\tilde{X}_i$  некоррелирована с компонентой ошибок  $v_i$ . Таким образом, при выполнении этого «лучшего предположения» в модели с ошибками измерения  $\hat{\beta}_1$  является состоятельной, но поскольку  $\text{var}(v_i) > \text{var}(u_i)$ , дисперсия  $\hat{\beta}_1$  больше, чем была бы при отсутствии ошибок измерения. Модель с ошибками измерения в предположении выполнения «лучшего предположения» рассматривается далее в упражнении 9.12.

Проблемы, возникающие из-за ошибок измерения, могут быть даже более сложными, если существует преднамеренный неправильный ответ. Например, предположим, что, отвечая на вопрос о налогооблагаемом доходе, респонденты умышленно занижают свой истинный налогооблагаемый доход, не включая наличные платежи. Если, например, все респонденты говорят лишь о 90 % своих доходов, то  $\tilde{X}_i = 0,90X_i$  и  $\hat{\beta}_1$  будет смещена вверх на 10 %.

Хотя результат, представленный в уравнении (9.2), является специфическим и относится к классическим ошибкам измерения, он иллюстрирует более общее утверждение о том, что если независимая переменная измерена неточно, тогда МНК-оценка является смещенной даже в больших выборках. Концепция смещения из-за ошибок измерения в переменных представлена во вставке «Основные понятия 9.4».

**Ошибки измерения в  $Y$ .** Эффект ошибок измерения в  $Y$  отличается от ошибок измерения в  $X$ . Если  $Y$  имеет классические ошибки измерения, тогда эти ошибки измерения увеличивают дисперсию регрессии и  $\hat{\beta}_1$ , но не вызывают смещения в  $\hat{\beta}_1$ . Чтобы увидеть это, предположим, что вместо  $Y_i$  мы наблюдаем  $\tilde{Y}_i$ , который равен истинному  $Y_i$  плюс случайная ошибка измерения  $w_i$ . Тогда модель регрессии оценивается как  $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$ , где  $v_i = w_i + u_i$ . Если ошибка  $w_i$  действительно случайна, то  $w_i$  и  $X_i$  независимо распределены, так что  $E(w_i | X_i) = 0$  и  $E(v_i | X_i) = 0$ , и поэтому  $\hat{\beta}_1$  является несмещенной. Однако так как  $\text{var}(v_i) > \text{var}(u_i)$ , дисперсия  $\hat{\beta}_1$  больше, чем была бы без ошибок измерения. В примере зависимости результатов тестов от размеров класса предположим, что результаты теста имеют чисто случайные ошибки, связанные с процедурой оценивания экзамена, которые не зависят от регрессоров; тогда классическая модель с ошибкой измерения, рассмотренная в этом параграфе, применяется к  $\tilde{Y}_i$ , и  $\hat{\beta}_1$  является несмещенной. В общем случае ошибки измерения в  $Y$ , которые имеют нулевое условное среднее относительно объясняющих переменных, не будут вызывать смещение в МНК-оценках коэффициентов.

**Методы решения проблемы смещения из-за ошибок измерения переменных.** Наилучший способ решить проблему смещения из-за ошибок в переменных – это получить точную меру  $X$ . Тем не менее если это невозможно, некоторые эконометрические методы могут быть использованы для уменьшения смещения из-за ошибок в переменных.

Одним из таких методов является регрессия с инструментальными переменными. В основе его лежит использование другой («инструментальной») переменной, которая коррелирована с фактическим значением  $X_i$ , но не коррелирована с ошибками измерения. Этот метод изучается в главе 12.

Второй метод заключается в разработке математической модели с ошибками измерения и, если возможно, в использовании полученных формул для коррекции оценок. Например, если исследователь считает, что применяется классическая модель с ошибками измерения, и если отношение  $\sigma_w^2 / \sigma_X^2$  известно или может быть оценено, то выражение (9.2) может быть использовано для вычисления оценки  $\beta_1$ , скорректированной относительно смещения вниз. Поскольку такой подход требует специальных знаний о природе ошибок измерения, которые обычно являются специфическими для каждой конкретной базы данных и проблем измерения, свойственных этой базе данных, мы не будем рассматривать этот подход далее в учебнике.

### **Отсутствующие данные и отбор наблюдений**

Отсутствие данных является общей чертой экономических баз данных. Представляют ли угрозу для внутренней обоснованности отсутствующие данные,

зависит от того, почему данные отсутствуют. Рассмотрим три случая: когда данные отсутствуют совершенно случайно; когда отсутствуют данные по объясняющей переменной  $X$  и когда отсутствуют данные из-за процесса их отбора, связанного с объясняемой переменной  $Y$  и не зависящего от  $X$ .

Если данные отсутствуют совершенно случайно, то есть случайные причины не связаны со значениями  $X$  или  $Y$ , то в этом случае уменьшается размер выборки, но смещение в оценки не вносится. Например, предположим, что вы опрашиваете простым случайнным образом 100 школьников, а затем случайно теряете половину записей. Данная ситуация выглядит так, как если бы вы никогда не опрашивали школьников из потерянной половины. В этом случае вы бы остались с простой случайной выборкой из 50 школьников, так что случайная потеря записей не вносит ошибки.

Если данные отсутствуют из-за значения объясняющей переменной, то размер выборки также уменьшается, но оценки не будут смещеными. Например, в примере зависимости результатов тестов от соотношения учеников и учителей предположим, что мы использовали только округа, в которых соотношение учеников и учителей превышает 20. Несмотря на то что мы были бы не в состоянии сделать выводы о том, что случится, если размер класса станет  $STR \leq 20$ , это обстоятельство не внесло бы смещение в наш анализ влияния размеров класса для округов с  $STR > 20$ .

### Смещение из-за отбора наблюдений

Смещение из-за отбора наблюдений возникает, когда на процесс отбора влияет доступность данных и этот процесс связан с объясняемой переменной вне зависимости от регрессоров. Отбор наблюдений приводит к наличию корреляции между одним или несколькими регрессорами и компонентой ошибок, что влечет за собой смещение и несостоительность МНК-оценок.

**ОСНОВНЫЕ ПОНЯТИЯ**

9.5

В отличие от первых двух случаев, если данные отсутствуют из-за процесса отбора, который связан со значением зависимой переменной ( $Y$ ) вне зависимости от регрессоров ( $X$ ), то в результате такого процесса отбора может возникать корреляция между компонентой ошибок и регрессорами. Результирующее смещение МНК-оценок называется *смещением из-за отбора наблюдений*. Пример смещения из-за отбора наблюдений в голосовании был описан во вставке «Лэндон выигрывает!» в разделе 3.1. В этом примере метод отбора наблюдений (случайный выбор телефонных номеров автовладельцев) был связан с зависимой переменной (кого из кандидатов поддержит респондент на выборах в 1936 году), так как в 1936 году владельцы автомобилей с телефонами, скорее всего, были республиканцами. Проблема отбора наблюдений может быть приведена либо как следствие неслучайного отбора, либо как проблема отсутствия данных. В примере с голосованием в 1936 году

выборка представляла собой случайную выборку владельцев автомобилей с телефонами, но не была случайной выборкой избирателей. Кроме того, этот пример можно рассматривать как пример проблемы отсутствия данных при случайной выборке избирателей, но с отсутствующими данными для тех, кто не имеет машин и телефонов. Причина, из-за которой отсутствуют данные, связана с зависимой переменной, что приводит к смещению из-за отбора наблюдений.

Во вставке «Опережают ли акции паевых фондов рынок?» приведен пример смещения из-за отбора наблюдений в финансовой экономике. Основные идеи, связанные со смещением из-за отбора наблюдений, описаны во вставке «Основные понятия 9.5»<sup>1</sup>.

**Методы решения проблемы смещения из-за отбора наблюдений.** Методы, которые мы обсудили ранее, не могут устраниТЬ смещение из-за отбора наблюдений. Изучение методов, используемых для оценки моделей по данным с отбором наблюдений, выходит за рамки данной книги. Такие методы основываются на технике, рассмотренной в главе 11, в которой приведены все необходимые для дальнейшего ознакомления ссылки.



### ***Опережают ли акции паевых фондов рынок?***

Паевые инвестиционные фонды – это инвестиционные посредники, держащие портфели акций. Покупая акции в паевых фондах, мелкий инвестор может держать широко диверсифицированный портфель без хлопот и расходов (транзакционные издержки) на куплю-продажу акций отдельных компаний. Некоторые паевые фонды просто следят за рынком (например вкладывают лишь в акции в S&P 500), в то время как другие активно управляют своими портфелями, с тем чтобы получить прибыль, превышающую прибыль как рынка в целом, так и конкурирующих фондов. Но достигают ли этой цели активно управляемые паевые фонды? Существуют ли паевые фонды, прибыль которых от инвестиционных вложений устойчиво превышает прибыль конкурирующих фондов и рынка в целом?

Один из возможных способов ответа на эти вопросы заключается в том, чтобы сравнить будущие доходности паевых фондов, которые имеют наибольшие доходности по прошедшему году, с будущими доходностями других фондов и рынка в целом. Финансовые экономисты знают, что для проведения таких сравнений важно осторожно формировать выборку анализируемых паевых фондов. Однако эта задача не так проста, как кажется. В некоторых базах данных есть исторические данные о фондах, продолжающих работать и сейчас, но это означает, что самые неэффективные фонды не включаются в эти базы данных, потому что они вышли из бизнеса или были объединены в другие фонды. По этой причине результаты исследования с использованием данных по исторической эффективности

---

<sup>1</sup> В упражнении 18.16 даны задания, с помощью которых можно убедиться в формальной корректности обсуждаемых здесь утверждений.

существующих в настоящее время фондов будут смешены из-за отбора наблюдений: наблюдения отобраны в зависимости от значения объясняемой переменной, доходностей паевых фондов, потому что информация о фондах с низкими доходностями отсутствует в базах данных. Таким образом, реальная средняя доходность всех фондов (в том числе несуществующих) за десятилетний период будет меньше, чем средняя доходность тех фондов, которые все еще существуют в конце этих десяти лет, поэтому изучение только этих существующих фондов приведет к завышению оценок эффективности. Финансовые экономисты называют это смещение из-за отбора наблюдений «смещением вследствие выживаемости», поскольку только лучшие фонды выживают и оказываются в базе данных.

Когда финансовые эконометристы корректируют смещение вследствие выживаемости за счет включения данных о несуществующих фондах, результаты портят красивый портрет управляемых паевыми фондами. С поправкой на смещение вследствие выживаемости эконометрические оценки говорят о том, что активно управляемые портфели паевых фондов в среднем не опережают рынок и что высокая эффективность в прошлом не гарантирует высокую эффективность в будущем. Для дальнейшего ознакомления с экономикой паевых фондов и смещением вследствие выживаемости см. Malkiel (2003, Chapter 11) и Carhart (1997). Проблема смещения вследствие выживаемости также возникает в оценке деятельности хедж-фондов; для дальнейшего чтения по этому вопросу см. Aggarwal, Jorion (2010).



## **Одновременная причинность**

До сих пор мы предполагали, что причинность имеет однонаправленный характер от объясняющих переменных к зависимой переменной ( $X$  является причиной  $Y$ ). Но что случится, если причинные связи работают и в обратном направлении от зависимой переменной к одному или более регрессорам ( $Y$  является причиной  $X$ )? Если это так, то между переменными имеются причинные связи, работающие в обоих направлениях, то есть существует **одновременная причинность**. Если есть одновременная причинность, МНК-регрессия учитывает оба эффекта, поэтому МНК-оценка является смещенной и несостоятельной.

Например, наш анализ зависимости результатов тестов от уменьшения соотношения учеников и учителей основывается на предположении о том, что изменение соотношения учеников и учителей влияет на результаты тестов. Предположим, однако, что по инициативе правительства происходит субсидирование найма учителей в школьных округах с плохими результатами тестов. Если это так, то причинные связи будут двунаправленными: по причинам, связанным со спецификой учебного процесса, в округах с низким соотношением учеников и учителей будут высокие результаты тестов, но из-за наличия государственной

программы в округах с низкими результатами тестов будет уменьшаться соотношение учеников и учителей.

Одновременная причинность приводит к корреляции между регрессором и компонентой ошибок. В рассматриваемом выше примере предположим, что существует пропущенный фактор, приводящий к плохим результатам тестов; из-за государственной программы этот снижающий результаты тестов фактор, в свою очередь, приводит к снижению соотношения учеников и учителей. Таким образом, отрицательная компонента ошибок в теоретической регрессии зависимости результатов тестов от отношения учеников и учителей снижает результаты тестов, но из-за государственной программы она также приводит к уменьшению соотношения учеников и учителей. Другими словами, соотношение учеников и учителей положительно коррелирует с компонентой ошибок в теоретической регрессии. Это в свою очередь приводит к смещению из-за одновременной причинности и несостоительности МНК-оценки.

Эта корреляция между компонентой ошибок и регрессором может быть получена математически, если ввести дополнительное уравнение, которое описывает обратную причинную связь. Для удобства рассмотрим только две переменные,  $X$  и  $Y$ , и проигнорируем другие возможные регрессоры. Соответственно, есть два уравнения, в одном из которых  $X$  объясняет  $Y$ , а в другом  $Y$  объясняет  $X$ :

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (9.3)$$

$$\text{и } X_i = \gamma_0 + \gamma_1 Y_i + v_i. \quad (9.4)$$

В знакомом нам уравнении (9.3) коэффициент  $\beta_1$  отражает эффект влияния на переменную  $Y$  изменения переменной  $X$ , а  $u$  представляет другие факторы. Уравнение (9.4) представляет обратный причинный эффект от  $Y$  на  $X$ . Применительно к моделированию результатов тестов уравнение (9.3) представляет эффект влияния размера класса на результаты тестов, в то время как уравнение (9.4) представляет обратный причинный эффект влияния результатов тестов на размер класса, связанный с наличием соответствующей государственной программы.

Одновременная причинность приводит к корреляции между  $X_i$  и компонентой ошибок  $u_i$  в уравнении (9.3). Представим, что  $u_i$  отрицательна, из чего следует, что  $Y_i$  уменьшается. Однако это более низкое значение  $Y_i$  влияет на значение  $X_i$  через второе из этих уравнений, и если коэффициент  $\gamma_1$  положителен, то низкое значение  $Y_i$  будет приводить к низкому значению  $X_i$ . Таким образом, если коэффициент  $\gamma_1$  положителен,  $X_i$  и  $u_i$  будут положительно коррелированы<sup>1</sup>.

<sup>1</sup> Для того чтобы показать это математически, заметим, что уравнение (9.4) предполагает, что  $\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(v_i, u_i)$ . Предполагая, что  $\text{cov}(v_i, u_i) = 0$ , из уравнения (9.3) получаем, что  $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$ . Решая уравнение для  $\text{cov}(X_i, u_i)$ , получим  $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$ .

Одновременная причинность может быть выражена математически с использованием двух одновременных уравнений, поэтому смещение из-за одновременной причинности иногда называется *смещением одновременных уравнений*. Основные моменты, связанные со смещением из-за одновременной причинности, описаны во вставке «Основные понятия 9.6».

### Смещение из-за одновременной причинности

Смещение из-за одновременной причинности, также называемое смещением одновременных уравнений, возникает в регрессии  $Y$  на  $X$ , если помимо интересующего нас причинного влияния переменной  $X$  на переменную  $Y$  существует причинное влияние переменной  $Y$  на переменную  $X$ . Эта обратная причинность влечет коррелированность переменной  $X$  с компонентой ошибок в интересующей нас теоретической регрессии.

## ОСНОВНЫЕ ПОНЯТИЯ

9.6

**Методы решения проблемы смещения из-за одновременной причинности.** Существует два способа уменьшения смещения, возникающего из-за одновременной причинности. Один из них заключается в использовании регрессии с инструментальными переменными и рассматривается в главе 12. Второй – в разработке и реализации случайного управляемого эксперимента, в котором канал обратной причинности не действует, и такие эксперименты обсуждаются в главе 13.

### Источники несостоятельности стандартных ошибок МНК

Несостоятельные стандартные ошибки представляют собой другую угрозу для внутренней обоснованности. Даже если МНК-оценка является состоятельной, а выборка велика, несостоятельные стандартные ошибки будут порождать тестиирование гипотез с размером критерия, отличающимся от желаемого уровня значимости, и «95 %-е» доверительные интервалы, которые не содержат истинного значения в 95 % повторяющихся выборок.

Существуют две главные причины несостоятельности стандартных ошибок: гетероскедастичность случайных ошибок регрессии и корреляция компоненты ошибок между наблюдениями.

**Гетероскедастичность.** Как обсуждалось в разделе 5.4, в силу исторических причин некоторые эконометрические программные пакеты дают информацию о стандартных ошибках коэффициентов в предположении гомоскедастичности ошибок регрессии. Однако если ошибки регрессии гетероскедастичны, некорректно тестировать гипотезы и строить доверительные интервалы с использованием полученных таким образом

стандартных ошибок. Решение этой проблемы заключается в том, чтобы использовать стандартные ошибки, устойчивые к гетероскедастичности, и строить  $F$ -статистики, используя оценку дисперсии, устойчивую к гетероскедастичности. Расчет устойчивых к гетероскедастичности стандартных ошибок предусмотрен в качестве опции в современных программных пакетах.

**Корреляция компоненты ошибок между наблюдениями.** В некоторых случаях ошибки теоретической регрессии могут быть коррелированы наблюдениями. Это не может произойти, если данные получены при помощи случайной выборки из генеральной совокупности, поскольку случайность процесса отбора данных гарантирует, что ошибки распределены независимо от одного наблюдения к другому. Однако иногда выборка носит лишь частично случайный характер. Наиболее распространенный случай – это наблюдение за одним и тем же объектом в разные моменты времени, например наблюдение за одними и теми же школьными округами в различные годы. Если пропущенные переменные, которые включаются в ошибки регрессии, являются постоянными (например демографические характеристики школьных округов), в ошибках регрессии возникает «серийная» корреляция во времени. Серийная корреляция в компоненте ошибок может возникнуть в панельных данных (данные по нескольким округам за несколько лет) и во временных рядах (данные по одному округу за несколько лет).

Еще одна ситуация, при которой ошибки могут быть коррелированы между наблюдениями, возникает, когда отбор данных основан на географических единицах. Если существуют пропущенные переменные, которые отражают влияние географических факторов, то такие пропущенные переменные могут привести к корреляции ошибок регрессии для наблюдений, являющихся соседними с географической точки зрения.

Корреляция ошибок регрессии между наблюдениями не делает МНК-оценку смещенной или несостоительной, но нарушает второе предположение метода наименьших квадратов из вставки «Основные понятия 6.4». Последствия этого заключаются в том, что стандартные МНК-ошибки – и при наличии гомоскедастичности, и устойчивые к гетероскедастичности – некорректны в том смысле, что они не позволяют рассчитать доверительные интервалы с желаемым уровнем значимости.

Во многих случаях эта проблема может быть решена, если использовать альтернативную формулу для расчета стандартных ошибок. Мы приводим формулы для вычисления стандартных ошибок, которые устойчивы и к гетероскедастичности, и к серийной корреляции в главе 10 (панельная регрессия) и в главе 15 (регрессия временных рядов).

Во вставке «Основные понятия 9.7» суммируются угрозы для внутренней обоснованности, которые могут возникнуть при оценке множественной регрессии.

### **Множественная регрессия и угрозы для внутренней обоснованности**

Существует пять основных проблем, из-за которых может нарушаться внутренняя обоснованность исследования, основанного на оценке множественной регрессии:

1. Пропущенные переменные.
2. Неправильная спецификация функциональной формы.
3. Ошибки измерения переменных (регрессоров).
4. Отбор наблюдений.
5. Одновременная причинность.

Любая из этих причин приводит к нарушению первого предположения метода наименьших квадратов, то есть  $E(u_i | X_{1i}, \dots, X_{ki}) \neq 0$ , что в свою очередь означает, что МНК-оценка будет смещенной и несостоительной.

Некорректное вычисление стандартных ошибок коэффициентов также представляет собой угрозу для внутренней обоснованности. Стандартные ошибки коэффициентов, вычисленные в предположении гомоскедастичности случайных ошибок регрессии, являются некорректными, если имеет место гетероскедастичность. Если переменные не являются независимыми между наблюдениями, что нередко случается в панельных данных и во временных рядах, то необходима дальнейшая корректировка формулы для вычисления стандартных ошибок коэффициентов с целью получения их корректных значений.

Проверка отсутствия всех перечисленных проблем при оценке множественной регрессии представляет собой систематический способ оценки внутренней обоснованности проведенного исследования.

### **ОСНОВНЫЕ ПОНЯТИЯ**

**9.7**

## **9.3. Внутренняя и внешняя обоснованность при прогнозировании по модели регрессии**

До сих пор наше обсуждение множественного регрессионного анализа было сосредоточено на оценке причинных эффектов. Однако модели регрессии могут быть использованы и для других целей, включая прогнозирование. При использовании моделей регрессии для прогнозирования наличие внешней обоснованности очень важно, а вот наличие несмещенностии оценок причинных эффектов – нет.

### **Использование моделей регрессии для прогнозирования**

Глава 4 начинается с рассмотрения ситуации с окружным школьным инспектором, которая хочет знать, насколько улучшатся результаты тестов, если она

уменьшит размеры классов в своем школьном округе, то есть окружной школьный инспектор хочет знать причинный эффект влияния изменений в размерах классов на результаты тестов. Соответственно, в главах 4–8 мы знакомимся с тем, как можно использовать методы регрессионного анализа для оценки причинных эффектов, используя наблюдаемые данные.

Рассмотрим теперь другую проблему. Переезжая в крупный город, родители основывают свой выбор места проживания, в частности, исходя из качества местных школ. Родители хотели бы знать, как ученики различных школьных округов пишут стандартизованные тесты. Предположим, однако, что данные по результатам тестов недоступны (возможно они конфиденциальны), но есть данные о размерах классов. В этой ситуации родители должны были бы догадаться, насколько хорошо пишут стандартизованные тесты ученики из различных школьных округов на основе такой ограниченной информации. То есть проблема родителей заключается в прогнозировании средних результатов тестов в конкретном округе на основе информации, относящейся к результатам тестов, – в частности на основе информации о размерах классов.

Каким образом родители могут построить прогноз? Вспомним регрессию зависимости результатов тестов от соотношения учеников и учителей (*STR*) из главы 4:

$$\widehat{\text{TestScore}} = 698,9 - 2,28 \times \text{STR} . \quad (9.5)$$

Мы сделали вывод, что эта регрессия не является полезной для окружного школьного инспектора: МНК-оценка углового коэффициента смещена из-за пропущенных переменных, таких как индивидуальные особенности ученика и наличие возможностей для его внешкольного обучения.

Однако уравнение (9.5) может быть полезным для родителей, которые пытаются выбрать место проживания. Более точно, размер класса не является единственным фактором, определяющим качество выполнения тестов, но с точки зрения родителя важно, является ли он надежным предиктором качества выполнения теста. Родителям, заинтересованным в получении прогнозов результатов тестов, все равно, оценивает ли коэффициент в уравнении (9.5) причинный эффект влияния размера класса на результаты тестов. Скорее всего, родители просто хотят, чтобы такая регрессия объясняла большую часть вариации результатов тестов по округам и была стабильной, то есть была адекватна по отношению к школьным округам, рассматриваемым родителями как потенциальное место проживания. Хотя смещение из-за пропущенной переменной делает уравнение (9.5) бесполезным с точки зрения ответа на вопрос о причинности, но оно все еще может быть полезным для целей прогнозирования.

В более общем случае регрессионные модели могут давать надежные прогнозы, даже если их коэффициенты не имеют причинной интерпретации. Это обстоятельство лежит в основе использования модели регрессии для целей прогнозирования.

## ***Оценка обоснованности моделей регрессии для прогнозирования***

Поскольку проблемы, стоящие перед окружным школьным инспектором и родителями, концептуально различны, требования к обоснованности использования регрессии также различны для соответствующих проблем. Чтобы получить заслуживающие доверия оценки причинных эффектов, мы должны исследовать наличие/отсутствие проблем, угрожающих внутренней обоснованности, которые описаны во вставке «Основные понятия 9.7».

В противоположность этому, если мы хотим получить заслуживающие доверия прогнозы, оцененная регрессия должна иметь хорошую объясняющую силу, ее коэффициенты должны быть оценены точно и она должна быть стабильной в том смысле, что регрессия, оцененная на одном наборе данных, может быть использована для прогноза с применением других данных. Когда модель регрессии используется для прогнозирования, первостепенной задачей является выполнение условия внешней обоснованности модели – она должна быть стабильной и количественно применимой к ситуации, при которой сделан прогноз. В части IV мы возвращаемся к проблеме оценки обоснованности модели регрессии для прогнозирования будущих значений в моделях временных рядов.

### **9.4. Пример: результаты тестов и размеры классов**

Понятия внутренней и внешней обоснованности дают нам возможность критически взглянуть на то, что мы изучили (и что не изучили) при анализе данных по результатам тестов в Калифорнии.

#### ***Внешняя обоснованность***

Может ли анализ по Калифорнии быть обобщен, то есть является ли он внешне обоснованным, зависит от генеральной совокупности и условий, при которых обобщение сделано. В данном разделе мы попытаемся понять, могут ли результаты наших оценок быть обобщены для других стандартизованных тестов в других округах начальных школ в Соединенных Штатах.

В разделе 9.1 отмечалось, что наличие более одного исследования по одной и той же теме предоставляет возможность оценить внешнюю обоснованность обоих исследований, сравнивая их результаты. В случае анализа зависимости результатов тестов от размеров классов нам доступны другие сопоставимые базы данных. В этом разделе мы изучаем другую базу данных по результатам стандартизованных тестов для четвероклассников в 220 школьных округах штата Массачусетс в 1998 году. И в Массачусетсе, и в Калифорнии тесты являются распространенным измерителем знаний школьников и их академических навыков, хотя некоторые детали и отличаются. Кроме того, организация учебных занятий во многом схожа на уровне начальной школы в двух штатах

(что, вообще говоря, верно для большинства американских округов начальных школ), хотя особенности финансирования начальной школы и учебные программы отличаются. Таким образом, получение похожих результатов относительно эффекта влияния соотношения учеников и учителей на результаты тестов в Калифорнии и штате Массачусетс могло бы свидетельствовать о внешней обоснованности результатов исследования по Калифорнии. И наоборот, если полученные результаты были бы различны в двух штатах, то это подняло бы вопрос о внутренней или внешней обоснованности, по крайней мере, одного из исследований.

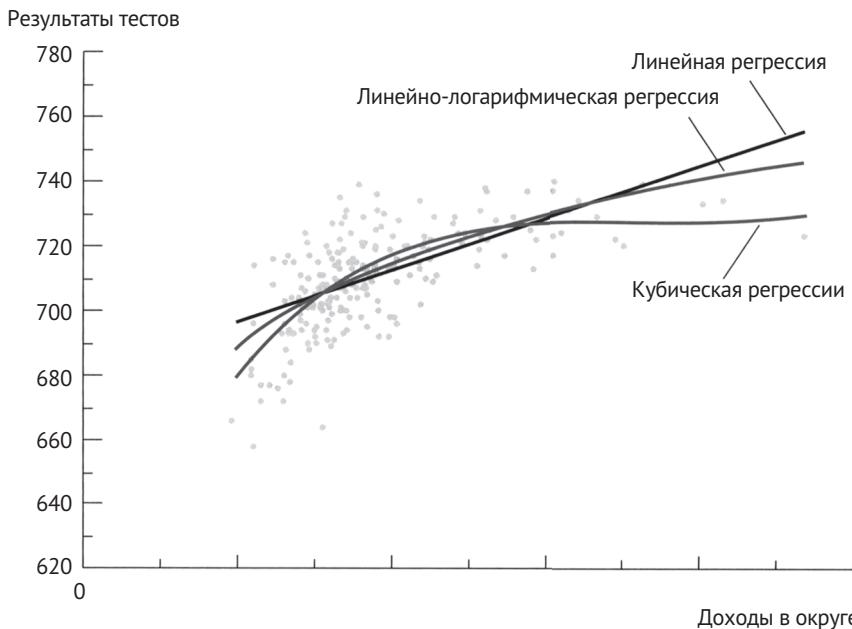
**Сравнение данных по Калифорнии и штату Массачусетс.** Как и данные по Калифорнии, данные по штату Массачусетс рассматриваются на уровне школьного округа. Определения переменных в данных по штату Массачусетс те же самые, что и в данных по Калифорнии, или почти такие же. Более подробная информация о данных по штату Массачусетс, в том числе определения переменных, приведена в приложении 9.1.

Таблица 9.1 представляет сводные статистические характеристики для выборок Калифорнии и штата Массачусетс. Результаты тестов в среднем выше в штате Массачусетс, но поскольку тесты в штатах различны, прямое сравнение результатов не является корректным. Среднее соотношение учеников и учителей выше для Калифорнии (19,6 против 17,3). Средний доход на 20 % выше в штате Массачусетс, но стандартное отклонение дохода больше в Калифорнии, то есть существует более высокий спред в средних доходах в Калифорнии, чем в штате Массачусетс. Средний процент учеников, все еще изучающих английский язык, и средний процент учеников, имеющих право на субсидированные обеды, в школьных округах Калифорнии выше, чем в штате Массачусетс.

Таблица 9.1

**Сводные статистические характеристики данных  
по результатам тестов для Калифорнии и штата Массачусетс**

	Калифорния		Штат Массачусетс	
	Среднее значение	Стандартное отклонение	Среднее значение	Стандартное отклонение
Результаты тестов	654,1	19,1	709,8	15,1
Соотношение учеников и учителей	19,6	1,9	17,3	2,3
% изучающих английский язык школьников	15,8 %	18,3 %	1,1 %	2,9 %
% учеников, имеющих право на субсидированный обед	44,7 %	27,1 %	15,3 %	15,1 %
Средний подушевой доход (долл.)	15 317 долл.	7226 долл.	18 747 долл.	5808 долл.
Число наблюдений	420		220	
Год	1999		1998	



**Рисунок 9.1. Результаты тестов и доходы для данных по штату Массачусетс**

Оцененная функция линейной регрессии не учитывает нелинейность соотношения между доходом и результатами тестов в данных по штату Массачусетс. Оцененные линейно-логарифмическая и кубическая функции регрессии близки для доходов в округе, расположенных в интервале между 13 тыс. долл. и 30 тыс. долл., в котором содержится наибольшее количество наблюдений.

**Результаты тестов и средний доход по округу.** Для экономии места мы не представляем диаграммы рассеяния всех данных по штату Массачусетс. Однако интересно было бы изучить взаимосвязь между результатами тестов и средним доходом в округе для штата Массачусетс, поскольку мы уделили этому много внимания в главе 8. Диаграмма рассеяния представлена на рисунке 9.1. Общий вид этой диаграммы рассеяния схож с диаграммой рассеяния для данных по Калифорнии, изображенной на рисунке 8.2: соотношение между доходом и результатами тестов оказывается крутым для низких значений доходов и более плоским для высоких значений. Очевидно, что линейная регрессия, приведенная на рисунке, не учитывает эту нелинейность. Функции кубической и логарифмической регрессий также изображены на рисунке 9.1. Функция кубической регрессии имеет несколько более высокий  $\bar{R}^2$ , чем логарифмическая спецификация (0,486 против 0,455). Сравнение рисунков 8.7 и 9.1 показывает, что общая картина нелинейности, наблюдающаяся для данных по доходам и результатам тестов в Калифорнии, также имеет место и для данных по штату Массачусетс. Однако точные функциональные формы, лучше всего описывающие эту нелинейность, отличаются: для данных по штату Массачусетс лучше подходит кубическая спецификация, а для Калифорнии – линейно-логарифмическая спецификация.

**Результаты оценки множественных регрессий.** Результаты оценки множественных регрессий для данных по штату Массачусетс представлены в таблице 9.2. Первая регрессия, представленная в столбце (1), включает в качестве

регрессора только соотношение учеников и учителей. Угловой коэффициент отрицателен ( $-1,72$ ), и гипотеза о том, что он равен нулю, может быть отвергнута на 1%-м уровне значимости ( $t=-1,72/0,50=-3,44$ ).

Таблица 9.2

**Оценки множественных регрессий зависимости результатов тестов от соотношения учеников и учителей: данные по штату Массачусетс**

Зависимая переменная: средняя оценка за комбинированный тест по английскому языку, математике и общественным наукам в школьном округе в четвертом классе; 220 наблюдений						
Объясняющая переменная	(1)	(2)	(3)	(4)	(5)	(6)
Соотношение учеников и учителей ( <i>STR</i> )	$-1,72^{**}$ (0,50)	$-0,69^*$ (0,27)	$-0,64^*$ (0,27)	12,4 (14,0)	$-1,02^{**}$ (0,37)	$-0,67^*$ (0,27)
<i>STR</i> <sup>2</sup>				$-0,680$ (0,737)		
<i>STR</i> <sup>3</sup>				0,011 (0,013)		
% школьников, изучающих английский язык		$-0,411$ (0,306)	$-0,437$ (0,303)	$-0,434$ (0,300)		
Бинарная переменная, равная единице, если процент школьников, изучающих английский язык, превышает медиану ( <i>HiEL</i> )					$-12,6$ (9,8)	
<i>HiEL</i> $\times$ <i>STR</i>					0,80 (0,56)	
% школьников, имеющих право на субсидированный обед		$-0,521^{**}$ (0,077)	$-0,582^{**}$ (0,097)	$-0,587^{**}$ (0,104)	$-0,709^{**}$ (0,091)	$-0,653^{**}$ (0,72)
Логарифм среднего подушевого дохода в округе		16,53** (3,15)				
Средний подушевой доход в округе			$-3,07$ (2,35)	$-3,38$ (2,49)	$-3,87^*$ (2,49)	$-3,22$ (2,31)
Средний подушевой доход в округе <sup>2</sup>			0,164 (0,085)	0,174 (0,089)	0,184* (0,090)	0,165 (0,085)
Средний подушевой доход в округе <sup>3</sup>			$-0,002\ 2^*$ (0,001 0)	$-0,002\ 3^*$ (0,001 0)	$-0,002\ 3^*$ (0,001 0)	$-0,002\ 2^*$ (0,001 0)
Константа	739,6** (8,6)	682,4** (11,5)	774,0** (21,3)	665,5** (81,3)	759,9** (23,2)	747,4** (20,3)
<i>F</i> -статистики и <i>p</i> -значения для проверки совместных гипотез						
Все переменные <i>STR</i> и компоненты взаимодействия равны нулю				2,86 (0,038)	4,01 (0,020)	
<i>STR</i> <sup>2</sup> , <i>STR</i> <sup>3</sup> =0				0,45 (0,641)		
Доход <sup>2</sup> , Доход <sup>3</sup> =0			7,74 (<0,001)	7,75 (<0,001)	5,85 (0,003)	6,55 (0,002)

## Окончание таблицы 9.2

<b>Зависимая переменная: средняя оценка за комбинированный тест по английскому языку, математике и общественным наукам в школьном округе в четвертом классе; 220 наблюдений</b>						
<b>Объясняющая переменная</b>	(1)	(2)	(3)	(4)	(5)	(6)
<i>HiEL, HiEL × STR = 0</i>					1,58 (0,208)	
<i>SER</i>	14,64	8,69	8,61	8,63	8,63	8,64
$\bar{R}^2$	0,063	0,670	0,676	0,675	0,675	0,674

*Примечание.* Регрессии оценены на данных по школьным округам штата Массачусетс, описанных в приложении 9.1. В скобках под коэффициентами приведены стандартные ошибки; в скобках под F-статистиками приведены соответствующие им p-значения. Коэффициенты значимы на \*5 %-м или \*\*1 %-м уровнях значимости.

В оставшихся столбцах приведены результаты расчетов после включения дополнительных переменных, учитывающих индивидуальные характеристики учеников и нелинейности в оцененной функции регрессии. Контролируя регрессию на процент изучающих английский язык школьников, процент учеников, имеющих право на бесплатный обед, и средний доход в округе, мы уменьшаем оцененный коэффициент при соотношении учеников и учителей на 60 %, с -1,72 в регрессии (1) до -0,69 в регрессии (2) и -0,64 в регрессии (3).

Сравнение  $\bar{R}^2$ -ов регрессии (2) и (3) указывает на то, что кубическая спецификация (3) лучше для моделирования зависимости между результатами тестов и доходом, чем логарифмическая спецификация (2), даже если мы считаем соотношение учеников и учителей постоянным. Не существует статистически значимого свидетельства наличия нелинейной зависимости между показателями результатов тестов и соотношения учеников и учителей: F-статистика в регрессии (4), проверяющая гипотезу о том, что коэффициенты при  $STR^2$  и  $STR^3$  равны нулю, имеет p-значение, равное 0,641. Аналогично, не обнаружено свидетельства того, что уменьшение соотношения учеников и учителей оказывает различное влияние в округах с большим и малым числом изучающих английский язык школьников [ $t$ -статистика при  $HiEL \times STR$  в регрессии (5) равна  $0,80 / 0,56 = 1,43$ ]. Наконец, регрессия (6) показывает, что оцененный коэффициент при соотношении учеников и учителей не изменяется существенно, если исключить из модели процент изучающих английский язык школьников [который является незначимым в регрессии (3)]. В целом результаты регрессии (3) не чувствительны к изменениям функциональной формы, и различные ее спецификации рассматриваются в регрессиях (4)–(6) в таблице 9.2. Таким образом, можно считать регрессию (3) базовой для оценки эффекта влияния соотношения учеников и учителей на результаты тестов в штате Массачусетс.

**Сравнение результатов по штату Массачусетс и Калифорнии.** Анализируя данные по Калифорнии, мы пришли к следующим выводам:

1. Включение переменных, которые учитывают индивидуальные характеристики социального положения школьников, уменьшает коэффициент

при соотношении учеников и учителей на 68 % с  $-2,28$  [таблица 7.1, регрессия (1)] до  $-0,73$  [таблица 8.3, регрессия (2)].

2. Гипотеза о том, что истинный коэффициент при соотношении учеников и учителей равен нулю, отвергается на 1%-м уровне значимости даже после включения переменных, которые учитывают характеристики социального положения учеников и экономические характеристики округа.
3. Эффект от сокращения соотношения учеников и учителей не имеет сильной зависимости от процента изучающих английский язык школьников в округе.
4. Существует некоторое свидетельство того, что зависимость между результатами тестов и соотношением учеников и учителей нелинейна.

Получаем ли мы те же самые результаты для данных по штату Массачусетс?

Для выводов (1), (2) и (3) ответ положителен. Включение в регрессию дополнительных контрольных переменных уменьшает коэффициент при соотношении учеников и учителей на 60 % с  $-1,72$  [таблица 9.2, регрессия (1)] до  $-0,69$  [таблица 9.2, регрессия (2)]. Коэффициенты при соотношении учеников и учителей остаются значимыми и после включения контрольных переменных. Эти коэффициенты значимы только на 5 %-м уровне значимости для данных по штату Массачусетс, в то время как они являются значимыми на 1 %-м уровне значимости для данных по Калифорнии. Однако в данных по Калифорнии почти вдвое больше наблюдений, так что неудивительно, что оценки по Калифорнии являются более точными. Как и в данных по Калифорнии, в данных по штату Массачусетс отсутствует статистически значимое свидетельство взаимодействия между соотношением учеников и учителей и бинарной переменной, характеризующей факт наличия в школьном округе большого числа школьников, изучающих английский язык.

Вывод (4), однако, неверен для данных по штату Массачусетс: гипотеза о том, что зависимость между соотношением учеников и учителей и результатами тестов является линейной, не может быть отвергнута на 5 %-м уровне значимости при использовании в качестве альтернативы кубической регрессии.

Поскольку стандартизованные тесты в штатах различаются, коэффициенты регрессий не могут сравниваться непосредственно: один балл оценки за тест в штате Массачусетс неэквивалентен баллу оценки за тест в Калифорнии. Однако если в качестве оценок результатов тестов использовать одинаковые единицы, то оцененные эффекты влияния изменения размеров классов можно сравнивать. Один из способов сделать это – преобразовать результаты тестов, стандартизировав их: нужно вычесть выборочное среднее и разделить на стандартное отклонение, так чтобы преобразованная переменная имела нулевое среднее и единичную дисперсию. Тогда угловые коэффициенты в регрессии с преобразованной зависимой переменной равны угловым коэффициентам исходной регрессии, деленным на стандартное отклонение показателя результатов тестов. Таким образом, коэффициент при соотношении учеников и учителей, деленный на стандартное отклонение результатов тестов, сопоставим для двух наборов данных.

Таблица 9.3

**Соотношение учеников и учителей и результаты тестов: сравнение оценок по данным Калифорнии и штата Массачусетс**

<b>Оцененный эффект от уменьшения соотношения учеников и учителей на 2, в единицах измерения:</b>				
	МНК-оценка $\hat{\beta}_{STR}$	Стандартное отклонение показателя результатов тестов в округе	Баллы за тест	Стандартизованные единицы
<b>Калифорния</b>				
Линейная регрессия: таблица 8.3 (2)	–0,73 (0,26)	19,1	1,46 (0,52)	0,076 (0,027)
Кубическая регрессия: таблица 8.3 (7) Снижение $STR$ с 20 до 18	–	19,1	2,93 (0,70)	0,153 (0,037)
Кубическая регрессия: таблица 8.3 (7) Снижение $STR$ с 22 до 20	–	19,1	1,90 (0,69)	0,099 (0,036)
<b>Штат Массачусетс</b>				
Линейная регрессия: таблица 9.2 (3)	–0,64 (0,27)	15,1	1,28 (0,54)	0,085 (0,036)

Примечание. В скобках приведены стандартные ошибки.

Результаты такого сравнения приводятся в таблице 9.3. В первом столбце даны МНК-оценки коэффициентов при соотношении учеников и учителей в регрессии, в которую в качестве контрольных переменных включены переменные, характеризующие процент изучающих английский язык школьников, процент учеников, получающих субсидированные обеды, и средний доход в округе. Во втором столбце приводятся значения стандартных отклонений результатов тестов среди округов конкретного штата. В последних двух столбцах приведены значения оцененных эффектов влияния уменьшения соотношения учеников и учителей на двух учеников на учителя на результаты тестов (предложение нашего школьного инспектора), первый – в единицах баллов за тест и второй – в стандартизованных единицах. Для линейной спецификации МНК-оценка коэффициента в регрессии, оцененной по данным Калифорнии, равна –0,73, поэтому уменьшение соотношения учеников и учителей на 2 оценивается улучшением результатов тестов в округе на  $-0,73 \times (-2) = 1,46$  балла. Поскольку стандартное отклонение результатов тестов равно 19,1 баллов, это соответствует  $1,46 / 19,1 = 0,076$  стандартизованных результатов тестов среди округов штата. Стандартная ошибка этой оценки равна  $0,26 \times 2 / 19,1 = 0,027$ . Оцененные эффекты для нелинейных моделей и их стандартные ошибки вычисляются с использованием метода, описанного в разделе 8.1.

На основе линейной модели, оцененной с использованием данных по Калифорнии, уменьшение соотношения учеников и учителей на 2 оценивается

улучшением результатов тестов на 0,076 стандартизованных единиц со стандартной ошибкой 0,027. Нелинейные модели для данных Калифорнии показывают несколько больший размер улучшения результатов тестов со специфическим эффектом, зависящим от начального значения соотношения учеников и учителей. На основе данных по штату Массачусетс этот оцененный эффект составляет 0,085 стандартизованных единиц со стандартной ошибкой 0,036.

Эти оценки практически совпадают. Уменьшение соотношения учеников и учителей предсказывает рост результатов тестов, но предсказанное улучшение мало. В данных по Калифорнии, например, разница в результатах тестов между медианным округом и округом, находящимся на 75 процентиле, составляет 12,2 баллов (таблица 4.1), или 0,64 ( $=12,2/19,1$ ) стандартизованных единиц. Оцененный по линейной модели эффект составляет чуть более одной десятой этой величины; другими словами, согласно полученной оценке, сокращение соотношения учеников и учителей на 2 передвинет школу из округа с медианным результатом тестов по направлению к округу с результатом тестов, находящимся на 75 % процентиле, на расстояние чуть большее одной десятой расстояния между медианным округом и округом, находящимся на 75 % процентиле по результатам тестов. Уменьшение соотношения учеников и учителей в округе на 2 является большим изменением для округа, но оцененный выигрыш, приведенный в таблице 9.3, является довольно меленьkim, хотя и ненулевым.

## ***Внутренняя обоснованность***

Сходство результатов исследований, проведенных для Калифорнии и штата Массачусетс, не гарантирует их внутреннюю обоснованность. В разделе 9.2 перечисляются пять возможных угроз внутренней обоснованности, которые могут вызывать смещение в оценке влияния изменения размеров классов на результаты тестов. Рассмотрим эти проблемы.

***Пропущенные переменные.*** Во множественных регрессиях, представленных в этой и предыдущей частях, мы осуществляли контроль над индивидуальными характеристиками учеников (процент изучающих английский язык), экономическими характеристиками семьи (процент учащихся, получающих субсидированные обеды) и более широкой мерой благосостояния округа (средний доход в округе).

Если эти контрольные переменные адекватны, то для целей регрессионного анализа это равносильно предположению о том, что соотношение учеников и учителей случайным образом распределено между районами с одинаковыми значениями этих контрольных переменных, и в этом случае предположение о независимости условного среднего выполняется. Однако могут существовать некоторые пропущенные факторы, для которых рассматриваемые три переменные не являются контрольными. Например, если соотношение учеников и учителей коррелировано с показателями, характеризующими квалификацию учителей, даже в округах с одинаковой долей иммигрантов и одинаковыми социоэкономическими характеристиками (возможно, потому, что лучших учителей привлекают в школы с меньшим соотношением учеников и учителей),

и если квалификация учителей влияет на результаты тестов, то пропуск показателя, характеризующего квалификацию учителей, может привести к смещению коэффициента при соотношении учеников и учителей. Аналогично, среди округов с одинаковыми социоэкономическими характеристиками семьи в округах с низким соотношением учеников и учителей могут быть более привержены к обучению своих детей на дому. Такие пропущенные факторы могут привести к смещению из-за пропущенной переменной.

Один из способов устранения смещения из-за пропущенных переменных, по крайней мере в теории, заключается в проведении эксперимента. Например, ученики могут быть случайно отобраны в классы различных размеров, и впоследствии будут сравниваться результаты выполнения ими стандартизованных тестов. Такое исследование действительно проводилось в штате Теннесси, и мы изучаем его в главе 13.

**Функциональная форма.** Здесь и в главе 8 были изучены различные функциональные формы регрессии. Мы обнаружили, что некоторые из возможных видов нелинейностей, которые мы проанализировали, не были статистически значимыми. Те же оценки, что мы получили, существенно не меняли эффект влияния уменьшения соотношения учеников и учителей. Несмотря на то что может быть проведен дальнейший анализ функциональной формы регрессии, полученные результаты говорят о том, что основные выводы нашего исследования вряд ли могут быть чувствительны к использованию различных спецификаций нелинейной регрессии.

**Ошибки в переменных.** Среднее соотношение учеников и учителей в округе является широкой и потенциально неточной мерой размера класса. Например, поскольку ученики переезжают из округа в округ, соотношение учеников и учителей не может точно представлять фактические размеры класса, которые определяются исходя из количества школьников, которые писали тесты, что, в свою очередь, может приводить к смещению в сторону нуля оценки влияния размера класса. Другой переменной с потенциальной ошибкой измерения является доход в округе. Эти данные взяты из переписи населения 1990 года, в то время как все остальные данные выборки относятся к 1998 году (в штате Массачусетс) или к 1999 году (в Калифорнии). Если экономическая ситуация в округе существенно изменилась за 1990-е годы, то имеющиеся данные были бы неточной мерой фактического среднего дохода в округе.

**Отбор наблюдений.** Данные по Калифорнии и штату Массачусетс относятся к округам государственных начальных школ в штате, что удовлетворяет ограничению на минимальный размер выборки, поэтому у нас нет оснований полагать, что в данном случае мы сталкиваемся с проблемой отбора наблюдений.

**Одновременная причинность.** Одновременная причинность будет возникать, если результаты стандартизованных тестов будут влиять на соотношение учеников и учителей. Такая ситуация может возникнуть, например, если существует бюрократический или политический механизм для увеличения финансирования незэффективных школ или округов, которые, в свою очередь, в результате будут нанимать большее число учителей. Известно, что в штате Массачусетс не было никаких механизмов для выравнивания финансирования школ в рассматриваемый период

времени. В Калифорнии ряд судебных дел привел к некоторому выравниванию финансирования, но это перераспределение средств не было основано на успеваемости учащихся. Таким образом, ни в исследовании по штату Массачусетс, ни в исследовании по Калифорнии не возникает проблема одновременной причинности.

**Гетероскедастичность и корреляция ошибок между наблюдениями.** Во всех расчетах, приведенных здесь и ранее, используются стандартные ошибки коэффициентов, устойчивые к гетероскедастичности ошибок регрессии, поэтому гетероскедастичность не является угрозой для внутренней обоснованности. Корреляция компоненты ошибок между наблюдениями, однако, может угрожать состоятельности оценок стандартных ошибок, поскольку не используется простая случайная выборка (обе выборки состоят из данных по всем округам начальных школ в штате). Несмотря на существование альтернативных формул для расчета стандартных ошибок коэффициентов, которые могут быть использованы в этой ситуации, обсуждение довольно сложных и специфических деталей этого мы оставляем для более углубленных работ.

## ***Обсуждение и применение***

Сходство между результатами по штату Массачусетс и Калифорнии показывает, что эти исследования внешне обоснованы, в том смысле что основные выводы могут быть обобщены на результаты стандартизованных тестов в других округах начальных школ в Соединенных Штатах.

Некоторые из самых важных потенциальных угроз для внутренней обоснованности были решены путем включения в регрессии переменных, характеризующих индивидуальные особенности учеников, экономическое положение их семей и благосостояние округа, и путем проверки нелинейности функции регрессии. Тем не менее некоторые потенциальные угрозы для внутренней обоснованности все еще остаются. Главным кандидатом здесь является смещение из-за пропущенных переменных, возникающее, возможно, потому, что контрольные переменные не учитывают некоторые другие характеристики школьных округов или возможность внешкольного обучения.

На основании данных как по Калифорнии, так и по штату Массачусетс, мы можем ответить на вопрос школьного инспектора из раздела 4.1: после учета экономического положения семьи, индивидуальных характеристик ученика и благосостояния округа, а также после оценки нелинейных функций регрессии снижение соотношения учеников и учителей на два ученика на учителя предсказывает улучшение результатов тестов примерно на 0,08 единиц стандартизованных результатов тестов по округам. Этот эффект является статистически значимым, но достаточно малым. Такой небольшой оцененный эффект согласуется с результатами многих исследований, в которых изучалось влияние сокращения размеров классов на результаты тестов<sup>1</sup>.

Теперь окружной школьный инспектор может использовать полученные оценки, чтобы принять решение об уменьшении размеров классов в школах

---

<sup>1</sup> Если вы заинтересованы в дальнейшем изучении зависимости между размерами классов и результатами тестов, см. обзоры Ehrenberg et al. (2001a, 2001b).

округа. Принимая это решение, она должна будет взвесить расходы на предполагаемое сокращение размеров классов и получаемые при этом выгоды. Расходы состоят из зарплат учителей и расходов на оборудование дополнительных классов. Выгоды включают в себя улучшение успеваемости, которое мы измеряем результатами стандартизованных тестов, но есть и другие потенциальные преимущества, которые мы не изучали, в том числе снижение числа не выполнивших тест учеников и повышение будущих доходов. Оцененный эффект влияния этого предложения на результаты стандартизованных тестов является одним из важных моментов при расчете расходов и выгод.

## 9.5. Заключение

Понятия внутренней и внешней обоснованности составляют структуру для оценки того, что мы узнали из эконометрического исследования.

Исследование на основе множественной регрессии является внутренне обоснованным, если оцененные коэффициенты являются несмещеными и состоятельными и если оценки стандартных ошибок коэффициентов являются состоятельными. Проблемы, угрожающие внутренней обоснованности такого исследования, включают пропущенные переменные, неправильную спецификацию функциональной формы регрессии (нелинейность), неточное измерение независимых переменных (ошибки в переменных), отбор наблюдений и одновременную причинность. Каждая из них приводит к корреляции между независимой переменной и компонентой ошибок, которая, в свою очередь, делает МНК-оценки смещенными и несостоятельными. Если ошибки коррелированы между наблюдениями, что часто встречается во временных рядах, или если они гетероскедастичны, а стандартные ошибки коэффициентов вычисляются с использованием предположения об их гомоскедастичности, то внутренняя обоснованность находится под угрозой, потому что оценки стандартных ошибок будут несостоятельными. Эти проблемы могут быть решены путем более корректного вычисления стандартных ошибок коэффициентов.

Исследование, использующее регрессионный анализ, как и любое статистическое исследование, является внешне обоснованным, если его результаты могут быть обобщены за пределы исследуемой генеральной совокупности и имеющихся условий. Иногда может быть полезным сравнить несколько исследований по рассматриваемой теме. Однако независимо от того, существуют или нет другие исследования по теме, оценка внешней обоснованности требует вынесения суждений о сходствах изучаемой генеральной совокупности и имеющихся условий, а также целевой генеральной совокупности и соответствующих условий, то есть генеральной совокупности и условий, на которые обобщаются результаты исследования.

В следующих двух частях учебника изучаются методы устранения угроз для внутренней обоснованности, которые не могут быть решены только путем оценки множественной регрессии. Часть III расширяет модель множественной регрессии рассмотрением способов, направленных на решение проблем потенциального смещения МНК-оценок для всех пяти источников этого смещения;

также в части III обсуждается другой подход получения внутренней обоснованности – случайные управляемые эксперименты. В части IV рассматриваются методы анализа временных рядов, а также использование временных рядов для оценки так называемых динамических причинных эффектов, которые являются причинными эффектами, меняющимися во времени.

## **Выводы**

1. Чтобы оценить статистическое исследование, необходимо задать вопрос, является ли оно внутренне и внешне обоснованным. Исследование является внутренне обоснованным, если статистические выводы о причинных эффектах обоснованы для исследуемой генеральной совокупности. Исследование является внешне обоснованным, если его выводы и заключения могут быть обобщены с имеющейся генеральной совокупности и имеющихся условий на другие генеральные совокупности и условия.
2. При регрессионной оценке причинных эффектов существует два типа угроз для внутренней обоснованности. Во-первых, МНК-оценки будут смещеными и несостоительными, если регрессоры и компонента ошибок коррелированы. Во-вторых, доверительные интервалы и результаты тестирования гипотез не будут надежными, если стандартные ошибки коэффициентов рассчитаны некорректно.
3. Объясняющие переменные и компоненты ошибок могут быть коррелированы, если существуют пропущенные переменные, используется некорректная функциональная форма, один или более регрессоров измеряются с ошибкой, наблюдения выбраны неслучайно из генеральной совокупности или есть одновременная причинность между регрессорами и зависимыми переменными.
4. Стандартные ошибки некорректны, если ошибки гетероскедастичны и используемый для расчетов эконометрический пакет рассчитывает стандартные ошибки коэффициентов только в предположении гомоскедастичности или если компонента ошибок коррелирована между различными наблюдениями.
5. Если модели регрессии используются исключительно для прогнозирования, то условие несмещенности оценок коэффициентов регрессии не является необходимым. В этом случае важно, однако, чтобы модель регрессии была внешне обоснованной для прогнозирования.

## **Основные понятия**

Изучаемая генеральная совокупность (с. 320).

Внутренняя обоснованность (с. 320).

Внешняя обоснованность (с. 320).

Целевая генеральная совокупность (с. 320).

Неправильно специфицированная функциональная форма (с. 326).

- Смещение из-за ошибок в переменных (с. 327).
- Классическая модель с ошибками измерения (с. 328).
- Смещение из-за отбора наблюдений (с. 331).
- Одновременная причинность (с. 333).
- Смещение одновременных уравнений (с. 335).

### ***Вопросы для повторения и закрепления основных понятий***

- 9.1. Каково различие между понятиями внутренней и внешней обоснованности? Между изучаемой генеральной совокупностью и целевой генеральной совокупностью?
- 9.2. Во вставке «Основные понятия 9.2» описана проблема выбора объясняющих переменных в терминах компромисса между смещением и дисперсией. В чем заключается этот компромисс? Почему включение дополнительного регрессора могло бы уменьшить смещение? Увеличить дисперсию?
- 9.3. Экономические переменные часто измеряются с ошибками. Означает ли это, что регрессионный анализ ненадежен? Поясните.
- 9.4. Предположим, что в штате было проведено бесплатное тестирование для всех третьеклассников и что эти данные были использованы для исследования влияния размера класса на успеваемость учащихся. Объясните, как смещение из-за отбора наблюдений может привести к тому, что результаты исследования окажутся несостоительными.
- 9.5. Исследователь оценивает эффект влияния расходов на полицию на уровень преступности, используя статистику по городам. Объясните, как одновременная причинность может привести к тому, что результаты исследования окажутся несостоительными.
- 9.6. Исследователь оценивает регрессию, используя два различных программных пакета. В первом пакете стандартные ошибки коэффициентов рассчитываются в предположении гомоскедастичности. Во втором – используя устойчивые к гетероскедастичности стандартные ошибки. Полученные стандартные ошибки сильно отличаются. Какую формулу должен использовать исследователь? Почему?

### ***Упражнения***

- 9.1. Предположим, что вы только что ознакомились с результатами тщательно проведенного статистического исследования влияния рекламы сигарет на спрос на них. На основе данных 1970-х годов по Нью-Йорку в исследовании был сделан вывод, что реклама, расположенная на автобусах и в метро, была более эффективной, чем печатная реклама. Используйте понятие внешней обоснованности, чтобы определить, можно ли распространить эти результаты на Бостон в 1970-х годах; Лос-Анджелес в 1970-х годах; Нью-Йорк в 2010-х годах.

- 9.2. Рассмотрим модель парной регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$  и предположим, что она удовлетворяет предположениям из вставки «Основные понятия 4.3». Предположим, что переменная  $Y_i$  измерена с ошибками, поэтому у нас есть данные в виде  $\tilde{Y}_i = Y_i + w_i$ , где  $w_i$  является ошибкой измерения, которая является i.i.d. и не зависит от  $Y_i$  и  $X_i$ . Рассмотрим теоретическую регрессию  $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$ , где  $v_i$  – ошибка регрессии, оцененной с использованием неправильно измеренной зависимой переменной  $\tilde{Y}_i$ .
- Покажите, что  $v_i = u_i + w_i$ .
  - Покажите, что регрессия  $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$  удовлетворяет предположениям из вставки «Основные понятия 4.3». (Предположим, что  $w_i$  не зависит от  $Y_j$  и  $X_j$  для всех значений  $i$  и  $j$  и имеет конечный четвертый момент.)
  - Являются ли МНК-оценки состоятельными?
  - Может ли доверительный интервал строиться обычным способом?
  - Прокомментируйте утверждения: «Ошибка измерения в  $X$ -ах является серьезной проблемой. Ошибка измерения в  $Y$  – нет».
- 9.3. Экономисты, изучающие факторы, влияющие на доходы женщин, получили загадочный эмпирический результат. Используя случайную выборку, состоящую из работающих женщин, они оценивали регрессию доходов женщин на количество детей у них и множество контрольных переменных (возраст, образование, занятость и так далее). Было обнаружено, что женщины с большим количеством детей имели более высокие зарплаты, даже если учитывались контрольные переменные. Объясните, как отбор наблюдений может быть причиной такого результата. (Подсказка: Заметьте, что неработающие женщины пропущены в выборке.) [Эта эмпирическая загадка, мотивированная исследованием Джеймса Хекмана по проблеме отбора наблюдений, привела его к получению Нобелевской премии по экономике в 2000 году. См. Heckman (1974).]
- 9.4. Используя регрессии из столбца (2) таблицы 8.3 и столбца (2) таблицы 9.2, постройте таблицу, аналогичную таблице 9.3, для сравнения оцененных эффектов влияния от 10 %-го увеличения доходов в округе на результаты тестов в Калифорнии и штате Массачусетс.
- 9.5. Уравнение спроса на товары имеет вид:  $Q = \beta_0 + \beta_1 P + u$ , где  $Q$  обозначает количество товара,  $P$  – его цену и  $u$  обозначает факторы, отличные от цены и определяющие спрос. Уравнение предложения товаров имеет вид:  $Q = \gamma_0 + \gamma_1 P + v$ , где  $v$  обозначает факторы, отличные от цены товара и определяющие предложение. Предположим, что  $u$  и  $v$  имеют нулевое среднее, дисперсии  $\sigma_u^2$  и  $\sigma_v^2$  и взаимно некоррелированы.
- Решите два одновременных уравнения, чтобы показать, каким образом  $Q$  и  $P$  зависят от  $u$  и  $v$ .
  - Выполните формулы средних  $P$  и  $Q$ .
  - Выполните формулы дисперсий  $P$ ,  $Q$  и ковариации между  $Q$  и  $P$ .
  - У нас есть случайная выборка наблюдений  $(Q_i, P_i)$  и строится регрессия  $Q_i$  на  $P_i$  (т.е.  $Q_i$  является зависимой переменной, а  $P_i$  – регрессором). Предположим, что выборка очень велика.

- (i) Используйте ответы из пунктов (б) и (в), чтобы получить формулы для значений регрессионных коэффициентов. [Подсказка: используйте уравнения (4.7) и (4.8).]
- (ii) Исследователь использует угловой коэффициент такой регрессии как оценку углового коэффициента функции спроса ( $\beta_1$ ). Является ли оцененный угловой коэффициент слишком большим или слишком маленьким? [Подсказка: вспомните, что кривая спроса имеет отрицательный наклон, а кривая предложения — положительный наклон.]
- 9.6. Предположим, что у нас есть выборка из  $n = 100$  независимых одинаково распределенных наблюдений ( $X_i, Y_i$ ), для которой мы получаем следующие оценки регрессии:

$$\hat{Y} = 32,1 + 66,8 X, \text{SER} = 15,1, R^2 = 0,81.$$

(15,1)      (12,2)

Другой исследователь оценивает такую же регрессию, но он делает ошибку при вводе данных в компьютер: он вводит каждое наблюдение дважды, поэтому у него 200 наблюдений (наблюдение 1 введено дважды, наблюдение 2 введено дважды и так далее).

- а) Какие оценки он получит, используя эти 200 наблюдений? (Подсказка: запишите «некорректные» значения выборочных средних, дисперсий и ковариации  $Y$  и  $X$  как функций от «корректных» значений. Используйте их для определения регрессионных статистик.)

$$\hat{Y} = \frac{\underline{\quad}}{\underline{\quad}} + \frac{\underline{\quad}}{\underline{\quad}} X, \text{SER} = \underline{\quad}, R^2 = \underline{\quad}.$$

- б) Какие условия внутренней обоснованности нарушаются (если такие имеются)?

- 9.7. Верны или ложны следующие утверждения? Объясните ваши ответы.
- а) «МНК-регрессия  $Y$  на  $X$  будет внутренне несостоительной, если переменная  $X$  коррелирована с компонентой ошибок».
- б) «Каждая из пяти основных угроз для внутренней обоснованности предполагает, что переменная  $X$  коррелирована с компонентой ошибок».
- 9.8. Будет ли регрессия в уравнении (9.5) полезной для прогнозирования результатов тестов в школьном округе в штате Массачусетс? Почему да или почему нет?
- 9.9. Рассмотрим линейную регрессию *TestScore* на *Income*, изображенную на рисунке 8.2, и нелинейную регрессию из уравнения (8.18). Будет ли любая из этих регрессий давать надежную оценку влияния доходов на результаты тестов? Будет ли любая из этих регрессий надежной с точки зрения прогнозирования результатов тестов? Объясните.
- 9.10. Прочтите вставку «Отдача от образования и гендерный разрыв» из раздела 8.3. Обсудите внутреннюю и внешнюю обоснованность оценки влияния уровня образования на доходы.
- 9.11. Прочтите вставку «Спрос на экономические журналы» из раздела 8.3. Обсудите внутреннюю и внешнюю обоснованность оценки влияния цены цитаты на подписку на журналы.

9.12. Рассмотрим парную регрессию  $Y_i = \beta_0 + \beta_1 X_i + u_i$  и предположим, что она удовлетворяет предположениям из вставки в «Основные понятия 4.3». Регрессор  $X_i$  пропущен, но данные по связанной с ним переменной  $Z_i$  доступны, и значение  $X_i$  оценено на основе  $\tilde{X}_i = E(X_i | Z_i)$ . Пусть  $w_i = \tilde{X}_i - X_i$ .

- a) Покажите, что  $\tilde{X}_i$  является наилучшей оценкой  $X_i$ , полученной с использованием  $Z_i$ . То есть покажите, что если  $\hat{X}_i = g(Z_i)$  – некоторое другое предположение о  $X_i$  на основе  $Z_i$ , то  $\text{var}(\hat{X}_i - X_i) \geq \text{var}(\tilde{X}_i - X_i)$ . (Подсказка: посмотрите упражнение 2.27.)
- b) Покажите, что  $E(w_i | \tilde{X}_i) = 0$ .
- c) Предположим, что  $E(u_i | Z_i) = 0$  и  $\tilde{X}_i$  используется в качестве регрессора вместо  $X_i$ . Покажите, что  $\hat{\beta}_1$  является состоятельной. Является ли  $\hat{\beta}_0$  состоятельной?

9.13. Предположим, что модель регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$  удовлетворяет предположениям из вставки «Основные понятия 4.3» из раздела 4.4. Вы и ваш друг собирают случайную выборку из 300 наблюдений  $Y$  и  $X$ .

- a) Ваш друг сообщает, что он случайно перепутал  $X$ -ы для 20 % выборки. Для этих перепутанных наблюдений значение  $X$  соответствует не значению  $X_i$  для  $i$ -го наблюдения, а значению  $X$  для некоторого другого наблюдения. В обозначениях из раздела 9.2 измеренное значение регрессора  $\tilde{X}_i$  равно  $X_j$  для 80 % наблюдений, но равно случайно выбранным  $X_j$  для оставшихся 20 % наблюдений. Вы оцениваете регрессию  $Y_i$  на  $\tilde{X}_i$ . Покажите, что  $E(\hat{\beta}_1) = 0,8\beta_1$ .
- b) Объясните, как вы можете построить несмешенную оценку  $\beta_1$ , используя МНК-оценку из пункта (a).
- c) Предположим, что ваш друг говорит вам, что  $X$ -ы были перепутаны для первых 60 наблюдений, но оставшиеся 240 наблюдений корректны. Вы оцениваете  $\beta_1$ , оценивая регрессию  $Y$  на  $X$  с использованием только корректно измеренных 240 наблюдений. Будет ли такая оценка  $\beta_1$  лучше, чем оценка, которую вы предложили в пункте (б)? Объясните.

## Компьютерные упражнения

E9.1. Используя базу данных *CPS08*, описанную в Е4.1, ответьте на следующие вопросы:

- a) Обсудите внутреннюю обоснованность регрессий, которые вы использовали для ответа на вопросы в упражнении Е8.1 (a). Обсудите, в том числе, проблемы возможного смещения из-за: пропущенных переменных, неправильно специфицированной функциональной формы регрессии, ошибок в переменных, отбора наблюдений, одновременной причинности и несостоятельности стандартных ошибок МНК.
- b) База данных *CPS92\_08*, описанная в Е3.1, включает данные за 2008 и 1992 годы. Используйте эти данные для исследования (временной) внешней обоснованности выводов, которые вы сделали в Е8.1 (a).

[Замечание: не забудьте скорректировать инфляцию, как объясняется в Е3.1 (б).]

- Е9.2. Комитет по совершенствованию обучения студентов в колледже нуждается в вашей помощи, прежде чем дать соответствующие рекомендации в деканат. Комитету интересно ваше мнение как эконометриста, касающееся необходимости учитывать физические характеристики преподавателей при приеме их на работу. (Это законно до тех пор, пока при подготовке таких рекомендаций не отдается предпочтение определенной расе, религии, возрасту и полу.) У вас нет времени, чтобы собрать свои собственные данные, поэтому ваши рекомендации должны быть основаны на выводах из анализа базы данных *TeachingRatings*, описанной в Е 4.2, которая послужила основой для нескольких эмпирических упражнений в части II книги. Что вы можете посоветовать на основе анализа этих данных? Обоснуйте свой совет, исходя из тщательной и полной оценки внутренней и внешней обоснованности регрессий, которые вы оценили в компьютерных упражнениях в предыдущих главах.
- Е9.3. Используя базу данных *CollegeDistance*, описанную в Е4.3, ответьте на следующие вопросы:
- Обсудите внутреннюю обоснованность регрессий, которые вы использовали для ответа на вопрос упражнения Е8.3 (i). Обсудите, в том числе, проблемы возможного смещения из-за: пропущенных переменных, неправильно специфицированной функциональной формы регрессии, ошибок в переменных, отбора наблюдений, одновременной причинности и несостоительности стандартных ошибок МНК.
  - В базе данных *CollegeDistance* нет данных о студентах из западных штатов; данные по таким студентам включены в базу данных *CollegeDistanceWest*. Используйте эти данные для исследования (географической) внешней обоснованности выводов, которые вы сделали в упражнении Е8.3 (i).

## Приложения

### **Приложение 9.1. Данные по тестам в начальных школах штата Массачусетс**

Данные по штату Массачусетс представляют собой средние значения показателей по округам государственных начальных школ штата в 1998 году. Результаты тестов взяты из Системы всесторонней оценки знаний штата Массачусетс (MCAS)<sup>1</sup>, которые проводились для всех четвероклассников государственных школ штата Массачусетс весной 1998 года. Проведение тестов финансируется Департаментом образования штата Массачусетс и является обязательным для

---

<sup>1</sup> Massachusetts Comprehensive Assessment System (MCAS).

## Часть II. Основы регрессионного анализа

---

всех государственных школ. Данные, анализируемые здесь, являются общей суммой баллов трех тестов: по английскому языку, математике и общенаучного теста.

Данные о соотношении учеников и учителей, проценте учеников, получающих субсидированный обед, и проценте учеников, все еще изучающих английский язык, являются средними для каждого округа начальных школ в 1997–1998 учебном году и были получены в Департаменте образования штата Массачусетс. Данные о средних доходах в округе были взяты из переписи населения США, проведенной в 1990 году.

Часть III

РЕГРЕССИОННЫЙ  
АНАЛИЗ:  
ДОПОЛНИТЕЛЬНЫЕ  
ГЛАВЫ



# **Глава 10. Регрессионный анализ панельных данных**

Множественная регрессия является мощным средством контроля за различными эффектами влияния переменных, которые присутствуют в используемых данных. Однако если какие-либо данные для нескольких переменных отсутствуют, то они не могут быть включены в уравнение регрессии, а МНК-оценки регрессионных коэффициентов могут быть смещены из-за наличия пропущенных переменных.

В данном разделе описан метод, позволяющий учитывать (контролировать) некоторые виды пропущенных переменных, при которых не требуется наличия явных наблюдений. Для использования данного метода необходимо наличие специфического типа данных, называющихся панельными данными, в которых наблюдения за объектом (или некой сущностью) представлены за два или более временных периода. Изучая то, как изменяется зависимая переменная во времени, можно избавиться от эффекта пропущенных переменных, который может быть разным для различных объектов, но является постоянным во времени.

В качестве эмпирических примеров в данном разделе будут рассматриваться приложения, связанные с проблемой вождения в нетрезвом виде, а именно каково влияние налогов на алкогольные напитки и законодательства, связанного с вождением в нетрезвом виде, на дорожно-транспортные происшествия с летальным исходом. Для изучения данного вопроса будут использоваться данные по количеству дорожно-транспортных происшествий с летальным исходом, статистика по налогам на алкогольные напитки, а также данные по законодательству, связанному с вождением в нетрезвом виде, собранные для 48 отдельных штатов США за период с 1982 по 1988 год. Такая выборка, состоящая из панельных данных, позволяет контролировать ненаблюдаемые переменные, которые принимают различные значения для разных штатов, например, различные культурные аспекты отношения к употреблению спиртных напитков и вождению транспортных средств в нетрезвом виде, которые не меняются или слабо меняются со временем. Также с помощью данной выборки можно контролировать переменные, которые изменяются с течением времени, но не различаются между штатами (например улучшение параметров безопасности новых автомобилей).

В разделе 10.1 представлены описание и структура используемой базы данных. Оценка регрессий с фиксированными эффектами – один из основных методов анализа панельных данных – представляет собой расширение

модели множественной регрессии, которое позволяет использовать специфику панельных данных для исследования переменных, которые принимают различные значения для разных объектов, но не изменяются с течением времени. В разделах 10.2 и 10.3 представлено описание регрессии с фиксированными эффектами для двух и более периодов. В разделе 10.4 описанная методология расширена для возможности учета так называемых временных фиксированных эффектов, которые позволяют учесть ненаблюдаемые переменные, не изменяющиеся между различными объектами наблюдения, но изменяющиеся во времени. В разделе 10.5 представлены основные предположения регрессионного анализа панельных данных, а также основные сложности и ошибки, возникающие в рамках данного метода. В разделе 10.6 описанные ранее методы используются для изучения влияния налогов на алкогольную продукцию и законодательства, связанного с вождением в нетрезвом виде, на количество ДТП с летальным исходом.

## ОСНОВНЫЕ ПОНЯТИЯ

### 10.1

#### Обозначения для панельных данных

Панельные данные состоят из наблюдений над одними и теми же  $n$  объектами в течение двух или более временных периодов  $T$ , как показано в таблице 1.3. Если данные содержат наблюдения для переменных  $X$  и  $Y$ , то для панельных данных приняты такие обозначения:

$$(X_{it}, Y_{it}), i = 1, \dots, n, t = 1, \dots, T, \quad (10.1)$$

где первый индекс  $i$  определяет номер объекта наблюдения, а второй индекс  $t$  – номер периода.

## 10.1. Панельные данные

Представленное в разделе 1.3 определение говорит о том, что *панельные данные* (их иногда также называют лонгитюдными данными) представляют собой наблюдения для  $n$  различных объектов в течение  $T$  различных временных периодов. Используемая в данном разделе статистика по ДТП с летальным исходом представляет собой панельные данные. В ней представлены наблюдения для  $n = 48$  объектов (штатов) в течение  $T = 7$  временных периодов (каждый период представляет собой один из годов в промежутке с 1982 по 1988 г.), всего  $7 \times 48 = 336$  наблюдений.

Вспомним, что при описании межъобъектных данных удобно использовать индексы для обозначения номера рассматриваемого объекта. Например,  $Y_i$  представляет собой значение переменной  $Y$  для  $i$ -го объекта (сущности). Для описания панельных данных, в свою очередь, необходимо использовать дополнительную нумерацию для одновременного описания номеров объекта и временного периода: индекс  $i$  определяет номер объекта наблюдения, а второй индекс  $t$  – номер периода наблюдения. Таким

образом,  $Y_{it}$  выражает значение переменной  $Y$  для  $i$ -го объекта в  $t$ -м из  $T$  периодов. Это обозначение представлено во вставке «Основные понятия 10.1».

Ряд дополнительных определений для панельных данных позволяет описывать ситуации, когда некоторые наблюдения пропущены. В *сбалансированной панели* представлены все наблюдения, то есть все переменные имеют наблюдения для каждого объекта в каждом временном периоде. Панель, в которой имеются пропущенные данные для хотя бы одного объекта в одном из периодов, называется *несбалансированной панелью*. Упоминаемая выше статистика по количеству ДТП с летальным исходом содержит наблюдения для всех 48 штатов США в течение всех семи лет. Таким образом, эта выборка представляет собой сбалансированную панель. Тем не менее если бы некоторые данные в этой выборке отсутствовали (например отсутствовали бы данные для одного или нескольких штатов в 1983 г.), то такая панель являлась бы несбалансированной. В данном разделе будут представлены методы, предназначенные для оценки сбалансированных панельных данных. Тем не менее они также могут применяться и для несбалансированных панельных данных, хотя конкретные практические ходы зависят от используемого программного обеспечения.

### **Пример: смертность в ДТП и налоги на алкоголь**

На автомагистралях США за год происходит приблизительно 40 тыс. дорожно-транспортных происшествий с летальным исходом. Около четверти аварий со смертельным исходом происходит при участии нетрезвых водителей, причем эта доля увеличивается в отдельные периоды. В работе Левитта и Портера (Levitt, Porter, 2001) показано, что около 25 % водителей, находившихся за рулем в промежутке с 01:00 до 03:00 часов ночи, пребывали в состоянии алкогольного опьянения, а также то, что вероятность попадания в ДТП со смертельным исходом для водителя, который был признан прибывающим в состоянии алкогольного опьянения, в 13 раз больше, чем для водителя, который не пил.

В этой главе будет исследовано, насколько эффективны различные действия правительства, направленные на предотвращение вождения в нетрезвом виде, в сфере сокращения количества ДТП со смертельным исходом. Выборка, представляющая собой набор панельных данных, содержит переменные, связанные с ДТП и употреблением алкоголя водителями, в том числе количество ДТП с летальным исходом в каждом штате за год, тип законов, ограничивающих вождение в нетрезвом виде в каждом штате за каждый год, и уровень налога на пиво в каждом штате. В качестве меры смертности в ДТП используется уровень смертности, который представляет собой количество ДТП с летальным исходом на 10 тыс. человек населения в штате. В качестве меры налогов на алкоголь используются «реальные» налоги на пиво, которые представляют собой налоги на пиво, рассчитываемые в долларах 1988 года.

с учетом инфляции<sup>1</sup>. Более подробное описание данных представлено в приложении 10.1.

На рисунке 10.1а представлена диаграмма рассеяния для переменных (для 1982 г.) – уровня аварийности и реального налога на пиво. Каждая точка на данной диаграмме представляет собой значения вышеуказанных переменных в 1982 году в одном штате США. С помощью оценки регрессии методом МНК получена регрессионная линия (также изображена на рисунке), уравнение которой может быть записано в таком виде:

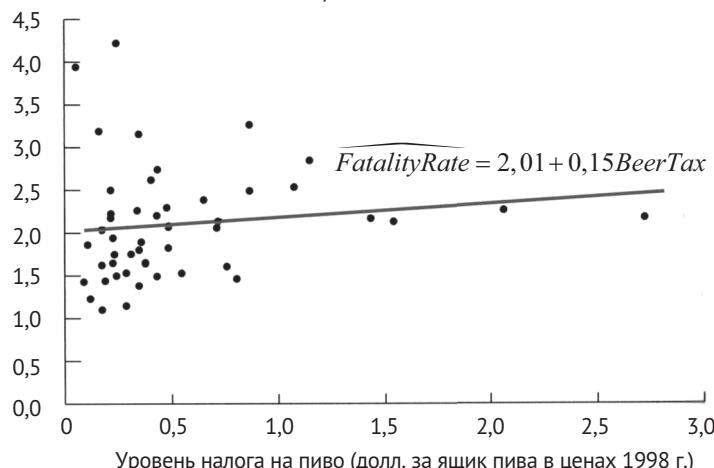
$$\widehat{\text{FatalityRate}} = 2,01 + 0,15 \text{BeerTax}. \quad (10.2)$$

Оценка коэффициента перед показателем реального налога на пиво является положительной, но не является статистически значимой на уровне значимости 10 %.

Поскольку в наличии имеются данные более чем за один год, то возможно провести дополнительную оценку этого выражения на данных другого года. Результаты такой оценки (рис. 10.1б) представляют собой аналогичную диаграмму рассеяния, но построенную уже на данных за 1988 год. Уравнение регрессионной линии (МНК-оценка) для этих данных выглядит следующим образом:

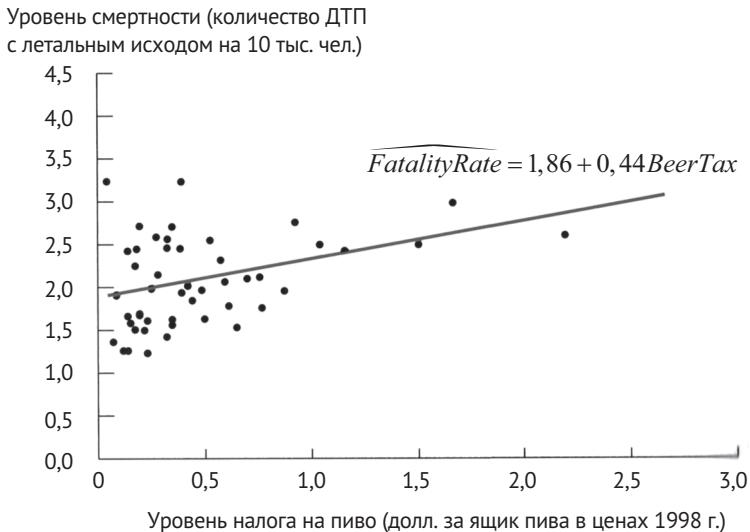
$$\widehat{\text{FatalityRate}} = 1,86 + 0,44 \text{BeerTax}. \quad (10.3)$$

Уровень смертности (количество ДТП с летальным исходом на 10 тыс. чел.)



(a) Данные 1982 года

<sup>1</sup> Для того чтобы налоги были сопоставимы с течением времени, они пересчитаны в ценах 1988 года (в долларах) с помощью индекса потребительских цен (ИПЦ). Например, из-за инфляции налог 1 долл. в 1982 году соответствует налогу 1,23 долл. в ценах 1988 года.



(б) Данные 1988 года

**Рисунок 10.1. Уровень смертности в ДТП и налоги на пиво**

На диаграмме (а) представлена диаграмма рассеяния для переменных уровня смертности и реального налога на пиво (в ценах 1988 г.) в 48 штатах США в 1982 году. На диаграмме (б) представлена диаграмма рассеяния для переменных уровня смертности и реального налога на пиво (в ценах 1988 г.) в 48 штатах США в 1988 году. Обе диаграммы показывают положительную зависимость между уровнем смертности и уровнем налога на пиво

В отличие от полученных результатов для данных 1982 года, оценка коэффициента при уровне реального налога на пиво является статистически значимой на уровне значимости 1 % ( $t$ -статистика равна 3,43). Интересным является тот факт, что оценки коэффициентов на данных 1982 года и на данных 1988 года являются положительными. То есть повышение уровня налогов на пиво приводит к увеличению, а не снижению уровня смертности в ДТП.

Является ли вывод о том, что повышение уровня налогов на пиво оказывается причиной роста уровня смертности в ДТП, однозначным? Не обязательно, потому что полученные оценки могут быть существенно смешены из-за наличия пропущенных переменных. На уровень смертности в ДТП может оказывать влияние множество других факторов, в том числе качество автомобилей, которые используются для передвижения по автодорогам того или иного штата, состояние автомобильных дорог, распределение транспортных потоков (наибольшая плотность в сельской или городской местности), плотность движения на отдельных участках автомобильных дорог, уровень социальной неприязни по отношению к вождению в нетрезвом виде. Любой из этих факторов может быть связан с уровнем налогов на алкоголь, и если такие зависимости имеют место, то могут приводить к смешению получаемых оценок. Одним из подходов, который может применяться в таких случаях (при наличии потенциальных источников смешения из-за пропущенных переменных), может быть сбор данных по всем этим переменным и добавление в общую годовую выборку

межобъектных данных с последующей оценкой выражений (10.2) и (10.3). К сожалению, некоторые из этих переменных, такие как культурный уровень не-приятия вождения в нетрезвом состоянии, достаточно сложно или даже невозможно измерить.

Если эти факторы остаются неизменными во времени в том или ином штате, то возможно применение другого способа. С помощью использования панельных данных можно учесть эти факторы как некие константы, хотя и без возможности их точного измерения. Для этого используются МНК-оценки регрессии с фиксированными эффектами.

## 10.2. Панельные данные с наличием двух периодов: сравнения «до и после»

При наличии данных для каждого штата для  $T = 2$  временных периодов оказывается возможным сравнить значения зависимой переменной во втором периоде с ее значениями в первом периоде. С помощью такого анализа (сравнение «до и после» или «разности»), который концентрируется на изменениях значений зависимой переменной, можно выделить эффекты (ненаблюдаемые переменные), которые принимают различные значения для разных штатов, но для каждого отдельного штата не изменяются во времени.

Пусть  $Z_i$  представляет собой переменную, которая в некоторой степени определяет уровень смертности в ДТП в  $i$ -м штате, но не изменяется во времени (поэтому индекс  $t$  опущен). Например,  $Z_i$  может представлять собой уровень культурной или социальной неприязни к употреблению спиртных напитков за рулем или вождению в нетрезвом виде, который изменяется во времени достаточно медленно и, таким образом, может рассматриваться как постоянный в рамках временного периода с 1982 по 1988 год. Тогда уравнение линейной регрессии, связывающее уровень смертности в ДТП,  $Z_i$  и уровень налогов на пиво, может быть записано в следующем виде:

$$\text{FatalityRate}_{it} = \beta_0 + \beta_1 \text{BeerTax}_{it} + \beta_2 Z_i + u_{it}, \quad (10.4)$$

где  $u_{it}$  – ошибки,  $i = 1, \dots, n$  и  $t = 1, \dots, T$ .

Поскольку в уравнении регрессии (10.4) переменная  $Z_i$  не изменяется с течением времени, то она не будет приводить к изменению уровня смертности в ДТП в период с 1982 по 1988 год. Таким образом, влияние переменной  $Z_i$  может быть устранено с помощью анализа изменения уровня смертности между двумя вышеуказанными периодами. Для того чтобы проследить это формально (математически), необходимо рассмотреть уравнение (10.4) для каждого из двух периодов (1982 и 1988 гг.):

$$\text{FatalityRate}_{i1982} = \beta_0 + \beta_1 \text{BeerTax}_{i1982} + \beta_2 Z_i + u_{i1982}, \quad (10.5)$$

$$\text{FatalityRate}_{i1988} = \beta_0 + \beta_1 \text{BeerTax}_{i1988} + \beta_2 Z_i + u_{i1988}. \quad (10.6)$$

Вычитая уравнение (10.5) из уравнения (10.6), как было указано выше, можно устранить влияние переменной  $Z_i$ :

$$\begin{aligned} \text{FatalityRate}_{i1988} - \text{FatalityRate}_{i1982} &= \\ &= \beta_1 (\text{BeerTax}_{i1988} - \text{BeerTax}_{i1982}) + u_{i1988} - u_{i1982}. \end{aligned} \quad (10.7)$$

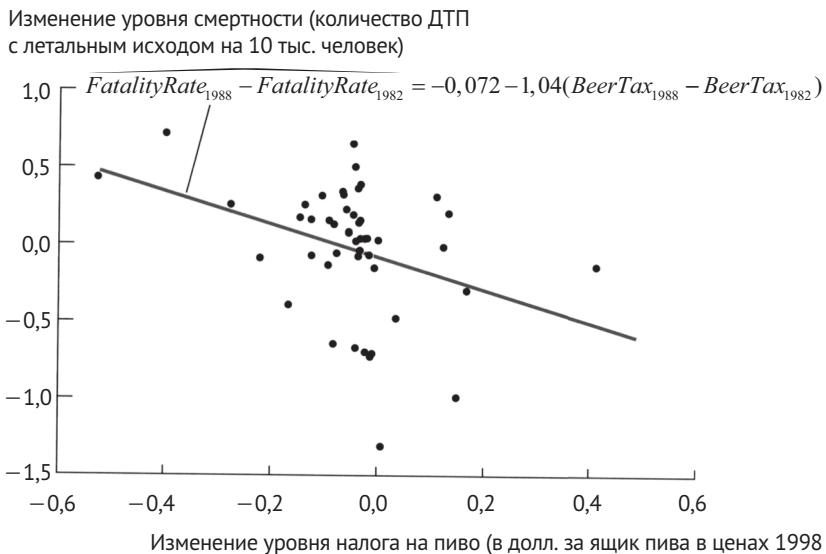
Данная спецификация имеет интуитивно понятную интерпретацию. Культурное или социальное отношение к вождению в нетрезвом состоянии оказывает влияние на количество водителей, находящихся за рулем в состоянии алкогольного опьянения и, следовательно, на уровень смертности в ДТП в каждом отдельном штате. Если, однако, значения этой переменной не изменились в период между 1982 и 1988 годами, то она не оказывает никакого влияния на изменение уровня смертности в ДТП в том или ином штате. Скорее всего, изменение уровня смертности в ДТП с течением времени должно было возникнуть из-за других причин. В формуле (10.7) эти другие источники изменений представляют собой колебания уровня налога на пиво и изменения в остаточном члене (отражающем изменения в других факторах, которые могут оказывать влияние на уровень смертности в ДТП).

Записывая уравнение регрессии в разностях в виде (10.7), можно избавиться от эффекта ненаблюдаемых переменных  $Z_i$ , которые являются неизменными во времени. Другими словами, анализируя изменения  $Y$  и  $X$ , можно учесть переменные, которые являются постоянными во времени, и, таким образом, устранить данный источник смещения в получаемых оценках.

На рисунке 10.2 представлена диаграмма рассеяния для изменения уровня смертности в ДТП в период с 1982 по 1988 год и изменения уровня налогов на пиво в период с 1982 по 1988 год для 48 штатов США, имеющихся в выборке. Каждая точка на рисунке 10.2 представляет собой изменение уровня смертности и изменение уровня реальных налогов на пиво в период с 1982 по 1988 год в одном штате. Уравнение регрессионной линии, полученной с помощью МНК-оценки и изображенной на данной диаграмме, может быть записано в таком виде:

$$\begin{aligned} \text{FatalityRate}_{i1988} - \text{FatalityRate}_{i1982} &= \\ &= -0,072 - 1,04 \times (\text{BeerTax}_{i1988} - \text{BeerTax}_{i1982}). \end{aligned} \quad (10.8)$$

Добавление константы в уравнение (10.8) оставляет возможность для изменения среднего значения уровня смертности в ДТП, которое может иметь место в отсутствии изменений в уровне реального налога на пиво. Например, полученная отрицательная оценка константы (-0,072) может отражать повышение уровня безопасности в используемых автомобилях в период с 1982 по 1988 год, которое позволило снизить средний уровень смертности в ДТП в вышеуказанный период.



**Рисунок 10.2. Изменение уровня смертности в ДТП и налога на пиво в 1982–1988 годах**

На рисунке представлена диаграмма рассеяния для изменения уровня смертности и изменения реального налога на пиво (в ценах 1988 г.) в 48 штатах США в период с 1982 по 1988 год. Полученные результаты говорят о наличии отрицательной зависимости между изменениями уровня смертности в ДТП и изменениями налога на пиво.

В отличие от оценок, полученных на основе межъектных данных, эти результаты показывают наличие отрицательного влияния изменений в уровне налога на пиво на уровень смертности в ДТП, как и следует из экономической теории. Гипотеза о том, что коэффициент наклона в уравнении регрессии равен нулю, отвергается на 5 %-м уровне значимости. В соответствии с полученной оценкой коэффициента увеличение налога на пиво на 1 долл. снижает уровень смертности в ДТП на 1,04 погибших на 10 тыс. человек. Величина полученного эффекта очень велика: средний уровень смертности в ДТП в используемой для получения оценок выборке составляет приблизительно два погибших на 10 тыс. человек в год. Таким образом, полученные оценки показывают, что количество ДТП с летальным исходом может быть сокращено в 2 раза только за счет увеличения реальных налогов на пиво на 1 долл.

Изучая изменение уровня смертности в ДТП с течением времени с помощью регрессии, описываемой уравнением (10.8), можно учесть такие неизменные факторы, как культурные различия в отношении к вождению в нетрезвом состоянии. Но есть много иных факторов, влияющих на безопасность дорожного движения, и если они изменяются с течением времени и коррелированы с уровнем реального налога на пиво, то их отсутствие будет приводить к смещению в получаемых оценках. В разделе 10.5 будет проведен более тщательный анализ, который позволит учитывать несколько таких факторов, поэтому пока лучше воздержаться от каких-либо существенных выводов о возможном влиянии налога на пиво на количество ДТП с летальным исходом.

Описанный выше анализ вида «сравнение до и после» может применяться, когда наблюдаемые данные содержат два периода. Но стоит отметить, что ис-

пользуемая в данном разделе выборка содержит данные для семи различных лет, и было бы неосмотрительно отбрасывать эти потенциально полезные данные. Однако метод «сравнения до и после» не применяется в явном виде при  $T > 2$ . Для анализа всех имеющихся в выборке данных будет использоваться регрессия с фиксированными эффектами.

### 10.3. Регрессия с фиксированными эффектами

Оценка регрессии с фиксированными эффектами является методом, позволяющим учесть влияние пропущенных в уравнении регрессии переменных, которые различаются между объектами наблюдения (штатами), но не изменяются с течением времени. В отличие от подхода «сравнение до и после», описанного в разделе 10.2, регрессия с фиксированными эффектами может использоваться в случае наличия двух и более временных периодов.

В регрессионной модели с фиксированными эффектами используются  $n$  различных констант по одной для каждого объекта наблюдения. Эти константы могут быть представлены в виде набора бинарных (или индикаторных) переменных, в которых будет в той или иной степени учтено влияние всех пропущенных в регрессионном уравнении переменных, которые различны для разных объектов наблюдения, но неизменны во времени.

#### *Модель регрессии с фиксированными эффектами*

Рассмотрим регрессионную модель, описанную выражением (10.4), где зависимая переменная (*FatalityRate*) и наблюдаемая независимая переменная (*BeerTax*) обозначены как  $Y_{it}$  и  $X_{it}$ , соответственно:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \quad (10.9)$$

где  $Z_i$  представляет собой ненаблюданную переменную, которая принимает различные значения для отдельных штатов США, но не изменяется с течением времени (например,  $Z_i$  может описывать некие культурные или социальные эффекты, которые могут оказывать влияние на употребление спиртных напитков за рулем). Как было описано выше, интерес представляет получение оценки коэффициента  $\beta_1$ , описывающего влияние  $X$  на  $Y$ , при фиксированной ненаблюданной переменной  $Z$ .

Поскольку  $Z_i$  принимает различные значения для разных штатов, но не изменяется с течением времени, то, сгруппировав слагаемые в виде  $\alpha_i = \beta_0 + \beta_2 Z_i$ , можно интерпретировать получившееся выражение, как то что оно имеет  $n$  различных для каждого отдельного штата констант. Тогда выражение (10.9) можно переписать в таком виде:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}. \quad (10.10)$$

Выражение (10.10) описывает модель регрессии с фиксированными эффектами, где  $\alpha_1, \dots, \alpha_n$  рассматриваются как неизвестные константы, которые необходимо оценить (одна для каждого штата). Интерпретация данных констант может быть получена при рассмотрении правой части уравнения (10.10). Коэффициент

наклона в полученной оценке регрессионной линии  $\beta_1$  одинаков для всех штатов. Константы же варьируются в зависимости от штата.

Поскольку  $\alpha_i$  в выражении (10.10) может быть рассмотрено как некое «специфическое свойство»  $i$ -го объекта наблюдения (в примере, приведенном в данном разделе, объектами наблюдения являются штаты), константы  $\alpha_1, \dots, \alpha_n$  иногда отождествляют с фиксированными эффектами. Изменения этих констант («фиксированных эффектов») могут быть вызваны изменением пропущенных переменных, например  $Z_i$  в уравнении (10.9).

Описанные выше константы, специфичные для объектов наблюдения, могут также моделироваться с помощью бинарных переменных, используемых для обозначения каждого отдельного объекта наблюдения (в данном разделе – штата). В разделе 8.3 был рассмотрен случай, когда наблюдения относились к одной из двух групп, а полученная оценка регрессионной линии имела одинаковый наклон в обеих группах, но разные свободные члены (рис. 8.8а). В рассмотренном примере уравнение регрессионной линии математически было записано с использованием бинарной переменной, являющейся индикатором одной из двух групп (случай № 1 во вставке «Основные понятия 8.4»). Если бы выборка, используемая в текущем разделе, включала данные лишь по двум штатам, то описанный выше метод с использованием бинарных переменных мог бы быть применен и сейчас. Однако поскольку выборка содержит данные по большему числу штатов, то необходимо использовать дополнительные бинарные переменные, для того чтобы учесть все возможные особенности штатов («фиксированные эффекты», описанные в выражении (10.10)).

Для построения регрессионной модели с фиксированными эффектами, которая использовала бы бинарные переменные, предположим, что  $D1_i$  – бинарная переменная, которая равна единице в случае  $i=1$  и равна нулю в ином случае,  $D2_i$  – бинарная переменная, которая равна единице в случае  $i=2$  и равна нулю в ином случае, и так далее. Однако включение в уравнение регрессии всех  $n$  бинарных переменных одновременно с общим свободным членом привело бы к появлению совершенной мультиколлинеарности (данный случай «ловушки фиктивных переменных» был рассмотрен в разделе 6.7). Поэтому необходимо произвольно опустить одну из бинарных переменных, например, для первой группы наблюдений –  $D1_i$ . Тогда уравнение регрессии с фиксированными эффектами, эквивалентное (10.10), может быть записано в таком виде:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \dots + \gamma_n Dn_i + u_{it}, \quad (10.11)$$

где  $\beta_0, \beta_1, \gamma_2, \dots, \gamma_n$  – неизвестные коэффициенты, которые необходимо оценить. Для получения соотношения между коэффициентами в уравнении (10.11) и коэффициентами в уравнении (10.10) необходимо сравнить правые части этих выражений. В уравнении (10.10) выражение для линии регрессии для первого штата можно записать как  $\beta_0 + \beta_1 X_{it}$ , тогда  $\alpha_1 = \beta_0$ . Для второго и последующих штатов оно принимает вид:  $\beta_0 + \beta_1 X_{it} + \gamma_i$ , следовательно,  $\alpha_i = \beta_0 + \gamma_i$  для  $i \geq 2$ .

Таким образом, для построения модели регрессии с фиксированными эффектами можно использовать два эквивалентных способа, которые представле-

ны в виде уравнений (10.10) и (10.11). В уравнении (10.10) модель записана с помощью  $n$  различных констант, величины которых зависят от особенностей того или иного штата. В уравнении (10.11) модель построена с помощью одной общей константы и  $n-1$  бинарной переменной. В обеих формулировках угловые коэффициенты при  $X$  принимают одни и те же значения для всех штатов. Стоит отметить, что специфические константы в уравнении (10.10) и бинарные переменные в уравнении (10.11) имеют по сути один и тот же источник – ненаблюдаемую переменную  $Z_i$ , которая принимает различные значения для разных штатов, но не изменяется с течением времени.

**Расширение модели до нескольких переменных  $X$ .** Если существуют иные наблюдаемые факторы, которые влияют на  $Y$ , коррелированы с  $X$  и изменяются с течением времени, то такие факторы должны быть включены в уравнение регрессии для того, чтобы избежать смещения, вызванного пропущенными переменными. Результат добавления дополнительных переменных в модель регрессии с фиксированными эффектами представлен во вставке «Основные понятия 10.2».

### Модель регрессии с фиксированными эффектами

Модель регрессии с фиксированными эффектами может быть представлена в таком виде:

$$Y_{it} = \beta_1 X_{1,it} + \dots + \beta_k X_{k,it} + \alpha_i + u_{it}, \quad (10.12)$$

где  $i = 1, \dots, n; t = 1, \dots, T$ ,  $X_{1,it}$  – значение первого регрессора для объекта  $i$  в период  $t$ ,  $X_{2,it}$  – значение второго регрессора для объекта  $i$  в период  $t$  и так далее;  $\alpha_1, \dots, \alpha_n$  – константы, зависящие от объекта наблюдения.

Модель может быть эквивалентно записана с помощью общей константы, независимых переменных  $X$  и  $n-1$  бинарных переменных:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \dots + \beta_k X_{k,it} + \gamma_2 D2_i + \gamma_3 D3_i + \dots + \gamma_n Dn_i + u_{it}, \quad (10.13)$$

где  $D2_i$  – бинарная переменная, которая равна единице в случае  $i = 2$  и равна нулю в ином случае и так далее.

## ОСНОВНЫЕ ПОНЯТИЯ

10.2

### Оценка модели и статистические выводы

В принципе оценка спецификации регрессионной модели с фиксированными эффектами, построенная на основе бинарных переменных, может быть оценена с помощью метода наименьших квадратов (далее МНК). Такая регрессия будет иметь  $k+n$  регрессоров (среди них  $k$  переменных  $X$ ,  $n-1$  бинарных переменных и общая константа). Таким образом, на практике такую МНК-регрессию оценить достаточно сложно, а в некоторых статистических пакетах

и вовсе невозможно, если количество объектов достаточно велико. Эконометрические статистические пакеты имеют специальные алгоритмы, которые применяются для построения МНК-оценок моделей с фиксированными эффектами. Эти специальные алгоритмы эквивалентны использованию МНК для оценки регрессии с бинарными переменными, но работают несколько быстрее за счет того, что используют упрощающие предположения, которые возникают в регрессиях с фиксированными эффектами.

**МНК-алгоритм для «центрированных на внутриобъектное среднее»<sup>1</sup>.** Эконометрические программные пакеты в большинстве случаев получают оценки моделей с фиксированными эффектами с помощью МНК в два шага. На первом шаге для каждой переменной вычисляется среднее для каждого объекта значение, а затем это среднее вычитается из значений для каждой переменной. На втором шаге проводится оценка уравнения регрессии, но уже с использованием скорректированных на величину среднего значения переменных. Рассмотрим, например, случай одного регрессора в модели с фиксированными эффектами, описываемой уравнением (10.10), тогда  $\bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_i$ , где  $\bar{Y}_i = (1/T) \sum_{t=1}^T Y_{it}$ ;  $\bar{X}_i$  и  $\bar{u}_i$  определяются аналогично. Тогда после вычитания вышеуказанного выражения уравнение (10.10) может быть переписано в виде:  $Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)$ . Пусть  $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$ ;  $\tilde{X}_{it} = X_{it} - \bar{X}_i$ ;  $\tilde{u}_{it} = u_{it} - \bar{u}_i$ . Тогда выражение

можно переписать так:

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}. \quad (10.14)$$

Таким образом, оценка коэффициента  $\beta_1$  может быть получена с помощью МНК при оценке влияния  $\tilde{X}_{it}$  на  $\tilde{Y}_{it}$  в выражении (10.14). Фактически эта оценка равносильна МНК-оценке  $\beta_1$ , полученной при оценке модели с фиксированными эффектами.

**Регрессия в разностях («до и после») и спецификация с бинарными переменными.** Несмотря на то что уравнение (10.11), в котором используются бинарные переменные, по виду отличается от регрессии в разностях («до и после») в уравнении (10.7), в случае двух периодов МНК-оценка  $\beta_1$  в рамках этих двух спецификаций не будет отличаться, если в регрессии в разностях не будет учитываться свободный член. Таким образом, в случае двух периодов  $T = 2$  существуют три способа получения МНК-оценки коэффициента  $\beta_1$ : оценка уравнения (10.7) (метод «до и после», описанный ранее) без константы, спецификация с бинарными переменными в уравнении (10.11) и центрированная на внутриобъектное среднее спецификация, представленная уравнением (10.14). Указанные три метода являются эквивалентными и дают одинаковые МНК-оценки коэффициента  $\beta_1$  (см. упражнение 10.11).

---

<sup>1</sup> Такую оценку часто называют внутригрупповой оценкой или просто внутри-оценкой. – Примеч. науч. ред. перевода.

**Выборочное распределение, стандартные ошибки и статистические выводы.**

В случае множественной регрессии для межобъектных данных, если выполняются четыре основных предположения из вставки «Основные понятия 6.4», то выборочное распределение МНК-оценки является нормальным. Дисперсия выборочного распределения может быть оценена с помощью имеющихся данных. Квадратный корень, извлеченный из полученной оценки дисперсии, будет являться стандартной ошибкой и может использоваться для тестирования гипотез с помощью  $t$ -статистики и построения доверительных интервалов на ее основе.

Аналогично, в случае оценки регрессии на панельных данных, если выполнен ряд предположений, которые иногда называют «предположениями модели регрессии с фиксированными эффектами», то выборочное распределение полученных МНК-оценок в модели с фиксированными эффектами является нормальным для больших выборок, а дисперсия может быть оценена с помощью имеющейся выборки; корень из дисперсии будет давать стандартную ошибку, а стандартная ошибка может быть использована для построения  $t$ -статистик и доверительных интервалов. При заданной стандартной ошибке тестирование гипотез (включая тестирование гипотез с помощью  $F$ -статистики) и построение доверительных интервалов проводится полностью аналогично случаю множественной регрессии, оцениваемой на межобъектных данных.

Предположения модели регрессии с фиксированными эффектами и стандартные ошибки для данной модели будут описаны далее в разделе 10.5.

**Пример: модель числа летальных исходов в ДТП**

МНК-оценка модели регрессии с фиксированными эффектами, описывающей взаимосвязь между реальным налогом на пиво и количеством ДТП с летальным исходом, полученная на основе данных за 7 лет (336 наблюдений), может быть записана в таком виде:

$$\widehat{\text{FatalityRate}} = -0,66 \text{ BeerTax} + \text{StateFixedEffects}, \quad (10.15)$$

где оценки констант, отвечающих за фиксированные эффекты, опущены в целях экономии места, поскольку они не представляют интереса в рамках данного примера.

Как и в спецификации в разностях, представленной в уравнении (10.8), полученная оценка углового коэффициента в уравнении (10.15) является отрицательной, то есть, как и было предсказано экономической теорией, повышение налогов на пиво приводит к снижению количества ДТП с летальным исходом, что прямо противоположно тому, что было получено в исходных уравнениях (10.2) и (10.3), оценка которых производилась на межобъектных данных. Однако стоит заметить, что эти вышеупомянутые регрессии неидентичны, поскольку для регрессии «в разностях» в уравнении (10.8) используются данные лишь за 1982 и 1988 годы (в частности, используется разность между указанными годами). В то же время уравнение (10.15) регрессии с фиксированными эффектами использует данные за все семь лет. Как раз из-за использования дополнительных данных стандартные ошибки в уравнении (10.15) ниже, чем в уравнении (10.8).

Включение фиксированных эффектов в регрессию для количества ДТП с летальным исходом позволяет исключить смещение, вызванное пропущенными переменными, такими как различные культурные аспекты отношений к употреблению спиртных напитков и вождению транспортных средств в нетрезвом виде, которые могут различаться для штатов, но оставаться постоянными во времени. Тем не менее могут оставаться подозрения о том, что существуют неучтенные переменные, пропуск которых может приводить к смещению оценок. Например, в течение вышеуказанного периода машины постепенно становились более безопасными, а водители стали чаще использовать ремни безопасности. Если реальный налог на пиво в среднем рос в течение 1980-х годов, то *BeerTax* может также содержать в себе эффект роста общего уровня автомобильной безопасности. Но если уровень безопасности повышался с течением времени равномерно для всех штатов, то его влияние можно устранить с помощью оценки регрессии с фиксированными эффектами.

## 10.4. Модель регрессии с фиксированными временными эффектами

Модель с фиксированными эффектами можно использовать не только для учета эффектов, которые меняются для различных объектов, но постоянны во времени. Ее можно также использовать для учета переменных, которые принимают одинаковые значения для различных объектов, но изменяются во времени.

Например, поскольку улучшения в новых машинах появляются на общенациональном уровне, то они позволяют снизить количество ДТП с летальным исходом во всех штатах. Таким образом, можно рассматривать общий уровень автомобильной безопасности как одну из пропущенных переменных, которые изменяются во времени, но принимают одинаковые значения для всех штатов. Уравнение регрессии (10.9) может быть модифицировано для учета эффекта изменения уровня автомобильной безопасности в явном виде (обозначим его как  $S_t$ ):

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}, \quad (10.16)$$

где  $S_t$  – ненаблюдаемая переменная, а индекс  $t$  отражает то, что уровень безопасности изменяется с течением времени, но одинаков для всех штатов. Поскольку  $\beta_3 S_t$  отражает переменные, которые определяют  $Y_{it}$ , если  $S_t$  коррелирована с  $X_{it}$ , то отсутствие  $S_t$  в уравнении регрессии может привести к смещению, вызванному пропущенными переменными.

### **Только временные эффекты**

Представим на некоторое время, что переменные  $Z_i$  отсутствуют, то есть  $\beta_2 Z_i$  в уравнении (10.16) можно опустить. Слагаемое  $\beta_3 S_t$  по-прежнему сохраняется. Необходимо оценить  $\beta_1$ , при этом учитывая тем или иным образом влияние  $S_t$ .

Несмотря на то что  $S_t$  является ненаблюдаемой переменной, существует возможность устраниить эффект влияния переменной  $Z_i$ , которая меняется между штатами, но не изменяется во времени. В модели с фиксированными эффектами

для объектов наличие  $Z_i$  приводит к модели с фиксированными эффектами, представленной в уравнении (10.10), в которой каждому штату соответствует отдельная константа (или фиксированный эффект). Аналогично, поскольку  $S_t$  изменяется с течением времени, а не между штатами, то наличие  $S_t$  в уравнении регрессии приводит к тому, что необходимо ввести в уравнение регрессии константы, каждая из которых соответствует одному временному периоду.

Тогда модель с фиксированными временными эффектами с одним регрессором  $X$  может быть записана следующим образом:

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}, \quad (10.17)$$

В этой модели каждому временному периоду соответствует отдельная константа  $\lambda_t$ . Эту константу  $\lambda_t$  в рамках уравнения (10.17) можно рассматривать как отражающую некий эффект влияния года  $t$  на  $Y$  (в общем случае временного периода  $t$ ). Таким образом, слагаемые  $\lambda_1, \dots, \lambda_T$  представляют собой *временные фиксированные эффекты*. Изменения во временных фиксированных эффектах зачастую вызваны пропущенными переменными, как, например,  $S_t$  в уравнении (10.16), которые изменяются во времени, но не меняются между объектами.

Модель с индивидуальными фиксированными эффектами может быть представлена с помощью  $n-1$  бинарных переменных. Аналогично модель с временными фиксированными эффектами может быть записана с помощью  $T-1$  бинарных переменных:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \dots + \delta_T BT_t + u_{it}, \quad (10.18)$$

где  $\delta_2, \dots, \delta_T$  являются неизвестными коэффициентами,  $B2_t = 1$ , если  $t = 2$ , и  $B2_t = 0$  в ином случае и так далее. Как и в регрессии с фиксированными индивидуальными эффектами, представленной уравнением (10.11), в вышеуказанной модели с фиксированными временными эффектами присутствует свободный член, но исключена первая бинарная переменная  $B1$ , в целях предотвращения возникновения совершенной мультиколлинеарности.

При необходимости дополнительные наблюдаемые регрессоры « $X$ » добавляются в уравнение (10.17) и таким же образом в уравнение (10.18).

Построенная спецификация для модели, объясняющей количество ДТП с летальным исходом, позволяет избежать возникновения смещений в оценках, вызванных наличием пропущенных переменных, таких как общенациональное внедрение стандартов автомобильной безопасности, которые могут изменяться с течением времени, но одинаковы для всех штатов.

### **Модель с индивидуальными и временными фиксированными эффектами**

Если часть пропущенных переменных является постоянной во времени, но меняется между штатами (например культурные нормы), а другая часть принимает одинаковые значения для различных штатов, но изменяется с течением времени (например общенациональные стандарты автомобильной безопасности), то оптимальным является использование модели с включением

одновременно индивидуальных (по объектам наблюдения) и временных фиксированных эффектов.

Модель с индивидуальными и временными фиксированными эффектами можно записать в таком виде:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}, \quad (10.19)$$

где  $\alpha_i$  представляют собой индивидуальные фиксированные эффекты,  $\lambda_t$  – временные фиксированные эффекты. Эта модель может быть эквивалентным образом записана с использованием  $n-1$  бинарной переменной, соответствующей объектам, и  $T-1$  бинарной переменной, соответствующей временными периодам, и одного общего свободного члена:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \dots + \gamma_n Dn_i + \delta_2 B2_t + \dots + \delta_T BT_t + u_{it}, \quad (10.20)$$

где  $\beta_0, \beta_1, \gamma_2, \dots, \gamma_n$  и  $\delta_2, \dots, \delta_T$  – неизвестные коэффициенты.

При наличии дополнительных наблюдаемых переменных « $X$ », они добавляются одновременно в уравнения (10.19) и (10.20).

Модель, совмещающая в себе индивидуальные и временные фиксированные эффекты, позволяет избежать возникновения смещений в оценках, вызванных пропущенными переменными, которые могут быть постоянны как во времени, так и для различных объектов наблюдения (штатов).

**Оценка.** Вышеописанные спецификации модели с фиксированными эффектами, включающие только временные фиксированные эффекты, либо одновременно индивидуальные и временные фиксированные эффекты, являются всего лишь вариациями стандартной модели множественной регрессии. Таким образом, коэффициенты в рамках данных моделей могут быть оценены с помощью МНК при добавлении дополнительных бинарных переменных, соответствующих временным периодам. С другой стороны, при использовании сбалансированной панели коэффициенты при  $X$  могут быть оценены с помощью вычитания из  $Y$  и  $X$  их средних по объектам наблюдения и по временным периодам значений с последующей оценкой регрессии для  $Y$  и  $X$  в отклонениях от среднего. Данный алгоритм, достаточно часто применяющийся в эконометрических программных пакетах, устраниет необходимость использования полного набора бинарных переменных, которые представлены в уравнении (10.20). В рамках альтернативного подхода можно вычислить отклонения от среднего по объектам наблюдения (но не по времени) для  $Y$ ,  $X$  и констант, соответствующих временным периодам, а затем оценить  $k+T$  коэффициентов в рамках множественной регрессии в отклонениях от среднего для  $Y$  на  $X$  и ряда временных индикаторных (бинарных) переменных. Наконец, если  $T = 2$ , модель с фиксированными индивидуальными и временными эффектами может быть оценена с помощью подхода «сравнение до и после», который был описан в разделе 10.2, при добавлении свободного члена в регрессию. Таким образом, в модели, представленной уравнением (10.8), в которой проводится оценка влияния изменений *BeerTax* на изменения *FatalityRate* в период с 1982 по 1988 год в рамках подхода «сравнение до и после», были получены те же оценки угловых коэффициентов, что и при оценивании с помощью МНК-модели регрессии с фиксированными

индивидуальными и временными эффектами для влияния изменений *BeerTax* на изменения *FatalityRate* только лишь для 1982 и 1988 годов (т.е. двух лет).

**Пример: модель числа летальных исходов в ДТП.** Включение временных и индивидуальных фиксированных эффектов в модель регрессии приводит к следующим оценкам МНК:

$$\widehat{\text{FatalityRate}} = -0,64 \text{ BeerTax} + \text{StateFixedEffects} + \text{TimeFixedEffects}. \quad (10.21)$$

(0,36)

Эта спецификация включает в себя налоги на пиво, 47 бинарных переменных (фиксированные эффекты, соответствующие штатам), 6 бинарных переменных (временные фиксированные эффекты, соответствующие тому или иному году) и свободный член, что в сумме дает  $1+47+6+1=55$  переменных в правой части регрессии. Оценки коэффициентов, отвечающих за временные и индивидуальные фиксированные эффекты, а также оценка свободного члена не представлены, поскольку они не представляют первостепенного интереса.

Включение временных фиксированных эффектов не дает значительного улучшения результатов оценок для коэффициента при переменной, отражающей уровень реальных налогов на пиво (сравните уравнения (10.15) и (10.21)). Несмотря на то что оценка коэффициента становится менее точной при добавлении временных фиксированных эффектов, коэффициент по-прежнему остается значимым, но уже на уровне значимости 10 %, а не на уровне значимости 5 % ( $t = -0,64 / 0,36 = -1,78$ )

Эта оценка соотношения между реальным налогом на пиво и количеством ДТП с летальным исходом застрахована от возникновения смещений в оценках коэффициентов, вызванных наличием пропущенных переменных, принимающих постоянные во времени или между штатами значения. Тем не менее многие важные детерминанты количества ДТП с летальным исходом не попадают в эту категорию, так что эта спецификация еще может быть подвержена смещению, вызванному пропущенными переменными. Поэтому в разделе 10.6 будет представлено более полное эмпирическое исследование влияния налогов на пиво, законов, направленных непосредственно на устранение вождения в нетрезвом виде, а также многих других факторов. Перед тем как обратиться к этому исследованию, необходимо обсудить предположения, лежащие в основе регрессии панельных данных, а также построение стандартных ошибок для оценок фиксированных эффектов.

## 10.5. Предположения модели регрессии с фиксированными эффектами и стандартные ошибки модели регрессии с фиксированными эффектами

В панельных данных ошибка регрессии может быть коррелирована во времени для каждого объекта. Аналогично случаю гетероскедастичности ошибок эта корреляция не приводит к смещению в оценках модели регрессии

с фиксированными эффектами, но она влияет на дисперсию оценки фиксированных эффектов и, следовательно, на стандартные ошибки. Стандартные ошибки в регрессии с фиксированными эффектами, которые приводятся в этой главе, являются так называемыми кластеризованными стандартными ошибками и устойчивы как к наличию гетероскедастичности, так и к корреляции наблюдений по объекту во времени. Если в выборке много объектов (т.е.  $n$  велико), то тестирование гипотез и построение доверительных интервалов может быть осуществлено с использованием критических значений для асимптотического нормального и  $F$ -распределений.

В данном разделе описываются кластеризованные стандартные ошибки. Сначала мы рассматриваем предположения регрессии с фиксированными эффектами, которые являются расширением предположений метода наименьших квадратов на случай панельных данных; при выполнении этих предположений оценка фиксированных эффектов является асимптотически нормально распределенной при больших  $n$ . Для того чтобы сохранить обозначения настолько простыми, насколько это возможно, мы рассматриваем в данном разделе только модель с индивидуальными фиксированными эффектами из раздела 10.3, в которой отсутствуют временные эффекты.

### ***Предположения регрессии с фиксированными эффектами***

Четыре предположения модели регрессии с фиксированными эффектами приведены во вставке «Основные понятия 10.3». Эти предположения являются расширениями четырех предположений метода наименьших квадратов для межобъектной регрессии из вставки «Основные понятия 6.4» на случай панельных данных.

Первое предположение говорит о том, что ошибка регрессии имеет нулевое условное среднее относительно значений  $X$  данного объекта во все моменты времени  $T$ . Это предположение играет ту же роль, что и первое предположение метода наименьших квадратов из вставки «Основные понятия 6.4» для межобъектных данных, и предполагает, что в оценках нет смещения из-за пропущенных переменных. Требование о том, что условное среднее  $u_{it}$  не зависит от любых значений  $X$  для данного объекта – прошлых, настоящих или будущих – добавляет важную тонкость, лежащую за пределами предположений метода наименьших квадратов для межобъектных данных. Это предположение нарушается, если текущее значение  $u_{it}$  коррелирует с прошлыми, настоящими или будущими значениями  $X$ .

Второе предположение заключается в том, что для каждого объекта переменные распределены одинаково и независимо от переменных для другого объекта, то есть переменные являются i.i.d. для объекта для любого  $i=1, \dots, n$ . Аналогично второму предположению метода наименьших квадратов из вставки «Основные понятия 6.4» второе предположения для регрессии с фиксированными эффектами выполняется, если объекты отобраны из генеральной совокупности простым случайным образом.

**Предположения модели с фиксированными эффектами**

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

где

1.  $u_{it}$  – имеет нулевое условное среднее  $E(u_{it} | X_{i1}, X_{i2}, \dots, X_{iT}) = 0$ .
2.  $(X_{i1}, X_{i2}, \dots, X_{iT}, u_{i1}, u_{i2}, \dots, u_{iT})$ ,  $i = 1, \dots, n$  – являются независимыми и одинаково распределенными (i.i.d.).
3. Большие выбросы (отклонения от среднего) маловероятны:  $(X_{it}, u_{it})$  имеют ненулевые конечные четвертые моменты.
4. Отсутствует совершенная мультиколлинеарность.

В случае нескольких регрессоров  $X_{it}$  необходимо заменить на их полный перечень  $X_{1,it}, X_{2,it}, \dots, X_{k,it}$ .

**ОСНОВНЫЕ ПОНЯТИЯ  
10.3**

Третье и четвертое предположения в вышеописанной модели с фиксированными эффектами аналогичны третьему и четвертому предположениям в рамках МНК для межобъектных данных из вставки «Основные понятия 6.4».

В рамках предположений, лежащих в основе метода наименьших квадратов для панельных данных, изложенных во вставке «Основные понятия 10.3», оценки фиксированных эффектов являются состоятельными и нормально распределенными в случае больших  $n$ . Более подробно этот случай рассматривается в приложении 10.2.

Важным различием между основными предположениями модели регрессии для панельных данных из вставки «Основные понятия 10.3» и предположениями модели регрессии для межобъектных данных из вставки «Основные понятия 6.4» является предположение 2. В рамках этого предположения для межобъектных данных каждое наблюдение является независимым, что может иметь место уже в случае простой случайной выборки. С другой стороны, предположение 2 для панельных данных говорит о том, что переменные являются независимыми между различными объектами, но не делает таких ограничений в рамках одного объекта наблюдения. Например, предположение 2 позволяет  $X_{it}$  быть коррелированной во времени в рамках одного объекта наблюдения.

Если  $X_{it}$  коррелирована с  $X_{is}$  для различных значений  $s$  и  $t$ , то есть  $X_{it}$  коррелирована во времени в рамках одного объекта наблюдения, то  $X_{it}$  называется *автокоррелированной* (коррелированной с самой собой в различные моменты времени) или *серийно коррелированной*. Автокорреляция является распространенной особенностью временных рядов. То, что происходит в одном году, имеет тенденцию быть связанным с тем, что происходит в следующем году. В примере с количеством ДТП с летальным исходом переменная  $X_{it}$ , которая описывает уровень налогов на пиво в штате  $i$  в году  $t$ , является автокоррелированной, поскольку большую часть времени законодательный орган не меняет уровень налогов на пиво. Поэтому, если он высок в каком-то году по сравнению со средним уровнем для данного штата  $i$ , то, скорее всего, он будет высок также

и в следующем году. Аналогичным образом можно предположить, что  $u_{it}$  могут быть автокоррелированы. Вспомним, что  $u_{it}$  состоит из изменяющихся во времени факторов, которые являются детерминантами  $Y_{it}$ , но не включены в качестве регрессоров, и некоторые из этих пропущенных факторов могут быть автокоррелированы. Например, спад местной экономики может приводить к увольнениям, что, в свою очередь, уменьшит автомобильный трафик, тем самым уменьшая количество ДТП с летальным исходом в течение последующих двух и более лет. Кроме того, крупные проекты по улучшению дороги могут снизить количество дорожно-транспортных происшествий не только в год завершения реконструкции, но и в последующие годы. Такие пропущенные переменные (факторы), влияние которых сохраняется на протяжении нескольких лет, могут приводить к ошибкам в регрессионных оценках, вызванным наличием автокорреляции. Однако не все пропущенные факторы будут приводить к наличию автокорреляции в  $u_{it}$  — например, тяжелые зимние условия оказывают сильное влияние на изменение числа ДТП с летальным исходом, но если погодные условия зимой для данного штата независимо распределены из года в год, то эта компонента остаточного члена не будет являться автокоррелированной. В целом, однако, если некоторые пропущенные факторы автокоррелированы, то  $u_{it}$  также будут автокоррелированы.

### ***Стандартные ошибки в модели регрессии с фиксированными эффектами***

Если ошибки регрессии автокоррелированы, то использование обычных устойчивых к гетероскедастичности формул для стандартных ошибок для построенной на межобъектных данных регрессии [уравнения (5.3) и (5.4)] не является корректным. Один из способов убедиться в этом — провести аналогию с гетероскедастичностью. В регрессии с межобъектными данными, если имеет место гетероскедастичность в ошибках, то (как обсуждалось в разделе 5.4) обычные формулы для стандартных ошибок не являются допустимыми, поскольку они были получены в рамках неверного в данном случае предположения о гомоскедастичности. Точно так же, если ошибки автокоррелированы, то обычные формулы для стандартных ошибок не будут корректны, потому что они были получены при неверном предположении об отсутствии автокорреляции.

Формулы для стандартных ошибок, которые являются действительными при потенциальной возможности наличия гетероскедастичности и автокорреляции во времени в рамках одного объекта наблюдения в  $u_{it}$ , называются *устойчивыми к гетероскедастичности и автокорреляции стандартными ошибками (НАС)*. Стандартные ошибки, используемые в этой главе, являются одним из типов НАС-стандартных ошибок, а именно — *клUSTERизованными стандартными ошибками*. Термин *клusterный* возникает, поскольку эти стандартные ошибки позволяют ошибкам регрессии иметь произвольную корреляцию в пределах кластера или группы. Однако далее будем предполагать, что ошибки регрессии не коррелируют между кластерами. В рамках панельных данных каждый кластер состоит из нескольких объектов. Таким образом, клusterизованные стандартные ошибки

допускают возможность наличия гетероскедастичности и автокорреляции для одного объекта, но возможность корреляции ошибок между субъектами отсутствует. То есть использование кластеризованных стандартных ошибок допускает возможность наличия гетероскедастичности и автокорреляции таким образом, чтобы это соответствовало второму предположению в рамках регрессионной модели с фиксированными эффектами из вставки «Основные понятия 10.3».

Как и устойчивые к наличию гетероскедастичности стандартные ошибки в регрессиях межобъектных данных, кластеризованные стандартные ошибки могут использоваться как в случае наличия гетероскедастичности или автокорреляции (или одновременного наличия), так и в случае отсутствия. Если количество субъектов  $n$  велико, то использование стандартных ошибок такого вида может сопровождаться использованием (в том числе для статистических выводов) обычных для больших выборок нормальных критических значений для  $t$ -статистик и  $F_{q,\infty}$  критических значений для  $F$ -статистик, используемых для тестирования  $q$  ограничений.

На практике же могут существовать большие различия между кластеризованными стандартными ошибками и стандартными ошибками, которые не допускают наличия автокорреляции в  $u_i$ . Например, обычная (для межобъектных данных) оценка устойчивой к гетероскедастичности стандартной ошибки для коэффициента *BeerTax* в уравнении (10.21) составляет 0,25, что значительно меньше, чем аналогичная оценка при использовании кластеризованных стандартных ошибок, которая равна 0,36. Соответствующие  $t$ -статистики для тестирования гипотезы  $\beta_1 = 0$  равны -2,51 и -1,78. Причина, по которой в этом разделе приводятся кластеризованные стандартные ошибки, заключается в том, что они допускают наличие автокорреляции  $u_i$  в рамках одного субъекта, в то время как обычные устойчивые к гетероскедастичности стандартные ошибки – нет. Формулы для кластеризованных стандартных ошибок приведены в приложении 10.2.

## **10.6. Количество ДТП с летальным исходом и законы, направленные на сокращение случаев вождения в нетрезвом виде**

Введение налогов на алкогольную продукцию является лишь одним из способов воспрепятствовать вождению в нетрезвом состоянии. Штаты различаются по степени тяжести наказания за вождение в нетрезвом виде, и количество случаев вождения в нетрезвом состоянии в том или ином штате может быть снижено путем ужесточения соответствующих законов, а также повышения налогов. Тогда отсутствие переменной, характеризующей законодательство в отношении вождения в нетрезвом виде, в уравнении регрессии может привести к возникновению смещений (смещений, вызванных пропущенными переменными) в МНК-оценках влияния налогов на пиво на количество ДТП с летальным исходом даже в модели регрессии с временными и индивидуальными фиксированными эффектами. Кроме того, поскольку использование транспортных средств частично зависит от наличия у водителей работы, а также поскольку изменения уровня налогов

могут отражать состояние экономики (дефицит бюджета штата может привести к повышению налогов), то отсутствие переменной, отражающей экономические условия, также может привести к смещениям в оценках. В этом разделе проведенный ранее анализ детерминант ДТП с летальным исходом будет расширен за счет включения переменных, отражающих законодательство, направленное на предотвращение вождения в нетрезвом виде, и экономические условия.

Результаты приведены в таблице 10.1. Формат таблицы такой же, как и в таблицах, в которых представлены результаты регрессионного анализа в главах 7–9. Каждый столбец отвечает за отдельную регрессию, а в строках представлены оценки коэффициентов и стандартные ошибки,  $F$ -статистика и  $p$ -значение, либо другая информация о результатах оценки регрессии.

В столбце (1) в таблице 10.1 представлены результаты для МНК-оценки регрессии уровня смертности в ДТП от уровня налогов на пиво без индивидуальных (по штатам) и временных фиксированных эффектов. Как и в регрессии на межобъектных данных для 1982 и 1988 годов [уравнения (10.2) и (10.3)], оценка коэффициента перед переменной уровня реального налога на пиво является положительной и равна (0,36). В соответствии с этой оценкой повышение налогов на пиво приводит к увеличению количества ДТП с летальным исходом. Тем не менее результаты оценки регрессии, представленные в столбце (2) [ранее рассматривалась как уравнение (10.15)], включающие в себя индивидуальные фиксированные эффекты, показывают, что положительный коэффициент в регрессии (1) является результатом смещения, вызванного пропущенными переменными (оценка коэффициента при переменной уровня налогов на пиво равна  $-0,66$ ).  $R^2$  регрессии резко увеличился с 0,091 до 0,889 при включении фиксированных эффектов. По-видимому, индивидуальные фиксированные эффекты отвечают за большую часть дисперсии данных.

При включении временных фиксированных эффектов возникают лишь небольшие изменения в полученных оценках, что показано в столбце (3) [рассматривалось ранее как уравнение (10.21)], за исключением того, что оценка коэффициента перед переменной уровня налогов на пиво становится менее точной. Результаты, представленные в столбцах (1) – (3), соответствуют тому, что пропущенные фиксированные факторы – исторические и культурные факторы, общие дорожные условия, плотность населения, отношение к вождению в нетрезвом состоянии и так далее – являются важными детерминантами изменения уровня ДТП с летальным исходом в разных штатах.

Таблица 10.1

**Результаты регрессионного анализа влияния законодательства на ДТП со смертельным исходом**

Зависимая переменная: уровень смертности в ДТП (количество погибших на 10 тыс. населения)							
Регрессор	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Уровень налогов на пиво	0,36** (0,05)	-0,66* (0,29)	-0,64*** (0,36)	-0,45 (0,30)	-0,69* (0,35)	-0,46 (0,31)	-0,93** (0,34)

Окончание таблицы 10.1

Зависимая переменная: уровень смертности в ДТП (количество погибших на 10 тыс. населения)							
Регрессор	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Минимальный возраст употребления спиртных напитков 18 лет				0,028 (0,070)	-0,010 (0,083)		0,037 (0,102)
Минимальный возраст употребления спиртных напитков 19 лет				-0,018 (0,050)	-0,076 (0,068)		-0,065 (0,099)
Минимальный возраст употребления спиртных напитков 20 лет				0,032 (0,051)	-0,100* (0,56)		-0,113 (0,125)
Минимальный возраст употребления спиртных напитков						-0,0023 (0,021)	
Обязательное тюремное заключение или обязательные общественные работы				0,038 (0,103)	0,085 (0,112)	0,039 (0,103)	0,089 (0,164)
Средний автомобильный пробег на одного водителя				0,008 (0,007)	0,017 (0,011)	0,009 (0,007)	0,124 (0,049)
Уровень безработицы				-0,063** (0,013)		-0,063** (0,013)	-0,091** (0,021)
Логарифм реального дохода на душу населения				1,82** (0,64)		1,79** (0,64)	1,00 (0,68)
Временной период	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	Только 1992 и 1988
Индивидуальные эффекты	Нет	Да	Да	Да	Да	Да	Да
Временные эффекты	Нет	Нет	Да	Да	Да	Да	Да
Кластеризованные стандартные ошибки	Нет	Да	Да	Да	Да	Да	Да
<i>F</i> -статистики и <i>p</i> -значения для тестирования исключения групп переменных							
Временные эффекты отсутствуют			4,22 (0,002)	10,12 (< 0,001)	3,48 (0,006)	10,28 (< 0,001)	37,49 (< 0,001)
Коэффициенты при переменных минимального возраста употребления спиртных напитков =0				0,35 (0,786)	1,41 (0,253)		0,42 (0,738)
Уровень безработицы =0, реальный доход на душу населения =0				29,62 (< 0,001)		31,96 (< 0,001)	25,20 (< 0,001)
$\bar{R}^2$	0,091	0,889	0,891	0,926	0,893	0,926	0,899

Примечание. Данные оценки регрессий были получены на основе панельных данных по 48 штатам США. В регрессиях с (1) по (6) использовались данные за период с 1982 по 1988 год, в регрессии (7) использовались данные исключительно за 1982 и 1988 год. Используемая выборка описана в приложении 10.1. Стандартные ошибки приведены в скобках под оценками коэффициентов, *p*-значения (*p*-values) представлены в скобках под значениями *F*-статистик. Статистическая значимость отдельных коэффициентов обозначена для \*\*\*10%-го, \*\*5%-го или \*1%-го уровней значимости.

Следующие четыре регрессии в таблице 10.1 включают в себя дополнительные переменные, потенциально влияющие на уровень смертности в ДТП, наряду с индивидуальными (по штатам) и временными эффектами. Базовая спецификация, результаты оценки которой представлены в столбце (4), включает переменные, отражающие влияние законов в сфере вождения в нетрезвом виде, а также переменные, отвечающие за объем автомобильного движения, общие экономические условия в каждом из штатов. Первыми используемыми переменными, отражающими законодательство, являются переменные, отвечающие за минимальный возраст, в котором разрешено употребление спиртных напитков, представленные тремя бинарными переменными, соответствующими минимальному возрасту в 18, 19 и 20 лет (таким образом, опущена группа с минимальным возрастом от 21 года и старше). Другой переменной, отвечающей за законодательство, является степень наказания, положенного за первый установленный случай вождения в нетрезвом состоянии,— либо обязательное тюремное заключение, либо обязательные общественные работы (в данном случае пропущена группа менее суровых наказаний). Также используются три переменные, отражающие общие условия в автомобильной сфере и общие экономические условия, — средний уровень пробега (в милях) автомобиля на одного водителя, уровень безработицы и логарифм реального (в долларах 1988 г.) располагаемого дохода на душу населения (использование логарифма позволяет интерпретировать коэффициент перед исследуемой переменной в терминах процентного изменения; см. раздел 8.2). Последняя регрессия, представленная в таблице 10.1, использует подход «сравнение до и после», описанный в разделе 10.2, основываясь только на данных за 1982 и 1988 годы. Таким образом, регрессия (7) расширяет регрессию, представленную в уравнении (10.8), посредством включения дополнительных регрессоров.

Оценки регрессии, представленные в столбце (4), показывают четыре интересных результата:

1. Включение дополнительных переменных уменьшает предполагаемое влияние налогов на пиво с  $-0,64$  в столбце (3) до  $-0,45$  в столбце (4). Чтобы оценить величину этого коэффициента, можно представить себе штат, в котором налог на пиво в 2 раза больше по сравнению со средним уровнем налогов на пиво. Поскольку средний уровень налогов на пиво в используемых данных составляет около 0,50 долл. за ящик (в долларах 1988 г.), то это влечет за собой увеличение налога на 0,50 долл. за ящик пива. Тогда эффект от такого повышения на уровень смертности в ДТП можно оценить как  $0,45 \times 0,50 = 0,23$  погибших на 10 тыс. населения. Полученная оценка показывает, что этот эффект значителен, поскольку средний уровень смертности в ДТП составляет два погибших на 10 тыс. человек, таким образом, снижение на 0,23 соответствует снижению трафика смертей почти на одну восьмую. Это говорит о том, что полученная оценка является весьма неточной. Поскольку стандартная ошибка этого коэффициента равна 0,30, то доверительный интервал для него равен  $-0,45 \times 0,50 \pm 1,96 \times 0,30 \times 0,50 = (-0,52; 0,07)$ . Полученный 95 %-й-процентный доверительный интервал включает в себя

- ноль, поэтому гипотеза о том, что уровень налогов на пиво не оказывает значимого эффекта, не может быть отвергнута на 5 %-м уровне значимости.
2. Полученные оценки достаточно точно показывают, что минимальный возраст, в котором разрешено употребление алкогольных напитков, оказывает небольшое влияние на уровень смертности в ДТП. В соответствии с результатами, представленными в столбце (4), 95 %-й доверительный интервал для увеличения смертности ДТП в штате с минимальным возрастом для употребления алкоголя, установленным на уровне 18 лет, равен  $(-0,11; 0,17)$ . Совместная гипотеза о том, что коэффициенты перед переменными, отражающими минимальный установленный возраст для употребления спиртных напитков, равны нулю, не может быть отвергнута на уровне значимости 10 %, поскольку  $F$ -статистика для тестирования этой гипотезы равна 0,35 при  $p$ -значении, равном 0,786.
  3. Оценка коэффициента при первой переменной, отвечающей за степень наказания за вождение в нетрезвом виде, также мала и не отличается от нуля на уровне значимости 10 %.
  4. Оценки показывают, что переменные, отражающие общие экономические условия, обладают значительной объясняющей силой для уровня смертности в ДТП. Высокий уровень безработицы приводит к уменьшению количества смертельных случаев в ДТП: увеличение уровня безработицы на один процентный пункт, как было оценено, приводит к снижению уровня смертности в ДТП на 0,063 погибших на 10 тыс. человек. С другой стороны, высокие значения реальных доходов на душу населения приводят к высокому количеству ДТП с летальным исходом: полученная оценка коэффициента составляет 1,82, таким образом, увеличение реального дохода на душу населения на 1% приводит к росту уровня смертности в ДТП на 0,0182 погибших за 10 тыс. человек (см. пример 1 во вставке «Основные понятия 8.2» для интерпретации этого коэффициента). Согласно этим оценкам, хорошие экономические условия приводят к большему количеству погибших в ДТП, возможно, из-за более высокой плотности движения, когда уровень безработицы низок, или большего потребления алкоголя, когда уровень доходов высок. Эти две экономические переменные являются совместно статистически значимыми на уровне значимости 0,1% ( $F$ -статистика равна 29,62).

В столбцах (5) – (7) таблицы 10.1 представлены результаты оценок регрессий, в которых проверяется чувствительность представленных выше выводов к изменениям в базовой спецификации регрессии. В столбце (5) показаны результаты оценки регрессии, в которой опущены переменные, отражающие общие экономические условия. Они показывают увеличение оценки влияния налогов на пиво, которое становится значимым на уровне значимости 5 %, но значимые изменения в оценках других коэффициентов отсутствуют. Чувствительность полученной оценки коэффициента перед переменной уровня налогов

на пиво к включению в уравнение регрессии дополнительных переменных, отражающих общие экономические условия, в сочетании со статистической значимостью коэффициентов при этих переменных в столбце (4) показывает, что переменные, отражающие общие экономические условия, должны оставаться в базовой спецификации регрессии. Результаты оценки регрессии, представленные в столбце (6), показывают, что результаты в столбце (4) не чувствительны к изменению функциональной формы в том случае, когда три индикаторные переменные, отвечающие за минимальный возраст для употребления алкоголя, заменены на одну переменную, непосредственно отражающую минимальный возраст для употребления алкоголя. Когда коэффициенты оцениваются на основе изменений переменных с 1982 по 1988 год [столбец (7)], как в разделе 10.2, выводы, полученные при анализе результатов, представленных в столбце (4), в значительной степени остаются неизменными, за исключением того, что коэффициент при переменной уровня налогов на пиво становится больше, а также становится значимым на уровне значимости 1 %.

Сильная сторона этого анализа заключается в том, что включение индивидуальных (для штатов) и временных фиксированных эффектов снижает угрозу возникновения смещений, вызванных пропущенными переменными, возникающими из-за ненаблюдаемых переменных, которые либо не изменяются с течением времени (например, общественное отношение к употреблению алкоголя за рулем), либо не принимают различные значения для разных штатов (например, инновации в сфере автомобильной безопасности). Тем не менее важно не забывать о возможных угрозах корректности полученных результатов. Одним из потенциальных источников возникновения смещений в оценках, вызванных пропущенными переменными, может являться то, что мера уровня налогов на алкоголь, используемая в данном исследовании – реальный налог на пиво, – может изменяться в зависимости от уровней других налогов на алкогольную продукцию, что позволяет более широко интерпретировать полученные результаты, нежели как просто относящиеся к налогам на пиво. Существует небольшая вероятность того, что скачкообразные повышения уровня реальных налогов на пиво могут быть связаны с просветительскими кампаниями для населения. Если это действительно так, то изменения в уровнях налога на пиво могут также включать в себя более «широкие» эффекты различных мероприятий (например, как указано выше, государственных кампаний), направленных на снижение вождения в нетрезвом виде.

Вместе все эти результаты дают довольно провокационную картину влияния мер, направленных на контроль вождения в нетрезвом виде, и количества ДТП с летальным исходом. Согласно представленным оценкам, ни жесткие наказания, ни увеличение минимального возраста для употребления алкоголя не оказывают существенного влияния на уровень смертности в ДТП. С другой стороны, есть некоторые доказательства того, что увеличение налогов на алкоголь, если судить по реальным налогам на пиво, снижает уровень смертности в ДТП, предположительно за счет снижения потребления алкоголя. Некая неточность полученных оценок коэффициентов при переменной уровня налогов на пиво означает, по-видимому, что, основываясь на данном анализе, необходимо до-

статочно осторожно подходить к изменению существующих правил и политик, а также что необходимы дополнительные исследования<sup>1</sup>.

## 10.7. Заключение

Проведенный в этой главе анализ показал, как наблюдения, полученные для одного и того же субъекта в течение определенного периода, могут быть использованы для учета ненаблюдаемых переменных, которые принимают различные значения для разных объектов, но постоянны во времени. Ключевой особенностью является то, что если ненаблюданная переменная не изменяется во времени, то любые изменения зависимой переменной должны быть вызваны влиянием иных (отличных от этой) фиксированных характеристик. Если общественное или культурное отношение к употреблению алкоголя за рулем не менялось значительно в течение семи лет в том или ином штате, то необходимо искать другие объяснения изменений уровня смертности в ДТП в течение этих семи лет.

Чтобы использовать этот факт, необходимо наличие данных, которые содержат наблюдения для одного и того же субъекта в течение двух или более периодов времени, то есть необходимы панельные данные. С использованием панельных данных модель множественной регрессии, представленная в части II, может быть расширена посредством включения в нее полного набора бинарных переменных, каждая из которых соответствует одному объекту. Такая модификация модели может быть оценена с помощью МНК. Дальнейшим развитием модели с фиксированными эффектами является добавление временных фиксированных эффектов, которые позволяют учесть ненаблюдаемые переменные, которые изменяются с течением времени, но принимают одинаковые значения для различных субъектов. Индивидуальные и временные фиксированные эффекты могут быть включены в регрессию для того, чтобы учитывать переменные, которые различаются для разных объектов, но являются постоянными во времени, а также переменные, которые меняются с течением времени, но постоянны по объектам.

Несмотря на вышеперечисленные достоинства, модель с индивидуальными и временными фиксированными эффектами не может учесть влияния пропущенных переменных, которые меняются как по субъектам, так и по времени. И, очевидно, использование методов анализа панельных данных подразумевает и требует наличия самих данных такого рода (панельных данных), которые часто являются недоступными. Таким образом, сохраняется необходимость в способе построения оценок, который помог бы устраниТЬ влияние ненаблюдаемых пропущенных переменных в тех случаях, когда методы панельных данных не могут быть применены. Достаточно мощным и общим методом для решения

<sup>1</sup> Для дальнейшего анализа этих данных см. Ruhm (1996). Проведенный в современных работах мета-анализ 112 работ, посвященных эффектам влияния цен на алкогольную продукцию и налогам на потребление алкоголя, показал, что эластичности равны  $-0,46$  для пива,  $-0,69$  для вина и  $-0,80$  для алкоголя в целом. Анализ показал, что налоги на алкоголь оказывают значительный эффект, направленный на снижение его потребления, по сравнению с другими программами [см. Wagenaar, Salois, Komro (2009)]. Для получения более подробной информации относительно экономических эффектов вождения в нетрезвом виде и употребления алкоголя, а также об экономике алкогольной сферы в целом, см. Cook, Moore (2000), Chaloupka, Grossman, Saffer (2002), Young, Bielinska-Kwapisz (2006), Dang (2008).

указанной проблемы является модель регрессии с использованием инструментальных переменных, которая будет рассмотрена в главе 12.

## **Выводы**

1. Панельные данные состоят из наблюдений для нескольких ( $n$ ) объектов – штатов, фирм, людей и так далее, – где наблюдение за каждым объектом велось в течение двух или более периодов времени ( $T$ ).
2. Регрессии с индивидуальными фиксированными эффектами позволяют учесть ненаблюдаемые переменные, значения которых различны для разных объектов, но остаются неизменными во времени.
3. При наличии данных лишь по двум временным периодам модель с фиксированными эффектами может быть оценена с помощью подхода «сравнение до и после», в рамках которого оценивается регрессия изменений в  $Y$  между периодами в зависимости от соответствующих изменений  $X$ .
4. Модель с индивидуальными фиксированными эффектами может быть оценена с помощью включения в нее бинарных переменных для  $n-1$  субъекта, наблюдаемых независимых переменных ( $X$ ) и свободного члена.
5. Временные фиксированные эффекты помогают учитывать влияние ненаблюдаемых переменных, которые принимают одинаковые значения для всех субъектов, но меняются с течением времени.
6. Модель регрессии с индивидуальными и временными фиксированными эффектами может быть оценена посредством добавления в регрессию  $n-1$  бинарной переменной для субъектов,  $T-1$  бинарных переменных для временных периодов, наблюдаемых независимых переменных ( $X$ ) свободного члена.
7. В панельных данных переменные, как правило, являются автокоррелированными, то есть коррелированы во времени в рамках одного субъекта (объекта наблюдения). Поэтому стандартные ошибки должны по возможности учитывать как наличие автокорреляции, так и потенциальное наличие гетероскедастичности. Одним из вариантов решения данной проблемы является использование кластеризованных стандартных ошибок.

## **Основные понятия**

Панельные данные (с. 360).

Сбалансированная панель (с. 361).

Несбалансированная панель (с. 361).

Модель регрессии с фиксированными эффектами (с. 367).

Индивидуальные фиксированные эффекты (с. 373).

Модель регрессии с временными фиксированными эффектами (с. 373).

Временные фиксированные эффекты (с. 374).

Модель регрессии с индивидуальными и временными фиксированными эффектами (с. 374).

Автокорреляция, серийная корреляция (с. 377).

Устойчивые к гетероскедастичности и автокорреляции стандартные ошибки (HAC) (с. 378).

Кластеризованные стандартные ошибки (с. 378).

### **Вопросы для повторения и закрепления основных понятий**

- 10.1. Почему для описания панельных данных необходимо использовать два индекса – индекс  $i$  и индекс  $t$ ? Что означает индекс  $i$ ? Что означает индекс  $t$ ?
- 10.2. Исследователь использует панельные данные с наблюдениями для  $n = 1000$  работников в течение  $T = 10$  лет (с 2001 по 2010 г.), которые содержат данные по доходам работников, их полу, образованию и возрасту. Исследователь пытается оценить влияние образования на уровень заработной платы. Приведите примеры ненаблюдаемых и специфичных по объектам переменных, которые связаны (коррелированы) одновременно с уровнем образования и уровнем доходов. Приведите также примеры изменяющихся во времени переменных, которые могут быть связаны с уровнем образования и уровнем доходов. Как можно учесть влияние этих индивидуальных и временных эффектов в модели регрессии с панельными данными?
- 10.3. Может ли регрессия, которая использовалась для ответа на вопрос 10.2, использоваться для оценки влияния пола рабочего на уровень его заработной платы? Может ли эта регрессия быть использована для оценки влияния общего уровня безработицы на уровень заработной платы? Приведите пояснения.
- 10.4. В контексте регрессии, предложенной в рамках ответа на вопрос 10.2, поясните, почему регрессионные ошибки для данного индивида могут быть автокоррелированы?

### **Упражнения**

- 10.1. В данном упражнении используются результаты оценки регрессий, представленные в таблице 10.1.
  - а) Население Нью-Джерси составляет 8,1 млн человек. Предположим, что в Нью-Джерси налог на ящик пива увеличился на 1 долл. (в долларах 1988 г.). Используя результаты из столбца (4), покажите, как изменится количество погибших в ДТП в течение следующего года. Постройте 95%-й доверительный интервал для предложенного ответа.
  - б) Минимальный возраст для употребления алкогольных напитков в Нью-Джерси составляет 21 год. Предположим, что минимальный возраст для употребления алкогольных напитков в Нью-Джерси снижен до 18 лет.

Используя результаты из столбца (4), покажите изменение количества погибших в ДТП в следующем после изменений году. Постройте 95%-й доверительный интервал для предложенного ответа.

- в) Предположим, что реальные доходы на душу населения в Нью-Джерси увеличатся на 1% в следующем году. Используя результаты из столбца (4), покажите изменение количества погибших в ДТП в следующем после изменений году. Постройте 90%-й доверительный интервал для предложенного ответа.
- г) Должны ли временные эффекты быть включены в уравнение регрессии? Аргументируйте свой ответ.
- д) Исследователь предполагает, что безработица имеет различное влияние на уровень смертности в ДТП происшествиях в западных штатах и иных штатах. Как вы предложили бы проверить эту гипотезу? (Предложите конкретную спецификацию регрессионной модели и статистический тест, которые необходимо использовать.)
- 10.2. Рассмотрим версию модели с фиксированными эффектами с включением бинарных переменных, представленную в уравнении (10.11), в которую добавлены дополнительные регрессоры  $D1_i$ :
- $$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_1 D1_i + \gamma_2 D2_i + \dots + \gamma_n Dn_i + u_{it}.$$
- а) Предположим, что  $n = 3$ . Покажите, что бинарные переменные и свободный член являются совершенно мультиколлинеарными, то есть переменные  $D1_i, D2_i, D3_i$  и  $X_{0,it}$  могут быть записаны в виде линейной функции, где  $X_{0,it} = 1$  для всех  $i$  и  $t$ .
- б) Запишите результаты из пункта (а) для произвольного  $n$ .
- в) Что произойдет в случае попытки оценки коэффициентов регрессии с помощью МНК?
- 10.3. В разделе 9.2 был представлен список из пяти потенциальных угроз для внутренней обоснованности регрессионного анализа. Используя этот перечень применительно к эмпирическому анализу, проведенному в разделе 10.6, дайте оценку сделанным на его основе выводам.
- 10.4. Используя модель регрессии, представленную в уравнении (10.11), оцените величину коэффициентов (углового коэффициента и свободного члена) для:
- а) субъекта 1 в периоде 1;
  - б) субъекта 1 в периоде 3;
  - в) субъекта 3 в периоде 1;
  - г) субъекта 3 в периоде 3.
- 10.5. Рассмотрим модель с одним регрессором:  $Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$ . Эта модель может быть переписана в таком виде:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \delta_2 B2_t + \dots + \delta_T BT_t + \gamma_2 D2_i + \dots + \gamma_n Dn_i + u_{it}$$

где  $B2_i = 1$ , если  $t=2$ ,  $B2_i = 0$  – в ином случае;  $D2_i = 1$ , если  $i = 2$ ,  $D2_i = 0$  – в ином случае и так далее. Как коэффициенты  $(\beta_0, \delta_2, \dots, \delta_T, \gamma_2, \dots, \gamma_n)$  могут быть связаны с коэффициентами  $\alpha_1, \dots, \alpha_n, \dots, \lambda_1, \dots, \lambda_T$ ?

- 10.6. Подразумевают ли основные предположения модели регрессии с фиксированными эффектами, представленные во вставке «Основные понятия 10.3», что  $\text{cov}(\tilde{\nu}_{it}, \tilde{\nu}_{is}) = 0$  для  $t \neq s$  в уравнении (10.28)?
- 10.7. Исследователь считает, что количество ДТП с летальным исходом увеличивается при наличии гололеда на дорогах. Таким образом, в штатах с более снежными погодными условиями уровень смертности в ДТП будет выше, чем в других штатах. Прокомментируйте следующие методы, предназначенные для оценки влияния снежной погоды на уровень смертности в ДТП:
- Исследователь собирает данные по среднему уровню выпадения снега в каждом штате и добавляет дополнительный регрессор ( $AverageSnow_i$ ) в регрессии, представленные в таблице 10.1.
  - Исследователь собирает данные по уровню выпадения снега в каждом штате за год в течение нескольких лет и добавляет дополнительный регрессор ( $Snow_{it}$ ) в регрессии, представленные в таблице 10.1.
- 10.8. Рассмотрим наблюдения  $(Y_{it}, X_{it})$  в рамках линейной регрессионной модели для панельных данных  $Y_{it} = X_{it}\beta_1 + \alpha_i + \lambda_i t + u_{it}$ , где  $t = 1, \dots, T$ ;  $i = 1, \dots, N$ ;  $\alpha_i + \lambda_i t$  – ненаблюдаемый специфичный для субъектов временной тренд. Как можно оценить коэффициент  $\beta_1$ ?
- 10.9. а) Являются ли оценки индивидуальных фиксированных эффектов  $\alpha_i$  в модели регрессии с фиксированными эффектами состоятельными при  $n \rightarrow \infty$  при фиксированных  $T$ ? (Подсказка: проанализируйте аналогичную модель в отсутствии регрессоров  $X$ :  $Y_{it} = \alpha_i + u_{it}$ )  
б) Если  $n$  принимает достаточно большие значения (например  $n = 200$ ), а  $T$ , в свою очередь, маленькие (например,  $T = 4$ ), то являются ли оценки  $\alpha_i$  нормально распределенными? Поясните свой ответ (Подсказка: проанализируйте модель  $Y_{it} = \alpha_i + u_{it}$ )
- 10.9. В исследовании влияния уровня образования на уровень заработной платы на основании панельных данных о годовом доходе для большого числа работников исследователь с помощью регрессии с фиксированными эффектами оценивает зависимость заработной платы в определенном году от возраста, образования, профсоюзного статуса и уровня заработной платы работника в предшествующем году. Окажутся ли полученные в результате оценки коэффициентов при регрессорах (возраст, уровень образования, профсоюзный статус и уровень заработной платы работника предшествующем году) достаточно надежными? Поясните свой ответ. (Подсказка: проверьте основные предположения регрессионной модели с фиксированными эффектами, представленные в разделе 10.5.)
- 10.10. Пусть  $\hat{\beta}_1^{DM}$  – оценка (центрированная внутри группы – внутргрупповая) коэффициента  $\beta_1$  в уравнении (10.22), а  $\hat{\beta}_1^{BA}$  – оценка коэффициента  $\beta_1$  с помощью подхода «сравнение до и после» (в регрессии без свободного члена). Тогда  $\hat{\beta}_1^{BA} = \left[ \sum_{i=1}^n (X_{i2} - \bar{X}_{i1}) \right] \left[ (Y_{i2} - \bar{Y}_{i1}) \right] / \left[ \sum_{i=1}^n (X_{i2} - \bar{X}_{i1})^2 \right]$ . Покажите, что если  $T = 2$ , то  $\hat{\beta}_1^{DM} = \hat{\beta}_1^{BA}$  [Подсказка: используйте определение  $\tilde{X}_{it}$ ,

предшествующее уравнению (10.22), чтобы показать, что  $\tilde{X}_{i1} = -1/2(X_{i2} - X_{i1})$  и  $\tilde{X}_{i2} = 1/2(X_{i2} - X_{i1})$ ].

### **Компьютерные упражнения**

E10.1. В некоторых штатах США были принятые законы, позволяющие гражданам осуществлять скрытое ношение оружия. Эти законы известны как законы *shall-issue*, потому что они поручают местным властям выдачу разрешений на скрытое ношение оружия всем заявителям, которые являются гражданами, психически вменяемыми, не имеют судимостей за уголовные преступления (в некоторых штатах имеют место некоторые дополнительные ограничения). Сторонники этих законов утверждают, что если бы больше людей имели возможность скрыто носить оружие, то уровень преступности бы снизился, поскольку преступники воздерживаются от нападения на других людей. Противники утверждают, что уровень преступности будет увеличиваться из-за случайного или самопроизвольного использования оружия. В этом упражнении будет необходимо проанализировать влияние законов о скрытом ношении оружия на количество преступлений, связанных с насилием над личностью. На интернет-сайте учебника [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) представлен файл с данными под названием *Guns*, который содержит сбалансированную панель для 50 штатов США и округа Колумбия за период с 1977 по 1999 год<sup>1</sup>. Подробное описание данных приведено в файле *Guns\_Description*, доступном на веб-сайте.

- a) Оцените (1) регрессию  $\ln(vio)$  на *shall* и (2) регрессию  $\ln(vio)$  на *shall, incare\_rate, density, avginc, pop, pb1064, pw1064* и *pm1029*.
  - (i) Проинтерпретируйте оценку коэффициента перед переменной *shall* в регрессии (2). Является ли эта оценка значительной (или незначительной) по сравнению с реальностью?
  - (ii) Влияет ли включение контрольных переменных в регрессию (2) на изменение оценки эффекта наличия закона, разрешающего скрытое ношение оружия в регрессии (1), в плане статистической значимости? А в смысле соответствия ситуации в «реальном мире»?
  - (iii) Предложите переменную, которая изменяется в зависимости от конкретного штата, но практически не изменяется (или не изменяется совсем) с течением времени, и которая может привести к возникновению смещений, вызванных пропущенными переменными, в регрессии (2).
- б) Изменится ли полученный результат при включении в регрессию индивидуальных фиксированных эффектов (для штатов)? Если измене-

<sup>1</sup> Эти данные были предоставлены профессором Стэнфордского университета Джоном Донохью (John Donohue) и были использованы в совместной с Айаном Эйресом (Ian Ayres) работе «Shooting Down the 'More Guns Less Crime' Hypothesis», Stanford Law Review, 2003, 55: 1193–1312.

ния имеют место, то какой набор результатов регрессионного анализа является более надежным и почему?

- в) Изменится ли полученный результат при включении временных фиксированных эффектов (для периодов наблюдения)? Если изменения имеют место, то какой набор результатов регрессионного анализа является более надежным и почему?
  - г) Повторите анализ, используя  $\ln(\text{rob})$  и  $\ln(\text{mur})$  вместо  $\ln(\text{vio})$ .
  - д) Каковы, на ваш взгляд, наиболее важные остающиеся угрозы для внутренней обоснованности этого регрессионного анализа?
  - е) Какие выводы могут быть сделаны на основании проведенного вами анализа относительно влияния законов о скрытом ношении оружия на уровень преступности?
- E10.2. Дорожно-транспортные происшествия являются одной из главных причин смерти среди американцев в возрасте от 5 до 32 лет. С помощью различных политик распределения средств правительство США заставило штаты принять законы об обязательном использовании ремней безопасности в целях сокращения числа погибших и получивших серьезные травмы в ДТП. В этом упражнении будет необходимо исследовать эффективность законов принятых в сфере повышения использования ремней безопасности и снижения уровня смертности в ДТП. На интернет-сайте учебника [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) можно найти файл под названием Seatbelts, в котором представлены панельные данные для 50 американских штатов и округа Колумбия за период с 1983 по 1997 год<sup>1</sup>. Подробное описание данных приведено в файле Seatbelts\_Description, доступном на том же веб-сайте.
- а) Оцените величину эффекта от введения обязательного использования ремней безопасности с помощью оценки регрессии  $FatalityRate$  зависимости от  $sb\_useage, speed65, speed70, ba08, drinkage21, \ln(\text{income})$  и  $age$ . Показывают ли результаты регрессии, что использование ремней безопасности снижает уровень смертности в ДТП?
  - б) Изменяются ли результаты оценки регрессии при включении индивидуальных фиксированных эффектов (для штатов)? Приведите интуитивное объяснение того, почему изменяются результаты.
  - в) Изменяются ли результаты оценки регрессии при включении дополнительно к индивидуальным фиксированным эффектам (для штатов) временных фиксированных эффектов?
  - г) Какая из спецификаций регрессии – (а), (б) или (в) – является наиболее надежной? Поясните свой ответ.
  - д) Используя результаты регрессии в пункте (в), обсудите величину коэффициента при  $sb\_useage$ . Является ли этот коэффициент значительным или нет? Насколько снизится уровень смертности при увеличении использования ремней безопасности с 52 до 90%?

<sup>1</sup> Данные были предоставлены профессором Стэнфордского университета Лираном Эинавом (Liran Einav), а также использовались в его совместной работе с Альмой Коен (Alma Cohen) «The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities», The Review of Economics and Statistics, 2003, 85 (4): 828–843.

- е) Существует два способа заставить людей использовать ремни безопасности. «Основной» – полицейский может остановить машину и оштрафовать водителя, если сотрудник полиции заметил, что пассажир не пристегнут ремнем безопасности. «Вторичный» способ заключается в том, что сотрудник полиции может выписать штраф, если пассажир не пристегнут ремнем безопасности, но должны быть дополнительные причины, чтобы остановить автомобиль. В используемой выборке присутствует *primary* – бинарная переменная обозначения «основного» способа, *secondary* – бинарная переменная для обозначения «вторичного» способа. Оцените с помощью регрессии влияние связи между *sb\_useage* и *primary*, *secondary*, *speed65*, *speed70*, *ba08*, *drinkage21*, *ln(income)* и *age* включением в регрессию индивидуальных и временных фиксированных эффектов. Приводит ли применение описанных выше «основного» и «вторичного» способов к более интенсивному использованию ремней безопасности?
- ж) В 2000 году в Нью-Джерси перешли от «основного» к «вторичному» способу контроля за применением ремней безопасности. Оцените снижение уровня смертности в ДТП в год, в котором были сделаны данные изменения.

## Приложения

### Приложение 10.1. Данные по уровню смертности в ДТП в штатах США

В используемой выборке представлены ежегодные данные для 48 штатов США (кроме Аляски и Гавайских островов) за период с 1982 по 1988 год. Уровень смертности в ДТП представляет собой количество погибших в дорожно-транспортных происшествиях на 10 тыс. человек в течение одного года. Данные по уровню смертности в ДТП были получены с помощью Системы сбора информации о ДТП с летальным исходом Министерства транспорта США (U.S. Department of Transportation Fatal Accident Reporting System). Данные по уровню налогов на пиво (налог на ящик пива) были получены из Beer Institute's *Brewers Almanac*. Переменные, отражающие минимальный возраст для употребления алкогольных напитков, представленные в таблице 10.1, являются бинарными переменными. Бинарная переменная, отражающая степень наказания, представленная в таблице 10.1, описывает минимальные требования в том или ином штате к степени наказания за установленный факт вождения в нетрезвом виде. Эта переменная равна 1, если в качестве наказания в штате установлено тюремное заключение или общественные работы, и равна 0 в противном случае (мягкое наказание). Данные об общей ежегодной протяженности автомобильного пробега в каждом штате были получены с помощью Министер-

ства транспорта США. Данные по уровню личного дохода были получены с помощью Бюро экономического анализа США, а данные по уровню безработицы были получены с помощью Бюро статистики труда США.

Все эти данные были любезно предоставлены профессором кафедры экономики Университета Северной Каролины Кристофером Дж. Румом (Christopher J. Ruhm).

### **Приложение 10.2. Стандартные ошибки в модели регрессии с фиксированными эффектами**

В данном приложении представлены формулы для стандартных ошибок в модели с фиксированными эффектами с одним регрессором. Представленные формулы расширены на случай нескольких регрессоров в упражнении 18.15.

#### **Асимптотическое распределение оценок в модели с фиксированными эффектами при больших значениях $n$**

**Оценки в модели с фиксированными эффектами.** Оценка коэффициента  $\beta_1$  в модели с фиксированными эффектами является МНК-оценкой коэффициента, полученной в рамках регрессии после центрирования переменных внутри каждого объекта, представленной в уравнении (10.14), в которой оценивается регрессия  $\tilde{Y}_{it}$  на  $\tilde{X}_{it}$ , где  $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$ ,  $\tilde{X}_{it} = X_{it} - \bar{X}_i$ ,  $\bar{Y}_i = T^{-1} \sum_{t=1}^T Y_{it}$  и  $\bar{X}_i = T^{-1} \sum_{t=1}^T X_{it}$ .

Формула для МНК-оценки может быть получена с помощью замены  $X_i - \bar{X}$  на  $\tilde{X}_{it}$  и  $Y_i - \bar{Y}$  на  $\tilde{Y}_{it}$  в уравнении (4.7) и замены простого суммирования в уравнении (4.7) на двойное суммирование – по субъектам ( $i = 1, \dots, n$ ), по временным периодам ( $t = 1, \dots, T$ )<sup>1</sup>. Тогда выражение можно записать в таком виде:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it} \tilde{Y}_{it}}{\sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it}^2}. \quad (10.22)$$

Вывод выборочного распределения  $\hat{\beta}_1$  аналогичен выводу выборочного распределения для МНК-оценок для межобъектных данных, представленному в приложении 4.3. На первом шаге необходимо подставить  $\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$  (уравнение (10.14)) в числитель уравнения (10.22), получив, таким образом, аналог уравнения (4.30) для панельных данных:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it} \tilde{u}_{it}}{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it}^2}. \quad (10.23)$$

<sup>1</sup> Двойное суммирование представляет собой суммирование по второму индексу после суммирования по первому:

$$\begin{aligned} \sum_{i=1}^n \sum_{t=1}^T X_{it} &= \sum_{i=1}^n \left( \sum_{t=1}^T X_{it} \right) = \sum_{i=1}^n (X_{i1} + X_{i2} + \dots + X_{iT}) = \\ &= (X_{11} + X_{12} + \dots + X_{1T}) + (X_{21} + X_{22} + \dots + X_{2T}) + \dots + (X_{n1} + X_{n2} + \dots + X_{nT}). \end{aligned}$$

Затем необходимо преобразовать данное выражение и умножить обе его части на  $\sqrt{nT}$ , чтобы получить:

$$\sqrt{nT}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i}}{\hat{Q}_{\tilde{X}}}, \quad (10.24)$$

где  $\eta_i = \sqrt{\frac{1}{T} \sum_{t=1}^T \tilde{X}_{it} \tilde{u}_{it}}$  и  $\hat{Q}_{\tilde{X}} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it}^2$ .

Нормирующий множитель  $nT$  в уравнении (10.24) представляет собой полное число наблюдений.

**Распределение и стандартные ошибки при больших значениях  $n$ .** В большинстве случаев применения панельных данных  $n$  принимает значительно большие значения, чем  $T$ , что позволяет аппроксимировать выборочное распределение, устремляя  $n \rightarrow \infty$  при фиксированном  $T$ . В рамках основных предположений модели с фиксированными эффектами, представленных во вставке «Основные понятия 10.3»,  $\hat{Q}_{\tilde{X}} \rightarrow Q_{\tilde{X}} = ET^{-1} \sum_{i=1}^T \tilde{X}_{it}^2$  при  $n \rightarrow \infty$ . Так же  $\eta_i$  являются независимыми и одинаково распределенными переменными (i.i.d.) для  $i = 1, \dots, n$  (в рамках предположения 2) с нулевым средним значением (по предположению 1) и дисперсией  $\sigma_\eta^2$  (принимает конечные значения по предположению 3). Тогда по центральной предельной теореме  $\sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i} \xrightarrow{d} N(0, \sigma_\eta^2)$ .

Из уравнения (10.24) следует:

$$\sqrt{nT}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} -N\left(0, \frac{\sigma_\eta^2}{Q_{\tilde{X}}^2}\right). \quad (10.25)$$

Из уравнения (10.25) дисперсия  $\hat{\beta}_1$  (в больших выборках) равна:

$$\text{var}(\hat{\beta}_1) = \frac{1}{nT} \frac{\sigma_\eta^2}{Q_{\tilde{X}}^2}. \quad (10.26)$$

С помощью формулы для кластеризованных стандартных ошибок необходимо заменить их аналог в уравнении (10.26) на выборочные значения:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{nT} \frac{s_\eta^2}{\hat{Q}_{\tilde{X}}^2}}, \quad (10.27)$$

где  $s_\eta^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \hat{\eta}_i - \bar{\hat{\eta}}_i \right)^2 = \frac{1}{n-1} \sum_{i=1}^n \hat{\eta}_i^2$  (кластеризованные стандартные ошибки),  $\hat{\eta}_i = \sqrt{\frac{1}{T} \sum_{t=1}^T \tilde{X}_{it} \hat{u}_{it}}$  – выборочный аналог для  $\eta_i$  ( $\hat{\eta}_i$  представляет собой  $\eta_i$  в уравнении (10.24) после замены  $\tilde{u}_{it}$  на остатки регрессии с фиксированными эффектами  $\hat{u}_{it}$ ),  $\bar{\hat{\eta}}_i = \frac{1}{n} \sum_{i=1}^n \hat{\eta}_i$ . Итоговое уравнение (10.27) получено, поскольку  $\hat{\eta}_i = 0$ , так как остатки и регрессоры не являются коррелированными [уравнение (4.34)]. Стоит отметить,

что  $s_{\eta}^2$  представляет собой выборочную дисперсию  $\hat{\eta}_i$  [см. уравнение (3.7)]. Оценка  $s_{\eta}^2$  является состоятельной оценкой  $\sigma_{\eta}^2$  при  $n \rightarrow \infty$  даже при наличии гетероскедастичности или автокорреляции (см. упражнение 18.15). Таким образом, кластеризованные стандартные ошибки в уравнении (10.27) являются устойчивыми к наличию гетероскедастичности или автокорреляции. Поскольку кластеризованные стандартные ошибки являются состоятельными, то  $t$ -статистика для тестиования гипотезы  $\beta_1 = \beta_{1,0}$  имеет стандартное нормальное распределение в рамках нулевой гипотезы при  $n \rightarrow \infty$ .

Все вышеизложенные результаты сохраняются и при наличии нескольких регрессоров. Кроме того, если  $n$  велико, то  $F$ -статистика для тестиования  $q$  ограничений (вычисленная с использованием кластеризованной формулы для дисперсии) имеет свое обычное асимптотическое распределение  $F_{q,\infty}$ .

**Почему обычная устойчивая к наличию гетероскедастичности оценка из главы 5 неприменима для панельных данных?** Существуют две причины для этого. Наиболее важной из них является то, что устойчивая к наличию гетероскедастичности оценка из главы 5 не может быть использована при наличии автокорреляции в рамках кластера. Дисперсия для двух случайных переменных  $U$  и  $V$  равна:  $\text{var}(U+V) = \text{var}(U) + \text{var}(V) + 2\text{cov}(U,V)$ . Тогда дисперсия  $\eta_i$  в уравнении (10.24) может быть записана как сумма дисперсий и ковариаций. Пусть  $\tilde{v}_{it} = \tilde{X}_{it}\tilde{u}_{it}$ , тогда

$$\begin{aligned} \text{var}(\eta_i) &= \text{var}\left(\sqrt{\frac{1}{T} \sum_{t=1}^T \tilde{v}_{it}}\right) = \frac{1}{T} \text{var}(\tilde{v}_{i1} + \tilde{v}_{i2} + \dots + \tilde{v}_{iT}) = \\ &= \frac{1}{T} \left[ \text{var}(\tilde{v}_{i1}) + \text{var}(\tilde{v}_{i2}) + \dots + \text{var}(\tilde{v}_{iT}) + 2\text{cov}(\tilde{v}_{i1}, \tilde{v}_{i2}) + \right. \\ &\quad \left. + 2\text{cov}(\tilde{v}_{iT-1}, \tilde{v}_{iT}) \right]. \end{aligned} \quad (10.28)$$

В устойчивой к наличию гетероскедастичности формуле из главы 5 упущены ковариации в последней части уравнения (10.28). Поэтому при наличии автокорреляции такая оценка не будет являться состоятельной.

Вторая причина заключается в том, что если  $T$  не принимает больших значений, то оценка фиксированных эффектов привносит смещение в устойчивую к гетероскедастичности оценку дисперсии из главы 5. Однако этой проблемы не возникает в случае использования межобъектных данных.

Одним из случаев, когда обычные устойчивые к наличию гетероскедастичности стандартные ошибки могут быть использованы для панельных данных, является оценка модели с фиксированными эффектами в случае  $T = 2$ . Тогда оценки в модели с фиксированными эффектами эквивалентны оценкам регрессии в разностях в рамках подхода «сравнение до и после», представленного в разделе 10.2, а обычные устойчивые к наличию гетероскедастичности и кластеризованные стандартные ошибки эквивалентны.

Эмпирические примеры, показывающие важность использования кластеризованных стандартных ошибок в панельных данных, представлены

в работе Бертрана, Дафло и Муллаинатана [Bertrand, Duflo, Mullainathan (2004)].

**Стандартные ошибки в случае наличия корреляции в  $u_{it}$  между субъектами.**

В некоторых случаях  $u_{it}$  могут быть коррелированы между субъектами. Например, при исследовании доходов предположим, что при составлении выборки та или иная семья выбирается случайным образом, но затем выбираются все братья и сестры в семье. Поскольку пропущенные факторы, которые входят в остаточный член, могут иметь общие элементы для братьев и сестер, то нецелесообразно предполагать, что ошибки являются независимыми для братьев и сестер (даже если они являются независимыми между семьями).

В вышеприведенном примере про братьев и сестер каждая семья представляет собой некий естественный кластер (или группу) наблюдений, в рамках которого могут коррелировать  $u_{it}$  (но не между кластерами). Вывод уравнения (10.27) может быть модифицирован таким образом, чтобы допустить наличие кластеров для субъектов (например семей) или для субъектов и времени при наличии большого количества кластеров.

### **Распределение и стандартные ошибки при малых значениях $n$**

Если  $n$  мало и  $T$  велико, то возможность использовать кластеризованные стандартные ошибки по-прежнему сохраняется. Однако  $t$ -статистики необходимо сравнивать с таблицами критических значений  $t_{n-1}$ , а  $F$ -статистику для тестирования  $q$  ограничений необходимо сравнивать с критическим значением  $F_{q,n-q}$ , умноженным на  $(n-1)/(n-q)$ . Эти распределения имеют место в рамках предположений из вставки «Основные понятия 10.3», а также некоторых дополнительных предположений о совместном распределении  $X_{it}$  и  $u_{it}$  в рамках одного субъекта с течением времени. Хотя наличие  $t$ -распределения в регрессии для межобъектных данных предполагает наличие нормальности и гомоскедастичности ошибок регрессии (раздел 5.6), какие-либо требования для использования  $t$ -распределения с кластеризованными стандартными ошибками в панельных данных при больших значениях  $T$  отсутствуют.

Чтобы понять, почему кластеризованные  $t$ -статистики имеют  $t_{n-1}$ -распределение при малых значениях  $n$  и больших значениях  $T$ , даже если  $u_{it}$  не являются нормально распределенными и гомоскедастичными, в первую очередь необходимо помнить о том, что  $T$  принимает большие значения. Тогда при дополнительных предположениях  $\eta_i$  в уравнении (10.24) будут удовлетворять условиям центральной предельной теоремы и, таким образом,  $\eta_i \rightarrow N(0, \sigma_\eta^2)$ . (Дополнительные предположения, необходимые для получения данного результата, являются значительными, но носят в большей степени технический характер. Поэтому они вынесены за пределы текущего раздела для дальнейшего рассмотрения в рамках изучения временных рядов в главе 14.)

Таким образом, если  $T$  принимает большие значения, то  $\sqrt{nT}(\hat{\beta}_1 - \beta_1)$  в уравнении (10.24) является нормированным средним для  $n$  нормально распределенных случайных величин  $\eta_i$ . Кроме того, кластеризованная формула для  $s_\eta^2$  в уравнении (10.27) примет вид обычной формулы для выборочной дисперсии,

и, если она может быть вычислена с помощью  $\eta_i$ , то  $(n-1)s_\eta^2 / \sigma_\eta^2$  будет иметь  $\chi_{n-1}^2$ -распределение. Тогда  $t$ -статистика будет иметь  $t_{n-1}$ -распределение (см. раздел (3.6)). Использование полученных остатков для расчета  $\hat{\eta}_i$  и  $s_\eta^2$  не меняет представленных выводов. В случае нескольких регрессоров аналогичные рассуждения приводят к выводу о том, что  $F$ -статистика для тестирования  $q$  ограничений, вычисленная с использованием кластеризованной формулы для дисперсии, имеет распределение  $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$ . (Например, 5 %-е критическое значение для такой  $F$ -статистики при  $n=10$  и  $q=4$  равно  $\left(\frac{10-1}{10-4}\right) \times 4,53 = 6,80$ , где 4,53 – 5 %-е критическое значение для  $F_{4;6}$ -распределения, представленного в таблице 5Б приложения.) Необходимо отметить, что при увеличении  $n$  распределения  $t_{n-1}$  и  $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$  стремятся к стандартному нормальному и  $F_{q,\infty}$ -распределениям<sup>1</sup>.

Если  $n$  и  $T$  одновременно принимает малые значения, то в общем случае  $\hat{\beta}_1$  не будет иметь стандартного нормального распределения, из-за чего кластеризованные стандартные ошибки не будут давать надежных результатов.

---

<sup>1</sup> Не все программные пакеты позволяют использовать кластеризованные стандартные ошибки с применением распределений  $t_{n-1}$  и  $\left(\frac{n-1}{n-q}\right)F_{q,n-q}$  в случае малых значений  $n$ . Поэтому необходима проверка используемых в ваших программных пакетах формул для кластеризованных стандартных ошибок.

# **Глава 11. Регрессии с бинарными зависимыми переменными<sup>1</sup>**

Предположим, что два схожих внешне человека приходят в банк для получения ссуды, достаточной для того, чтобы каждый из них мог приобрести дом (дома также схожи между собой). Рассматривает ли банк их запросы одинаковым образом? С равной ли вероятностью их заявки могут быть одобрены? По закону они должны получить ссуду с одинаковой вероятностью. Но будет ли это действительно так, в большей степени зависит от банковских работников.

Займы могут быть одобрены (или в выдаче займа может быть отказано) в рамках многих законных оснований. Например, если предлагаемые платежи по кредиту будут «съедать» большую часть или весь ежемесячный доход заявителя, то сотрудник, рассматривающий выдачу ссуд, может обоснованно отказать в ссуде. Кроме того, сотрудники, занимающиеся выдачей ссуд, являются обычными людьми и иногда могут совершать ошибки, поэтому отказ в ссуде для одного заявителя нельзя рассматривать как некое проявление дискриминации. В значительном количестве работ проводился поиск статистических свидетельств наличия дискриминации, то есть доказательств, содержащихся в больших сборниках данных, показывающих, что к различным группам населения относятся по-разному при рассмотрении их заявок на получение ипотечного кредита.

Но как именно следует искать статистические свидетельства наличия дискриминации на рынке ипотечного кредитования? Во-первых, необходимо сравнить доли заявителей с белым цветом кожи (далее – белых) и из представителей национальных меньшинств, которым было отказано в выдаче ипотечных кредитов. В данных, рассмотренных в этой главе, по поданным заявлениям на получение ипотечных кредитов в 1990 году в Бостоне, штат Массачусетс, 28% черных заявителей было отказано в выдаче ипотечного кредита. При этом доля белых заявителей, которым было отказано в выдаче ипотечного кредита, составила лишь 9%. Однако это сравнение в полной мере не отвечает на вопрос, который был задан вначале, поскольку черные заявители и белые заявители не являются в полной мере «одинаковыми в рамках их расовой принадлежности». Поэтому необходимо использовать другой метод, который позволял бы сравнивать количество отказов при сохранении других характеристик заявителей постоянными.

Основываясь на озвученных требованиях, можно предположить, что необходимо использование множественного регрессионного анализа – это действи-

---

<sup>1</sup> В русском языке часто используется название «модель бинарного выбора». – Примеч. науч. ред. перевода.

тельно так, но с некоторыми особенностями. Различия заключаются в том, что зависимая переменная, отражающая факт одобрения заявления на получение кредита, является бинарной. В части II бинарные переменные регулярно использовались в качестве регрессоров, что не вызывало дополнительных сложностей. Но в том случае, когда бинарной является зависимая переменная, все несколько иначе. Не совсем ясно, что означает построение регрессионной линии для зависимой переменной, которая может принимать только два значения – 0 и 1?

Ответ на указанный вопрос заключается в интерпретации функции регрессии как некоей предсказанной вероятности. Эта интерпретация будет рассмотрена в разделе 11.1, где будет показана возможность применения модели множественной регрессии из части II для бинарных зависимых переменных. В разделе 11.1 рассматривается линейная вероятностная модель (linear probability model). Однако интерпретация предсказанной вероятности также предполагает, что альтернативные нелинейные модели регрессии могут быть использованы для лучшего моделирования этих вероятностей. Такие методы, называемые пробит- и логит-регрессиями, будут рассмотрены в разделе 11.2. В дополнительном разделе 11.3 рассматривается метод, используемый для оценки коэффициентов пробит- и логит-моделей регрессии – метод максимального правдоподобия. В разделе 11.4 вышеуказанные методы применяются к данным о выдаче ипотечных кредитов в Бостоне с целью поиска свидетельств наличия расовой дискриминации в ипотечном кредитовании.

Бинарные зависимые переменные, рассматриваемые в этой главе, являются примером зависимой переменной с ограниченным диапазоном значений, то есть *ограниченной зависимой переменной*. Модели для других типов ограниченных зависимых переменных, например зависимых переменных, которые могут принимать диапазон из нескольких дискретных значений, рассматриваются в приложении 11.3.

## 11.1. Бинарные зависимые переменные и линейная вероятностная модель

Одобрение или отклонение заявки на получение ипотечного кредита является одним из примеров бинарной переменной. Многие другие важные вопросы также затрагивают бинарные переменные. Например, каков эффект получения субсидии на обучение на решение человека о поступлении в колледж? Что определяет, начнет ли подросток курить или нет? Каким образом определяется, получит ли страна гуманитарную помощь? Каким образом определяется, принимают ли того или иного соискателя на работу? Во всех этих примерах результат может быть описан с помощью бинарной переменной: студент может пойти или не пойти учиться в колледж, подросток начинает или не начинает курить, страны получают или не получают иностранную гуманитарную помощь, соискатель получит работу или нет.

В данном разделе будут рассмотрены отличия регрессии с бинарной зависимой переменной от регрессии с непрерывной зависимой переменной, а затем

представлена простая модель, в рамках которой могут быть использованы бинарные зависимые переменные, – линейная вероятностная модель.

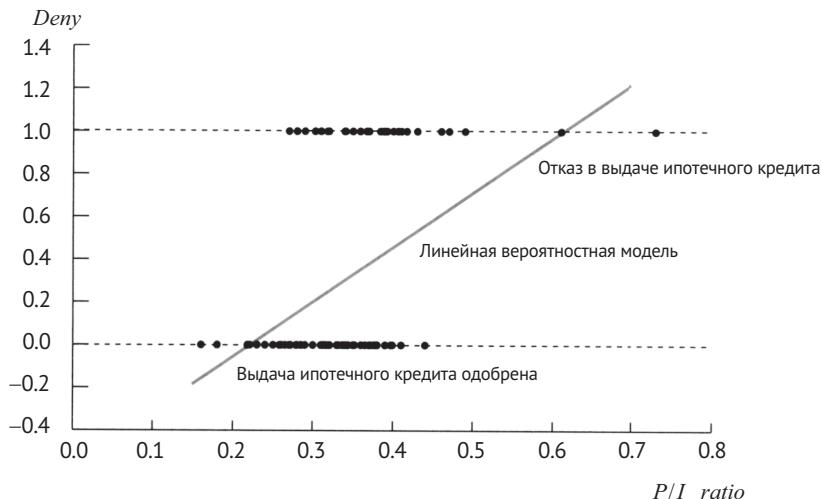
### **Бинарные зависимые переменные**

В данной главе будут рассмотрены практические аспекты выдачи ипотечных кредитов, а именно – оказывает ли расовая принадлежность влияние на принятие решения о выдаче кредита. В данном случае двоичной зависимой переменной является окончательное решение – было ли заявление на выдачу ипотечного кредита одобрено или отклонено. Используемые в этой главе данные являются частью более подробной базы данных, собранной исследователями Федерального резервного банка Бостона (Federal Reserve Bank of Boston) в рамках Закона о раскрытии сведений, касающихся закладных под жилье (Home Mortgage Disclosure Act, HMDA), и представляют собой сведения об ипотечных заявках, поданных в Бостоне, штат Массачусетс, в 1990 году. Описание данных из Бостона в рамках HMDA представлено в приложении 11.1.

Заявления на получение ипотечного кредита достаточно сложны, как и процесс, в рамках которого сотрудник банка, отвечающий за ипотечное кредитование, принимает решение о выдаче кредита. Сотрудник банка должен оценить (спрогнозировать), сможет ли заявитель совершать регулярные платежи для выплаты выданного кредита. Важным аспектом является информация о соотношении размера необходимых платежей по кредиту и доходов заявителя. Это известно любому, кто хотя бы раз брал кредит или ссуду: гораздо проще осуществлять платежи, которые составляют 10% от вашего дохода, платежи, забирающие его половину! Поэтому прежде всего необходимо рассмотреть соотношение между двумя переменными: бинарной зависимой переменной *deny*, которая равна 1, если заявление на выдачу ипотечного кредита было отклонено, и равна 0, если оно было одобрено, и непрерывной переменной *P/I ratio*, которая отражает отношение общих предполагаемых ежемесячных платежей заявителя по кредиту к его или ее ежемесячному доходу.

На рисунке 11.1 представлена диаграмма рассеяния для *deny* и *P/I ratio* для 127 из 2380 наблюдений из используемой выборки (диаграмма рассеяния имеет более понятный вид для указанной подвыборки). Эта диаграмма рассеяния отличается от диаграмм рассеяния, представленных в части II, поскольку переменная *deny* является бинарной. Тем не менее она отражает связь между переменными *deny* и *P/I ratio*. Лишь некоторым заявителям с соотношением платежей к доходам менее 0,3 было отказано в выдаче кредитов, но большинству заявителей с отношением платежей к доходам более 0,4 также было отказано в выдаче кредита.

Эта положительная связь между *deny* и *P/I ratio* (чем выше *P/I ratio*, тем больше доля отказов) показана на рисунке 11.1 в виде МНК-линии регрессии, оцененной на основе этих 127 наблюдений. Эта линия показывает предсказанное значение уровня *deny* как функцию регрессора (т.е. *P/I ratio*). Например, если *P/I ratio*=0,3, то *deny* составляет 0,20. Но что именно означает то, что предсказанное значение бинарной переменной *deny* составляет 0,20?



**Рисунок 11.1. Диаграмма рассеяния отказов в выдаче ипотечных кредитов и отношения месячных платежей к уровню доходов**

Сискатели с высоким отношением платежей к уровню доходов ( $P/I ratio$ ) с большей вероятностью получают отказ в получении ипотечного кредита ( $deny = 1$  в случае отказа,  $deny = 0$  в ином случае). Линейная модель вероятности использует линейную функцию для моделирования вероятности отказа при заданном уровне  $P/I ratio$ .

Ключ к ответу на этот вопрос и, что более важно, к лучшему пониманию модели регрессии с бинарными зависимыми переменными заключается в интерпретации регрессии как моделирования вероятности того, что зависимая переменная равна 1. Таким образом, прогнозируемое значение 0,20 интерпретируется как то, что когда  $P/I ratio$  составляет 0,3, то вероятность отказа в получении кредита оценивается в 20%. Иначе говоря, если было подано значительное количество заявлений на получение кредита с  $P/I ratio = 0,3$ , то 20% из них будет отклонено.

Такая интерпретация следует из двух фактов. Во-первых, в части II было показано, что теоретическая функция регрессии представляет собой ожидаемое значение  $Y$  при заданных значениях регрессоров, то есть  $E(Y|X_1, \dots, X_k)$ . Во-вторых, в разделе 2.2 было показано, что если  $Y$  является бинарной переменной, принимающей значения 0 или 1, то ее ожидаемое (или среднее) значение представляет собой вероятность того, что  $Y = 1$ , то есть  $E(Y) = 0 \times \Pr(Y = 0) + 1 \times \Pr(Y = 1) = \Pr(Y = 1)$ . В контексте регрессии ожидаемое значение условно зависит от значений регрессоров, поэтому вероятность условно зависит от  $X$ . Таким образом, для бинарной переменной  $E(Y|X_1, \dots, X_k) = \Pr(Y = 1|X_1, \dots, X_k)$ . То есть для бинарной зависимой переменной предсказанное значение с помощью теоретической функции регрессии представляет собой вероятность того, что  $Y = 1$  при заданном значении  $X$ .

Линейная модель множественной регрессии в случае бинарной зависимой переменной называется линейной вероятностной моделью: «линейная» — потому что используется линейная модель, «вероятностная модель» — потому что она моделирует вероятность того, что зависимая переменная равна 1, в нашем примере — это вероятность получения отказа в выдаче ипотечного кредита.

## Линейная вероятностная модель

*Линейная вероятностная модель*, по сути, представляет собой модель множественной регрессии из части II в случае, когда зависимая переменная является бинарной, а не непрерывной. Поскольку зависимая переменная  $Y$  является бинарной, теоретическая функция регрессии соответствует вероятности того, что зависимая переменная равна 1 при заданном значении  $X$ . Оценка коэффициента  $\beta_1$  при регрессоре  $X$  представляет собой *изменение вероятности* того, что  $Y=1$  при единичном изменении  $X$ . Аналогично МНК-оценка  $\hat{Y}$ , полученная с помощью оцененной функции регрессии, является оценкой вероятности того, что зависимая переменная равна 1, а МНК-оценка  $\hat{\beta}_1$  оценивает изменение вероятности того, что  $Y=1$  при единичном изменении  $X$ .

Почти все инструменты в части II могут быть использованы в рамках линейной вероятностной модели. Коэффициенты модели могут быть оценены с помощью МНК. Доверительные интервалы (95%-е доверительные интервалы) могут быть построены с помощью  $\pm 1,96$  стандартных ошибок, гипотезы, касательно нескольких коэффициентов могут быть проверены с помощью  $F$ -статистики, рассмотренной в главе 7, а взаимодействия между переменными могут быть смоделированы с помощью методов из раздела 8.3. Поскольку ошибки в линейной вероятностной модели всегда являются гетероскедастичными (см. упражнение 11.8), для формирования статистических выводов важно использование устойчивых к гетероскедастичности стандартных ошибок.

### Линейная вероятностная модель

Линейной вероятностной моделью называется модель множественной регрессии с бинарной зависимой переменной  $Y_i$ :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i. \quad (11.2)$$

Поскольку  $Y$  является бинарной переменной, то  $E(Y|X_1, X_2, \dots, X_k) = \Pr(Y=1|X_1, \dots, X_k)$ , и, таким образом, для линейной вероятностной модели можно записать:

$$\Pr(Y=1|X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}.$$

Регрессионный коэффициент  $\beta_1$  представляет собой изменение вероятности того, что  $Y=1$  при единичном изменении  $X_1$  и фиксированных других регрессорах. Аналогично определяются коэффициенты  $\beta_2, \dots, \beta_k$ . Оценки коэффициентов регрессии могут быть получены с помощью МНК, а стандартные ошибки (устойчивые к наличию гетероскедастичности) могут быть использованы для построения доверительных интервалов и тестирования статистических гипотез.

## ОСНОВНЫЕ ПОНЯТИЯ 11.1

Единственным инструментом, который не может быть использован наравне с остальными, является  $R^2$ . В случае непрерывной зависимой переменной мож-

но представить себе ситуацию, в которой  $R^2$  равен 1 – все данные лежат именно на регрессионной линии. Это невозможно, когда зависимая переменная является бинарной, если только регрессоры также будут бинарными. Соответственно,  $R^2$  не несет какой-либо полезной информации в данной модели. Иные меры оценки качества получаемых оценок (качества «подгонки») будут рассмотрены в следующем разделе. Описание линейной модели вероятности представлено во вставке «Основные понятия 11.1».

**Применение к данным из Бостона в рамках HMDA.** МНК-регрессия, в которой оценивается зависимость бинарной переменной *deny* от отношения ежемесячных платежей к уровню доходов (*P/I ratio*) была оценена с помощью выборки, состоящей из 2380 наблюдений. Результаты оценки могут быть записаны в таком виде:

$$\widehat{deny} = -0,080 + 0,604 P/I \ ratio . \quad (11.1)$$

(0,032)                    (0,098)

Полученная оценка коэффициента при *P/I ratio* положительна и статистически значимо отличается от нуля на уровне значимости 1% (*t*-статистика равна 6,13). Таким образом, для заявителей с более высокой долей долговых платежей в доходах вероятность отказа в получении кредита более значительна. Этот коэффициент может быть использован для расчета прогнозируемого изменения вероятности отказа при заданном изменении регрессора. Например, в соответствии с уравнением (11.1), если *P/I ratio* увеличивается на 0,1, то вероятность отказа в получении кредита возрастает на  $0,604 \times 0,1 \approx 0,060$ , то есть на 6,0 процентных пункта.

Оценка линейной вероятностной модели в уравнении (11.1) может быть использована для вычисления вероятности получения отказа как функции *P/I ratio*. Например, если прогнозируемые выплаты по долгам составляют 30% доходов заявителя, то есть *P/I ratio* составляет 0,3, то прогнозируемое значение из уравнения (11.1) равно  $-0,080 + 0,604 \times 0,3 = 0,101$ . То есть, согласно линейной модели вероятности, для заявителя, прогнозируемые долговые платежи которого составляют 30% дохода, вероятность отклонения заявления на получение ипотечного кредита составляет 10,1%. Этот результат отличается от вероятности 20%, полученной с помощью линии регрессии на рисунке 11.1, потому что регрессионная линия была оценена с использованием только 127 из 2380 наблюдений, используемых для оценки уравнения (11.1).

Каково же влияние расовой принадлежности на вероятность отказа в получении кредита при постоянном уровне *P/I ratio*? Чтобы не усложнять, будем ориентироваться на различия между «черными» и «белыми» (представителями негроидной и европеоидной рас) заявителями. Для оценки эффекта, который оказывает раса при постоянном уровне *P/I ratio*, в уравнение (11.1) будет добавлен бинарный регрессор, который равен 1, если заявителем является черный, и равен 0, если заявителем является белый. Оценка линейной модели вероятности в данной спецификации может быть записана в следующем виде:

$$\widehat{deny} = -0,091 + 0,559 P/I ratio + 0,177 black . \quad (11.3)$$

Оценка коэффициента при переменной *black* (равна 0,177) показывает, что для заявителя, являющегося афроамериканцем, вероятность отказа в получении ипотечного кредита на 17,7% выше, чем для белого заявителя, при одинаковом уровне *P/I ratio*. Этот коэффициент является значимым на уровне значимости 1% (*t*-статистика равна 7,11).

Таким образом, эта оценка показывает, что существует влияние расовых предрассудков на принятие решений в ипотечной сфере, но такой вывод может быть преждевременным. Несмотря на то что показатель отношения платежей к уровню доходов играет важную роль в рамках принятия решения для сотрудника, отвечающего за выдачу ипотечных кредитов, важны и многие другие факторы, такие как потенциальный доход заявителя, а также его кредитная история. Если какая-либо из этих переменных коррелирует с регрессорами *black* или *P/I ratio*, то их отсутствие в уравнении (11.3) приведет к смещениям в оценках, вызванным пропущенными переменными. Таким образом, необходимо воздерживаться от любых выводов о дискриминации в области ипотечного кредитования, пока не будет завершен более тщательный анализ в разделе 11.3.

**Недостатки линейной вероятностной модели.** Линейность, которая делает линейную вероятностную модель простой в применении, является также главным ее недостатком. Поскольку вероятность не может превышать 1, влияние на вероятность того, что  $Y = 1$ , заданного изменения в  $X$ , должно быть нелинейным. Изменение *P/I ratio* с 0,3 до 0,4 может иметь большое влияние на вероятность отказа в получении кредита, но когда *P/I ratio* принимает такие большие значения, что в выдаче кредита, скорее всего, будет отказано, дальнейшее увеличение *P/I ratio* не будет иметь существенного эффекта. В линейной модели вероятности, напротив, влияние данного изменения *P/I ratio* является постоянным, что приводит к тому, что предсказанная вероятность, показанная на рисунке 11.1, падает ниже 0 при низких значениях *P/I ratio* и превышает 1 при больших значениях *P/I ratio*. Такие результаты не представляются разумными, поскольку вероятность не может быть меньше 0 или больше 1. Эти в некоторой степени бессмысленные результаты являются неизбежным следствием модели линейной регрессии. Чтобы решить эту проблему, будут рассмотрены новые нелинейные модели, специально предназначенные для бинарных зависимых переменных, а именно пробит- и логит-модели регрессии.

## 11.2. Пробит- и логит-модели регрессии

Пробит- и логит-модели<sup>1</sup> представляют собой нелинейные регрессионные модели, специально предназначенные для бинарных зависимых переменных. Поскольку регрессии с бинарной зависимой переменной  $Y$  моделируют вероятность того, что  $Y = 1$ , то имеет смысл использовать нелинейную постановку

---

<sup>1</sup> Произносится как про-бит и ло-гит.

модели, которая приводит к тому, что прогнозируемые значения лежат в диапазоне от 0 до 1. Поскольку интегральные функции распределения вероятностей (c.d.f) дают значения вероятностей, расположенные между 0 и 1 (раздел 2.1), то и они используются в логит- и пробит-моделях регрессии. В пробит-модели используется стандартная нормальная функция распределения вероятностей. В логит-модели, которая также называется *логистической регрессией*, используется логистическая функция распределения.

## Пробит-модель

**Пробит-модель с одним регрессором.** Пробит-модель регрессии с единственным регрессором  $X$  может быть записана в таком виде:

$$\Pr(Y = 1 | X) = \Phi(\beta_0 + \beta_1 X), \quad (11.4)$$

где  $\Phi$  представляет собой (интегральную) функцию распределения для стандартного нормального распределения (табулированные значения данной функции распределения представлены в таблице 1 приложения).

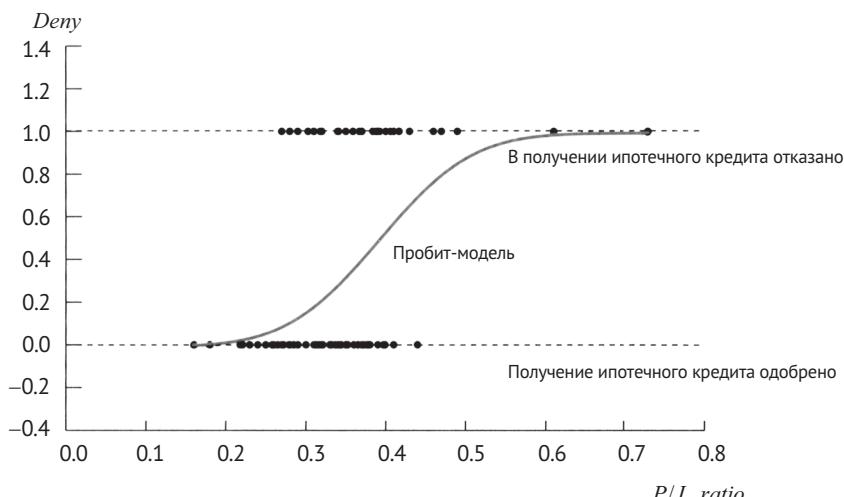
Например, предположим, что  $Y$  является бинарной переменной *deny* (отражает решение о выдаче ипотечного кредита), регрессор  $X$  представляет собой отношение платежей по кредиту к уровню доходов *P/I ratio*,  $\beta_0 = -2$  и  $\beta_1 = 3$ . Какова тогда вероятность получения отказа в выдаче кредита при  $P/I ratio = 0,4$ . В соответствии с уравнением (11.4) эта вероятность равна  $\Phi(\beta_0 + \beta_1 P/I ratio) = \Phi(-2 + 3 \times 0,4) = \Phi(-0,8)$ . В соответствии с табличными значениями стандартной нормальной функции распределения (представлены в таблице 1 приложения)  $\Phi(-0,8) = \Pr(Z \leq -0,8) = 21,2\%$ . То есть если  $P/I ratio = 0,4$ , то предсказанная с помощью пробит-модели с коэффициентами  $\beta_0 = -2$  и  $\beta_1 = 3$  вероятность того, что заявление на получение ипотечного кредита будет отклонено, равна 21,2%.

В пробит-модели слагаемое  $\beta_0 + \beta_1 X$  играет роль « $z$ » в таблице для кумулятивной функции стандартного нормального распределения в таблице 1 приложения. Таким образом, вычисления, представленные в предыдущем параграфе, могут быть выполнены с вычислением « $z$ -значения»,  $z = \beta_0 + \beta_1 X = -2 + 3 \times 0,4 = -0,8$  с последующим поиском значения функции распределения для  $z = -0,8$ , которое равно 21,2%.

Коэффициент  $\beta_1$  пробит-модели, представленной в уравнении (11.4), отражает изменение  $z$ -значения при единичном изменении  $X$ . Если  $\beta_1$  является положительным, то увеличение  $X$  приводит к увеличению  $z$ -значения, увеличивая, таким образом, вероятность того, что  $Y = 1$ . Несмотря на то что влияние  $X$  на  $z$  является линейным, влияние этой переменной на изменение вероятности является нелинейным. Таким образом, наиболее простой практический способ интерпретации коэффициентов в пробит-модели – это вычисление предсказанной вероятности для одного или нескольких значений регрессоров. В случае наличия лишь одного регрессора предсказанная вероятность может быть построена на графике как функция  $X$ .

На рисунке 11.2 представлены (в виде диаграммы рассеяния) оценки регрессионной функции, полученной с помощью регрессии пробит-модели *deny*

от  $P/I ratio$  для 127 наблюдений. Оценка функции регрессии для пробит-модели имеет вид растянутой буквы S. Она практически равна 0 и является «плоской» при малых значениях  $P/I ratio$ , начинает расти для промежуточных значений и затем снова становится более «плоской» при значениях, близких к 1. При малых значениях платежей к доходу вероятность отказа мала. Например, для  $P/I ratio = 0,2$  вероятность отказа, основанная на оценке пробит-модели, представленной на рисунке 11.2, равна:  $\Pr(deny = 1 | P/I ratio = 0,2) = 2,1\%$ . Когда  $P/I ratio = 0,3$ , оценка вероятности отказа составляет 16,1%. Когда  $P/I ratio = 0,4$ , вероятность отказа в получении кредита резко возрастает до 51,9%, а при  $P/I ratio = 0,6$  вероятность отказа равна 98,3%. Согласно этим оценкам пробит-модели для заявителей с высоким отношением платежей к доходу вероятность отказа составляет почти 1.



**Рисунок 11.2. Пробит-модель вероятности отказа в получении ипотечного кредита при заданном уровне  $P/I ratio$**

В пробит-модели используется функция стандартного нормального распределения для оценки вероятности отказа в получении ипотечного кредита при заданном уровне  $P/I ratio$  или (в более общем случае) для моделирования  $\Pr(Y=1|X)$ . В отличие от линейной модели вероятности, условные вероятности в пробит-модели всегда лежат в пределах от 0 до 1.

**Пробит-модель с несколькими регрессорами.** Во всех регрессионных моделях, которые рассматривались до сих пор, пропуск одной из детерминант  $Y$ , коррелированной с включенными в модель регрессорами, приводит к смещениям в оценках, вызванным наличием пропущенных переменных. Пробит-модель не является исключением. В линейной регрессии для решения проблемы необходимо включать в модель дополнительные переменные в качестве регрессоров. Данный подход также является решением в случае пропущенных переменных в пробит-модели.

Пробит-модель с несколькими регрессорами является расширением стандартной модели с одним регрессором с помощью включения регрессоров для вычисления  $z$ -значения. Соответственно, пробит-модель с двумя регрессорами  $X_1$   $X_2$  может быть записана в таком виде:

$$\Pr(Y=1 | X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2). \quad (11.5)$$

Например, предположим, что  $\beta_0 = -1,6$ ,  $\beta_1 = 2$ ,  $\beta_2 = 0,5$ . Тогда, если  $X_1 = 0,4$  и  $X_2 = 1$ , то  $z$ -значение равно  $z = -1,6 + 2 \times 0,4 + 0,5 \times 1 = -0,3$ . Таким образом, вероятность того, что  $Y = 1$  при заданных  $X_1 = 0,4$  и  $X_2 = 1$ , равна  $\Pr(Y = 1 | X_1 = 0,4, X_2 = 1) = \Phi(-0,3) = 38\%$ .

**Влияние изменений  $X$ .** В общем случае влияние изменений  $X$  на  $Y$  представляет собой ожидаемое изменение  $Y$ , следующее из изменений  $X$ . Если  $Y$  является бинарной переменной, ее условное математическое ожидание – это условная вероятность того, что она равна 1, поэтому ожидаемые изменения  $Y$ , вытекающие из изменений  $X$ , представляют собой изменение вероятности того, что  $Y = 1$ . В разделе 8.1 было показано, что в случае когда теоретическая функция регрессии является нелинейной функцией  $X$ , то это ожидаемое изменение может быть оценено в три этапа. Во-первых, необходимо вычислить предсказанное значение при исходном значении  $X$ , используя оценку функции регрессии. Затем необходимо вычислить предсказанное значение для значения  $(X + \Delta X)$ . На последнем шаге необходимо вычислить разницу между двумя предсказанными значениями  $Y$ . Данная процедура резюмируется во вставке «Основные понятия 8.1». Как подчеркивается в разделе 8.1, этот метод всегда работает для вычислений предсказанных последствий изменений  $X$  независимо от того, насколько сложной является нелинейная модель. Применение в рамках пробит-модели метода, представленного во вставке «Основные понятия 8.1», дает оценку эффекта влияния изменения  $X$  на вероятность того, что  $Y = 1$ .

Регрессионная пробит-модель, предсказанные вероятности и оцененные эффекты представлены во вставке «Основные понятия 11.2».

### Пробит-модель, предсказанные вероятности и оцененные эффекты

Теоретическая пробит-модель с несколькими регрессорами может быть записана в таком виде:

$\Pr(Y = 1 | X_1, X_2, \dots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$ . (11.6), где зависимая переменная  $Y$  является бинарной,  $\Phi$  представляет собой стандартную нормальную функцию распределения, а  $X_1, X_2, \dots, X_k$  – регрессоры. Модель может быть наилучшим образом проинтерпретирована с помощью вычисления предсказанных вероятностей и эффектов изменений регрессоров.

Предсказанные вероятности того, что  $Y = 1$  при заданных значениях  $X_1, X_2, \dots, X_k$  вычисляются с помощью расчета  $z$ -значения,  $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ , с последующим поиском соответствующих вероятностей для полученного  $z$ -значения в таблице распределения (таблица 1 приложения).

Коэффициент  $\beta_i$  может рассматриваться как изменение  $z$ -значения при изменении  $X_i$  при заданном постоянном уровне  $X_2, \dots, X_k$ .

Влияние изменений регрессоров на предсказанную вероятность вычисляется с помощью (1) вычисления предсказанной вероятности для начального значения регрессоров, (2) вычисления предсказанной вероятности для новых значений регрессоров, (3) вычисления их разности.

## ОСНОВНЫЕ ПОНЯТИЯ

### 11.2

**Применение к данным по ипотечным кредитам.** В качестве иллюстративного примера описанная выше пробит-модель будет оценена на использовавшейся ранее базе данных по отказам (*deny*) в получении ипотечных кредитов в зависимости от отношения платежей к уровню доходов *P/I ratio*, содержащей 2380 наблюдений:

$$\widehat{\Pr(deny = 1 | P/I ratio)} = \Phi\left(-2,19 + 2,97 P/I ratio\right). \quad (11.7)$$

Полученные оценки коэффициентов,  $-2,19$  и  $2,97$ , достаточно сложно интерпретировать непосредственно, поскольку они оказывают влияние на вероятность отказа в получении кредита через  $z$ -значение. В самом деле, единственное, что можно легко заключить из полученных регрессионных оценок пробит-модели в уравнении (11.7), это то, что переменная, отражающая отношение платежей к уровню доходов, положительно связана с вероятностью отказа в получении кредита (оценка коэффициента при переменной *P/I ratio* является положительной) и что эта зависимость является статистически значимой.

Каково же изменение вероятности того, что заявление будет отклонено, при изменении *P/I ratio* с  $0,3$  до  $0,4$ ? Чтобы ответить на этот вопрос, необходимо следовать процедуре, описанной во вставке «Основные понятия 8.1»: вычислить вероятность отказа для  $P/I ratio = 0,3$ , затем для  $P/I ratio = 0,4$ , а затем вычислить разность. Вероятность отказа в получении кредита при  $P/I ratio = 0,3$  равна:  $\Phi(-2,19 + 2,97 \times 0,3) = \Phi(-1,30) = 0,097$ . Вероятность отказа в получении кредита при  $P/I ratio = 0,4$  равна:  $\Phi(-2,19 + 2,97 \times 0,4) = \Phi(-1,00) = 0,159$ . Оценка изменения вероятности отказа равна  $0,159 - 0,097 = 0,062$ . Таким образом, увеличение *P/I ratio* с  $0,3$  до  $0,4$  приводит к увеличению вероятности отказа на  $6,2\%$ , с  $9,7$  до  $15,9\%$ .

Поскольку регрессионная функция в пробит-модели является нелинейной, то эффект от изменений  $X$  зависит от начального значения  $X$ . Например, если  $P/I ratio = 0,5$ , то предсказанная вероятность отказа в получении кредита, основанная на уравнении (11.7), равна:  $\Phi(-2,19 + 2,97 \times 0,5) = \Phi(-0,71) = 0,239$ . Таким образом, при увеличении *P/I ratio* с  $0,4$  до  $0,5$  предсказанная вероятность отказа в получении кредита увеличивается на  $(0,239 - 0,159)$ , или  $8,0\%$ , что больше, чем на  $6,2\%$  в случае роста *P/I ratio* с  $0,3$  до  $0,4$ .

Каково влияние расовой принадлежности на вероятность отказа в получении ипотечного кредита при постоянном уровне *P/I ratio*? Для оценки этого эффекта проведена оценка пробит-модели с двумя регрессорами *P/I ratio* и *black*:

$$\begin{aligned} \widehat{\Pr(deny = 1 | P/I ratio, black)} &= \\ &= \Phi\left(-2,26 + 2,74 P/I ratio + 0,71 black\right). \end{aligned} \quad (11.8)$$

Опять же, значения коэффициентов трудно интерпретировать, но можно сделать некоторые выводы об их знаках и статистической значимости. Оценка коэффициента при переменной *black* положительна и показывает, что

заявитель – афроамериканец – имеет более высокую вероятность отказа в получении кредита, чем белые заявители, при одинаковом отношении платежей к уровню доходов. Этот коэффициент является статистически значимым на уровне значимости в 1% ( $t$ -статистика для коэффициента при данной переменной равна 8,55). Таким образом, для белого заявителя с  $P/I\ ratio = 0,3$  вероятность отказа в получении ипотечного кредита равна 7,5%, в то время как для черного заявителя с  $P/I\ ratio = 0,3$  этот показатель составляет 23,3%. Разность вероятностей отказа этим двум гипотетическим заявителям составляет 15,8%.

**Оценка коэффициентов в пробит-модели.** Коэффициенты пробит-модели, представленные выше, были получены с помощью метода максимального правдоподобия, который дает эффективные (с минимальной дисперсией) оценки в рамках широкого спектра приложений, в том числе в регрессиях с бинарной зависимой переменной. Оценки максимального правдоподобия являются состоятельными и нормально распределенными в больших выборках, таким образом,  $t$ -статистики и доверительные интервалы для коэффициентов могут быть построены обычным способом.

В эконометрических программных пакетах для оценки пробит-моделей зачастую используется метод максимального правдоподобия, так что вышеописанный метод достаточно просто применить на практике. Стандартные ошибки, полученные при помощи таких программ, могут быть использованы так же, как обычные стандартные ошибки коэффициентов регрессии. Например, 95%-й доверительный интервал для истинного значения коэффициента в пробит-модели может быть построен как  $\pm 1,96 \times$ (стандартная ошибка). Аналогично  $F$ -статистика, полученная с помощью метода максимального правдоподобия, может быть использована для тестирования гипотез о совместной значимости коэффициентов. Метод максимального правдоподобия рассмотрен далее в разделе 11.3, дополнительная информация приведена в приложении 11.2.

### Логит-модель

Логит-модель регрессии с бинарной зависимой переменной  $Y$  и с несколькими регрессорами может быть записана в таком виде:

$$\Pr(Y = 1 | X_1, X_2, \dots, X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k) = \\ = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}. \quad (11.9)$$

Логит-модель похожа на пробит-модель. Различие заключается в том, что в логит-модели используется иная функция распределения.

**ОСНОВНЫЕ ПОНЯТИЯ**

**11.3**

### Логит-модель

Логит-модель похожа на пробит-модель. Различие заключается в том, что в логит-модели вместо стандартной нормальной функции распределения  $\Phi$ ,

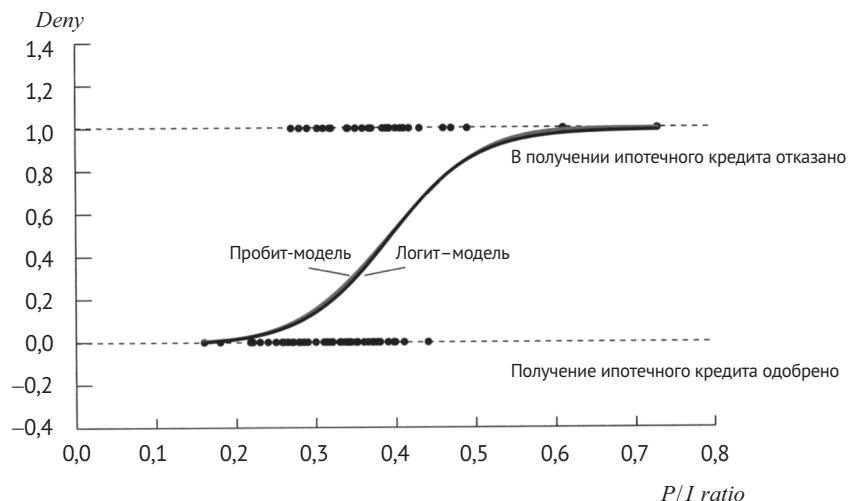
описанной в уравнении (11.6), используется стандартная функция логистического распределения, которая обозначена как  $F$ . Основные положения логит-модели представлены во вставке «Основные понятия 11.3». Логистическая функция распределения имеет особую функциональную форму, описанную в терминах экспоненциальной функции, которая представлена в уравнении (11.9).

Как и в случае пробит-модели, оценки коэффициентов логит-модели проще всего интерпретировать с помощью вычисления предсказанных вероятностей и их разностей.

Оценки коэффициентов логит-модели могут быть получены с помощью метода максимального правдоподобия. Оценки максимального правдоподобия являются состоятельными и нормально распределенными в больших выборках, таким образом,  $t$ -статистики и доверительные интервалы для коэффициентов могут быть построены обычным способом.

Как уже упоминалось ранее, логит- и пробит-модели похожи. Это отчетливо видно на рисунке 11.3, на котором изображены пробит- и логит-регрессионные функции для зависимой переменной *deny* и регрессора *P/I ratio*, полученные с помощью метода максимального правдоподобия с использованием тех же 127 наблюдений, как и для рисунков 11.1 и 11.2. Можно видеть, что различия между этими двумя функциями малы.

Исторически сложилось, что основной мотивацией для использования логит-модели было то, что значения логистической функции распределения могли быть вычислены быстрее, чем значения нормальной функции распределения. С появлением более производительных компьютеров это различие уже перестало быть важным.



**Рисунок 11.3. Пробит-модель и логит-модель для вероятности отказа в получении ипотечного кредита при заданном уровне *P/I ratio***

Представленные пробит- и логит-модели дают практически идентичные результаты для вероятности того, что заявление о выдаче ипотечного кредита будет отклонено при заданном уровне *P/I ratio*.

**Применение к данным из Бостона в рамках HMDA.** Оценка логит-модели для зависимой переменной *deny* и регрессоров *P / I ratio* и *black* из базы данных, содержащей 2380 наблюдений, дает следующий результат:

$$\begin{aligned} \Pr(deny = 1 | P / I ratio, black) &= \\ &= F\left(-4,13 + 5,37 P / I ratio + 1,27 black\right). \end{aligned} \quad (11.10)$$

Оценка коэффициента при переменной *black* положительна и статистически значима на уровне значимости в 1% (*t*-статистика равна 8,47). Предсказанная вероятность отказа в получении кредита для белого заявителя с *P/I ratio* = 0,3 равна  $1/(1+e^{(-4,13+5,37\times0,3+1,27\times0)}) = 1/(1+e^{2,52}) = 0,074$ , или 7,4%. Предсказанная вероятность получения отказа в выдаче кредита для афроамериканца с *P/I ratio* = 0,3 составляет  $1/(1+e^{1,25}) = 0,222$ , или 22,2%. Таким образом, разность в вероятности получения отказа составляет 14,8%.

### **Сравнение линейной вероятностной, пробит- и логит-моделей**

Все три модели – линейная вероятностная модель, пробит- и логит-модели – представляют собой всего лишь приближение (аппроксимацию) неизвестной теоретической функции регрессии  $E(Y|X) = \Pr(Y=1|X)$ . Линейная вероятностная модель является наиболее простой в использовании и интерпретации, но она не может учесть нелинейный характер истинной теоретической функции регрессии. Пробит- и логит-модели регрессии позволяют учесть эту нелинейность в вероятностях, но их регрессионные коэффициенты сложнее интерпретировать. Какую же модель можно использовать на практике?

Единого правильного ответа на поставленный выше вопрос не существует, и разные исследователи используют разные модели. Пробит- и логит-регрессии зачастую приводят к аналогичным результатам. Например, в соответствии с оценкой пробит-модели в уравнении (11.8) разность в вероятности получения отказа для «черного» и «белого» заявителей с *P/I ratio* = 0,3 составляет 15,8%, в то время как логит-оценка этой разности, основанная на уравнении (11.10), составляет 14,9%. Однако с практической точки зрения эти две оценки очень похожи. Одним из способов, который позволяет сделать выбор между логит- и пробит-моделями, является выбор того метода, который проще всего использовать в имеющемся статистическом программном пакете.

Линейная вероятностная модель дает наименее разумное приближение нелинейной теоретической регрессионной функции. Несмотря на это, в некоторых наборах данных может быть несколько экстремальных значений регрессоров, и в этом случае линейная вероятностная модель по-прежнему может давать адекватную степень приближения данных. В регрессии, описываемой уравнением (11.3), оценка разности в вероятности отказа в получении ипотечного

кредита для черного и белого заявителей, полученная с помощью линейной модели вероятности, составляет 17,7 процентных пункта, что несколько больше, чем при оценках с помощью пробит- и логит-моделей, но все же результаты количественно схожи. Единственный способ оценить вышеуказанные различия состоит в том, чтобы оценить предсказанные вероятности с помощью линейной и нелинейных моделей и сравнить различия между ними.

### **11.3. Оценка логит- и пробит-моделей и проверка статистических гипотез<sup>1</sup>**

Нелинейные модели, рассмотренные в разделах 8.2 и 8.3, представляют собой нелинейные функции независимых переменных, но являются линейными функциями от неизвестных коэффициентов (параметров). Следовательно, неизвестные коэффициенты этих нелинейных функций регрессий могут быть оценены с помощью МНК. В отличие от вышесказанного, в пробит- и логит-моделях регрессии функции являются нелинейными функциями коэффициентов. То есть коэффициенты пробит-модели  $\beta_0, \beta_1, \dots, \beta_k$  в уравнении (11.6) фигурируют *внутри* стандартной нормальной функции распределения  $F$ , а коэффициенты логит-модели в уравнении (11.9) фигурируют *внутри* стандартной логистической функции распределения  $F$ . Поскольку теоретическая функция регрессии является нелинейной функцией коэффициентов  $\beta_0, \beta_1, \dots, \beta_k$ , то эти коэффициенты не могут быть оценены с помощью МНК.

Этот раздел представляет собой введение в стандартный метод оценки коэффициентов в пробит- и логит-моделях – метод максимального правдоподобия. Дополнительные математические подробности приведены в приложении 11.2. Поскольку возможность применения метода максимального правдоподобия встроена в большинство современных статистических программ, то получение оценок максимального правдоподобия в пробит-моделях легко осуществимо на практике. С теоретической точки зрения метод максимального правдоподобия, однако, является более сложным, чем метод наименьших квадратов. Поэтому перед переходом к методу максимального правдоподобия будет рассмотрен другой метод получения оценок – нелинейный метод наименьших квадратов.

#### **Нелинейный метод наименьших квадратов**

Нелинейный метод наименьших квадратов является общим методом для оценки неизвестных параметров регрессионной функции в тех случаях, когда эти параметры входят в регрессионную модель нелинейно (как, например, в случае пробит-модели). Оценка нелинейного метода наименьших квадратов, которая была представлена в приложении 8.1, расширяет стандартные МНК-оценки на случай нелинейных по параметрам регрессионных моделей. Как и МНК, не-

---

<sup>1</sup> Данный раздел содержит материал продвинутого уровня, в связи с чем может быть пропущен без потери информации, необходимой для изучения следующих разделов.

линейный метод наименьших квадратов позволяет найти значения параметров, которые минимизируют сумму квадратов ошибок в используемой модели.

Рассмотрим оценки параметров пробит-модели с помощью нелинейного метода наименьших квадратов. Условное математическое ожидание  $Y$  при заданном  $X$  равно:  $E(Y|X_1, \dots, X_k) = \Pr(Y=1|X_1, \dots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)$ . Оценка нелинейным методом наименьших квадратов позволяет с помощью этой функции, которая является нелинейной по параметрам, приблизить значения зависимой переменной. То есть нелинейные оценки наименьших квадратов пробит-коэффициентов представляют собой такие значения  $b_0, \dots, b_k$ , которые минимизируют сумму квадратов ошибок:

$$\sum_{i=1}^n [Y_i - \Phi(b_0 + b_1 X_{1i} + \dots + b_k X_{ki})]^2. \quad (11.11)$$

Оценки, полученные с помощью нелинейного метода наименьших квадратов, сохраняют два ключевых свойства МНК-оценок в линейной регрессии. Во-первых, они являются состоятельными (вероятность того, что полученная оценка близка к истинному значению коэффициента, стремится к 1 при увеличении размера выборки). Во-вторых, оценки имеют нормальное распределение для больших выборок. Тем не менее существуют оценки, которые имеют меньшую дисперсию, чем оценки нелинейного метода наименьших квадратов, то есть оценки нелинейного метода наименьших квадратов не являются эффективными. По этой причине нелинейный метод наименьших квадратов достаточно редко применяется на практике для оценки коэффициентов пробит-модели, вместо него чаще используется метод максимального правдоподобия.

### **Оценки с помощью метода максимального правдоподобия**

**Функция правдоподобия** представляет собой совместную функцию распределения вероятностей данных, рассматриваемую как функцию неизвестных коэффициентов. **Оценки максимального правдоподобия** (ОМП) неизвестных коэффициентов дают значения коэффициентов, которые максимизируют функцию правдоподобия. Поскольку ОМП коэффициентов максимизируют функцию правдоподобия, которая, в свою очередь, является совместной функцией распределения вероятностей, то ОМП позволяют получить значения параметров, которые максимизируют вероятность реализации именно тех данных, которые действительно наблюдаются (содержатся в используемой выборке). В этом смысле ОМП являются значениями параметров, которые с «наибольшей вероятностью» производят имеющиеся данные.

Для иллюстрации получения оценок максимального правдоподобия рассмотрим два независимых одинаково распределенных наблюдения бинарной случайной величины  $Y_1$  и  $Y_2$  без регрессоров. Таким образом,  $Y$  является случайной величиной, распределенной по Бернуlli, и единственным неизвестным параметром, который необходимо оценить, является вероятность  $p$  того, что  $Y=1$ , которая также является средним значением  $Y$ .

Для получения оценок максимального правдоподобия необходимо записать выражение функции правдоподобия, которая, в свою очередь, требует выражения для функции совместного распределения данных. Совместная функция распределения двух наблюдений  $Y_1$  и  $Y_2$  равна:  $\Pr(Y_1 = y_1, Y_2 = y_2)$ . Поскольку  $Y_1$  и  $Y_2$  распределены независимо, то совместное распределение представляет собой произведение отдельных функций распределения для каждой переменной (уравнение (2.23)), то есть  $\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)$ . Распределение Бернулли может быть записано в виде:  $\Pr(Y = y) = p^y(1-p)^{1-y}$ . При  $y=1$   $\Pr(Y=1) = p^1(1-p)^0 = p$ , а если  $y=0$ , то  $\Pr(Y=0) = p^0(1-p)^1 = 1-p$ . Таким образом, совместная функция распределения  $Y_1$  и  $Y_2$  равна:  $\Pr(Y_1 = y_1, Y_2 = y_2) = [p^{y_1}(1-p)^{1-y_1}] \times [p^{y_2}(1-p)^{1-y_2}] = p^{(y_1+y_2)}(1-p)^{2-(y_1+y_2)}$ . Функция правдоподобия представляет собой совместную функцию распределения, рассматриваемую как функцию неизвестных коэффициентов. Для  $n=2$  независимых одинаково распределенных случайных величин, распределенных по Бернулли, функция правдоподобия равна:

$$f(p; Y_1, Y_2) = p^{(Y_1+Y_2)}(1-p)^{2-(Y_1+Y_2)}. \quad (11.12)$$

Оценка максимального правдоподобия  $p$  – это значение  $p$ , которое максимизирует функцию правдоподобия в уравнении (11.12). Как и во всех задачах максимизации или минимизации, это может быть сделано путем проб и ошибок, то есть можно попробовать вычислять значения функции правдоподобия  $f(p; Y_1, Y_2)$  для различных значений  $p$ , пока не будет достигнут максимум функции. В представленном примере максимизация функции правдоподобия дает простую формулу для ОМП:  $\hat{p} = \frac{1}{2}(Y_1 + Y_2)$ . Иными словами, ОМП для  $p$  представляет собой простое среднее арифметическое данных в используемой выборке. В самом деле, в общем случае при наличии  $n$  наблюдений ОМП вероятности  $p$  для случайных величин, распределенных по Бернулли, является выборочное среднее арифметическое  $\hat{p} = \bar{Y}$  (это показано в приложении 11.2). В данном примере ОМП представляет собой обычную оценку  $p$ , которая отражает долю  $Y_i = 1$  в выборке.

Представленный пример похож на задачу оценки неизвестных коэффициентов в пробит- и логит-моделях регрессии. В этих моделях вероятность «успеха»  $p$  не является постоянной, а скорее зависит от  $X$ . То есть это вероятность «успеха» при заданном  $X$  (условно по  $X$ ), которая задана в уравнении (11.6) для пробит-модели и в уравнении (11.9) для логит-модели. Таким образом, функции правдоподобия для пробит- и логит-моделей функции правдоподобия довольно похожи на функцию правдоподобия в уравнении (11.32), кроме того что вероятность «успеха» варьируется от одного наблюдения к другому (поскольку она зависит от  $X_i$ ). Выражения функций правдоподобия для пробит- и логит-моделей приведены в приложении 11.2.

Как и оценки нелинейного метода наименьших квадратов, ОМП являются состоятельными и нормально распределенными в больших выборках. Поскольку эконометрическое программное обеспечение обычно позволяет вычислить ОМП для коэффициентов пробит-модели, то этот метод достаточно просто использовать на практике. Все оценки коэффициентов пробит- и логит-моделей, представленные в данной главе, являются оценками максимального правдоподобия.

**Статистические выводы на основе ОМП.** Поскольку ОМП, как правило, являются нормально распределенными в больших выборках, формирование статистических выводов относительно оценок коэффициентов в пробит- и логит-моделях на основе метода максимального правдоподобия происходит таким же образом, как в случае оценок коэффициентов линейной регрессии на основе МНК. То есть тестирование гипотез проводится с помощью  $t$ -статистик и 95%-х доверительных интервалов, которые могут быть построены с помощью  $\pm 1,96 \times$ (стандартная ошибка). Проверка множественных гипотез о нескольких коэффициентах может использовать  $F$ -статистику в виде, подобном тому, который рассматривался в главе 7 для модели линейной регрессии. Все вышеперечисленное можно сделать аналогично формированию статистических выводов в модели линейной регрессии.

Важной практической особенностью является то, что некоторые статистические программы приводят результаты тестов совместных гипотез, построенных на основе  $F$ -статистики, в то время как другие программы используют хи-квадрат статистики. Статистика хи-квадрат представляет собой  $q \times F$ , где  $q$  – это количество тестируемых ограничений. Поскольку  $F$ -статистики в рамках нулевой гипотезы распределены как  $\chi^2_q / q$  в больших выборках, то  $q \times F$  будет распределена как  $\chi^2_q$  в больших выборках. Поскольку эти два подхода отличаются только лишь множителем  $q$ , то они дают идентичные статистические выводы, однако необходимо понимать, какой именно подход реализован в программном обеспечении, чтобы верно выбрать (или скорректировать) критические значения.

### Меры качества полученных оценок

В разделе 11.1 было отмечено, что  $R^2$  является плохой мерой качества подгонки для линейной вероятностной модели. Это также верно для пробит- и логит-моделей. Существуют две меры, которые пригодны для моделей с бинарной зависимой переменной, – «правильно предсказанная доля наблюдений»<sup>1</sup> и «псевдо- $R^2$ »<sup>2</sup>. Показатель «правильно предсказанной доли наблюдений» строится по следующим правилам: если  $Y_i = 1$  и при этом предсказанная вероятность превышает 50% или если  $Y_i = 0$  и предсказанная вероятность менее 50%, то  $Y_i$  называется правильно предсказанный.

В противном случае говорят, что  $Y_i$  предсказана некорректно. «Правильно предсказанная доля наблюдений» представляет собой долю из  $n$  наблюдений  $Y_1, \dots, Y_n$ , которая была правильно предсказана.

Преимущество этой меры заключается в том, что она достаточна проста для понимания. Недостатком меры является то, что она в достаточной мере не отражает качество подгонки. Например, если  $Y_i = 1$ , то наблюдение рассматривается как правильно предсказанное, если предсказанная вероятность составляет 51% или 90%.

<sup>1</sup> «Fraction correctly predicted». – Примеч. науч. ред. перевода.

<sup>2</sup> «Pseudo- $R^2$ ». – Примеч. науч. ред. перевода.

**Псевдо- $R^2$**  измеряет качество подгонки модели с помощью функции правдоподобия. Поскольку ОМП максимизируют функцию правдоподобия, то включение дополнительного регрессора в пробит- или логит-модели повышает значение уже максимизированной функции правдоподобия, так же как включение дополнительных регрессоров снижает сумму квадратов остатков в модели линейной регрессии, оцененной с помощью МНК. Это говорит в пользу измерения качества «подгонки» пробит-модели путем сравнения значений максимизированной функции правдоподобия со всеми регрессорами со значениями указанной функции без регрессоров. Это то, что на самом деле представляет собой псевдо- $R^2$ . Формула для расчета псевдо- $R^2$  представлена в приложении 11.2.

## 11.4. Применение к данным для Бостона

Регрессии, представленные в двух предыдущих разделах, показали, что вероятность получения отказа в выдаче ипотечного кредита выше для черных, чем для белых заявителей при одинаковом отношении уровня платежей к доходам. Сотрудники, отвечающие за выдачу кредитов, однако, правомерно учитывают множество факторов при принятии решения о возможности выдачи кредита, и если любые из этих других факторов систематически отличаются для людей разных рас, то рассматриваемые до сих пор оценки имеют смещения, вызванные наличием пропущенных переменных.

В данном разделе будет проведено более тщательное изучение вопроса о возможном наличии статистических свидетельств дискриминации в данных для Бостона в рамках HMDA. В частности, основная цель заключается в оценке влияния расовой принадлежности на вероятность отказа в получении кредита при постоянстве тех характеристик, которые сотрудник, отвечающий за выдачу кредита, может учитывать при принятии решения о возможности выдачи кредита.

Наиболее важные переменные, доступные для кредитных специалистов, согласно ипотечным формам в данных для Бостона в рамках HMDA, приведены в таблице 11.1. Именно эти переменные будут играть основную роль в эмпирической модели, описывающей принятие решений в сфере ипотечного кредитования. Первые две переменные являются прямыми мерами финансового бремени, налагаемого на заявителя посредством кредита, измеряемые в терминах его (ее) дохода. Первой из них является *P/I ratio*, вторая переменная представляет собой отношение связанных с жильем расходов к уровню доходов. Следующей переменной является размер кредита по отношению к оценочной стоимости дома: если это соотношение приблизительно равно 1, то у банка могут возникнуть проблемы возмещения полной суммы кредита; если заявитель не может совершать платежи по кредиту, банк будет вынужден изъять заложенное под ипотечный кредит имущество. Последние три финансовые переменные описывают кредитную историю заявителя. Если кредитная история заявителя была ненадежной (заявитель не закрывал взятые на себя долговые обязательства в прошлом), то кредитный специалист может правомерно беспокоиться о способности заявителя или его (ее) желании совершать платежи по кредиту в будущем. Три упомянутые переменные описывают различные типы кредит-

ных историй, каждую из которых кредитный специалист оценивает по-разному. Первая описывает имеющиеся у заявителя кредиты, такие как, например, задолженности на счетах кредитных карт, вторая – историю выплат по прошлым ипотечным кредитам (если таковые были), а третья – существенные проблемы в кредитной сфере, которые были отражены в форме публичных судебных актов, например в случае подачи заявления на банкротство.

В таблице 11.1 также представлены некоторые другие переменные, имеющие отношение к принятию решения для сотрудника, отвечающего за выдачу кредита. Иногда заявитель обязан также подать заявление на обеспечение частного ипотечного страхования<sup>1</sup>. Кредитный специалист будет знать, было ли удовлетворено заявление, и отказ будет оказывать отрицательное влияние на его решение относительно выдачи кредита. Следующие три переменные затрагивают степень занятости, семейное положение и уровень образования заявителя, описывая перспективную возможность заявителя погасить выданный кредит. В случае изъятия банком заложенного под ипотечный кредит имущества характеристики недвижимости также очень важны, поэтому следующая переменная отражает факт нахождения недвижимости в совместной собственности. Две последние переменные в таблице 11.1 отражают расовую принадлежность заявителя (черный или белый) и финальное решение – была ли одобрена выдача ипотечного кредита. В используемой базе данных 14,2% заявителей являются черными, и 12,0% заявлений были отклонены.

Таблица 11.1

**Переменные, включенные в модель, описывающую принятие решений  
в сфере ипотечного кредитования**

Переменная	Определение	Выборочное среднее
<b>Финансовые переменные</b>		
<i>P/I ratio</i>	Отношение ежемесячных платежей по кредиту к месячному уровню дохода	0,331
<i>housing expence – to – income ratio</i>	Отношение ежемесячных расходов на проживание к месячному уровню дохода	0,255
<i>loan – to – value ratio</i>	Отношение размера кредита к размеру залога (оценочной стоимости имеющейся собственности)	0,738
<i>consumer credit score</i>	1 – отсутствие «медленных» (задержка составляет более 30 дней) платежей по кредитам или непогашенных кредитов; 2 – наличие одного или двух «медленных» платежей по кредитам или непогашенных кредитов; 3 – наличие более двух «медленных» платежей; 4 – недостаточная кредитная история; 5 – наличие в кредитной истории невыплаченных кредитов с платежами, просроченными более 60 дней; 6 – наличие в кредитной истории невыплаченных кредитов с платежами, просроченными более 90 дней	2,1

<sup>1</sup> Страхование ипотечных кредитов представляет собой тип страхования, в рамках которого страховая компания осуществляет ежемесячные платежи банку в том случае, если клиент объявляет себя банкротом. В рамках проводимого исследования, если соотношение величины кредита к стоимости дома превышало 80 %, то заявителю, как правило, было необходимо приобретать страховку.

Окончание таблицы 11.1

Переменная	Определение	Выборочное среднее
<i>mortgage credit score</i>	1 – отсутствие задержек по выплатам ипотечных кредитов; 2 – отсутствие ипотечной кредитной истории; 3 – наличие одной или двух задержек в рамках выплат ипотечных кредитов; 4 – наличие более двух задержек в рамках выплат ипотечных кредитов	1,7
<i>public bad credit record</i>	1 – наличие публичных записей кредитных проблем (банкротство, списание кредита как безнадежного, коллекторские действия); 0 – иначе	0,074
<b>Дополнительные характеристики заявителя</b>		
<i>denied mortgage insurance</i>	1 – заявление на приобретение частной страховки отклонено; 0 – иначе	0,020
<i>self-employed</i>	1 – заявитель является самозанятым (имеет собственное предприятие); 0 – иначе	0,116
<i>single</i>	1 – заявитель не состоит в браке; 0 – иначе	0,393
<i>high school diploma</i>	1 – заявитель имеет среднее образование (окончил среднюю школу); 0 – иначе	0,984
<i>unemployment rate</i>	Уровень безработицы в отрасли, где работает заявитель, в штате Массачусетс в 1989 году	3,8
<i>condominium</i>	1 – если объект находится в совместной собственности; 0 – иначе	0,288
<i>black</i>	1 – заявитель является «черным»; 0 – заявитель является «белым»	0,142
<i>deny</i>	1 – заявление на получение ипотечного кредита отклонено; 0 – иначе	0,120

В таблице 11.2 представлены результаты регрессионного анализа на основе вышеописанных переменных. Базовая спецификация, представленная в столбцах (1) – (3), включает в себя финансовые переменные из таблицы 11.1 и переменные, отражающие отказ в приобретении частного ипотечного страхования и наличие у заявителя собственного предприятия (самозанятость). В 1990-х годах кредитные специалисты достаточно часто использовали некие пороговые или граничные значения для переменной, отражающей отношение размера кредита к стоимости залога, поэтому в базовой спецификации для этой переменной используются бинарные переменные для случаев, когда она принимает высокие значения ( $\geq 0,95$ ), средние значения (от 0,8 до 0,95) или низкие значения ( $< 0,8$ ; этот случай опущен, чтобы избежать возникновения совершенной мультиколлинеарности). Регрессоры в первых трех столбцах аналогичны тем, что содержались в базовой спецификации регрессий, построенных исследователями Федерального резервного банка Бостона на первоначальном этапе анализа этих данных<sup>1</sup>. Регрессии в столбцах (1) – (3) различаются лишь в том, как именно

<sup>1</sup> Различия между регрессорами, представленными в столбцах (1) – (3), и регрессорами, представленными в рамках аналогичного исследования в таблице 2 (1) в работе Маннелла и др. [Munnell et al. (1996)], заключают-

моделируется вероятность отказа в получении ипотечного кредита — с помощью линейной модели вероятности, логит-модели и пробит-модели, соответственно.

Таблица 11.2

**Оценка регрессий, описывающих вероятность получения отказа  
в выдаче ипотечного кредита (на основе данных для Бостона в рамках HMDA)**

Регрессионная модель	Зависимая переменная: $deny = 1$ , если в выдаче ипотечного кредита было отказано, $deny = 0$ , если выдача ипотечного кредита была одобрена; 2380 наблюдений						
	Линейная модель вероятностная (1)	Логит-модель (2)	Пробит-модель (3)	Пробит-модель (4)	Пробит-модель (5)	Пробит-модель (6)	
<i>black</i>	0,084** (0,023)	0,688** (0,182)	0,389** (0,098)	0,371** (0,099)	0,363** (0,100)	0,246 (0,448)	
<i>P/I ratio</i>	0,449** (0,114)	4,76** (1,33)	2,44** (0,61)	2,46** (0,60)	2,62** (0,61)	2,57** (0,66)	
<i>housing expence-to-income ratio</i>	-0,048 (,110)	-0,11 (1,29)	-0,18 (0,68)	-0,30 (0,68)	-0,50 (0,70)	-0,54 (0,74)	
<i>medium loan-to-value ratio</i> ( $0,80 \leq loan-to-value ratio \leq 0,95$ )	0,031* (0,013)	0,46** (0,16)	0,21** (0,08)	0,22** (0,08)	0,22** (0,08)	0,22** (0,08)	
<i>high loan-to-value ratio</i> ( $loan-to-value ratio > 1 > 0,95$ )	0,189** (0,050)	1,49** (0,32)	0,79** (0,18)	0,79** (0,18)	0,84** (0,18)	0,79** (0,18)	
<i>consumer credit score</i>	0,031** (0,005)	0,29** (0,04)	0,15** (0,02)	0,16** (0,02)	0,34** (0,11)	0,16** (0,02)	
<i>mortgage credit score</i>	0,021 (0,011)	0,28* (0,14)	0,15* (0,07)	0,11 (0,08)	0,16 (0,10)	0,11 (0,08)	
<i>public bad credit record</i>	0,197** (0,035)	1,23** (0,20)	0,70** (0,12)	0,70** (0,12)	0,72** (0,12)	0,70** (0,12)	
<i>denied mortgage insurance</i>	0,702** (0,045)	4,55** (0,57)	2,56** (0,30)	2,59** (0,29)	2,59** (0,30)	2,59** (0,29)	
<i>self-employed</i>	0,060** (0,021)	0,67** (0,21)	0,36** (0,11)	0,35** (0,11)	0,34** (0,11)	0,35** (0,11)	
<i>single</i>				0,23** (0,08)	0,23** (0,08)	0,23** (0,08)	
<i>high school diploma</i>				-0,61** (0,23)	-0,60* (0,24)	-0,62** (0,23)	
<i>unemployment rate</i>				0,03 (0,02)	0,03 (0,02)	0,03 (0,02)	
<i>condominium</i>					-0,05 (0,09)		

ся в том, что в данной работе используются дополнительные индикаторы для местоположения дома и определения личности заемщика, которые не содержатся в открытом доступе; индикатор для многоквартирного дома, который здесь неуместен, поскольку наше исследование фокусируется на домах для одной семьи; переменная, отражающая размер чистых активов, опущена в нашем исследовании, потому что в используемой выборке имеется несколько очень больших положительных и отрицательных значений и, таким образом, это приводит к возникновению риска возникновения зависимости результатов от нескольких конкретных наблюдений-выбросов.

Окончание таблицы 11.2

Регрессионная модель	Линейная модель вероятностная (1)	Логит-модель (2)	Пробит-модель (3)	Пробит-модель (4)	Пробит-модель (5)	Пробит-модель (6)
<i>black × P/I ratio</i>						-0,58 (1,47)
<i>black × housing expense-to-income ratio</i>						1,23 (1,69)
<i>additional credit rating indicator variables</i>	нет	нет	нет	нет	да	нет
<i>constant</i>	-0,183** (0,028)	-5,71** (0,48)	-3,04** (0,23)	-2,57** (0,34)	-2,90** (0,39)	-2,54** (0,35)
<i>F</i> -статистики и <i>p</i> -значения для тестирования исключения групп переменных						
<i>applicant single; high school diploma; industry unemployment rate</i>				5,85 (<0,001)	5,22 (0,001)	5,79 (<0,001)
<i>additional credit rating indicator variables</i>					1,22 (0,291)	
<i>race interactions and black</i>						4,96 (0,002)
<i>race interactions only</i>						0,27 (0,766)
<i>difference in predicted probability of denial, white vs. black (в %)</i>	8,4%	6,0%	7,1%	6,6%	6,3%	6,5%

Примечание. Представленные регрессии были оценены с использованием  $n=2380$  наблюдений в рамках базы данных, описанной в приложении 11.1. Линейная вероятная модель оценивается с помощью МНК, пробит- и логит-модели были оценены с помощью метода максимального правдоподобия. Стандартные ошибки приведены в скобках под оценками коэффициентов, *p*-значения приведены в скобках под *F*-статистиками. Переменная, отражающая изменение предсказанной вероятности, представленная в последней строке, была вычислена для гипотетического заявителя, для которого значения регрессоров (за исключением расовой принадлежности) равны средним выборочным значениям. Отдельные оценки коэффициентов являются статистически значимыми на уровне значимости \*5% и \*\*1%.

Поскольку столбец (1) соответствует линейной вероятностной модели, то оценки коэффициентов представляют собой оценки изменений предсказанной вероятности при единичном изменении независимой переменной. Соответственно, оценки показывают, что увеличение *P/I ratio* на 0,1 приводит к увеличению вероятности отказа на 4,5% (оценка коэффициента при *P/I ratio* в колонке (1) составляет 0,449, а  $0,449 \times 0,1 \cong 0,045$ ). Аналогично при высоком уровне отношения размера кредита к стоимости залога увеличивается вероятность получения отказа: если отношение размера кредита к стоимости залога превышает 95%, то это приводит к приросту вероятности получения отказа в выдаче кредита на 18,9% (коэффициент равен 0,189) по сравнению со случаем, когда отношение размера кредита к стоимости залога составляет менее 80% при постоянных уровнях остальных переменных, представленных в столбце (1).

Кандидатам с плохим кредитным рейтингом также сложнее получить ипотечный кредит при прочих равных условиях, несмотря на то что коэффициент при переменной, отражающей потребительское кредитование, является статистически значимым, а коэффициент при переменной, отражающей ипотечное кредитование, не является статистически значимым. Кандидаты, у которых имеются публичные записи, свидетельствующие о наличии кредитных проблем, таких как процедура банкротства, сталкиваются с гораздо большими трудностями при получении кредита: при прочих равных условиях, согласно полученным оценкам, плохая кредитная история увеличивает вероятность получения отказа на 0,197 (или на 19,7%). Оценки показывают, что переменная, отражающая результат покупки частной страховки, оказывает решающее влияние на вероятность получения ипотечного кредита: оценка коэффициента составляет 0,702, что означает, что получение отказа в приобретении частной страховки увеличивает вероятность получения отказа в выдаче ипотечного кредита на 70,2% при прочих равных условиях. Из девяти переменных (кроме переменной, отражающей расовую принадлежность), использованных в регрессии, оценки коэффициентов практически для всех из них (кроме двух) являются статистически значимыми на уровне значимости 5%, что согласуется с гипотезой о том, что сотрудники, рассматривающие заявления на получение ипотечных кредитов, принимают решение на основе многих факторов.

Оценка коэффициента при переменной *black* в регрессии (1) равна 0,084 и свидетельствует о том, что разница в вероятности отказа для черных и белых заявителей составляет на 8,4% при фиксированных остальных переменных в регрессии. Эта оценка является статистически значимой на 1%-м уровне значимости.

Оценки логит- и пробит-моделей, представленные в столбцах (2) и (3), позволяют сделать схожие выводы. В логит-и пробит-регрессионных моделях восемь из девяти коэффициентов при переменных (кроме переменных, отражающих расовую принадлежность) статистически значимо отличаются от нуля на уровне значимости 5%, а коэффициент при переменной *black* является статистически значимым на уровне 1%. Как отмечалось ранее в разделе 11.2, поскольку эти модели являются нелинейными, то для вычисления разницы в оценке вероятности получения отказа для «белых» и «черных» заявителей необходимо использовать специфические значения всех регрессоров. Как правило, для этого рассматривают «усредненного» заявителя, значения всех регрессоров (кроме переменной, отражающей расовую принадлежность) для которого принимают равным их средним выборочным (средние для используемой выборки) значениям. В последней строке таблицы 11.2 представлена оценка разности вероятности получения отказа в выдаче кредита для этого «усредненного» заявителя. Оценки расходления вероятности получения кредита, вызванные расовой принадлежностью, достаточно схожи друг с другом: 8,4% для линейной модели вероятности [столбец (1)], 6,0% для логит-модели [столбец (2)] и 7,1% для пробит-модели [столбец (3)]. Эти оценочные эффекты, отражающие расовую

принадлежность, и оценки коэффициентов при переменной *black* меньше, чем в регрессиях из предыдущих разделов, в которых в качестве регрессоров использовались только *P/I ratio* и *black*, что говорит о наличии в последних смещений, вызванных пропущенными переменными.

Регрессии в столбцах (4) – (6) показывают чувствительность результатов, представленных в столбце (3), к изменениям в спецификации регрессий. Столбец (4) отличается от столбца (3) включением в модель дополнительных характеристик заявителя. Эти характеристики помогают предсказать, будет ли отказано в выдаче кредита. Например, наличие по меньшей мере диплома о среднем образовании снижает вероятность отказа (оценка отрицательна, коэффициент является статистически значимым на уровне 1%). Тем не менее включение в модель этих индивидуальных характеристик не меняет оценку коэффициента при переменной *black* или оценку различия в вероятности отказа в выдаче кредита (6,6%).

Столбец (5) показывает оценки для шести категорий потребительских кредитов и четырех категорий ипотечных кредитов для проверки нулевой гипотезы о том, что эти две переменные входят линейно. В эту регрессию также включена переменная, указывающая на то, является ли залоговая собственность кондоминиумом. Нулевая гипотеза о том, что переменная кредитного рейтинга входит в пробит-модель линейно, не отвергается, как и гипотеза о том, что бинарная переменная, характеризующая кондоминиум, является значимой на уровне значимости 5%. Самое главное заключается в том, что оценка разности в вероятности выдачи кредита (6,3%), вызванной расовыми различиями, практически та же, что и в столбцах (3) и (4).

Столбец (6) анализирует наличие различий в стандартных подходах, применяемых к оценке отношения ежемесячного платежа к уровню дохода (*payment-to-income ratio*) и отношения величины жилищных расходов к доходу (*housing expense-to-income ratio*) для черных заявителей по сравнению с белыми заявителями. Результаты, представленные в столбце (6), показывают отсутствие таких различий: переменные не являются совместно статистически значимыми на уровне 5%. Тем не менее переменная расовой принадлежности по-прежнему имеет значительный эффект, поскольку является совместно статистически значимой с остальными переменными на уровне значимости 1%. Опять же, оценка различий вероятности отказа в выдаче кредита (6,5%), вызванных расовыми различиями, не отличается значительно от результатов в других пробит-регрессиях.

Во всех шести спецификациях влияние расовой принадлежности на вероятность отказа в получении кредита при фиксированных значениях прочих характеристик заявителя является статистически значимым на уровне 1%. Оценка различия в вероятности отказа для черного и белого заявителей колеблется от 6,0 до 8,4%. Один из способов оценить, насколько эта разница велика или мала, заключается в том, чтобы вернуться к вопросу, поставленному в начале этой главы. Предположим, два человека подают заявления на получение ипотечного кредита, один белый и один черный, но в остальном с одинаковыми значениями других независимых переменных в регрессии (3), а именно – кро-

ме расовой принадлежности значения других переменных в регрессии (3) равны выборочному среднему на множестве данных HMDA. Белый заявитель сталкивается с вероятностью отказа в 7,4%, а черный заявитель – с 14,5%-й вероятностью отказа. Это предполагаемое различие в вероятности отказа в получении кредита величиной 7,1%, вызванное расовой принадлежностью, означает, что для черного заявителя почти в 2 раза больше шансов получить отказ, чем для белого заявителя.

Результаты в таблице 11.2 (и в оригинальном исследовании ФРС Бостона) представляют собой статистические доказательства наличия расовой дискриминации в области ипотечного кредитования, чего не должно быть по закону. Эти результаты играют важную роль в стимулировании изменений в политике банковских регуляторов<sup>1</sup>. Однако экономисты любят спорить, и не удивительно, что эти результаты также стимулировали бурные дебаты.

Поскольку было озвучено предположение о том, что в кредитовании существует (или существовала) расовая дискриминация, то необходимо кратко рассмотреть некоторые моменты вышеупомянутой дискуссии. При этом для удобства стоит использовать представленную ранее в рамках главы 9 структуру, то есть рассмотреть внутреннюю и внешнюю обоснованность результатов из таблицы 11.2, которые представляют собой описание предыдущих исследований бостонских данных HMDA. Ряд критических замечаний, представленных в рамках исходного исследования, проведенного Федеральным резервным банком Бостона, затрагивает вопросы внутренней обоснованности: возможные ошибки в данных, альтернативные нелинейные функциональные формы, дополнительные взаимодействия между переменными и так далее. Исходные данные были подвергнуты тщательной проверке, некоторые ошибки были найдены, а представленные здесь результаты (а также и в конечном варианте опубликованного исследования ФРС Бостона) основаны на «очищенном» наборе данных. Оценка других спецификаций, различных функциональных форм и / или включение дополнительных регрессоров также дают оценку различий, вызванных расовой принадлежностью, сопоставимых с таковыми в таблице 11.2. Потенциально более сложная проблема внутренней обоснованности заключается в возможности существования несвязанной с расовой принадлежностью финансовой информации, полученной в ходе очных интервью перед получением кредита и незарегистрированной в самом заявлении на получение кредита, которая коррелирует с расовой принадлежностью. Если такая информация существует, результаты, представленные в таблице 11.2, могут иметь смещения, вызванные наличием пропущенных переменных. Наконец, некоторые ставят под сомнение внешнюю обоснованность: даже если расовая дискриминация в Бостоне в 1990 году имела место, неверным было бы считать, что аналогичная ситуация имеет место повсюду в настоящее время. Кроме того, проявление расовой дискриминации менее

---

<sup>1</sup> Такие изменения в политике включают в себя изменения в направлении более справедливых и честных механизмов предоставления кредитования, реализованные федеральными банковскими регуляторами, изменения в запросах, реализованные Министерством юстиции США, а также расширение образовательных программ для банков и других компаний, предоставляющих кредиты на покупку недвижимости.

вероятно при использовании современных интернет-приложений, поскольку заявление на получение ипотечного кредита может быть одобрено или отклонено без непосредственного личного контакта. Единственный способ решить вопрос о внешней обоснованности заключается в рассмотрении данных для других городов/штатов и за иные периоды времени<sup>1</sup>.

## 11.5. Заключение

В случае когда зависимая переменная  $Y$  является бинарной, теоретическая функция регрессии представляет собой вероятность того, что  $Y=1$  при фиксированных значениях прочих регрессоров. Оценка этой функции регрессии влечет за собой нахождение функциональной формы, что оправдывает ее вероятностную интерпретацию, оценки неизвестных параметров этой функции и интерпретацию результатов. Полученные предсказанные значения представляют собой вероятностные оценки, а оценка эффекта от изменения регрессора  $X$  может быть оценена как изменение вероятности того, что  $Y=1$ , вызванное изменением  $X$ .

Естественным способом моделирования вероятности  $Y=1$  при заданных значениях регрессоров является использование функции распределения, где аргументы функции распределения зависят от регрессоров. Пробит-модели используют нормальные функции распределения в качестве регрессионных функций, а логит-модели используют логистические функции распределения. Поскольку данные модели можно рассматривать в качестве нелинейных функций от неизвестных параметров, то эти параметры оценить сложнее, нежели получить оценки коэффициентов линейной регрессии. Стандартным методом получения оценок является метод максимального правдоподобия. На практике формирование статистических выводов на основе оценок максимального правдоподобия происходит так же, как и в случае линейной множественной регрессии. Например, 95%-е доверительные интервалы для коэффициентов могут быть построены как оценка коэффициента  $\pm 1,96$  стандартной ошибки.

Несмотря на свою «внутреннюю» нелинейность, иногда теоретическая функция регрессии может быть достаточно хорошо аппроксимирована с помощью линейной модели вероятности, то есть фактически с помощью линейной множественной регрессии. Линейная вероятностная модель, пробит-модель и логит-модель – все они дают схожие «итоговые» результаты при их использовании для данных HMDA для Бостона. Все три метода показывают наличие существенных различий в уровнях отказа в получении ипотечных кредитов для заявителей с различной расовой принадлежностью.

---

<sup>1</sup> Хорошим началом для дальнейшего ознакомления с данной темой является изучение вопросов симпозиума по вопросам расовой дискриминации и экономики, прошедшего весной 1998 года, представленных в номере *Journal of Economic Perspectives*. В работе Ладда (Ladd, 1998) освещены основные свидетельства и вопросы расовой дискриминации в ипотечном кредитовании. Более подробное рассмотрение дано в работе Геринга и Винка (Goering, Wienk, 1996). Ипотечный рынок США значительно изменился со временем проведения исследования ФРС Бостона, изменения затронули в том числе ослабление стандартов кредитования, существование пузыря цен на жилье, финансовый кризис 2008–2009 годов, возвращение к ужесточению стандартов кредитования. Для ознакомления с изменениями, произошедшими на рынках ипотечного кредитования, см. Green, Wachter, 2007.

Бинарные зависимые переменные являются наиболее распространенным примером ограниченных зависимых переменных, то есть зависимых переменных с ограниченным набором значений. В последней четверти XX века были сделаны важные достижения в области эконометрических методов анализа других видов ограниченных зависимых переменных (см. вставку «Нобелевские лауреаты Джеймс Хекман и Дэниел Макфадден»). Некоторые из этих методов рассматриваются в приложении 11.3.



### ***Нобелевские лауреаты Джеймс Хекман и Дэниел Макфадден***

В 2000 году Нобелевская премия по экономике была присуждена двум эконометристам, Джеймсу Дж. Хекману (James J. Heckman) из Чикагского университета и Дэниелу Л. Макфаддену (Daniel L. McFadden) из Университета Калифорнии в Беркли за фундаментальный вклад в анализ данных для индивидов и фирм. Значительная часть их работы посвящена трудностям, которые возникают при использовании ограниченных зависимых переменных. Хекман был удостоен премии за разработку инструментов, учитывающих выборочный отбор. Как обсуждалось ранее в разделе 9.2, смещения в получаемых оценках, вызванные особенностями формирования выборки (смещение из-за отбора наблюдений), проявляются тогда, когда имеющиеся в наличии данные находятся под влиянием процесса отбора и при этом связаны со значениями зависимой переменной. Например, предположим, что есть необходимость оценить взаимосвязь между доходами и некоторым регрессором  $X$ , используя случайную выборку из генеральной совокупности. Если вы оцениваете регрессии с использованием подвыборки занятых работников, то есть тех, кто в отчетности указывает положительный уровень заработка, то оценка МНК может быть смещена из-за специфики отобранных данных. Решение, предложенное Хекманом, состояло в том, чтобы задать предварительное уравнение с бинарной зависимой переменной, описывающее, является ли работник трудоустроенным («в» или «вне» подвыборки), а затем рассматривать это уравнение и уравнение, описывающее уровень доходов, в качестве системы одновременных уравнений. Этот подход был использован для решения аналогичных проблем, возникающих при формировании выборки во многих областях, начиная от экономики труда и теории отраслевых рынков и заканчивая сферой финансов.

Макфадден был удостоен премии за разработку моделей для анализа дискретных данных, сформированных на основе осуществления выбора (какое решение принимает выпускник средней школы: идет служить в армию, поступает в колледж или устраивается на работу). Он начал с рассмотрения задачи отдельных индивидов по максимизации ожидаемой полезности при каждом из возможных вариантов выбора, что может зависеть от наблюдаемых переменных (таких как заработка плата, характеристики работы, семейного положения). Затем он сформировал модели для индивидуальных вероятностей выбора с неизвестными коэффициентами, которые, в свою очередь, могут быть оценены с помощью метода максимального правдоподобия. Эти модели и их развитие оказались очень полезными при анализе дискретных данных во многих областях, в том числе в сфере экономики труда, экономики здравоохранения, экономики транспорта. Для получения дополнительной информации

касательно вышеупомянутых и других лауреатов Нобелевской премии в области экономики можно посетить веб-сайт Нобелевского фонда: [www.nobel.se/economics](http://www.nobel.se/economics)



## **Выходы**

1. В случае когда  $Y$  является бинарной переменной, линейная модель множественной регрессии называется линейной вероятностной моделью. Линия теоретической регрессии показывает вероятность того, что  $Y = 1$  при заданных значениях регрессоров  $X_1, X_2, \dots, X_k$ .
2. Пробит- и логит-модели регрессии представляют собой нелинейные регрессионные модели, используемые в том случае, когда  $Y$  является бинарной переменной. В отличие от линейной вероятностной модели, в пробит и логит-моделях предсказанная вероятность того, что  $Y = 1$ , находится между 0 и 1 для всех значений  $X$ .
3. В пробит-модели используется стандартная нормальная функция распределения. В логит-модели используется логистическая функция распределения. Коэффициенты в логит и пробит-моделях оцениваются с помощью метода максимального правдоподобия.
4. Значения коэффициентов в пробит и логит-моделях сложно интерпретировать прямо. Изменение вероятности того, что  $Y = 1$ , связанное с изменением одного или нескольких регрессоров, может быть вычислено с помощью общей процедуры для нелинейных моделей, представленной во вставке «Основные понятия 8.1».
5. Тестирование гипотез относительно коэффициентов в линейной вероятностной модели, логит и пробит-моделях проводится с помощью стандартных  $t$ - и  $F$ -статистик.

## **Основные понятия**

- Ограниченные зависимые переменные (с. 399).  
Линейная вероятностная модель (с. 402).  
Пробит-модель (с. 404).  
Логит-модель (с. 404).  
Логистическая регрессия (с. 405).  
Функция правдоподобия (с. 413).  
Оценка максимального правдоподобия (с. 413).  
Правильно предсказанная доля наблюдений (с. 415).  
Псевдо- $R^2$  (с. 416).

## **Вопросы для повторения и закрепления основных понятий**

- 11.1. Предположим, что линейная вероятностная модель дает предсказанное значение  $Y$ , равное 1,3. Поясните, почему данное значение не имеет смысла.

- 11.2. В таблице 11.2 оценка коэффициента при переменной *black* равна 0,084 в столбце (1), 0,688 в столбце (2) и 0,389 в столбце (3). Несмотря на эти существенные различия, все модели дают схожие оценки предельной величины влияния расовой принадлежности на вероятность получения отказа в выдаче кредита. Поясните данные результаты.
- 11.3. Один из ваших друзей использует микроданные для изучения факторов, влияющих на курение в вашем университете. Какую модель необходимо использовать – пробит-модель, логит-модель или линейную вероятностную модель? Поясните свой выбор.
- 11.4. Почему коэффициенты в пробит- и логит-модели оцениваются с помощью метода максимального правдоподобия вместо МНК?

### Упражнения

Упражнения 11.1–11.5 основываются на следующих предположениях. Случайным образом были опрошены четыреста претендентов на получение водительского удостоверения. Каждому из них был задан вопрос о том, сдали ли они свой экзамен на получение водительского удостоверения ( $Pass_i = 1$ ) или провалили ( $Pass_i = 0$ ). Также дополнительно были собраны данные о половой принадлежности респондентов ( $Male_i = 1$ , если респондент является мужчиной, и  $Male_i = 0$ , если респондент является женщиной) и их водительском опыте ( $Experience_i$  – водительский стаж в годах). В приведенной ниже таблице представлены результаты оценок нескольких моделей.

Зависимая переменная – «Тест сдан»							
	Пробит (1)	Логит (2)	Линейная вероятностная модель (3)	Пробит (4)	Логит (5)	Линейная вероятностная модель (6)	Пробит (7)
<i>Experience</i>	0,031 (0,009)	0,040 (0,016)	0,006 (0,002)				0,041 (0,156)
<i>Male</i>				-0,333 (0,161)	-0,622 (0,303)	-0,071 (0,034)	-0,174 (0,259)
<i>Male</i> × <i>Experience</i>							-0,015 (0,019)
<i>Constant</i>	0,712 (0,126)	1,059 (0,221)	0,774 (0,034)	1,282 (0,124)	2,197 (0,242)	0,900 (0,022)	0,806 (0,200)

- 11.1. Используя результаты из столбца (1), ответьте на вопросы:

- Зависит ли вероятность успешной сдачи теста от *Experience*? Поясните свой ответ.
- Водительский стаж Мэтью составляет 10 лет. Какова вероятность того, что он успешно сдаст тест?
- Кристофер – начинающий водитель (не имеет опыта вождения). Какова вероятность того, что он сдаст тест?
- Выборка содержала значения *Experience* в интервале от 0 до 40 лет, но только четыре человека в выборке имели стаж вождения более 30 лет.

Джеду 95 лет, он водит автомобиль с тех пор, как ему исполнилось 15 лет. Какова, согласно модели, вероятность того, что Джед успешно сдаст тест? Считаете ли Вы, что это значение является надежным? Поясните свой ответ.

- 11.2. Используя таблицу, представленную выше, ответьте на следующие вопросы:
- Ответьте на пункты (а) – (в) упражнения 11.1, используя результаты, представленные в столбце (2);
  - Выпишите предсказанные вероятности на основе пробит- и логит-моделей из столбцов (1) и (2) для значений *Experience* в интервале от 0 до 60 лет. Похожи ли результаты, полученные на основе оценок пробит- и логит-моделей?
- 11.3. а) Ответьте на пункты (а) – (с) из упражнения 11.1, используя результаты столбца (3).
- Выпишите предсказанные вероятности на основе пробит- и логит-моделей из столбцов (1) и (3) как функции от *Experience* для значений *Experience* в интервале от 0 до 60 лет. Как вы считаете, можно ли использовать в данном случае линейную вероятностную модель? Поясните свой ответ.
- 11.4. Используя результаты столбцов (4) – (6):
- Вычислите оценки вероятностей успешной сдачи теста для мужчины и для женщины.
  - Различаются ли модели в столбцах (4) – (6)? Поясните свой ответ.
- 11.7. Используя результаты столбца (7):
- Акира – мужчина с 10-летним стажем вождения. Какова вероятность того, что он успешно сдаст тест?
  - Джейн – женщина с двухлетним стажем вождения. Какова вероятность того, что она успешно сдаст тест?
  - Зависит ли влияние водительского стажа на прохождение теста от пола испытуемого? Поясните свой ответ.
- 11.6. Используя оценки пробит-модели, представленной в уравнении (11.9), ответьте на следующие вопросы:
- Афроамериканец с  $P/I\ ratio = 0,35$  подает заявление на получение кредита. Какова вероятность того, что заявление будет отклонено?
  - Предположим, что заявитель снизил свой показатель  $P/I\ ratio$  до 0,30. Какой эффект это окажет на вероятность отказа в получении кредита?
  - Ответьте еще раз на вопросы из пунктов (а) и (б) для «белокожего» заявителя.
  - Зависит ли предельный эффект изменения  $P/I\ ratio$  на вероятность отказа в получении кредита от расовой принадлежности? Поясните свой ответ.
- 11.7. Повторите упражнение 11.6, используя логит-модель, представленную в уравнении (11.10). Однаковы ли результаты для логит- и пробит-моделей? Поясните свой ответ.

- 11.8. Рассмотрите линейную вероятностную модель  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , где  $\Pr(Y_i = 1 | X_i) = \beta_0 + \beta_1 X_i$ .
- Покажите, что  $E(u_i | X_i) = 0$ .
  - Покажите, что  $\text{var}(u_i | X_i) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]$ . (Подсказка: рассмотрите еще раз уравнение (2.7).)
  - Являются ли  $u_i$  гетероскедастичными? Поясните свой ответ.
  - Запишите функцию правдоподобия (необходимо рассмотреть раздел 11.3).
- 11.9. Используя оценки линейной вероятностной модели, представленные в столбце (1) таблицы 11.2, ответьте на следующие вопросы:
- Два заявителя, темнокожий и светлокожий, подали заявление на получение кредита. Они имеют одинаковые значения всех регрессоров, кроме расовой принадлежности. Насколько различаются вероятности отказа в получении кредита для данных заявителей?
  - Постройте 95%-й доверительный интервал для ответа, полученного в пункте (a).
  - Какая переменная может приводить к возникновению смещений в оценках, вызванных пропущенными переменными, в пункте (a)? Насколько сильным может быть смещение?
- 11.10. (Необходимо изучить материал раздела 11.3 и использовать дифференциальное исчисление.) Предположим, что случайная величина  $Y$  имеет распределение  $\Pr(Y = 1) = p$ ,  $\Pr(Y = 2) = q$  и  $\Pr(Y = 3) = 1 - p - q$ . Случайная выборка размера  $n$  формируется из генеральной совокупности с вышеуказанным распределением, случайные величины обозначаются как  $Y_1, Y_2, \dots, Y_n$ .
- Запишите функцию правдоподобия для параметров  $p$  и  $q$ .
  - Постройте формулы оценок максимального правдоподобия (ОМП) для параметров  $p$  и  $q$ .
- 11.11. (Требуется изучить приложение 11.3.) Какую модель вы использовали бы для:
- Исследования количества времени (в минутах), которое человек тратит на разговоры по мобильному телефону в месяц?
  - Исследования полученных оценок (от A до F) в большом классе?
  - Исследования потребительского выбора между колой, пепси и лимонадом?
  - Исследования количества мобильных телефонов, используемых в одной семье?

### **Компьютерные упражнения**

- E11.1. Было показано, что запрет курения на рабочем месте побуждает курильщиков избавляться от этой вредной привычки, поскольку ограничивает возможность курить. В этом задании необходимо оценить влияние запретов на курение на рабочем месте, используя базу данных по 10 тыс. офисным работникам в США в период с 1991 по 1993 год,

доступную на веб-сайте [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) и содержащуюся в файле *Smoking*. База данных содержит информацию о том, были ли введены по месту работы индивида запреты на курение на рабочем месте, является ли тот или иной индивид курильщиком, а также ряд других индивидуальных характеристик<sup>1</sup>. Детальное описание данных представлено в файле *Smoking\_Description*, который доступен на веб-сайте.

- a) Оцените вероятность того, что курят (i) все работники; (ii) работники, сталкивающиеся с запретами на курение на рабочем месте; (iii) работники, не сталкивающиеся с запретами на курение на рабочем месте.
  - б) Какова разница в вероятности курения для сотрудников, сталкивающихся и не сталкивающихся с запретами на курение на рабочем месте? Используйте линейную вероятностную модель, чтобы определить, является ли это различие статистически значимым.
  - в) Оцените линейную вероятностную модель с переменной *smoker* в качестве зависимой переменной и следующими регрессорами: *smkban*, *female*, *age*, *age*<sup>2</sup>, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black* и *hispanic*. Сравните полученные оценки влияния запретов на курение с ответом, полученным в пункте (б). Предложите причину, основанную на сущности этой регрессии, которая объясняла бы изменение в предполагаемом эффекте запрета на курение между пунктами (б) и (в).
  - г) Протестируйте гипотезу о том, что коэффициент при переменной *smkban* равен нулю в теоретической регрессии в пункте (в), против альтернативной гипотезы о том, что коэффициент не равен нулю, на уровне значимости 5%.
  - д) Протестируйте гипотезу о том, что вероятность курения не зависит от уровня образования в регрессии в пункте (в). Растет или снижается вероятность курения при росте уровня образования?
  - е) Основываясь на результатах регрессии из пункта (в), покажите, есть ли нелинейная зависимость между *age* и вероятностью курения? Начертите график зависимости между вероятностью курения и *age* в промежутке  $18 \leq age \leq 65$  для «белого», нелатиноамериканского выпускника колледжа мужского пола без запретов на курение на рабочем месте.
- E11.2. В этом упражнении используются те же данные, что и в эмпирическом упражнении Е 11.1.
- а) Оцените пробит-модель, используя те же регрессоры, что и в эмпирическом упражнении Е 11.1 (в).
  - б) Протестируйте гипотезу о том, что коэффициент при переменной *smkban* равен нулю в теоретической версии пробит-регрессии, против альтернативной гипотезы о том, что коэффициент не равен нулю,

---

<sup>1</sup> Данные были предоставлены профессором университета Мэриленд (University of Maryland) Уильямом Эвансом (Professor William Evans), данные использовались в его работе, написанной совместно с Мэттью Фарелли (Matthew Farrelly) и Эдвардом Монтгомери (Edward Montgomery) «Do Workplace Smoking Bans Reduce Smoking?», American Economic Review, 1999, 89 (4): 728–747.

на уровне значимости 5%. Сравните полученную *t*-статистику и соответствующие выводы с выводами, полученными в эмпирическом упражнении Е11.1 (г), основанными на линейной вероятностной модели.

- в) Протестируйте гипотезу о том, что курение не зависит от уровня образования в пробит-модели. Сравните полученные результаты с результатами эмпирического упражнения Е11.1 (д), полученными на основе линейной модели вероятности.
- г) Господин *A* – белый (нелатиноамериканец) в возрасте 20 лет, не окончивший среднюю школу. Используя пробит-регрессию из пункта (а) и предполагая, что господин *A* не сталкивался с запретами на курение на рабочем месте, вычислите вероятность того, что господин *A* курит. Проведите вычисления снова, предполагая, что он сталкивался с запретами на курение на рабочем месте. Каково воздействие мероприятий по запрету курения на рабочем месте на вероятность того, что господин *A* является курильщиком?
- д) Повторите вычисления пункта (г) для госпожи *B* – чернокожей женщины возрастом 40 лет, выпускницы колледжа.
- е) Повторите пункт (г) и (д), используя линейную вероятностную модель из эмпирического упражнения Е11.1 (в).
- ж) Основываясь на результатах пунктов (в) – (е), выясните, различаются ли результаты пробит-модели и линейной вероятностной модели. Если различия имеют место, то какие результаты можно считать более разумными? Насколько велика оценка эффекта по сравнению с реальной величиной данного эффекта?
- з) Существуют ли какие-либо серьезные угрозы для внутренней обоснованности результатов?
- Е11.3. В этом упражнении будет изучаться связь страхования в сфере здравоохранения, состояния здоровья и занятости на основе случайной выборки, содержащей данные по более чем 8000 работникам в США в 1996 году. Данные доступны для скачивания на веб-сайте [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) в файле Insurance<sup>1</sup>. Детальное описание данных представлено в файле Insurance\_Description, доступном на веб-сайте.
- а) Действительно ли работники, имеющие собственный бизнес (работающие «на себя»), с большей вероятностью имеют медицинскую страховку, чем работники, получающие фиксированную заработную плату? Если это так, то как это соотносится с реальной жизнью? Являются ли различия статистически значимыми?
- б) Возможно, имеют место систематические различия между «самозанятыми» и работниками с фиксированной заработной платой в их

<sup>1</sup> Данные были предоставлены профессором Принстонского университета Харви Розеном (Professor Harvey Rosen). Они использовались в его совместной с Крейгом Перри (Craig Perry) работе «The Self-Employed Are Less Likely Than Wage-Earners to Have Health Insurance. So What?» в Douglas Holtz-Eakin and Harvey S. Rosen (eds), Entrepreneurship and Public Policy (Cambridge, MA: MIT Press, 2004).

возрасте, уровне образования и так далее. После учета этих возможных различий по-прежнему ли «самозанятые» с большей вероятностью имеют медицинскую страховку?

- в) Как варьируется частота получения страховки в зависимости от возраста? Действительно ли «более возрастные» работники с большей вероятностью покупают страховку? Или с меньшей?
- г) Различается ли эффект «самозанятости» для работников старшего возраста и более молодых работников?
- д) Считается, что занятые собственным делом работники с меньшей вероятностью будут страховаться, но, несмотря на это, они столь же здоровы, как работники с фиксированной заработной платой (наемные работники). Действительно ли это является правильным? Справедливо ли это для молодых работников? Для работников старшего возраста? Существуют ли потенциальные проблемы, связанные с взаимной причинностью, которые могли бы подорвать внутреннюю обоснованность подобного статистического анализа?

## Приложения

### *Приложение 11.1. Бостонские данные HMDA*

Бостонская база данных HMDA была собрана исследователями из Федерального резервного банка Бостона. В базу данных включена информация о заявлениях на предоставление ипотечного кредита и последующего опроса банков и других кредитных учреждений, которые получили эти заявления на предоставление ипотечного кредита. Выборка содержит данные заявлений на предоставление ипотечного кредита, поданные в 1990 году в большом Бостоне. Полный набор данных содержит 2925 наблюдений, состоит из всех заявлений на предоставление ипотечного кредита, поданных афроамериканцами и латиноамериканцами, а также случайной выборки заявлений на предоставление ипотечного кредита белокожими заявителями.

Чтобы сузить объем анализа в этой главе, будет использоваться подмножество данных для домов, в которых может проживать только одна семья (исключаются данные по многоквартирным домам), и для темнокожих кандидатов и белых кандидатов (исключаются данные по кандидатам других национальных меньшинств). В итоге остается 2380 наблюдений. Описание переменных, используемых в этой главе, представлено в таблице 11.1.

Вышеописанные данные были любезно предоставлены Джеки Тутеллом (Geoffrey Tootell) из исследовательского отделения Федерального резервного банка Бостона. Дополнительная информация об используемой базе данных, как и выводы, к которым пришли исследователи Федерального резервного банка Бостона, представлены в: Alicia H. Munnell, Geoffrey M.B. Tootell, Lynne E. Browne,

and James McEneaney, «Mortgage Lending in Boston: Interpreting HMDA Data», American Economic Review, 1996. P. 25–53.

## Приложение 11.2. Оценка метода максимального правдоподобия

В данном приложении представлено краткое введение в методологию получения оценок методом максимального правдоподобия (ММП) в контексте бинарных моделей, обсуждаемых в данной главе. В первую очередь необходимо получить оценку максимального правдоподобия для вероятности успеха  $p$  для  $n$  независимых одинаково распределенных наблюдений для случайных переменных с распределением Бернулли. Затем рассматриваются пробит- и логит-модели, обсуждается псевдо- $R^2$  (pseudo- $R^2$ ). В конце будут рассмотрены стандартные ошибки для оценок вероятностей (предсказанных вероятностей). В двух частях приложения приводятся вычисления.

### ММП для $n$ i.i.d бернуlliевских случайных величин

На первом шаге для вычисления оценок максимального правдоподобия производится вычисление совместной функции распределения. Для  $n$  независимых одинаково распределенных наблюдений Бернуллиевских случайных величин совместная функция распределения представляет собой простое расширение случая  $n = 2$ , представленного в разделе 11.3, на случай произвольного  $n$ :

$$\begin{aligned} \Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) &= \\ &= \left[ p^{y_1} (1-p)^{(1-y_1)} \right] \times \left[ p^{y_2} (1-p)^{(1-y_2)} \right] \times \\ &\quad \times \dots \times \left[ p^{y_n} (1-p)^{(1-y_n)} \right] = p^{(y_1 + \dots + y_n)} (1-p)^{n-(y_1 + \dots + y_n)}. \end{aligned} \quad (11.13)$$

Функция правдоподобия представляет собой совместную функцию распределения, которая рассматривается как функция неизвестных коэффициентов.

Пусть  $S = \sum_{i=1}^n Y_i$ , тогда функция правдоподобия равна:

$$f_{\text{Bernoulli}}(p; Y_1, \dots, Y_n) = p^S (1-p)^{n-S}. \quad (11.14)$$

Оценка максимального правдоподобия  $p$  представляет собой значение  $p$ , которое максимизирует функцию правдоподобия, представленную выражением (11.14). Функция правдоподобия может быть максимизирована численно. Удобно максимизировать не саму функцию правдоподобия, а ее логарифм (поскольку логарифм является строго возрастающей функцией, максимизация самой функции и ее логарифма дает одинаковый результат). Логарифм приведенной выше функции правдоподобия равен:  $S \ln(p) + (n-S) \ln(1-p)$ , а его производная по  $p$  равна:

$$\frac{d}{dp} \ln[f_{\text{Bernoulli}}(p; Y_1, \dots, Y_n)] = \frac{S}{p} - \frac{n-S}{1-p}. \quad (11.15)$$

Приравнивая производную в выражении (11.15) к нулю и решая получившееся уравнение относительно  $p$ , можно получить оценку максимального правдоподобия  $\hat{p} = S / n = \bar{Y}$ .

### **Оценка максимального правдоподобия для пробит-модели**

В пробит-модели вероятность того, что  $Y_i = 1$  условно на  $X_{1i}, \dots, X_{ki}$ , равна:  $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})$ . Условная функция распределения для  $i$ -го наблюдения равна:  $\Pr[Y_i = y_i | X_{1i}, \dots, X_{ki}] = p_i^{y_i} (1 - p_i)^{1-y_i}$ . Предполагая, что  $(X_{1i}, \dots, X_{ki}, Y_i)$  являются одинаково независимо распределенными,  $i = 1, \dots, n$ , совместная функция распределения  $Y_1, \dots, Y_n$  при заданных значениях  $X$  равна:

$$\begin{aligned} \Pr(Y_1 = y_1, \dots, Y_n = y_n | X_{1i}, \dots, X_{ki}, i = 1, \dots, n) &= \\ &= \Pr(Y_1 = y_1 | X_{11}, \dots, X_{k1}) \times \dots \times \Pr(Y_n = y_n | X_{1n}, \dots, X_{kn}) = \\ &= (p_1^{y_1} (1 - p_1)^{1-y_1}) \times \dots \times (p_n^{y_n} (1 - p_n)^{1-y_n}). \end{aligned} \quad (11.16)$$

Функция правдоподобия представляет собой совместную функцию правдоподобия, рассматриваемую как функцию неизвестных коэффициентов. Удобно рассматривать логарифм функции правдоподобия. Соответственно, логарифм функции правдоподобия равен:

$$\begin{aligned} \ln [f_{\text{probit}}(\beta_0, \dots, \beta_k; Y_1, \dots, Y_n | X_{1i}, \dots, X_{ki}, i = 1, \dots, n)] &= \\ &= \sum_{i=1}^n Y_i \ln [\Phi(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})] + \\ &+ \sum_{i=1}^n (1 - Y_i) \ln [1 - \Phi(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})], \end{aligned} \quad (11.17)$$

где это выражение содержит формулу из пробит-модели для условной вероятности  $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})$ .

Оценка максимального правдоподобия для пробит-модели максимизирует функцию правдоподобия или ее логарифм, представленный в уравнении (11.17). Поскольку не существует простой формулы для оценки максимального правдоподобия, функция правдоподобия для пробит-модели может быть максимизирована с помощью вычислительных алгоритмов и использования компьютера.

При общих условиях оценки максимального правдоподобия являются состоятельными и имеют нормальное распределение в больших выборках.

### **Оценка максимального правдоподобия для логит-модели**

Оценка максимального правдоподобия для логит-модели может быть получена аналогично случаю пробит-модели. Единственное различие заключается в том, что условная вероятность  $p_i$  в логит-модели задана выражением (11.9). Соответственно, логарифм функции правдоподобия для логит-модели задан выражением (11.17), где  $\Phi(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})$  заменяется на  $[1 + e^{-(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})}]^{-1}$ .

Как и для пробит-модели, для логит-модели нет какой-либо простой формулы для оценки максимального правдоподобия коэффициентов. Поэтому логарифм функции правдоподобия необходимо максимизировать численно.

### Псевдо- $R^2$

Псевдо- $R^2$  сравнивает значения функции правдоподобия для оцененной модели и для модели, в которую не включается ни один из регрессоров  $X$ . В частности, псевдо- $R^2$  для пробит-модели равен:

$$\text{pseudo-}R^2 = 1 - \frac{\ln(f_{\text{probit}}^{\max})}{\ln(f_{\text{Bernoulli}}^{\max})}, \quad (11.18)$$

где  $f_{\text{probit}}^{\max}$  – максимальное значение функции правдоподобия для пробит-модели (в которую включены регрессоры  $X$ ), а  $f_{\text{Bernoulli}}^{\max}$  – максимальное значение функции правдоподобия для распределения Бернулли (пробит-модель, из которой исключены все регрессоры  $X$ ).

### Стандартные ошибки для предсказанных вероятностей

Для простоты рассмотрим случай одного регрессора в пробит-модели. Тогда предсказанная (оцененная) вероятность при заданных значениях регрессора  $x$  равна:  $\hat{p}(x) = \Phi(\hat{\beta}_0^{\text{MLE}} + \hat{\beta}_1^{\text{MLE}} x)$ , где  $\hat{\beta}_0^{\text{MLE}}$  и  $\hat{\beta}_1^{\text{MLE}}$  – оценки максимального правдоподобия для коэффициентов пробит-модели. Поскольку оценка вероятности зависит от оценок  $\hat{\beta}_0^{\text{MLE}}$  и  $\hat{\beta}_1^{\text{MLE}}$ , а эти оценки имеют выборочное распределение, то и сама оценка вероятности будет иметь выборочное распределение.

Дисперсия выборочного распределения  $\hat{p}(x)$  может быть вычислена с помощью аппроксимации функции  $\Phi(\hat{\beta}_0^{\text{MLE}} + \hat{\beta}_1^{\text{MLE}} x)$ , являющейся нелинейной функцией  $\hat{\beta}_0^{\text{MLE}}$  и  $\hat{\beta}_1^{\text{MLE}}$ , линейной функцией от  $\hat{\beta}_0^{\text{MLE}}$  и  $\hat{\beta}_1^{\text{MLE}}$ . В частности, пусть:

$$\hat{p}(x) = \Phi(\hat{\beta}_0^{\text{MLE}} + \hat{\beta}_1^{\text{MLE}} x) \cong c + a_0 (\hat{\beta}_0^{\text{MLE}} - \beta_0) + a_1 (\hat{\beta}_1^{\text{MLE}} - \beta_1), \quad (11.19)$$

где свободный член  $c$  и коэффициенты  $a_0$  и  $a_1$  зависят от  $x$  и могут быть рассчитаны численно. Выражение (11.19) представляет собой разложение уравнения Тейлора первого порядка,  $c = \Phi(\beta_0 + \beta_1 x)$ ,  $a_0$  и  $a_1$  – частные производные,  $a_0 = \frac{\partial \Phi(\beta_0 + \beta_1 x)}{\partial \beta_0} \Big|_{\hat{\beta}_0^{\text{MLE}}, \hat{\beta}_1^{\text{MLE}}} \text{ и } a_1 = \frac{\partial \Phi(\beta_0 + \beta_1 x)}{\partial \beta_1} \Big|_{\hat{\beta}_0^{\text{MLE}}, \hat{\beta}_1^{\text{MLE}}}$ . Дисперсия  $\hat{p}(x)$  теперь может быть вычислена с помощью аппроксимации в выражении (11.19) и формулы для дисперсии двух случайных переменных в выражении (2.31):

$$\begin{aligned} \text{var}[\hat{p}(x)] &\cong \text{var}\left[c + a_0 (\hat{\beta}_0^{\text{MLE}} - \beta_0) + a_1 (\hat{\beta}_1^{\text{MLE}} - \beta_1)\right] = \\ &= a_0^2 \text{var}(\hat{\beta}_0^{\text{MLE}}) + a_1^2 \text{var}(\hat{\beta}_1^{\text{MLE}}) + 2a_0 a_1 \text{cov}(\hat{\beta}_0^{\text{MLE}}, \hat{\beta}_1^{\text{MLE}}). \end{aligned} \quad (11.20)$$

Используя выражение (11.20), можно вычислить стандартную ошибку  $\hat{p}(x)$  с помощью оценок дисперсий и ковариации оценок максимального правдоподобия коэффициентов  $\hat{\beta}_0^{\text{MLE}}$  и  $\hat{\beta}_1^{\text{MLE}}$ .

### **Приложение 11.3. Прочие модели с ограниченными зависимыми переменными**

В этом приложении представлен краткий обзор некоторых моделей с ограниченными зависимыми переменными, отличных от наиболее часто встречающихся в экономической литературе моделей с бинарными зависимыми переменными. В большинстве случаев МНК-оценки параметров моделей с ограниченными зависимыми переменными являются несостоительными, и поэтому оценки зачастую получаются с помощью метода максимального правдоподобия. Для получения более детальной информации читатель может ознакомиться, например, со следующими работами: Ruud (2000) и Wooldridge (2002).

#### **Регрессионные модели с цензурированными или усеченными переменными**

Предположим, что имеются межобъектные данные по покупкам автомобилей частными лицами в каком-либо заданном году. Положительные расходы покупателей автомобилей можно рассматривать как непрерывные случайные переменные, но те клиенты, которые не купили автомобиль, потратили 0 долл. Таким образом, распределение автомобильных расходов – комбинация дискретного распределения (в нуле) и непрерывного распределения.

Нобелевский лауреат Джеймс Тобин предложил модель для зависимой переменной с частично непрерывным и частично дискретным распределением (Tobin, 1958). Тобин предложил моделировать  $i$ -е частное лицо в выборке как индивида, имеющего «желаемый» («потенциальный») уровень расходов  $Y_i^*$ , который связан с регрессорами (например размером семьи) согласно модели линейной регрессии. Таким образом, в модели парной регрессии желаемый уровень расходов равен:

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i, i = 1, \dots, n. \quad (11.21)$$

Если  $Y_i^*$  (объем средств, которые хочет потратить потребитель) превышает некое граничное значение, например минимальную цену автомобиля, то потребитель приобретает автомобиль, затрачивая при этом  $Y_i = Y_i^*$ , что является наблюдаемой величиной. Но если  $Y_i^*$  меньше этого граничного значения, то потребитель не приобретает автомобиль и наблюдается  $Y_i = 0$  вместо  $Y_i^*$ .

Когда выражение (11.21) оценивается при использовании наблюдаемых значений  $Y_i$  вместо  $Y_i^*$ , то МНК-оценки являются несостоительными. Тобин предложил решение этой проблемы через использование функции правдоподобия с дополнительным предположением о том, что  $u_i$  имеет нормальное распределение. Получившиеся оценки максимального правдоподобия широко применяются эконометристами для анализа многих экономических явлений. В честь Тобина выражение (11.20) совместно с предположением о нормальности остаточного члена было названо тобит-моделью регрессии. Тобит-модель представляет собой пример цензурированной регрессионной модели, которая имеет такое название из-за того, что зависимая переменная «ограничена» сверху или снизу неким граничным значением.

### **Модели выборочного отбора**

В цензурированной модели регрессии данные по индивидам, которые приобрели или не приобрели что-либо, получены путем случайного формирования выборки из всего объема взрослого населения. Если бы, однако, данные были получены на основе данных по налогам с продаж, то такая выборка включала бы только данные по индивидам, которые осуществили покупку, но не содержала бы данные о клиентах, которые ничего не приобрели. Данные, в которых наблюдения недоступны выше или ниже «порога» или граничного значения (например, данные, содержащие информацию только по индивидам, которые осуществили какую-либо покупку), называют усеченными данными. *Регрессионная модель с усеченными переменными* – модель, в которой используются данные, где часть наблюдений просто недоступна, когда зависимая переменная принимает значения выше или ниже определенного уровня.

Модель усеченной регрессии – это пример модели выборочного отбора, в которой механизм отбора наблюдений (то или иное частное лицо попадает в выборку только на основании покупки автомобиля) связан с величиной зависимой переменной (расходы на автомобиль). Как обсуждалось в разделе 11.4, одним из подходов к оценке моделей выборочного отбора является формирование двух выражений – одно из которых непосредственно описывает  $Y_i^*$ , а второе описывает, наблюдается ли  $Y_i^*$  или нет. Параметры модели могут тогда быть оценены с помощью метода максимального правдоподобия или с помощью пошаговой процедуры, в которой сначала оценивается выражение, описывающее формирование выборки, а затем выражение для  $Y_i^*$ . Для получения дополнительной информации можно ознакомиться со следующими работами: Ruud (2000. Chapter 28), Greene (2000. Section 20.4) или Wooldridge (2002. Chapter 17).

### **Счетные данные**

Счетные данные возникают в тех случаях, когда зависимая переменная представляет собой порядковое числительное, например, количество посещений ресторана потребителем за неделю. Когда подобные числа достаточно велики, переменная может рассматриваться как практически непрерывная, но когда они малы, непрерывное распределение неприменимо. Модель линейной регрессии, оцениваемая с помощью МНК, может использоваться для счетных данных даже в тех случаях, когда количество наблюдений невелико. Полученные с помощью регрессии оценки могут интерпретироваться как ожидаемые значения зависимой переменной при заданных значениях regressоров (условно на regressорах). Таким образом, когда зависимая переменная описывает количество посещений ресторана в неделю, то предсказанное значение величиной 1,7 означает, что индивид в среднем 1,7 раз в неделю посещает ресторан. Как и в случае модели бинарного выбора, МНК не учитывает всех особенностей структуры счетных данных и может давать неверные прогнозы например –0,2 посещения ресторана в неделю. Как пробит- или логит-модели устраняют подобные неточности в случае

бинарных зависимых переменных, так и существуют специальные модели для счетных данных. Две наиболее распространенных из них – пуассоновская модель регрессии и отрицательная биномиальная модель регрессии.

### ***Упорядоченные данные***

*Упорядоченные данные* возникают в тех случаях, когда существует естественный ряд взаимоисключающих качественных категорий. Например, получение школьного образования, получение образования в колледже (но без полного окончания) и окончание колледжа. Как и в случае счетных данных, упорядоченные данные имеют естественное упорядочивание, но, в отличие от счетных данных, не имеют численных значений.

Поскольку упорядоченные данные не имеют численных значений, то применение МНК является невозможным. Вместо этого подобного рода данные можно анализировать с помощью обобщения пробит-модели, которое называется *упорядоченной пробит-моделью*, в которой вероятность каждого из возможных исходов (например получения образования в колледже), при заданных значениях независимых переменных (например уровень доходов родителей), моделируется с помощью нормального распределения.

### ***Данные дискретного выбора***

Наблюдения в данных дискретного или множественного выбора могут принимать несколько неупорядоченных качественных значений. В качестве одного из примеров можно рассмотреть вид транспорта, который выбирает житель пригородной зоны, чтобы добраться на работу; он может выбрать поездку на метро, на автобусе, на личном автомобиле или своим ходом (на велосипеде или пешком). Если бы было необходимо проанализировать эти альтернативы, то зависимая переменная имела бы четыре возможных реализации (метро, автобус, машина и «свой ход»). Эти исходы не могут быть упорядочены каким-либо естественным способом. Вместо этого исходы можно рассматривать как выбор конечного набора качественных альтернатив.

В качестве эконометрической задачи можно рассматривать задачу моделирования вероятности выбора одной из альтернатив при заданных значениях нескольких регрессоров, таких как индивидуальные характеристики (например как далеко дом индивида находится от станции метро) и характеристик каждой альтернативы (например цена поездки на каждом виде транспорта). Как обсуждалось в разделе 11.3, модели для анализа данных дискретного выбора могут быть построены на основе принципов максимизации полезности. Вероятности выбора той или иной альтернативы могут быть выражены в логит- или пробит-формах, которые называются *множественными пробит-* и *множественными логит-моделями*, соответственно.

# Глава 12. Регрессии с инструментальными переменными

В главе 9 описан ряд проблем, включающий проблемы пропущенных переменных, ошибок в переменных и одновременной причинности, которые приводят к наличию корреляции между остаточным членом (ошибками регрессии) и регрессорами. Смещения, вызванные пропущенными переменными, в рамках модели множественной регрессии могут быть исключены просто путем непосредственного включения пропущенной переменной в уравнение регрессии, но это возможно только при наличии статистических данных для этой пропущенной переменной. А иногда, например, когда имеет место двусторонняя причинность (от  $X$  к  $Y$  и от  $Y$  к  $X$ ), в рамках множественной регрессии просто нельзя устранить данное смещение. Если прямое решение этих проблем является либо невозможным, либо недоступным, то нужно задуматься о каком-либо новом методе.

*Регрессия с инструментальными переменными*<sup>1</sup> (ИП) представляет собой общий способ получения состоятельных оценок неизвестных регрессионных коэффициентов в тех случаях, когда регрессор  $X$  коррелирует с ошибкой  $u$ . Чтобы понять, как устроена модель регрессии с инструментальными переменными, достаточно представить, что изменения  $X$  складываются из двух частей: одна часть, которая по какой-то причине коррелирует с  $u$  (это та часть, которая вызывает проблемы), и вторая часть, которая не коррелирует с  $u$ . При наличии дополнительной информации, которая позволяет изолировать вторую часть, можно было бы сосредоточить внимание на тех изменениях  $X$ , которые некоррелированы с  $u$ , и игнорировать изменения  $X$ , которые приводят к смещению МНК-оценок. Так на самом деле и «работает» регрессия с инструментальными переменными<sup>2</sup>. Информация об изменениях  $X$ , которые не коррелируют с  $u$ , может быть получена при помощи одной или нескольких дополнительных переменных, называемых *инструментальными переменными* или просто *инструментами*. В регрессии с инструментальными переменными эти дополнительные переменные используются в качестве «инструментов», чтобы «изолировать» изменения  $X$ , которые не коррелируют с  $u$ , что в свою очередь позволяет получить состоятельные оценки коэффициентов регрессии.

В первых двух разделах данной главы описываются «механика» и основные предположения IV-регрессии: как устроена IV-регрессия, что именно является правильным инструментом, и как осуществлять и интерпретировать наиболее распространенные случаи применения IV-регрессий, в том числе применение

<sup>1</sup> Instrumental variables (IV). – Прим. научн. ред. перевода.

<sup>2</sup> Далее наравне с полным наименованием может использоваться сокращенное название – ИП-регрессия или IV-регрессия. – Примеч. науч. ред. перевода.

двухшагового метода наименьших квадратов (2МНК). Ключ к успешному проведению эмпирического анализа с использованием инструментальных переменных находится в применении правильных «инструментов». В разделе 12.3 рассматривается вопрос о том, как оценить, является ли набор инструментов допустимым<sup>1</sup> (подходящим). В качестве примера в разделе 12.4 используется IV-регрессии для оценки эластичности спроса на сигареты. Наконец, раздел 12.5 посвящен наиболее трудному вопросу поиска правильных инструментальных переменных.

## 12.1. ИП оценки с одним регрессором и одним инструментом

Изучение регрессии с инструментальными переменными стоит начать с рассмотрения случая регрессии с одним регрессором  $X$ , который может быть коррелирован с ошибкой регрессии  $u$ . Если  $X$  и  $u$  коррелируют друг с другом, то МНК-оценки являются несостоительными, то есть они не стремятся к истинному значению регрессионных коэффициентов при увеличении размера выборки [см. уравнение (6.1)]. Как обсуждалось в разделе 9.2, данная корреляция между  $X$  и  $u$  может иметь различные источники, включая пропущенные переменные, ошибки в переменных (ошибки измерения в регрессорах) и взаимную причинность (когда  $X$  оказывает влияние на  $Y$  и наоборот). Вне зависимости от источника корреляции  $X$  и  $u$  наличие подходящей инструментальной переменной  $Z$  может помочь верно оценить влияние единичного изменения  $X$  на  $Y$ .

### *Модель регрессии с инструментальными переменными и ее основные предположения*

Теоретическая модель, связывающая зависимую переменную  $Y_i$  и регрессор  $X_i$ , может быть записана в таком виде:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \dots, n, \quad (12.1)$$

где  $u_i$  – остаточный член, представляющий пропущенные факторы, которые могли бы определять  $Y_i$ . Если  $X_i$  и  $u_i$  коррелируют между собой, то МНК-оценки являются несостоительными. В рамках IV-регрессий используется дополнительная «инструментальная» переменная  $Z$ , которая позволяет изолировать ту часть  $X$ , которая некоррелирована с  $u_i$ .

**Эндогенность и экзогенность.** В IV-регрессиях используются некоторые специальные термины, позволяющие отличить переменные, которые коррелируют с теоретической ошибкой (остаточным членом) регрессии  $u$ , от тех, которые не коррелированы с ней. Переменные, которые коррелированы с остаточным членом, называют *эндогенными переменными*, а переменные, которые не коррели-

<sup>1</sup> Отметим, что в русском языке для обозначения этого понятия часто используется слово «валидный», что является непосредственным аналогом английского слова «valid». Поскольку в данном случае терминология не является устоявшейся, мы будем использовать в качестве перевода слова «допустимые», «корректные», «подходящие» или «правильные» инструменты. – Примеч. науч. ред. перевода.

рут с остаточным членом, называют *экзогенными переменными*. Историческим источником возникновения этих терминов послужили системы эконометрических уравнений, в которых «эндогенная» переменная определяется в рамках модели, в то время как «экзогенная» переменная определяется вне модели. Например, в разделе 9.2 рассматривается возможность того, что если низкие результаты тестов приводят к уменьшению соотношения учеников и учителей из-за политического вмешательства и увеличения финансирования, при этом причинность может иметь место в обе стороны: от соотношения числа учеников и учителей к результатам тестов и наоборот. Это может быть представлено математически в виде системы двух одновременных уравнений [уравнения (9.3) и (9.4)] по одному для каждой причинно-следственной взаимосвязи. Как обсуждалось в разделе 9.2, поскольку результаты тестов и соотношение учеников и учителей определяются в рамках модели, то они коррелированы с теоретическим остаточным членом  $u$ . То есть в данном примере обе переменные являются эндогенными. В отличие от экзогенной переменной, которая определяется вне модели и не коррелирована с  $u$ .

**Два условия, при которых инструменты являются подходящими.** Подходящая (допустимая) инструментальная переменная («инструмент») должна удовлетворять двум условиям, известным как *условие релевантности инструмента* и *условие экзогенности инструмента*:

1. Условие релевантности инструмента:  $\text{corr}(Z_i, X_i) \neq 0$ .
2. Условие экзогенности инструмента:  $\text{corr}(Z_i, u_i) = 0$ .

Если инструмент является релевантным, то изменения инструмента связаны с изменениями  $X_i$ . Если инструмент также является экзогенной переменной, то часть изменений  $X_i$ , отражаемая инструментальной переменной, является экзогенной. Эти экзогенные изменения могут быть использованы для получения оценок коэффициента  $\beta_1$ .

Два описанных выше условия допустимости инструментальной переменной являются жизненно необходимыми для использования IV-регрессий, им (а также их аналогам в случае нескольких регрессоров и нескольких инструментальных переменных) будет уделяться значительное внимание в ходе изложения данной главы.

## Двухшаговый метод наименьших квадратов

Если инструментальная переменная  $Z$  удовлетворяет необходимым условиям релевантности и экзогенности, то существует возможность получить оценки коэффициента  $\beta_1$  с помощью *двухшагового метода наименьших квадратов*<sup>1</sup> (2МНК). Как можно понять из названия, оценки с помощью 2МНК получаются в два шага. На первом шаге мы раскладываем  $X$  на две составляющие: проблемную компоненту, которая может быть коррелирована с остаточным членом регрессии, и вторую («хорошую», «очищенную») компоненту, которая не коррелирована

<sup>1</sup> В англоязычной литературе используется термин «Two Stage Least Squares» (TSLS). – Примеч. науч. ред. перевода.

с остаточным членом. На втором шаге для получения оценок используется выделенная «хорошая» компонента.

На первом шаге оценивается регрессия, связывающая  $X$  и  $Z$ :

$$X_i = \pi_0 + \pi_1 Z_i + v_i, \quad (12.2)$$

где  $\pi_0$  – свободный член,  $\pi_1$  – коэффициент, определяющий угол наклона прямой,  $v_i$  – ошибка регрессии. Регрессия дает возможность провести декомпозицию  $X_i$ . Первая компонента –  $\pi_0 + \pi_1 Z_i$ , доля  $X_i$ , которая может быть объяснена с помощью  $Z_i$ . Поскольку  $Z_i$  является экзогенной переменной, то эта компонента  $X_i$  не коррелирована с  $u_i$ , остаточным членом в выражении (12.1). Другая компонента  $X_i$  – это  $v_i$ , проблемная компонента  $X_i$ , которая может быть коррелирована с  $u_i$ .

Идея, которая лежит в основе двухшагового метода наименьших квадратов, заключается в использовании «хорошей» компоненты  $X_i$ ,  $\pi_0 + \pi_1 Z_i$  вместо проблемной компоненты  $v_i$ . Единственная сложность заключается в том, что значения  $\pi_0$  и  $\pi_1$  неизвестны, из-за чего выражение  $\pi_0 + \pi_1 Z_i$  не может быть вычислено. Соответственно, первый шаг 2МНК заключается в применении МНК к выражению (12.2) и использовании полученных оценок  $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ , где  $\hat{\pi}_0$  и  $\hat{\pi}_1$  – МНК-оценки.

Второй шаг 2МНК прост: оценивается регрессия  $Y_i$  на  $\hat{X}_i$  с помощью МНК. В результате после второго шага 2МНК получаются оценки  $\hat{\beta}_0^{TSLS}$  и  $\hat{\beta}_1^{TSLS}$ .

## Почему IV-регрессия работает?

Два приведенных ниже примера помогут более глубоко понять, каким образом IV-регрессия помогает решать проблему корреляции между  $X_i$  и  $u_i$ .

**Пример 1: Задача Филиппа Райта.** Метод получения оценок с помощью инструментальных переменных был впервые опубликован в 1928 году в приложении к книге, написанной Филиппом Райтом (Wright, 1928), хотя основные идеи IV-регрессии, по всей видимости, были разработаны автором совместно с его сыном, Сьюэллом Райтом (Sewall Wright) (см. вставку ниже). Филипп Райт рассматривал важную экономическую проблему своего времени: как установить импортный тариф (налог на импортируемые товары) на животные и растительные масла и жиры, такие как масло и соевое масло. В 1920 году тарифы на импорт были основным источником налоговых поступлений для США. Ключом к пониманию экономического эффекта от изменения тарифов было получение количественных оценок кривых спроса и предложения для товаров. Напомним, что эластичность предложения представляет собой процентное изменение предложения товаров при увеличении цены на 1%, а эластичность спроса это процентное изменение величины спроса при увеличении цены на 1%. Филипп Райт хотел получить оценки этих эластичностей спроса и предложения.

Рассмотрим задачу получения оценок эластичности спроса на масло. Вспомним еще раз вставку «Основные понятия 8.2», в которой показано, что коэффи-

циент в линейном уравнении, связывающем  $\ln(Y_i)$  и  $\ln(X_i)$ , представляет собой эластичность  $Y$  по  $X$ . В задаче Райта это подразумевает функцию спроса следующего вида:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i, \quad (12.3)$$

где  $Q_i^{butter}$  –  $i$ -е наблюдение количества потребленного масла,  $P_i^{butter}$  – его цена,  $u_i$  – отражает остальные факторы, которые могут оказывать влияние на спрос, например, такие как уровень дохода и вкусовые предпочтения. В уравнении (12.3) 1%-е увеличение цены на масло приводит к увеличению спроса на  $\beta_1$ , то есть,  $\beta_1$  – это и есть эластичность спроса по цене.



### **Кто придумал IV-регрессию?**

Модель регрессии с инструментальными переменными впервые была предложена в качестве решения проблемы взаимной причинности в эконометрике в приложении к выпущенной в 1928 году книге Филиппа Г. Райта «*The Tariff on Animal and Vegetable Oils*». Если вы хотите знать, как животные и растительные масла производились, перевозились и продавались в начале XX века, первые 285 страниц книги – для вас. Эконометристам, однако, будет более интересно приложение В. В приложении представлены два вывода «метода введения внешних факторов», то есть метода, который мы сейчас называем методом инструментальных переменных. В нем также представлены результаты оценки с помощью IV-регрессии эластичности спроса на сливочное и льняное масло. Филипп был никому не известным экономистом с достаточно скучным интеллектуальным наследием, помимо упомянутого выше приложения, но его сын Сьюэлл впоследствии стал выдающимся генетиком и статистиком. Поскольку математический материал, представленный в приложении, в значительной степени отличается от остальной части книги, то многие эконометристы предполагают, что эта часть книги была написана сыном Филиппа, Сьюэллом Райтом, анонимно. Так кто же написал приложение В?

На самом деле, ни отец, ни сын, возможно, не являются авторами. Филипп Райт (Philip Wright) (1861–1934) получил степень магистра экономики в Гарвардском университете в 1887 году, преподавал математику и экономику (а также литературу и физическое воспитание) в небольшом колледже в штате Иллинойс. В обзоре книги [Wright (1915)] он использовал рисунок, похожий на рисунки 12.1а и 12.1б, чтобы показать, как регрессия, в которой оценивается зависимость количества товара от его цены, не будет, в общем случае, представлять собой оценку кривой спроса. Вместо этого он оценивает комбинацию кривых спроса и предложения. В начале 1920-х годов Сьюэлл Райт (1889–1988) исследовал статистический анализ нескольких уравнений с несколькими переменными в контексте генетики. Это исследование, в частности, привело к получению им в 1930 году звания профессора в Университете Чикаго.

Несмотря на то что сейчас уже слишком поздно, чтобы спросить Филиппа или Сьюэлла, кто из них написал приложение В, никогда не поздно провести большое расследование. Стилометрия является подразделом статистики, которую изобрели Фредерик Мостеллер и Дэвид Уоллес (Mosteller, Wallace, 1963) и использует неявные, подсознательные различия в письменных стилях для определения авторства спорного текста с использованием статистического анализа грамматических конструкций и выбора слов. Стилометрия имеет ряд подтвержденных успехов, например, Дональд Фостер (Donald Foster) в 1996 году установил авторство Джозефа Клейна (Joseph Klein) для политического романа «Primary Colors». При статистическом сравнении приложения В с достоверно написанными отдельно друг от друга Филиппом и Сьюэллом текстами было показано, что автором является Филипп Райт.

Означает ли это, что Филипп Райт изобрел методику использования инструментальных переменных в регрессиях? Не совсем. Недавно была опубликована переписка между Филиппом и Сьюэллом, которая имела место в середине 1920-х годов. Эта переписка показывает, что развитие метода с инструментальными переменными проходило при помощи совместной интеллектуальной работы отца и сына. Дополнительную информацию можно получить из работы Стокса и Требби (Stock, Trebbi, 2003).

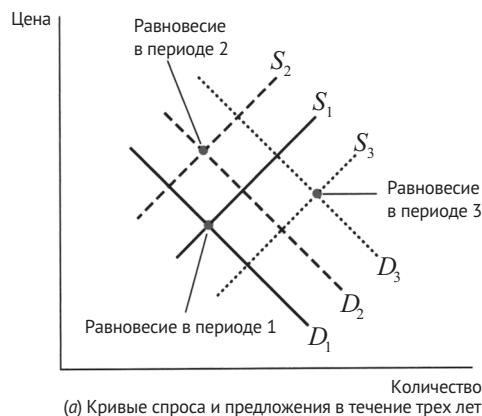


В распоряжении Филиппа Райта были данные об общем объеме годового потребления масла и его среднегодовой цене в США за период с 1912 по 1922 год. Было бы достаточно просто использовать эти данные для оценки эластичности спроса, применяя МНК к уравнению (12.3). Но существует один ключевой момент: из-за взаимодействия между спросом и предложением регрессор,  $\ln(P_i^{butter})$ , вероятно, будет коррелирован с ошибкой.

Убедиться в этом можно с помощью рисунка 12.1а, который показывает кривые рыночного спроса и предложения на масло для трех различных лет. Кривые спроса и предложения для первого года обозначаются соответственно  $D_1$  и  $S_1$ , равновесные для первого периода цена и количество товара определяются их пересечением. Для второго года спрос увеличивается с  $D_1$  до  $D_2$  (например из-за увеличения дохода), а предложение уменьшается от  $S_1$  до  $S_2$  (например из-за увеличения стоимости производства сливочного масла). Тогда равновесные цена и количество определяются пересечением новых кривых спроса и предложения. В третий год факторы, влияющие на спрос и предложение, изменяются снова, например, спрос увеличивается до  $D_3$ , предложение увеличивается до  $S_3$ , определяются новые равновесные значения цены и количества товара. На рисунке 12.1б представлены равновесные значения уровня цен и объемов масла для этих трех периодов, а также в течение восьми последующих лет, где каждый год сдвиги кривых спроса и предложения могут быть вызваны другими факторами (а не изменением

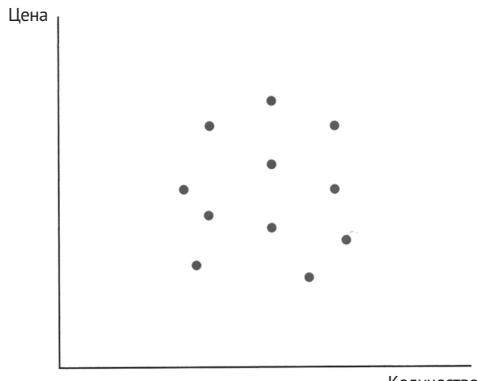
цены), которые влияют на рыночный спрос и предложение. Эти диаграммы рассеяния похожи на те, что мог бы увидеть Райт, если бы построил аналогичные графики на своих данных. Он считал, что «подгонка» прямой к этим точкам с помощью МНК не даст оценки ни кривой спроса, ни кривой предложения, так как точки были определены одновременными изменениями и спроса, и предложения.

Райт понял, что обойти эту проблему можно было бы с помощью какой-то третьей переменной, которая оказывает влияние на предложение, но при этом не влияет на спрос. Рисунок 12.1в показывает, что происходит, когда такая переменная сдвигает кривую предложения, но кривая спроса остается неподвижной. Теперь пары равновесных уровня цен и объема товара лежат на стабильной кривой спроса, и наклон кривой спроса можно легко оценить. Если сформулировать задачу Райта, используя терминологию метода инструментальных переменных, то эта третья инструментальная переменная коррелирует с ценой (она сдвигает кривую предложения, что приводит к изменению цен), но не коррелирует с  $u$  (кривая спроса остается неподвижной). Райт рассмотрел несколько потенциальных инструментальных переменных, в качестве одной из которых фигурировали погодные условия. Например, количество осадков ниже среднего уровня в животноводческом регионе может негативно отразиться на состоянии пастбищ и, тем самым, снизить производство сливочного масла при данной цене (это сдвигнет кривую предложения влево и увеличит равновесную цену). Таким образом, в аграрном регионе уровень осадков удовлетворяет условию релевантности инструментальной переменной. Но для такого региона уровень осадков не должен иметь прямого влияния на спрос на сливочное масло, то есть корреляция между уровнем осадков и  $u_i$  должна быть равна нулю, то есть уровень осадков удовлетворяет условию экзогенности инструментальной переменной.



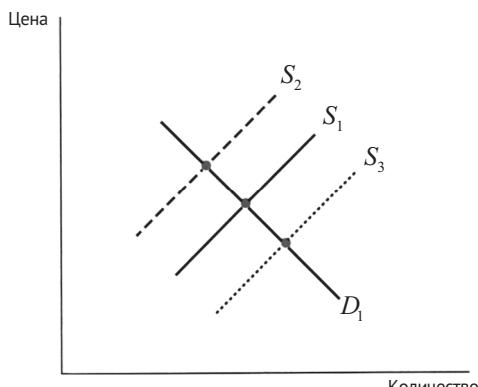
**Рисунок 12.1. Равновесие спроса и предложения**

(a) Равновесные значения цены и количества товара устанавливаются пересечением кривых спроса и предложения. Равновесие в первом периоде устанавливается на пересечении прямых  $D_1$  и  $S_1$ . Равновесие во втором периоде устанавливается на пересечении прямых  $D_2$  и  $S_1$ . Равновесие в третьем периоде устанавливается на пересечении прямых  $D_3$  и  $S_3$ .



(б) Равновесные значения уровня цен и физических объемов для 11 периодов

(б) На данной диаграмме рассеяния представлены равновесные значения цены и объема товара для 11 различных периодов времени. Кривые спроса и предложения не изображены. Возможно ли определить кривые спроса и предложения на основе точек, представленных на рисунке?



(в) Равновесные значения уровня цен и объема товара в случае сдвига кривой предложения

(в) Когда кривая предложения сдвигается из положения  $S_1$  в положение  $S_2$  или положение  $S_3$ , а кривая спроса остается в положении  $D_1$ , равновесные значения цены и объема блага лежат на кривой спроса.

**Пример 2: Оценка влияния размера класса на результаты тестов.** Несмотря на учет различных индивидуальных характеристик школьников и районов, где они обучаются, оценки влияния размера класса на оценки по тестам, представленные в части II, все же могут иметь смещения, вызванные пропущенными переменными, которые возникают из-за существования неучтенных переменных, таких как возможность обучения за пределами школы или «качество» преподавания того или иного учителя. Если данные по этим переменным недоступны, то вышеописанные смещения не могут быть устранены посредством добавления дополнительных регрессоров.

Регрессия с инструментальными переменными дает альтернативный подход к решению данной проблемы. Рассмотрим следующий гипотетический пример: некоторые калифорнийские школы вынуждены закрыться для восстановления

после летнего землетрясения. Наиболее близкие к эпицентру районы в большей степени подвержены влиянию землетрясения. Районы с несколькими закрытыми школами вынуждены распределить школьников, временно увеличивая размеры классов в других школах. Это означает, что расстояние от эпицентра удовлетворяет условию релевантности инструмента, поскольку коррелирует с размером учебного класса. Но если расстояние до эпицентра не будет связано с каким-либо из других факторов, влияющих на успеваемость учеников (например, продолжают ли школьники изучать английский язык), то тогда эта переменная может считаться экзогенной, поскольку не коррелирует с ошибкой регрессии. Таким образом, инструментальная переменная, расстояние до эпицентра землетрясения, могла бы использоваться, чтобы избавиться от смещения в оценках, вызванного пропущенными переменными, и чтобы оценить влияние размера класса на экзаменационные оценки школьников.

### **Выборочное распределение 2МНК-оценки**

Точное распределение 2МНК-оценки в маленьких выборках имеет достаточно сложный вид. Тем не менее, как и в случае МНК-оценок для больших выборок, это распределение имеет простой вид: 2МНК-оценки являются состоятельными и имеют нормальное распределение.

**Формула для 2МНК-оценок.** Несмотря на то что двухшаговая процедура оценки может показаться сложной, в случае одного регрессора  $X$  и одного инструмента  $Z$ , как и предполагается в данном разделе, существует простая формула для оценок 2МНК. Пусть  $s_{zy}$  – выборочная ковариация между  $Z$  и  $Y$ , а  $s_{zx}$  – выборочная ковариация между  $Z$  и  $X$ . Как показано в приложении 12.2, оценка 2МНК с одним инструментом равна:

$$\hat{\beta}_1^{TSLS} = \frac{s_{zy}}{s_{zx}}. \quad (12.4)$$

То есть 2МНК-оценка коэффициента  $\beta_1$  равна отношению выборочной ковариации между  $Z$  и  $Y$  к выборочной ковариации между  $Z$  и  $X$ .

**Выборочное распределение  $\hat{\beta}_1^{TSLS}$  в больших выборках.** Формула, представленная в выражении (12.4), может быть использована, чтобы показать, что  $\hat{\beta}_1^{TSLS}$  является состоятельной оценкой и в больших выборках имеет нормальное распределение. Основные аспекты доказательства представлены ниже, подробное математическое доказательство дано в приложении 12.3.

Состоятельность оценки  $\hat{\beta}_1^{TSLS}$  следует из предположения о том, что  $Z_i$  удовлетворяет условиям релевантности и экзогенности инструмента, а выборочные ковариации стремятся к своим теоретическим значениям при увеличении размера выборки. В первую очередь стоит заметить, что, поскольку  $Y_i = \beta_0 + \beta_1 X_i + u_i$  в выражении (12.1), то

$$\begin{aligned} \text{cov}(Z_i, Y_i) &= \text{cov}\left[Z_i, (\beta_0 + \beta_1 X_i + u_i)\right] = \\ &= \beta_1 \text{cov}(Z_i, X_i) + \text{cov}(Z_i, u_i), \end{aligned} \quad (12.5)$$

где второй знак равенства получается благодаря свойствам ковариаций [см. выражение (2.33)]. Согласно условию экзогенности инструмента  $\text{cov}(Z_i, u_i) = 0$  и условию релевантности инструмента,  $\text{cov}(Z_i, X_i) \neq 0$ . Тогда если инструмент является допустимым, то из выражения (12.5) можно получить:

$$\beta_1 = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)}. \quad (12.6)$$

Таким образом, теоретический коэффициент  $\beta_1$  равен отношению теоретической ковариации между  $Z$  и  $Y$  к теоретической ковариации между  $Z$  и  $X$ .

Как обсуждалось в разделе 3.7, выборочная ковариация является состоятельной оценкой теоретической ковариации, то есть  $s_{ZY} \xrightarrow{P} \text{cov}(Z_i, Y_i)$  и  $s_{ZX} \xrightarrow{P} \text{cov}(Z_i, X_i)$ . Тогда из выражений (12.4) и (12.6) следует, что 2МНК-оценка является состоятельной:

$$\hat{\beta}_1^{\text{TSLS}} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{P} \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)} = \beta_1. \quad (12.7)$$

Формула, представленная в выражении (12.4), может также использоваться для того, чтобы показать нормальность распределения  $\hat{\beta}_1^{\text{TSLS}}$  в больших выборках. Доказательство похоже на доказательство для любых других МНК-оценок, которые рассматривались ранее. 2МНК-оценка представляет собой некий вид среднего значения случайных переменных, и при увеличении размеров выборки, согласно центральной предельной теореме (ЦПТ), средние значения случайных переменных имеют нормальное распределение. В частности, числитель выражения для  $\hat{\beta}_1^{\text{TSLS}}$  в выражении (12.4) равен:  $s_{ZY} = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})$ , то

есть среднему значению  $(Z_i - \bar{Z})(Y_i - \bar{Y})$ . С помощью небольших преобразований, представленных в приложении 12.3, можно показать, что, согласно ЦПТ, в больших выборках  $\hat{\beta}_1^{\text{TSLS}}$  имеет выборочное распределение, которое приближенно можно рассматривать как  $N(\beta_1, \sigma_{\hat{\beta}_1^{\text{TSLS}}}^2)$ , где

$$\sigma_{\hat{\beta}_1^{\text{TSLS}}}^2 = \frac{1}{n} \frac{\text{var}[(Z_i - \mu_Z)u_i]}{\left[\text{cov}(Z_i, X_i)\right]^2}. \quad (12.8)$$

**Статистические выводы в больших выборках.** Дисперсия  $\sigma_{\hat{\beta}_1^{\text{TSLS}}}^2$  может быть оценена с помощью оценок дисперсии и ковариации, представленных в выражении (12.8). Квадратный корень из оценки  $\sigma_{\hat{\beta}_1^{\text{TSLS}}}^2$  – стандартная ошибка оценки, полученной с помощью оценки метода инструментальных переменных. Подобные вычисления производятся автоматически в различных эконометрических пакетах. Поскольку  $\hat{\beta}_1^{\text{TSLS}}$  имеет асимптотически нормальное распределение, то тестирование гипотез относительно  $\beta_1$  может проводиться с помощью  $t$ -статистик, а 95 %-е доверительные интервалы задаются с помощью  $\hat{\beta}_1^{\text{TSLS}} \pm 1,96SE(\hat{\beta}_1^{\text{TSLS}})$ .

## **Применение метода инструментальных переменных к исследованию спроса на сигареты**

Филипп Райт интересовался эластичностью спроса на масло, но сегодня другие товары, такие как сигареты, наиболее часто фигурируют в дебатах о государственной политике. Одним из способов сокращения болезней и смертельных случаев от курения, то есть издержек или внешних эффектов, является введение настолько высоких налогов на сигареты, чтобы текущее количество курильщиков сократилось, а потенциальные курильщики не стремились приобрести эту вредную привычку. Но каким в точности должно быть увеличение налогов, чтобы снизить потребление сигарет? Например, каким должно быть увеличение налогов, чтобы повышение цены на сигареты после уплаты всех налогов привело к снижению потребления сигарет на 20 %?

Ответ на этот вопрос зависит от эластичности спроса на сигареты. Если эластичность равна –1, то целевое снижение в 20 % может быть достигнуто посредством повышения цены на 20 %. Если эластичность равна –0,5, то для снижения спроса на 20 % необходимо повысить цену на 40 %. Конечно, никто точно не знает величину эластичности спроса на сигареты, поэтому ее необходимо оценить на основе данных по ценам на сигареты и по объемам их продаж. Но, как и в случае с маслом, из-за взаимодействия между спросом и предложением, эластичность спроса на сигареты не может быть состоятельно оценена с помощью МНК в регрессии логарифма объема продаж сигарет от логарифма их цены.

Таким образом, 2МНК будет использоваться для оценки эластичности спроса на сигареты на основе годовых данных для 48 штатов США в период с 1985 по 1995 год (описание данных представлено в приложении 12.1). Однако в данном разделе представлены результаты оценки регрессии для межобъектных данных для 1995 года, результаты оценок на данных за более ранние годы (панельные данные) представлены в разделе 12.4.

Инструментальная переменная  $SalesTax_i$  представляет собой объем налоговых поступлений с сигарет, выделенный из общей суммы налога с продаж и измеряемый в долларах с одной упаковки (в «реальных» долларах, дефлированных с помощью индекса потребительских цен). Объем потребления сигарет  $Q_i^{cigarettes}$  – это количество проданных упаковок сигарет на душу населения в определенном штате,  $P_i^{cigarettes}$  – реальная средняя цена за упаковку сигарет, включающая все налоги.

Перед использованием 2МНК важно проверить выполнение двух условий допустимости выбранной инструментальной переменной. Это будет детально проработано в разделе 12.3, где будут представлены некоторые статистические методы для упрощения этих действий. Но даже при наличии этих дополнительных методов очень важно не забывать об экономической сути проблемы, поэтому полезно подумать о том, удовлетворяет ли переменная, описывающая налог с продаж сигарет, двум необходимым условиям.

В первую очередь необходимо рассмотреть условие релевантности инструмента. Поскольку высокие налоги с продаж повышают цену  $P_i^{\text{cigarettes}}$ , которая устанавливается на товар после уплаты всех налогов, то переменная, описывающая подобный налог, в достаточной степени удовлетворяет условию релевантности инструментальной переменной.

Затем необходимо рассмотреть условие экзогенности инструмента. Чтобы переменная налога с продаж сигарет была экзогенной, необходимо, чтобы она была некоррелирована с остаточным членом в уравнении, описывающим кривую спроса. То есть налог с продаж должен влиять на спрос на сигареты только опосредованно через цену. Это требование, по-видимому, выполнимо: ставка общего налога с продаж разная по величине для различных штатов. Это происходит лишь потому, что правительство того или иного штата выбирает различные комбинации видов налогов (с продаж, с доходов, с собственности и т.д.) для финансирования своего бюджета. Такой выбор зачастую осуществляется на основе политического решения, а не факторов, которые влияют на спрос на сигареты. Правомерность этого утверждения будет более подробно рассмотрена в разделе 12.4, а сейчас это утверждение будет принято в качестве временной гипотезы.

В современных компьютерных статистических пакетах оценки первого шага 2МНК производятся автоматически, то есть нет необходимости оценивать эту регрессию самостоятельно, чтобы в итоге получить необходимые 2МНК-оценки. Тем не менее иногда полезно изучить результаты оценки регрессии на данном шаге оценивания. На основе имеющихся данных для 48 штатов в 1995 году были получены следующие результаты:

$$\widehat{\ln(P_i^{\text{cigarettes}})} = 4,63 + 0,031 \text{SalesTax}_i. \quad (12.9)$$

Как и ожидалось, более высокие налоги приводят к повышению цены. Коэффициент детерминации ( $R^2$ ) для этой регрессии равен 47%, то есть изменение налогов с продаж сигарет объясняют 47% изменения цены на сигареты в рассмотренных штатах.

На втором шаге 2МНК оценивается регрессия  $\widehat{\ln(Q_i^{\text{cigarettes}})}$  на  $\widehat{\ln(P_i^{\text{cigarettes}})}$  с помощью МНК. Результаты оценки могут быть записаны в таком виде:

$$\widehat{\ln(Q_i^{\text{cigarettes}})} = 9,72 - 1,08 \widehat{\ln(P_i^{\text{cigarettes}})}. \quad (12.10)$$

Данное выражение записано с помощью регрессора, который использовался на второй стадии —  $\widehat{\ln(P_i^{\text{cigarettes}})}$ . Однако удобнее и проще представлять результаты оценки регрессионной функции, используя  $\widehat{\ln(P_i^{\text{cigarettes}})}$ , чем  $\widehat{\ln(Q_i^{\text{cigarettes}})}$ . Оценки, полученные с помощью 2МНК, а также оценки устойчивых к наличию гетероскедастичности стандартных ошибок равны:

$$\widehat{\ln(Q_i^{\text{cigarettes}})} = 9,72 - 1,08 \widehat{\ln(P_i^{\text{cigarettes}})}. \quad (12.11)$$

Эти оценки показывают, что спрос на сигареты в удивительной степени эластичен по своей природе: увеличение цены на 1% приводит к снижению

потребления сигарет на 1,08 %. Однако вспоминая обсуждения экзогенности инструмента, возможно, эту оценку не стоит воспринимать слишком серьезно. Несмотря на то что эластичность была оценена с использованием инструментальных переменных, все еще могут существовать пропущенные переменные, которые коррелируют с переменной объемов налогов на сигареты. Одной из таких переменных, например, может быть уровень дохода: штаты с более высокими доходами могут в меньшей степени зависеть от налогов с продаж, но больше от налогов на прибыль, которые используются для финансирования государственного управления. Кроме того, спрос на сигареты, по-видимому, зависит от дохода. Таким образом, необходимо еще раз провести оценки уравнений спроса с добавлением переменной уровня дохода в качестве дополнительного регрессора. Для этого, однако, необходимо сначала расширить модель IV-регрессии для включения дополнительных регрессоров.

## 12.2. Обобщенная модель регрессии с инструментальными переменными

Обобщенная модель регрессии с инструментальными переменными содержит четыре типа переменных: зависимая переменная  $Y$ ; проблемные эндогенные регрессоры, например цена на сигареты, которые связаны с ошибкой регрессии и которые будут обозначаться как  $X$ ; дополнительные регрессоры, которые называются *включенными экзогенными переменными* и будут обозначаться как  $W$ , а также инструментальные переменные  $Z$ . В общем случае может быть несколько регрессоров  $X$ , несколько включенных экзогенных регрессоров  $W$  и несколько инструментальных переменных  $Z$ .

Для возможности оценки IV-регрессии необходимо, по крайней мере, столько инструментальных переменных  $Z$ , сколько имеется регрессоров  $X$ . В разделе 12.1 рассматривался случай с одним эндогенным регрессором и одним инструментом. Наличие (по крайней мере) одного инструмента для этого одного эндогенного регрессора имеет важное значение. Без инструмента не было бы возможности вычислить IV-оценки: не было бы первого шага оценки 2МНК.

Соотношение между количеством инструментов и количеством эндогенных регрессоров имеет свою терминологию. Говорят, что коэффициенты регрессии *точно определены*, если количество инструментов ( $m$ ) равно количеству эндогенных регрессоров ( $k$ ), то есть  $m = k$ . Коэффициенты называются *переопределеными*, если количество инструментов превышает количество эндогенных регрессоров, то есть  $m > k$ . Коэффициенты называют *недоопределенными*, если количество инструментов меньше, чем количество эндогенных регрессоров, то есть  $m < k$ . Коэффициенты должны быть точно определены или переопределены, если необходимо оценить их с помощью регрессии с инструментальными переменными.

Описание обобщенной модели регрессии с инструментальными переменными и ее терминология представлены во вставке «Основные понятия 12.1».

## ОСНОВНЫЕ ПОНЯТИЯ

### 12.1

#### Обобщенная модель регрессии с инструментальными переменными

Обобщенная модель регрессии с инструментальными переменными может быть записана в таком виде:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \dots \\ &\dots + \beta_{k+r} W_{ri}, i = 1, \dots, n, \end{aligned} \quad (12.12)$$

где

- $Y_i$  – зависимая переменная;
- $\beta_0, \beta_1, \dots, \beta_{k+r}$  – неизвестные регрессионные коэффициенты;
- $X_{1i}, \dots, X_{ki}$  –  $k$  эндогенных регрессоров, которые потенциально могут коррелировать с  $u_i$ ;
- $W_{1i}, \dots, W_{ri}$  –  $r$  включенных экзогенных регрессоров, которые не коррелируют с  $u_i$  или являются контрольными переменными;
- $u_i$  – остаточный член, который представляет собой ошибки измерения или пропущенные факторы;
- $Z_i$  –  $m$  инструментальных переменных.

Коэффициенты регрессии точно определены, если количество инструментов ( $m$ ) равно количеству эндогенных регрессоров ( $k$ ), то есть  $m = k$ . Коэффициенты называются переопределеными, если количество инструментов превышает количество эндогенных регрессоров, то есть  $m > k$ . Коэффициенты называют недопределенными, если количество инструментов ниже, чем количество эндогенных регрессоров, то есть  $m < k$ . Для использования модели регрессии с инструментальными переменными коэффициенты должны быть точно определены или переопределены.

**Включение экзогенных и контрольных переменных в IV-регрессию.** Все  $W$  переменные в уравнении (12.12) могут быть либо экзогенными переменными, для которых  $E(u_i|W_i) = 0$ , либо могут являться контрольными переменными, которые необязательно должны иметь некую интерпретацию, но добавлены, чтобы удостовериться, что инструментальные переменные не коррелируют с остаточным членом. Например, в разделе 12.1 поднимался вопрос о возможной корреляции налога с продаж и дохода, которые, как считается в экономической теории, определяют спрос на сигареты. Если корреляция действительно имеет место, то переменная налога с продаж будет коррелирована с ошибкой регрессии в уравнении спроса  $\ln(Q_i^{\text{cigarettes}}) = \beta_0 + \beta_1 \ln(P_i^{\text{cigarettes}}) + u_i$  и, таким образом, не будет являться экзогенным инструментом. Использование в регрессии переменной дохода устранит этот источник потенциальной корреляции между инструментом и остаточным членом. В общем случае, если  $W$  является эффективной контрольной переменной в IV-регрессии, то добавление  $W$  позволяет устранить корреляцию инструмента с  $u$ , и, таким образом, 2МНК-оценка коэффициента при  $W$  является состоятельной. Если  $W$  коррелирует с  $u$ , то 2МНК-оценка коэффициента при  $W$  имеет смещение, вызванное пропущенными переменными. Логика,

согласно которой контрольные переменные включаются в IV-регрессию, таким образом, сходна с логикой, по которой контрольные переменные включаются в МНК-регрессию, обсуждаемую в разделе 7.5.

Математическое условие эффективности  $W$  как контрольной переменной в IV-регрессии сходно с условиями для контрольных переменных в МНК-регрессиях, которые обсуждались в разделе 7.5. В частности, включение  $W$  должно привести к тому, что условное среднее значение  $u$  не зависит от  $Z$ , то есть выполняется условие независимости условного среднего  $E(u_i|Z_i, W_i) = E(u_i | W_i)$ . Для простоты понимания в данной главе будет сделан акцент на случае, в котором переменные  $W$  являются экзогенными, то есть  $E(u_i|W_i) = 0$ . В приложении 12.6 даются пояснения о том, как результаты данной главы могут быть обобщены на случаи, в которых  $W$  является контрольной переменной, а условие  $E(u_i|W_i) = 0$  заменено условием независимости условного среднего  $E(u_i|Z_i, W_i) = E(u_i | W_i)$ .

## **2МНК в обобщенной модели регрессии с инструментальными переменными**

**2МНК с одним эндогенным регрессором.** При наличии одного эндогенного регрессора  $X$  и нескольких дополнительных экзогенных переменных уравнение регрессии может быть записано в таком виде:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \dots + \beta_{1+r} W_{ri} + u_i, \quad (12.13)$$

где, как и ранее,  $X_i$  может быть коррелирован с ошибкой, а  $W_{1i}, \dots, W_{ri}$  не могут быть коррелированы с ошибкой регрессии.

На первом шаге 2МНК оценивается зависимость между  $X$  и экзогенными переменными, то есть  $W$  и инструментами  $Z$ :

$$X_i = \pi_0 + \pi_1 Z_{1i} + \dots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \dots + \pi_{m+r} W_i + v_i, \quad (12.14)$$

где  $\pi_0, \pi_1, \dots, \pi_{m+r}$  – неизвестные регрессионные коэффициенты,  $v_i$  – ошибка регрессии.

Уравнение (12.14) иногда называют *приведенной формой* уравнения для  $X$ . Оно связывает эндогенную переменную  $X$  со всеми доступными экзогенными переменными, которые используются в регрессии ( $W$  и инструментальные переменные  $Z$ ).

На первом шаге 2МНК неизвестные коэффициенты в уравнении (12.14) оцениваются с помощью МНК, а оценки значений, получаемые из оценки регрессии на этой стадии, записываются как  $\hat{X}_1, \dots, \hat{X}_n$ .

На втором шаге 2МНК с помощью МНК оценивается уравнение (12.13), в котором  $X_i$  заменяются их оценками, полученными на первом шаге. То есть с помощью МНК оценивается регрессия  $Y_i$  на  $\hat{X}_i, W_{1i}, \dots, W_{ri}$ . Получившиеся оценки коэффициентов  $\beta_0, \beta_1, \dots, \beta_{1+r}$  являются 2МНК-оценками.

**Расширение 2МНК на случай нескольких эндогенных регрессоров.** При наличии нескольких эндогенных регрессоров  $X_{1i}, \dots, X_{ki}$  алгоритм 2МНК в целом аналогичен описанному выше, за исключением того, что каждый из эндогенных регрессоров требует оценки своей собственной регрессии на первом шаге. Каждая

из этих регрессий имеет форму, аналогичную уравнению (12.14). Это означает, что зависимой переменной в них является тот или иной  $X$ , а регрессорами являются все инструментальные переменные  $Z$  и все используемые экзогенные переменные  $W$ . Первый шаг дает оценки значений всех эндогенных регрессоров.

На втором шаге 2МНК уравнение (12.12) оценивается с помощью МНК, за исключением того, что эндогенные регрессоры  $X$  заменяются на их оценки, полученные на первом шаге. Полученные на втором шаге оценки коэффициентов  $\beta_0, \beta_1, \dots, \beta_{k+r}$  являются 2МНК-оценками.

На практике оценки, полученные на каждом из двух шагов 2МНК, могут быть получены автоматически с помощью различных современных эконометрических программных пакетов. Двухшаговый метод наименьших квадратов кратко описан во вставке «Основные понятия 12.2».

## ОСНОВНЫЕ ПОНЯТИЯ

### 12.2

#### Двухшаговый метод наименьших квадратов

Оценки двухшагового метода наименьших квадратов (2МНК) в модели регрессии с несколькими инструментальными переменными (IV-регрессиях), описываемой уравнением (12.12), вычисляются при помощи двух шагов:

1. На *первом шаге* с помощью МНК (с включением свободного члена в спецификацию) проводится оценка регрессии  $X_{li}$  на инструментальные переменные ( $Z_{li}, \dots, Z_{mi}$ ) и используемые экзогенные переменные ( $W_{li}, \dots, W_{ri}$ ). Затем вычисляются оценки значений эндогенной переменной  $\hat{X}_{li}$ . Аналогичные действия нужно сделать для всех эндогенных регрессоров, получив подобным образом оценки всех эндогенных переменных  $\hat{X}_{li}, \dots, \hat{X}_{ki}$ .
2. На *втором шаге* с помощью МНК (с включением свободного члена в спецификацию) оценивается регрессия  $Y_i$  на оценки эндогенных переменных, полученные на первом шаге ( $\hat{X}_{li}, \dots, \hat{X}_{ki}$ ), и используемые экзогенные регрессоры ( $W_{li}, \dots, W_{ri}$ ), включая константу. Полученные оценки  $\hat{\beta}_0^{TSLS}, \dots, \hat{\beta}_{k+r}^{TSLS}$  представляют собой искомые 2МНК-оценки.

На практике оценки в рамках каждого из двух шагов 2МНК получаются автоматически с помощью современных эконометрических программных пакетов.

## Релевантность и экзогенность инструментов в обобщенной модели регрессии с инструментальными переменными

Условия релевантности и экзогенности инструментов в обобщенной модели регрессии с инструментальными переменными необходимо модифицировать.

При наличии одной эндогенной переменной, но нескольких инструментов, условие релевантности для инструмента заключается в том, что, по крайней мере, один регрессор  $Z$  может использоваться для оценок  $X$  при заданном  $W$ . При наличии нескольких эндогенных переменных это условие должно быть более сложным, поскольку мы должны исключить возможность возникновения совершенной мультиколлинеарности при оценке регрессии на втором шаге. Интуитивно понятно, что при наличии нескольких эндогенных переменных инструменты должны «давать» достаточно информации об экзогенных изменениях в этих переменных, чтобы выявить их влияние на  $Y$ .

Общий вид условий экзогенности инструмента заключается в том, что каждый инструмент не должен быть коррелирован с ошибкой регрессии  $u_i$ . Условия допустимости инструментов приведены во вставке «Основные понятия 12.3».

### Условия допустимости инструментов

Набор из  $m$  инструментов ( $Z_{1i}, \dots, Z_{mi}$ ) должен удовлетворять следующим условиям допустимости:

1. Условие релевантности инструментов

- В общем случае пусть  $\hat{X}_{1i}^*$  оценка регрессора  $X_{1i}$ , полученная с помощью теоретической регрессии  $X_{1i}$  на инструменты ( $Z$ 'ы) и используемые экзогенные регрессоры  $W$ . Пусть «1» обозначает константу, которая принимает значение 1 для всех наблюдений. Тогда  $(\hat{X}_{1i}^*, \dots, \hat{X}_{ki}^*, W_{1i}, \dots, W_{ri}, 1)$  не являются совершенно мультиколлинеарными.
- При наличии лишь одного регрессора  $X$  для выполнения упомянутого выше условия необходимо наличие хотя бы одного ненулевого коэффициента в регрессии  $X$  на  $Z$ 'ы и  $W$ 'ы.

2. Условие экзогенности инструментов

- Инструменты не должны коррелировать с остаточным членом, то есть  $\text{corr}(Z_{1i}, u_i) = 0, \dots, \text{corr}(Z_{mi}, u_i) = 0$ .

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**12.3**

## Предположения регрессии с инструментальными переменными и выборочное распределение 2МНК-оценок

В рамках предположений модели регрессии с инструментальными переменными 2МНК-оценки являются состоятельными и имеют выборочное распределение, которое в больших выборках можно считать приблизительно нормальным.

**Предположения ИП регрессии.** Основные предположения модели IV-регрессии представляют собой модификацию предположений МНК в случае множественной регрессии, представленных во вставке «Основные понятия 6.4».

Первое предположение в рамках IV-регрессии является модификацией предположения об условном среднем, представленном во вставке «Основные понятия 6.4» таким образом, чтобы оно применялось только для используемых экзогенных переменных. Так же, как и во втором предположении МНК

для множественной регрессии, в рамках второго предположения для модели IV-регрессии считается, что наблюдения являются независимыми случайно распределенными величинами, как если бы данные были получены случайным образом. Аналогично и третье предположение IV-регрессии говорит о том, что выбросы в используемой выборке маловероятны.

Четвертое предположение IV-регрессии заключается в том, что выполняются два условия допустимости инструментов, представленные во вставке «Основные понятия 12.3». Условие релевантности инструментов, представленное во вставке «Основные понятия 12.3», включает в себя четвертое предположение МНК во вставке «Основные понятия 4.6» (отсутствие совершенной мультиколлинеарности) посредством предположения о том, что регрессоры в регрессиях на второй стадии не являются коллинеарными. Основные предположения модели регрессии с инструментальными переменными представлены во вставке «Основные понятия 12.4».

**ОСНОВНЫЕ  
ПОНЯТИЯ**

**12.4**

**Предположения модели регрессии  
с инструментальными переменными**

Переменные и ошибка в модели регрессии с инструментальными переменными, представленной во вставке «Основные понятия 12.1», удовлетворяют следующим условиям:

1.  $E(u_i | W_{1i}, \dots, W_{ni}) = 0$ ; являются независимыми одинаково распределенными наблюдениями, полученными из их совместного распределения.
2.  $(X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ni}, Z_{1i}, \dots, Z_{mi}, Y_i)$  являются независимыми одинаково распределенными наблюдениями, полученными из их совместного распределения.
3. Большие выбросы маловероятны: переменные  $X, W, Z, Y$  имеют ненулевые конечные четвертые моменты.
4. Должны выполняться два условия допустимости инструмента, представленные во вставке «Основные понятия 12.3».

**Выборочное распределение 2МНК-оценок.** В рамках предположений IV-регрессии 2МНК-оценки являются состоятельными и нормально распределенными в больших выборках. Это было показано в разделе 12.1 (и приложении 12.3) для частного случая одного эндогенного регрессора и одного инструмента без использования дополнительных экзогенных переменных. Концептуально, рассуждения из раздела 12.1 повторяются для общего случая нескольких инструментов и нескольких эндогенных переменных. Уравнения в общем случае достаточно сложны и будут рассмотрены в главе 18.

**Статистические выводы с использованием 2МНК-оценок**

Поскольку выборочное распределение 2МНК-оценки является асимптотически нормальным, общие подходы к формированию статистических выводов

(проверка гипотез и построение доверительных интервалов) в регрессионных моделях подходят и для случая 2МНК. Например, 95%-е доверительные интервалы строятся как 2МНК-оценка  $\pm 1,96$  стандартной ошибки. Аналогично совместные гипотезы об истинных (теоретических) значениях коэффициентов могут быть проверены с помощью  $F$ -статистики, как описано в разделе 7.2.

**Вычисление стандартных ошибок 2МНК-оценок.** Есть два момента, о которых следует помнить при вычислении стандартных ошибок 2МНК-оценок. Во-первых, стандартные ошибки, которые можно вычислить при помощи МНК-оценки на втором шаге оценивания, неверны, поскольку они не учитывают, что это второй этап двухэтапного процесса. В частности, получаемые на втором этапе стандартные ошибки МНК не могут учесть использования на втором шаге 2МНК предсказанных значений используемых эндогенных переменных. Формулы для стандартных ошибок, в которых сделаны необходимые корректировки, включены (и автоматически используются) в стандартные эконометрические пакеты. Таким образом, все это не является проблемой на практике при использовании специализированных команд в эконометрических программных пакетах.

Во-вторых, как и всегда, ошибка регрессии  $\mu$  может быть гетероскедастична. Поэтому важно использовать устойчивые к наличию гетероскедастичности стандартные ошибки по той же самой причине, по которой важно использовать устойчивые к гетероскедастичности стандартные ошибки для МНК-оценок в модели множественной регрессии.

### **Приложение к изучению спроса на сигареты**

В разделе 12.1 оценивалась эластичность спроса на сигареты на основе данных о годовом потреблении сигарет в 48 штатах США в 1995 году с помощью 2МНК с одним регрессором (логарифм реальной цены за одну упаковку сигарет) и одного инструмента (реальный налог с продаж одной упаковки сигарет). Доход также влияет на спрос, поэтому он неявно представляет собой часть ошибки теоретической регрессии. Как отмечалось в разделе 12.1, если налог с продаж связан с уровнем доходов в штате, то он коррелирует с ошибкой уравнения спроса на сигареты, что является нарушением условия экзогенности инструмента. В этом случае IV-оценка в разделе 12.1 не является состоятельной. То есть в IV-регрессии имеет место некий вариант смещения оценок, вызванного пропущенными переменными. Для решения этой проблемы необходимо включить переменную дохода в уравнение регрессии.

Таким образом, рассмотрим альтернативную спецификацию модели спроса на сигареты, в которую дополнительно включен логарифм уровня дохода. В терминологии из вставки «Основные понятия 12.1» зависимой переменной  $Y$  является логарифм потребления  $\ln(Q_i^{\text{cigarettes}})$ , эндогенным регрессором  $X$  является логарифм реальной цены после налогообложения  $\ln(P_i^{\text{cigarettes}})$ , дополнительным экзогенным регрессором  $W$  является логарифм реального дохода на душу населения в том или ином штате  $\ln(Inc_i)$ , инструментальной переменной  $Z$  является реальный налог с продаж одной упаковки сигарет  $SalesTax_i$ . 2МНК-оценки и устойчивые к гетероскедастичности стандартные ошибки равны:

$$\widehat{\ln(Q_i^{\text{cigarettes}})} = 9,43 - 1,14 \ln(P_i^{\text{cigarettes}}) + 0,21 \ln(Inc_i). \quad (12.15)$$

В этой регрессии используется одна инструментальная переменная  $SalesTax_i$ , однако есть еще один кандидат на эту роль. В дополнение к общим налогам с продаж в некоторых штатах вводятся специальные виды налогов, которые применяются только к сигаретам и другим табачным изделиям. Эти налоги ( $CigTax_i$ ) представляют собой вторую возможную инструментальную переменную. Специфические налоги на сигареты увеличивают цену сигарет для потребителя и, таким образом, соответствуют условию релевантности инструмента. Если эта переменная не коррелирует с остаточным членом в уравнении спроса на сигареты, то ее можно рассматривать как экзогенный инструмент.

Теперь у нас есть две инструментальные переменные: реальный налог с продаж одной упаковки и реальный специфический налог на пачку сигарет в отдельном штате. Наличие двух инструментов и одного эндогенного регрессора означает, что показатель эластичности спроса по цене будет переопределен, так как число инструментов ( $SalesTax_i, CigTax_i, m = 2$ ) превышает число эндогенных переменных ( $P_i^{\text{cigarettes}}, k = 1$ ). Можно оценить эластичность спроса с использованием 2МНК, где на первом шаге в регрессию включены экзогенная переменная  $\ln(Inc_i)$  и оба инструмента.

Тогда 2МНК-оценки коэффициентов регрессии равны:

$$\widehat{\ln(Q_i^{\text{cigarettes}})} = 9,89 - 1,28 \ln(P_i^{\text{cigarettes}}) + 0,28 \ln(Inc_i). \quad (12.16)$$

Сравните уравнения (12.15) и (12.16): стандартная ошибка оценки эластичности по цене меньше на одну треть в уравнении (12.16) [0,25 в уравнение (12.16) по сравнению с 0,37 в уравнении (12.15)]. Причина, по которой стандартная ошибка в уравнении (12.16) стала меньше, заключается в том, что для получения этой оценки используется больше информации, чем в уравнении (12.15). В формуле (12.15) используется только один инструмент (налог с продаж), а в уравнении (12.16) используются два инструмента (налог с продаж и специфический налог на сигареты). Использование двух инструментов позволяет объяснить большую часть изменений цен на сигареты, чем при использовании только одного инструмента, и это отражается в меньших по величине стандартных ошибках оценки эластичности спроса по цене.

Являются ли эти оценки заслуживающими доверия? В конечном счете их достоверность зависит от того, удовлетворяет ли набор инструментальных переменных, в данном случае – две переменные, отражающие различные виды налогов, двум условиям допустимости инструментов. Поэтому очень важно проверить допустимость инструментов, что и будет сделано в следующем разделе.

### 12.3. Проверка допустимости инструментов

Степень полезности регрессии с инструментальными переменными в рамках того или иного приложения зависит от того, являются ли инструменты до-

пустыми. Неправильный выбор инструментов приводит к бессмысленным результатам. Поэтому важно оценить, является ли данный набор инструментов допустимым в данной конкретной ситуации.

### **Предположение № 1: релевантность инструмента**

Роль условия релевантности инструментов в модели IV-регрессии представляется собой достаточно узкое место. С одной стороны, можно считать, что значимость инструментов играет схожую роль с размером выборки: чем в большей степени релевантными являются инструменты, то есть чем больше изменений в  $X$  может быть объяснено с помощью инструментов, тем больше информации доступно для использования в IV-регрессии. Более релевантный инструмент дает более точную оценку, равно как и выборка большего размера позволяет получить более точные оценки. Кроме того, статистические выводы, полученные с помощью 2МНК, основываются на 2МНК-оценках, имеющих нормальное выборочное распределение, но в соответствии с центральной предельной теоремой нормальное распределение является хорошим приближением в больших, но необязательно в малых выборках. Если наличие более релевантного инструмента соответствует наличию выборки большего размера, то это говорит о том, что чем «более релевантным» является инструмент, тем лучше подходит нормальное распределение для приближения выборочного распределения 2МНК-оценок и  $t$ -статистики.

Инструменты, которые объясняют малую часть изменений  $X$ , называют *слабыми инструментами*. В приведенном выше примере с изучением спроса на сигареты расстояние штата от фабрики, где производятся сигареты, возможно, будет слабым инструментом. Несмотря на то что большее расстояние увеличивает стоимость доставки (и, таким образом, смещает кривую предложения и повышает равновесную цену), сигареты имеют небольшой вес, поэтому издержки на транспортировку составляют незначительную часть их цены. Таким образом, колебания цен, которые объясняются расходами на транспортировку и, следовательно, расстоянием до заводов-изготовителей, по-видимому, весьма малы.

В этом разделе будет рассмотрено, почему слабые инструменты могут приводить к некоторым сложностям, как выявить слабые инструменты и что делать при наличии слабых инструментов. Предполагается, что инструменты являются экзогенными.

**Почему слабые инструменты являются проблемой?** Если инструменты являются слабыми, то нормальное распределение является плохим приближением выборочного распределения 2МНК-оценки, даже если размер выборки достаточно велик. Таким образом, обычные методы построения статистических выводов не получают теоретического обоснования даже в больших выборках. В самом деле, если инструменты слабые, то 2МНК-оценки могут быть сильно смещены в направлении МНК-оценки. Кроме того, 95 %-й доверительный интервал, который строится как 2МНК-оценка  $\pm 1,96$  стандартной ошибки, может содержать истинное значение коэффициента гораздо меньше, чем в 95 % случаев. Короче говоря, если инструменты слабы, 2МНК-оценки больше не являются надежными.

Чтобы увидеть, что существуют проблемы с приближением выборочного распределения 2МНК-оценки нормальным распределением в больших выборках,

рассмотрим частный случай, который уже обсуждался в разделе 12.1 для одной эндогенной переменной, одного инструмента и отсутствия экзогенных регрессоров. Если инструмент не является слабым, то  $\hat{\beta}_1^{TSLS}$  является состоятельной, потому что выборочные ковариации  $s_{ZY}$  и  $s_{ZX}$  являются состоятельными, то есть

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{P} \text{cov}(Z_i, Y_i) / \text{cov}(Z_i, X_i) = \beta_1 \quad [\text{уравнение (12.7)}].$$

Теперь предположим,

что инструмент не только является слабым, но и не является релевантным, то есть  $\text{cov}(Z_i, X_i) = 0$ . Тогда  $s_{ZX} \xrightarrow{P} \text{cov}(Z_i, X_i) = 0$ , то есть формально знаменатель правой части приведенного выше равенства  $\text{cov}(Z_i, Y_i) / \text{cov}(Z_i, X_i)$  равен нулю. Таким образом, состоятельность  $\hat{\beta}_1^{TSLS}$  нарушается, если нарушается условие релевантности инструмента. Как показано в приложении 12.4, это нарушение приводит к тому, что 2МНК-оценки имеют выборочное распределение, отличное от нормального, даже если размер выборки очень велик. Фактически, когда инструмент не является релевантным, распределение  $\hat{\beta}_1^{TSLS}$  в больших выборках не совпадает с распределением нормальной случайной величины, но скорее совпадает с распределением *отношения* двух нормальных случайных величин.

Несмотря на то что подобные инструменты, полностью не являющиеся релевантными, не встречаются на практике, возникает вопрос о том, «насколько релевантными» должны быть инструменты, чтобы обеспечить хорошее их приближение нормальным распределением на практике? Дать ответ на этот вопрос в рамках обобщенной модели регрессии с инструментальными переменными достаточно сложно. К счастью, есть простое правило для наиболее распространенных на практике ситуаций в случае одного эндогенного регрессора.

**Тестирование слабых инструментов в случае одного эндогенного регрессора.** Одним из способов проверки на наличие слабых инструментов в случае одного эндогенного регрессора является вычисление на первом шаге 2МНК *F*-статистики для последующей проверки гипотезы о том, что все коэффициенты при инструментальных переменных равны нулю. Эта *F*-статистика, вычисленная на первом шаге, является «мерой информации», содержащейся в инструментах: чем больше информации, тем больше ожидаемое значение *F*-статистики. Существует одно простое эмпирическое правило – не следует беспокоиться о слабости инструментов, если на первом шаге *F*-статистика превышает 10. (Почему именно 10? См. приложение 12.5.) Приведенные выше утверждения приведены во вставке «Основные понятия 12.5».

## ОСНОВНЫЕ ПОНЯТИЯ 12.5

### Эмпирическое правило тестирования слабых инструментов

Построенная на первом этапе 2МНК *F*-статистика представляет собой *F*-статистику для проверки гипотезы о том, что коэффициенты при инструментах  $Z_{1i}, \dots, Z_{mi}$  равны нулю на первом шаге двухшагового метода наименьших квадратов. Если в случае одного эндогенного регрессора построенная на первом шаге *F*-статистика меньше 10, то инструменты являются слабыми, и в этом случае 2МНК-оценка подвержена смещениям (даже в больших выборках), а построенные в рамках 2МНК *t*-статистики и доверительные интервалы являются ненадежными.

**Что делать в случае слабых инструментов?** При наличии нескольких инструментов некоторые из них, вероятно, слабее, чем другие. В случае небольшого числа сильных инструментов и большого числа слабых инструментов лучше отбросить слабые инструменты и использовать наиболее подходящие сильные инструменты для анализа с помощью 2МНК. Стандартные ошибки 2МНК могут вырасти при отбрасывании слабых инструментов, однако необходимо помнить, что исходные стандартные ошибки не были бы значимыми в любом случае!

Если, однако, коэффициенты точно определены, нельзя отбросить слабые инструменты. Даже если коэффициенты переопределены, может быть недостаточно сильных инструментов для идентификации, и тогда отбрасывание некоторых слабых инструментов не принесет положительного результата. В этом случае есть две альтернативы. Первый способ заключается в поиске дополнительных более сильных инструментов. Но это легче сказать, чем сделать: этот способ требует глубокого знания проблемы и может повлечь за собой пересмотр набора данных и характера эмпирического исследования. Второй способ заключается в продолжении эмпирического анализа с использованием слабых инструментов, но с использованием отличных от 2МНК методов. Хотя в данной главе основной акцент сделан на 2МНК, существует ряд других методов анализа с помощью инструментальных переменных, которые менее чувствительны к слабым инструментам, чем 2МНК. Некоторые из этих методов будут представлены в приложении 12.5.

### **Предположение № 2: экзогенность инструмента**

Если инструменты не являются экзогенными, то 2МНК-оценки не являются состоятельными: 2МНК-оценка сходится по вероятности к величине, отличной от теоретического значения коэффициента регрессии. Кроме того, одна из идей модели регрессии с инструментальными переменными заключается в том, что инструмент содержит некую информацию об изменениях  $X_i$ , которые не связаны с остаточным членом  $u_i$ . Если инструмент не является экзогенным, то нельзя точно определить эти экзогенные изменения в  $X_i$ , и тогда IV-регрессия не в состоянии обеспечить состоятельность оценки. Математическое обоснование этого утверждения приведено в приложении 12.4.

**Можно ли статистически протестировать предположение о том, что инструмент является экзогенным?** И да и нет. С одной стороны, невозможно проверить гипотезу о том, что инструменты являются экзогенными, когда коэффициенты точно определены. С другой стороны, если коэффициенты переопределены, можно проверить так называемые сверхидентифицирующие ограничения, то есть проверить гипотезу о том, что «лишние» инструменты являются экзогенными в рамках тестируемой гипотезы о существовании достаточного количества допустимых инструментов для определения коэффициентов.

Сначала рассмотрим случай, когда коэффициенты точно определены, то есть инструментов столько же, сколько и эндогенных регрессоров. Тогда невозможно построить статистический тест для проверки гипотезы о том, что инструменты действительно являются экзогенными. То есть эмпирические

данные не могут быть использованы для ответа на вопрос о том, удовлетворяют ли эти инструменты условию экзогенности. В этом случае единственным способом оценки экзогенности этих инструментов является использование экспертных оценок (мнения экспертов) и собственных знаний о конкретном эмпирическом приложении и связанных с ним проблемах. Например, опыт Филиппа Райта в исследованиях спроса и предложения в сфере сельского хозяйства позволил ему выдвинуть предположение о том, что маленькие осадки (ниже среднего уровня), скорее всего, будут приводить к сдвигу кривой предложения сливочного масла, но не будут непосредственно влиять на кривую спроса.

Оценка того, являются ли инструменты экзогенными, обязательно требует проведения экспертной оценки на основе личных знаний и опыта в изучаемой сфере. Если, однако, имеется больше инструментов, чем эндогенных регрессоров, то существует статистический инструмент, который может помочь в поисках ответа – так называемый тест на сверхидентифицирующие ограничения.



### ***«Пугающая» регрессия***

Одним из способов оценки зависимости процентного увеличения доходов от каждого дополнительного года посещения школы («отдача от образования») является оценка регрессии логарифма дохода от числа лет обучения в школе на основе базы данных по физическим лицам. Но если более способные люди являются более успешными на рынке труда и посещают школу дольше (возможно, потому, что им легчедается обучение), то продолжительность обучения в школе будет коррелировать с пропущенной переменной, которая отражает врожденные способности человека, а МНК-оценки отдачи от образования будут смещены. Поскольку врожденные способности человека измерить чрезвычайно трудно, из-за чего они не могут быть использованы в качестве одного из регрессоров, некоторые экономисты обратились к IV-регрессии для оценки отдачи от образования. Но какая переменная коррелирует с числом лет обучения, но не коррелирует с ошибкой регрессии? Какую переменную можно использовать в качестве допустимого инструмента?

Экономисты Джошуа Энгрист (Joshua Angrist) и Аллан Крюгер (Alan Krueger) предложили рассматривать в качестве такой инструментальной переменной дату рождения. В соответствии с законом об обязательном школьном образовании дата рождения коррелирует с числом лет обучения. Если закон требует, чтобы вы посещали школу до 16-летнего возраста и вам исполнится 16 лет в январе, пока вы находитесь в десятом классе, то вы можете бросить обучение, но если вам исполнится 16 в июле, то вы уже должны будете закончить десятый класс. Таким образом, дата рождения удовлетворяет условию релевантности инструмента. Однако дата рождения в январе или июле не должна оказывать прямого влияния на уровень дохода (в отличие от продолжительности обучения), так что дата рождения удовлетворяет условию экзогенности инструмента. Данная идея была реализована при

помощи использования в качестве инструментальной переменной квартала (период в 3 месяца), в котором расположена дата рождения индивида. Для оценок была использована большая выборка данных из базы данных U.S. Census (в регрессии использовалось как минимум 329 тыс. наблюдений), и также рассматривались другие переменные, такие как возраст работника.

Однако еще один экономист, занимающийся экономикой труда, Джон Баунд (John Bound), был настроен достаточно скептически. Он знал, что использование слабых инструментов приводит к ненадежности 2МНК-оценок, и был обеспокоен тем, что, несмотря на чрезвычайно большой размер выборки, квартал, в котором находится дата рождения индивида, может быть слабым инструментом в некоторых спецификациях. Поэтому при встрече Баунд и Крюгер обсуждали вопрос о том, были ли используемые Энгристом и Крюгером инструменты слабыми. Крюгер считал, что это не так, и предложил оригинальный способ проверки. Почему бы не провести оценку регрессии еще раз с использованием явно нерелевантного инструмента, например заменив реальный квартал рождения индивида на неверный, случайно сгенерированный на компьютере квартал рождения, и сравнить результаты, полученные с использованием реальных и заведомо неверных инструментов? Были получены удивительные результаты: при использовании настоящего квартала рождения и заведомо неверного квартала рождения 2МНК дал одинаковые результаты!

Такой результат оказался довольно «пугающим» для экономистов, занимающихся исследованиями в сфере экономики труда. Стандартные ошибки 2МНК, полученные с использованием реальных данных, говорят о том, что оценки эффекта отдачи от образования точны, но тот же результат дают и стандартные ошибки, полученные с использованием заведомо неверных данных. Конечно, неверные данные не позволяют оценить эффект отдачи от образования точно, потому что неверный инструмент не является релевантным. Основная проблема заключается в том, что 2МНК-оценки, основанные на реальных данных, так же ненадежны, как те, которые основаны на поддельных данных.

Проблема в том, что используемые Энгристом и Крюгером в некоторых регрессиях инструменты действительно являются очень слабыми. В некоторых спецификациях построенная для первого этапа  $F$ -статистика меньше 2, что Энгрист гораздо меньше, чем эмпирический «уровень отсечки», равный 10. В других спецификациях Баунд и Крюгер имеют большие построенные на первом этапе  $F$ -статистики, и в этих случаях статистические выводы, полученные с помощью 2МНК, не являются следствием слабых инструментов. Стоит отметить, что в этих спецификациях оценки эффекта отдачи от образования составляют около 8%, что несколько больше, чем было получено с помощью МНК<sup>1</sup>.

<sup>1</sup> Исходные версии IV-регрессий представлены в работе Энгриста и Крюгера [Angrist, Krueger (1991)], а анализ с использованием заведомо неверных инструментов представлен в работе Баунда, Джегера и Бейкера [Bound, Jaeger, Baker (1995)].



**Тест на сверхидентифицирующие ограничения.** Предположим, что имеется один эндогенный регрессор и два инструмента. Тогда можно было бы вычислить две разные 2МНК-оценки: для вычисления одной используется первый инструмент, а для вычисления другой – второй инструмент. Эти две оценки не будут одинаковыми из-за изменчивости выборки, но если оба инструмента являются экзогенными, то оценки, как правило, будут достаточно близки друг к другу. Но что если эти два инструмента будут давать очень разные оценки? Можно было бы предположить, что что-то не в порядке с одним из инструментов или обоими одновременно. То есть было бы разумно заключить, что один из инструментов (или оба одновременно) не является экзогенным.

*Тест на сверхидентифицирующие ограничения* неявно проводит это сравнение. Неявно, поскольку проверка проводится без фактического расчета всех возможных IV-оценок. Его основная идея заключается в том, что экзогенность инструментов означает, что они не коррелируют с ошибками регрессии  $u_i$ . Из этого следует, что инструменты должны почти не коррелировать с остатками  $\hat{u}_i^{TSLS}$ , где  $\hat{u}_i^{TSLS} = Y_i - (\hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X_{1i} + \dots + \hat{\beta}_{k+r}^{TSLS} W_{ri})$  – оценка ошибки регрессии с помощью 2МНК-регрессии, оцененной с использованием всех инструментов. (Заметим, что эти остатки строятся с использованием истинных значений  $X$ -ов, а не их оценок, полученных на первом шаге). Соответственно, если инструменты действительно являются экзогенными, то все коэффициенты при инструментах в регрессии  $\hat{u}_i^{TSLS}$  на инструменты и используемые экзогенные переменные должны быть равны нулю, и эта гипотеза может быть протестирована.

Метод тестирования сверхидентифицирующих ограничений сформулирован во вставке «Основные понятия 12.6». Соответствующая статистика вычисляется с помощью построения устойчивой к наличию гетероскедастичности  $F$ -статистики и известна как  $J$ -статистика.

В больших выборках, если инструменты не являются слабыми, а ошибки являются гомоскедастичными, то в условиях нулевой гипотезы о том, что инструменты являются экзогенными,  $J$ -статистика имеет распределение хиквадрат с  $m-k$  степенями свободы. Важно помнить, что даже если количество тестируемых ограничений равно  $m$ , то количество степеней свободы в асимптотическом распределении  $J$ -статистики равно  $m-k$ . Модификация  $J$ -статистики на случай гетероскедастичных ошибок представлена в разделе 18.7.

Самым простым способом убедиться в том, что нельзя тестировать экзогенность регрессоров в случае, когда коэффициенты точно определены ( $m=k$ ), является изучение случая одного эндогенного регрессора. При наличии двух инструментов ( $k=1$ ) можно вычислить 2МНК оценки – по одной для каждого инструмента, а затем сравнить их на предмет близости. Но если имеется только один инструмент, то можно вычислить только одну 2МНК-оценку, которую не с чем будет сравнивать. Действительно, если коэффициенты точно определены ( $m=k$ ), то  $J$ -статистика для тестирования на сверхидентифицирующие ограничения в точности равна нулю.

**Тест на сверхидентифицирующие ограничения ( $J$ -статистика)**

Пусть  $\hat{u}_i^{TSLS}$  – остатки, полученные при помощи 2МНК-оценок регрессии (12.12). Используйте МНК для оценки коэффициентов регрессии:

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \dots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \dots + \delta_{m+r} W_{ri} + e_i, \quad (12.17)$$

где  $e_i$  – остаточный член. Пусть  $F$  обозначает расчитанную для случая гомоскедастичности ошибок  $F$ -статистику, построенную для тестирования ограничения  $\delta_1 = \dots = \delta_m = 0$ . Тогда статистика для тестирования на сверхидентифицирующие ограничения равна  $J = mF$ . В условиях нулевой гипотезы о том, что все инструменты являются экзогенными, и если  $e_i$  являются гомоскедастичными, то в больших выборках  $J$  имеет распределение  $\chi^2_{m-k}$ , где  $m - k$  представляет собой «степень сверхидентификации», то есть число инструментов за вычетом числа эндогенных регрессоров.

**ОСНОВНЫЕ ПОНЯТИЯ****12.6**

## 12.4. Приложение к изучению спроса на сигареты<sup>1</sup>

Предшествующие попытки оценить эластичность спроса на сигареты остановились на 2МНК-оценках, приведенных в уравнении (12.16), в котором уровень дохода использовался в качестве экзогенной переменной. Кроме того, в нем использовались две инструментальные переменные – общий налог с продаж и специфический налог на сигареты. Теперь можно провести более тщательную оценку этих инструментов.

Как и в разделе 12.1, логично предположить, что оба инструмента являются релевантными, поскольку налоги составляют значительную часть итоговой цены сигарет (после налогообложения) – это будет показано эмпирически. Сначала, однако, необходимо ответить на достаточно трудный вопрос о том, действительно ли эти две переменные являются экзогенными.

Первым шагом в проверке экзогенности инструмента является изучение аргументов о том, почему он может или не может быть экзогенным. Для этого необходимо понять, какие факторы входят в ошибку регрессии в уравнении спроса на сигареты, и определить, связаны ли эти факторы с инструментами.

Почему в некоторых штатах потребление сигарет на душу населения выше, чем в других? Одной из причин этого может быть различие в уровне доходов в разных штатах, но переменная, отражающая уровень доходов, добавлена в уравнение (12.16), так что не является частью остаточного члена. Другая причина может заключаться в наличии некоторых исторических факторов, влияющих на спрос. Например, в штатах, в которых исторически выращивался табак, имеет место более высокий уровень потребления табачных изделий, чем в большинстве других штатов. Может ли это быть связано с налогами? Вполне возможно –

<sup>1</sup> В данном разделе предполагается знание материала, представленного в разделах 10.1 и 10.2 – панельные данные с  $T = 2$  периодами.

если производство табака и сигарет является важной отраслью в каком-либо штате, то она может предпринимать меры по удержанию специфических налогов на сигареты на низком уровне. Это означает, что пропущенный фактор в уравнении спроса на сигареты в зависимости от того, выращивается ли в штате табак и производятся ли сигареты, может быть связан (может коррелировать) с уровнем специфических налогов на сигареты.

Одним из способов решения проблемы с возможной корреляцией между остаточным членом и инструментом может являться включение в регрессию переменной, отражающей информацию о размерах табачной и сигаретной промышленности в том или ином штате. Этот подход использовался при включении в качестве регрессора в уравнение спроса переменной, отражающей уровень дохода. Но поскольку в наличии имеются панельные данные, описывающие уровень потребления сигарет, возможен другой подход, в рамках которого не требуется использование этой информации. Как обсуждалось в главе 10, панельные данные позволяют исключить влияние переменных, которые различаются между объектами (в данном случае – между штатами), но не изменяются с течением времени – например, такие, как климат и исторические обстоятельства, которые привели к развитию табачной и сигаретной промышленности в штате. В главе 10 были приведены два метода: оценка регрессии в *разностях* (между двумя периодами) и оценка регрессии с фиксированными эффектами. Применить такой анализ в данном случае достаточно просто – первый подход будет использоваться при помощи оценки регрессий, описанных в разделе 10.2, основанных на построении разностей переменных между двумя временными периодами.

Временной промежуток между двумя различными годами влияет на интерпретацию оценки эластичности. Поскольку потребление сигарет вызывает привыкание, изменение уровня цен на них повлияет на потребление с некоторым лагом. Сначала увеличение цен на сигареты будет оказывать незначительное влияние на спрос. Однако со временем рост цен может способствовать возникновению у некоторых курильщиков желания избавиться от этой вредной привычки, а также, что не менее важно, может помочь в предотвращении пристрастия к курению у некурящих. Таким образом, реакция спроса на повышение цены может быть незначительной в краткосрочной перспективе, но значительной в долгосрочной перспективе. Иначе говоря, для продуктов, вызывающих привыкание – таких как сигареты, – спрос может быть неэластичным в краткосрочной перспективе, то есть он может иметь краткосрочную эластичность, близкую к нулю, но в то же время может быть более эластичным в долгосрочной перспективе.

В таком исследовании основной акцент делается на оценке долгосрочной ценовой эластичности на основе данных по изменениям уровня цен и спроса на сигареты в течение десятилетнего периода. В частности, в рамках рассмотренных ранее регрессий оценивается регрессия изменения логарифма числа потребляемых сигарет за десятилетний период  $\ln(Q_{i,1995}^{\text{cigarettes}}) - \ln(Q_{i,1985}^{\text{cigarettes}})$  от изменения логарифма цены за 10 лет  $\ln(P_{i,1995}^{\text{cigarettes}}) - \ln(P_{i,1985}^{\text{cigarettes}})$  и изменения логарифма уровня дохода за 10 лет  $\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$ . Используются две инструментальные переменные – изменение уровня налогов с продаж за 10 лет

$(SalesTax_{i,1995} - SalesTax_{i,1985})$  и изменение специфического налога на сигареты за 10 лет  $(CigTax_{i,1995} - CigTax_{i,1985})$ .

Результаты представлены в таблице 12.1. В столбцах указаны результаты оценок различных регрессий. Во всех регрессиях используются одинаковые регрессоры, коэффициенты оцениваются с помощью 2МНК. Единственная разница между представленными регрессиями заключается в различных наборах используемых инструментов. Результаты из столбца (1) соответствуют включению одного инструмента – налога с продаж, из столбца (2) – использованию одного инструмента – специфического налога на сигареты, в столбце (3) – оба вида налогов используются в качестве инструментальных переменных.

В IV-регрессиях надежность получаемых оценок коэффициентов тесно связана с тем, являются ли инструменты допустимыми, поэтому в первую очередь в таблице 12.1 необходимо рассмотреть диагностические статистики, отражающие допустимость инструментов.

Во-первых, являются ли инструменты релевантными? Необходимо рассмотреть построенную на первом шаге 2МНК F-статистику. Регрессия, оцениваемая на первом шаге 2МНК для столбца (1), выглядит следующим образом:

$$\begin{aligned} \ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes}) = & 0,53 - 0,22 \left[ \ln(Inc_{i,1995}) - \right. \\ & \left. - \ln(Inc_{i,1985}) \right] + 0,0255 (SalesTax_{i,1995} - \\ & - SalesTax_{i,1985}). \end{aligned} \quad (12.18)$$

Таблица 12.1

## 2МНК-оценки спроса на сигареты на основе панельных данных по 48 штатам США

Зависимая переменная	$\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$		
Регрессор	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	-0,94** (0,21)	-1,34** (0,23)	-1,20** (0,20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0,53 (0,34)	0,43 (0,30)	0,46 (0,31)
Свободный член	-0,12 (0,07)	-0,02 (0,07)	-0,05 (0,06)
Инструментальная переменная	Налог с продаж	Специфический налог на сигареты	Налог с продаж и специфический налог на сигареты
Построенная на первом этапе F-статистика (First-stage F-statistic)	33,70	107,20	88,60
J-статистика и p-значение для теста на сверхидентифицирующие ограничения	-	-	4,93 (0,026)

Примечание. Представленные результаты получены с помощью оценки регрессий на основе данных для 48 штатов США (48 наблюдений для 10-летних разностей). Данные описаны в приложении 12.1. J-тест на сверхидентифицирующие ограничения описан во вставке «Основные понятия 12.6» (его p-значение приведено в скобках), а построенная на первом этапе F-статистика описана во вставке «Основные понятия 12.5». Отдельные коэффициенты являются статистически значимыми на уровне значимости \*5 % и \*\*1 %.

Поскольку в этой регрессии используется лишь один инструмент, то построенная на первом шаге  $F$ -статистика равна квадрату  $t$ -статистики для тестирования гипотезы о том, что коэффициент при инструментальной переменной  $SalesTax_{i,1995} - SalesTax_{i,1985}$  равен нулю. То есть  $F = t^2 = (0,0255 / 0,0044)^2 = 33,7$ . Для регрессий, результаты которых представлены в столбцах (2) и (3), построенные для первой стадии  $F$ -статистики равны 107,2 и 88,6, то есть во всех трех случаях значение  $F$ -статистики превышает 10. Можно заключить, что инструменты не являются слабыми, то есть можно использовать стандартные статистические методы (тестирование гипотез и доверительные интервалы) с использованием 2МНК-оценок коэффициентов и стандартных ошибок.

Являются ли инструменты экзогенными? Поскольку в регрессиях, результаты которых представлены в столбцах (1) и (2), используется один инструмент и один эндогенный регрессор, то коэффициенты в этих регрессиях точно идентифицированы. Таким образом, нельзя провести  $J$ -тест в любой из этих регрессий. Коэффициенты регрессии в столбце (3) являются переопределеными, поскольку используются два инструмента и один эндогенный регрессор, поэтому есть одно ( $m - k = 2 - 1 = 1$ ) сверхидентифицирующее ограничение.  $J$ -статистика равна 4,93, имеет распределение  $\chi^2_1$ , так что 5 %-е критическое значение равно 3,84 (см. таблицу 3 приложения). Нулевая гипотеза о том, что оба инструмента являются экзогенными, отвергается на 5 %-м уровне значимости (этот вывод также может быть получен непосредственно из  $p$ -значения, равного 0,026 и представленного в таблице).

Причина, по которой нулевая гипотеза о том, что оба инструмента являются экзогенными, отвергается при использовании  $J$ -статистики, заключается в том, что эти два инструмента дают различные оценки коэффициентов. Когда единственным инструментом является налог с продаж [столбец (1)], оценка эластичности по цене равна  $-0,94$ , но когда в качестве инструмента используется только специфический налог на сигареты, то оценка эластичности равна  $-1,34$ . Стоит напомнить, что основная идея  $J$ -статистики заключается в следующем: если оба инструмента являются экзогенными, то обе 2МНК-оценки, полученные с помощью двух разных инструментов, являются состоятельными и отличаются друг от друга только из-за изменчивости случайной выборки. Если, однако, один из инструментов является экзогенным, а второй — нет, то оценка на основе эндогенного инструмента не является состоятельной, что определяется с помощью  $J$ -статистики. В данном случае различие между двумя полученными оценками эластичности достаточно велико, что вряд ли может быть результатом изменчивости выборки. Таким образом,  $J$ -статистика отвергает нулевую гипотезу о том, что оба инструмента являются экзогенными.

То, что  $J$ -статистика отвергает гипотезу, означает, что регрессия, результаты которой представлены в столбце (3), основана на недопустимых инструментах (нарушается условие экзогенности инструмента). Что это означает для оценок из столбцов (1) и (2)?  $J$ -статистика говорит о том, что по крайней мере один из инструментов является эндогенным, то есть существует три логических возможности. Во-первых, налог с продаж является экзогенной переменной, а специфический налог на сигареты не является, в этом случае результаты регрессии, представленные в столбце (1), являются надежными. Во-вторых, специфический налог на сига-

реты является экзогенной переменной, а налог с продаж – нет, поэтому результаты регрессии, представленные в столбце (2), являются надежными. В-третьих, ни один из видов налогов не является экзогенной переменной, то есть ни один из представленных результатов регрессий не является надежным. Статистические данные не могут предоставить информацию о том, какой из этих вариантов является верным, и, таким образом, необходимо использовать собственные экспертные оценки.

По-видимому, случай экзогенности общего налога с продаж «сильнее», чем случай экзогенности специфического налога на сигареты, потому что изменения в уровнях специфического налога на сигареты могут быть связаны с изменениями на рынке сигарет и табачных изделий или с изменениями законодательства в области употребления табачных изделий посредством некоего политического процесса. Например, если в некотором штате уровень потребления табака снижается, поскольку курение выходит из моды, то это будет приводить к снижению числа курильщиков, ослабляя лобби против специфических налогов на сигареты, что в свою очередь может привести к увеличению специфических налогов на сигареты. Таким образом, изменения во вкусах или предпочтениях (которые являются частью  $\mu$ ) могут быть связаны с изменениями в специфических налогах на сигареты (инструмент). Это говорит в пользу применения IV-оценок, которые получены с помощью переменной общего налога с продаж как инструмента, где оценка эластичности равна –0,94.

Оценка, равная –0,94, показывает, что потребление сигарет является достаточно эластичным: увеличение цены на 1 % приводит к снижению потребления на 0,94 %. Это может выглядеть несколько странным для такого вызывающего привыкание (или даже зависимость) продукта, как сигареты. Однако стоит помнить, что полученная оценка эластичности вычислена с помощью изменений показателей за 10-летний период, то есть это долгосрочная эластичность. Она показывает, что увеличение налогов может привести к существенному снижению потребления сигарет, по крайней мере в долгосрочном периоде.

При проведении оценок на основе пятилетних изменений показателей в период с 1985 по 1990 год (с использованием общего налога с продаж в качестве инструмента), а не на основе 10-летних разностей, как представлено в таблице 12.1, оценка эластичности равна –0,79, а для изменений за период с 1990 по 1995 год оценка эластичности равна –0,68. Эти оценки показывают, что спрос на сигареты является менее эластичным для временного промежутка в пять лет, чем для временного промежутка в 10 лет. Полученные оценки больших эластичностей для более длинных временных периодов в полной мере согласуются со значительным количеством работ, в которых анализировалась эластичность спроса на сигареты по цене. Полученные авторами работ оценки эластичности зачастую лежат в пределах от –0,3 до –0,5, но являются в большинстве своем краткосрочными эластичностями. В некоторых работах показано, что долгосрочные эластичности могут быть в 2 раза больше краткосрочных<sup>1</sup>.

---

<sup>1</sup> В работе Адда и Корнали [Adda, Cornaglia, 2006] предполагается, что курильщики компенсируют для себя более высокий уровень налогов посредством более интенсивного курения, таким образом, получая большее никотина из одной сигареты. Для получения дополнительной информации об экономике курения будут полезны следующие работы: Chaloupka, Warner, 2000; Gruber, 2001; Carpenter, Cook, 2008.



### **Экстерналии курения**

Курение приводит к затратам для окружающих, которые не полностью покрываются курильщиком, то есть оно создает экстерналии (внешнее влияние). Таким образом, одним из экономических обоснований налогообложения табачной продукции и сигарет является «интернализация» этого внешнего влияния (экстерналий). В теории налог с одной пачки сигарет должен быть равен размеру экстерналий в стоимостном выражении, которые возникают при выкушивании этой пачки. Но что именно является экстерналиями от курения?

В нескольких исследованиях для оценки экстерналий курения использовались эконометрические методы. К отрицательным экстерналиям можно отнести дополнительные издержки на медицинское обслуживание, которые вынуждены нести некурящие члены общества, затраты правительства на медицинское лечение больных курильщиков, а также ущерб от пожаров, вызванных сигаретами.

Но с чисто экономической точки зрения курение также создает и *положительные* экстерналии. Самая большая положительная экономическая экстерналия от курения заключается в том, что курильщики, как правило, делают намного большие отчисления в фонды социального страхования, чем они в итоге получают из этих же фондов. Имеет место также большая экономия в расходах на дома престарелых – продолжительность жизни курильщиков, как правило, ниже. Поскольку отрицательные внешние эффекты от курения возникают в то время, пока курильщик жив, а положительные – имеют место после его смерти, чистая приведенная стоимость экстерналий в расчете на одну пачку сигарет (стоимость чистых расходов за пачку, дисконтированная к текущему моменту) зависит от ставки дисконтирования.

Авторы различных исследований не сходятся во мнении относительно точной чистой долларовой стоимости экстерналий. Некоторые считают, что экстерналии при правильном выборе дисконтирования достаточно малы и по величине меньше, чем существующий уровень налогов. В самом деле, наиболее радикальные оценки показывают, что чистые экстерналии *положительны*, то есть курение должно субсидироваться! Другие исследования, в которых делались попытки учесть те расходы, которые, вероятно, важны, но которые трудно оценить количественно (например, уход за младенцами с плохим здоровьем из-за того, что их матери курят), позволяют предположить, что стоимостной эквивалент экстерналий может достигать 1 долл. за пачку (возможно, даже больше). Но все авторы исследований сходятся во мнении о том, что курильщики, для которых наибольший уровень смертности приходится на конец среднего возраста, платят гораздо больше налогов, чем они когда-либо получат обратно в течение своей довольно короткой пенсии<sup>1</sup>.

---

<sup>1</sup> Одно из первых исследований экстерналий курения представлено в работе Уилларда Г. Мэннинга и др. (Willard G. Manning et al., 1989). Расчеты, показывающие, что издержки на медицинское обслуживание вырастут, если все бросят курить, представлены в работе Барендрегта и др. (Barendregt et al., 1997). Обзор работ, посвященных изучению экстерналий курения, представлен в работе Чалупки и Уорнера (Chaloupka, Warner, 2000).



## 12.5. Где найти допустимые инструменты?

На практике наиболее сложным аспектом получения IV-оценок является поиск инструментов, которые одновременно являлись бы и релевантными, и экзогенными. Существуют два основных подхода, отражающих два различных направления в эконометрическом и статистическом моделировании.

Первый подход говорит о том, что для поиска инструментов необходимо использовать экономическую теорию. Например, понимание Филиппом Райтом экономики сельскохозяйственных рынков позволило ему найти инструменты, которые сдвигали бы кривую предложения, но не кривую спроса. Это, в свою очередь, привело его к рассмотрению погодных условий в сельскохозяйственных регионах. Одной из областей, в которой этот подход особенно продуктивен, является сфера финансовой экономики. Некоторые экономические модели поведения инвестора учитывают особенности того, как инвестор прогнозирует что-либо, что дает наборы переменных, которые не коррелируют с остаточным членом. Эти модели иногда являются нелинейными и по данным, и по параметрам, и в этом случае нельзя использовать IV-оценки, которые обсуждались в этой главе. Вместо этого в таких ситуациях обычно используется расширение IV-методов на случай нелинейных моделей, которое называется обобщенным методом моментов. Однако экономическая теория имеет некую степень абстракции и часто не учитывает отдельных нюансов и деталей, необходимых для анализа определенного набора данных. Поэтому данный подход не всегда работает.

Второй подход к построению инструментов заключается в поиске экзогенных источников изменения  $X$  в зависимости от того, что именно является случайным фактором, оказывающим влияние на эндогенный регрессор. Например, в гипотетическом примере из раздела 12.1 ущерб от землетрясений приводил к увеличению среднего размера классов в некоторых школьных округах, и это изменение в размерах классов не было связано с возможными пропущенными переменными, которые могли бы оказывать влияние на успеваемость учащихся. Данный подход обычно требует тщательного изучения рассматриваемой проблемы и пристального внимания к особенностям имеющихся данных, что лучше всего объяснить на примерах.

### *Три примера*

Далее будут рассмотрены три эмпирических приложения модели регрессии с инструментальными переменными, которые дают наглядный пример того, как разные исследователи используют свои экспертные знания для поиска инструментальных переменных в рамках своих эмпирических проблем.

#### *Снижают ли тюремные заключения преступников уровень преступности?*

Подобный вопрос может быть задан только экономистом. В конце концов преступник не может совершить преступление за пределами тюрьмы, находясь в тюрьме, кроме того, пойманные и помещенные под стражу преступники представляют собой пример, который служит для сдерживания других. Но оценка

величины совокупного эффекта, то есть изменение уровня преступности, вызванное увеличением количества заключенных на 1 %, является эмпириическим вопросом.

Одной из стратегий для оценки указанного выше эффекта является построение регрессии уровня преступности (количество преступлений на 100 тыс. человек населения) на количество заключенных на 100 тыс. человек населения на основе годовых данных на соответствующем уровне юрисдикции (например штаты США). Эта регрессия может включать некоторые контрольные переменные, отражающие экономические условия (преступность возрастает, если общие экономические условия ухудшаются), демографию (молодые люди совершают больше преступлений, чем пожилые) и так далее. Существует, однако, вероятность наличия смещений в оценках, вызванных проблемой взаимной причинности, что делает такой анализ некорректным. Если уровень преступности повышается и полиция делает свою работу хорошо, то заключенных станет больше. С одной стороны, увеличение количества заключенных снижает уровень преступности. С другой стороны, увеличение уровня преступности также приводит к росту количества заключенных в местах лишения свободы. Как и в приведенном ранее примере с маслом на рисунке 12.1, из-за проблемы взаимной причинности оценка регрессии уровня преступности от количества заключенных на 100 тыс. человек населения с помощью МНК будет давать некую оценку сложной комбинации этих двух эффектов. Эта проблема не может быть решена путем поиска более подходящих контрольных переменных.

Смещения, вызванные взаимной причинностью, тем не менее, могут быть устранены посредством поиска подходящих инструментальных переменных и применения 2МНК. Инструменты должны быть коррелированы с показателем количества заключенных (должны быть релевантными), но не должны коррелировать с остаточным членом в исследуемой зависимости (должны быть экзогенными). То есть они должны оказывать влияние на переменную, характеризующую количество заключенных, но не быть связанными с какими-либо ненаблюдаемыми факторами, которые определяют уровень преступности.

Как же найти подобные переменные, которые влияли бы на количество заключенных, но не имели прямого влияния на уровень преступности? Одной из таких переменных могло бы быть экзогенное изменение «емкости» существующих тюрем. Поскольку для увеличения емкости тюрьмы (максимально допустимого количества заключенных, которые могут в ней содержаться) требуется некоторое время, необходимое на непосредственное строительство, краткосрочные ограничения емкости могут приводить к увеличению количества досрочных освобождений или, что эквивалентно, снижению количества заключенных на 100 тыс. человек населения. На основе данной аргументации в работе (Levitt, 1996) было сделано предположение о том, что судебные иски, направленные против переполненности тюрем, могут служить инструментальной переменной. Проверка этой идеи была реализована на основе панельных данных для штатов США с 1972 по 1993 год.

Какие переменные, отражающие факт наличия судебных тяжб, касающихся борьбы против переполненности тюрем, можно было бы использовать в каче-

стве релевантных инструментов? Несмотря на то что в указанной выше работе не были представлены данные по вычисленной на первом шаге 2МНК F-статистике, ее результаты говорят о том, что борьба с переполненностью тюрем приводит к замедлению роста количества заключенных в используемых данных, показывая, что данный инструмент является релевантным. Кроме того, тяжбы, связанные с борьбой с переполненностью тюрем, обусловлены спецификой тюремных условий, но не уровнем преступности или его детерминантами, поэтому могут рассматриваться как экзогенный инструмент. Поскольку в работе (Levitt, 1996) законодательство, ограничивающее переполнение тюрем, было разбито на несколько типов и, следовательно, использовалось несколько инструментов, то это позволило провести тест на сверхидентифицирующие ограничения, который не отверг гипотезу об их наличии на основе полученной J-статистики.

Используя выбранные инструменты и 2МНК, Левитт показал, что влияние количества заключенных на уровень преступности является статистически значимым. Согласно полученным оценкам, величина данного эффекта в 3 раза больше, чем по оценкам, полученным на основе МНК, что свидетельствует о значительных смещениях в МНК-оценках, вызванных проблемой взаимной причинности.

**Повышает ли уменьшение размера класса успеваемость?** Как было показано в эмпирическом анализе в части II, школы с небольшими по количеству учеников учебными классами, как правило, богаче, а их ученики имеют доступ к большим возможностям для обучения, как в школе, так и вне ее. В части II для решения проблемы смещений, вызванных наличием пропущенных переменных, посредством учета различных характеристик студентов (таких как благосостояние, умение говорить на английском и других) была использована модель множественной регрессии. Тем не менее скептики могут усомниться в том, что этого было достаточно. Если было пропущено что-то важное, то полученные оценки влияния размера класса на качество образования будут по-прежнему смещены.

Эта потенциальная возможность наличия смещений, вызванных возможными пропущенными переменными, может быть устранена с помощью использования корректных контрольных переменных, но если данные по этим переменным отсутствуют (некоторые показатели, например, альтернативные возможности обучения, сложно измерить), то альтернативный подход заключается в использовании IV-регрессии. Эта модель требует использовать инструментальные переменные, которые коррелируют с размером учебного класса (релевантность инструментов), но не коррелируют с пропущенными детерминантами уровня успеваемости (оценки по тестам), которые входят в остаточный член. Например, к числу таких переменных можно отнести степень участия родителей в обучении своих детей, альтернативные возможности обучения за пределами учебного класса, качество преподавания и так далее (экзогенность инструментов).

Где же можно найти инструмент, который вызывает случайные, экзогенные изменения размера учебного класса, но не связан с другими детерминантами успеваемости в тестах? В работе Хоксби (Hoxby, 2000) было предложено

рассматривать сферу биологии. Из-за случайных флюктуаций в датах рождений размер групп в детских садах варьируется от года к году. Несмотря на то что фактическое число детей, идущих в детский сад, может быть эндогенным (последние новости о сфере школьного образования могут влиять на решение родителей об отправке детей в частные школы и детские сады), автор работы утверждает, что потенциальное количество детей, поступающих в детский сад (количество детей в возрасте четырех лет в данном районе), в значительной степени определяется случайными колебаниями дат рождения детей.

Является ли потенциальное количество школьников допустимым инструментом? Экзогенность данной переменной зависит от того, коррелирует ли она с ненаблюдаемыми детерминантами успеваемости в тестах (уровня успеваемости). Конечно, биологические колебания потенциального числа учащихся являются экзогенными, но эта переменная может колебаться из-за того, что родители с маленькими детьми могут решить переехать в район с «хорошей» школой или детским садом из района, где детские сады или школы «плохие». Если это так, то потенциальное количество учащихся может коррелировать с такими ненаблюдаемыми факторами, как уровень школьного менеджмента (качество управления школой), что делает этот инструмент недопустимым. Автором вышеуказанной работы было замечено, что рост или снижение количества школьников по этой причине будет происходить плавно в течение нескольких лет, в то время как случайные флюктуации периодов рождения будут производить к краткосрочным «пикам» числа потенциальных учащихся. Таким образом, в качестве инструмента было использовано не потенциальное количество учащихся, а его отклонение от долгосрочного тренда. Это отклонение удовлетворяет критерию релевантности инструмента (построенная на первом этапе F-статистика превышает 100). Вероятность того что этот инструмент является экзогенным, достаточно велика, но в рамках любого анализа с помощью модели IV-регрессии достоверность этого предположения достаточно условна.

Автором рассматриваемой работы были проведены оценки на основании панельных данных по ряду начальных школ в Коннектикуте в 1980 и 1990-х годах. Панельные данные позволили рассматривать фиксированные эффекты, которые отражают особенности той или иной школы, а также в дополнение к инструментальным переменным позволяют бороться с проблемой смещений в оценках, вызванной наличием пропущенных переменных на уровне отдельных школ. Полученные ею 2МНК-оценки показали, что влияние размера учебного класса на результаты тестов довольно мало. Большая часть оценок статистически значимо не отличалась от нуля.

**Продлевает ли жизнь агрессивное лечение сердечных приступов?** Агрессивное лечение при сердечных приступах (технически – острый инфаркт миокарда, AMI) обладает большим потенциалом для спасения жизней. Перед тем как новая медицинская процедура (в данном примере – катетеризация сердца<sup>1</sup>) получает одобрение для применения, она проходит продолжительные клини-

---

<sup>1</sup> Катетеризация сердца представляет собой процедуру, в рамках которой вводимый в кровеносный сосуд катетер (трубка) подводится к сердцу для получения информации о состоянии сердца и коронарных артерий. Может проводиться как в диагностических, так и в лечебных целях.

ческие испытания в серии случайных контролируемых экспериментов, предназначенных для измерения ее последствий и побочных эффектов. Но хорошие результаты в клинических испытаниях – это одно, а результаты, получаемые в реальной жизни, – это совсем другое.

Естественной отправной точкой для оценки реального влияния катетеризации сердца является сравнение пациентов, которые получали подобное лечение, и тех, которые не получали. Чтобы сделать это, нужно оценить регрессию продолжительности выживания пациента в зависимости от бинарной переменной, отражающей факт наличия подобного лечения (применялось ли к пациенту катетеризация сердца), и других контрольных переменных, которые влияют на смертность (возраст, вес, прочие характеристики состояния здоровья и т.д.). Теоретический коэффициент при бинарной переменной отражает увеличение продолжительности жизни пациента, как следствие вышеописанного лечения. К сожалению, МНК-оценки подвержены смещению: катетеризация сердца «просто так» не назначается пациентам. Данная процедура проводится, если врач и пациент совместно решают, что это будет эффективным. Если их решение основано частично на ненаблюдаемых факторах, связанных с характеристиками здоровья, не присутствующими в используемой выборке, то решение о назначении подобного лечения будет коррелировать с остаточным членом регрессии. Если наиболее здоровыми являются те пациенты, которым было назначено подобное лечение, то МНК-оценки будут смещены (решение о лечении коррелирует с пропущенной переменной), а лечение будет казаться более эффективным, чем есть на самом деле.

Эта потенциальная возможность смещений в оценках может быть устранена с помощью модели IV-регрессии с использованием допустимых инструментов. Инструмент должен быть коррелирован с переменной, отражающей назначение лечения (должен быть релевантным), но не должен коррелировать с пропущенными переменными, которые могут оказывать влияние на выживаемость пациентов после лечения (должен быть экзогенным).

Какие же факторы, отличные от состояния здоровья, могут оказывать влияние на лечение? В работе Макклеллана, Макнейла и Ньюхауса (McClellan, McNeil, and Newhouse, 1994) было предложено рассматривать географические факторы. Большинство лечебных учреждений в используемой выборке не специализируются на катетеризации сердца, поэтому многие пациенты территориально находятся ближе к «обычным» больницам, которые не предлагают подобный вид лечения, в отличие от лечебных учреждений, которые специализируются на катетеризации сердца. Поэтому Макклеллан, Макнейл и Ньюхаус использовали в качестве инструментальной переменной разность между расстоянием от дома пациента с сердечным приступом до ближайшей больницы, в которой осуществляется катетеризация сердца, и расстоянием до ближайшей больницы иного типа. Это расстояние равно нулю, если в ближайшей больнице осуществляется процедура катетеризации сердца, в противном случае оно является положительным. Если построенное подобным образом относительное расстояние влияет на вероятность получения лечения, то оно является релевантным

инструментом. Если оно имеет случайное распределение среди пациентов с сердечными приступами, то его можно считать экзогенным инструментом.

Является ли относительное расстояние до ближайшей больницы, где может быть проведена процедура катетеризации сердца, допустимым инструментом? Макклеллан, Макнейл и Ньюхаус не приводят построенную на первом шаге 2МНК *F*-статистику, однако представляют другие эмпирические доказательства того, что этот инструмент не является слабым. Можно ли считать эту переменную экзогенным инструментом? Авторы приводят два аргумента. Во-первых, опираясь на свой медицинский опыт и знания о системе здравоохранения, они утверждают, что расстояние до больницы действительно не коррелирует с любой из ненаблюдаемых переменных, определяющих случаи сердечных приступов. Во-вторых, у них имеются данные по некоторым дополнительным переменным, которые влияют на частоту сердечных приступов, например, таких, как вес пациента, и в используемой выборке построенная переменная расстояния не коррелирует с этими *наблюдаемыми* детерминантами выживаемости пациентов. Это, как считают авторы, повышает вероятность того, что данная переменная относительного расстояния не коррелирована с *ненаблюдаемыми* детерминантами в ошибке.

Используя выборку, состоящую из 205 021 наблюдений для американцев в возрасте старше 64 лет, у которых наблюдалась случаи сердечных приступов в 1987 году, Макклеллан, Макнейл и Ньюхаус получили удивительные результаты: полученные 2МНК-оценки показывают, что процедура катетеризации сердца имеет небольшое, возможно, нулевое влияние на состояние здоровья. То есть процедура катетеризации сердца значимо не увеличивает продолжительность жизни. С другой стороны, МНК-оценки свидетельствуют о наличии значительного положительного эффекта. Авторы интерпретируют данные различия как доказательство наличия смещений в МНК-оценках.

Метод инструментальных переменных, использованный Макклелланом, Макнейлом и Ньюхаусом, имеет интересную интерпретацию. При анализе с помощью МНК была использована переменная, отражающая факт проведения лечебной процедуры, а поскольку проведение процедуры само по себе является результатом совместного решения пациента и врача, то авторы утверждают, что данная переменная коррелирует с остаточным членом. С другой стороны, в рамках 2МНК использовалась переменная, отражающая предсказанную процедуру лечения, изменение которой возникает из-за изменений в используемой инструментальной переменной: пациенты, проживающие ближе к лечебным учреждениям, в которых проводится процедура катетеризации сердца, с большей вероятностью получат данный вид лечения.

Приведенная интерпретация позволяет сделать два вывода. Во-первых, в рамках регрессии с инструментальными переменными в действительности оценивается эффект проведения лечения не для «типовых» случайно выбранных пациентов, а для пациентов, для которых расстояние до больницы является важным фактором, влияющим на принятие решения о проведении лечения. Эффекты для подобных пациентов могут отличаться от эффектов для «типовых» пациентов, что дает одно из объяснений больших по величине оценок эффек-

тивности лечения в клинических исследованиях, чем в работе Макклеллана, Макнейла и Ньюхауса. Во-вторых, можно выделить общую стратегию для поиска инструментов в подобного рода исследованиях: найти инструмент, который влияет на вероятность назначения определенного вида лечения, но оказывает подобное влияние по причинам, которые не связаны с результатом лечения, кроме как через их влияние на вероятность лечения. Оба представленных вывода имеют непосредственное применение в экспериментальных и «квази-экспериментальных» исследованиях, рассмотренных в главе 13.

## 12.6. Заключение

Из малоизвестного способа оценки влияния роста цены масла на его потребление метод инструментальных переменных постепенно превратился в общий подход к оценке регрессий, который применяется в тех случаях, когда одна или несколько переменных коррелируют с ошибкой регрессии. В модели регрессии с инструментальными переменными инструменты используются для того, чтобы изолировать изменения эндогенных регрессоров, которые коррелируют с ошибкой оцениваемой регрессии — это первый шаг двухшагового метода наименьших квадратов. Все это, в свою очередь, позволяет проводить оценки исследуемых эффектов на втором этапе 2МНК.

Для получения корректных оценок с помощью IV-регрессии необходимо иметь допустимые инструменты, то есть инструменты, которые одновременно являются релевантными (не являются слабыми) и экзогенными. Если инструменты являются слабыми, то 2МНК-оценки могут быть смещены даже в больших выборках, а статистические выводы, основанные на полученных с помощью 2МНК *t*-статистиках и доверительных интервалах, могут быть неточны. К счастью, в случае одного эндогенного регрессора для того, чтобы проверить, является ли инструмент слабым, достаточно лишь посмотреть на построенную на первом этапе 2МНК *F*-статистику.

Если инструменты не являются экзогенными, то есть если один или несколько инструментов коррелированы с ошибкой, то 2МНК-оценки не являются со-стоятельными. Если число инструментов превышает количество эндогенных регрессоров, то их (инструментов) экзогенность может быть проверена с помощью *J*-статистики теста на сверхидентифицирующие ограничения. Тем не менее основное предположение о том, что экзогенных инструментов столько же, сколько и эндогенных регрессоров, не может быть проверено. Поэтому важно, чтобы исследователи и читатели использовали свои знаниями в рассматриваемой области для оценки разумности данного предположения.

Интерпретация IV-регрессии как способа использовать известные экзогенные изменения в эндогенных регрессорах может быть использована для определения конкретного направления поиска потенциальных инструментальных переменных в том или ином конкретном эмпирическом приложении. Подобная интерпретация лежит в основе эмпирического анализа в области оценок последствий принятия различных программ и программных документов, где для оценки последствий программ, проведения той или иной политики или других

мероприятий используются эксперименты или квазиэксперименты. В подобных эмпирических приложениях возникает множество дополнительных вопросов к интерпретации результатов построенных IV-оценок. Например, в рассматриваемой выше интерпретации результатов IV-регрессии при изучении последствий катетеризации сердца проведение одной и той же «процедуры» может иметь различные последствия для разных «пациентов». Эти и другие аспекты эмпирического анализа последствий различных программ рассматриваются в главе 13.

## **Выводы**

1. Регрессия с инструментальными переменными является способом оценки регрессионных коэффициентов в тех случаях, когда один или более регрессоров коррелированы с ошибкой.
2. Эндогенные переменные коррелированы с ошибкой рассматриваемой регрессии, а экзогенные переменные не коррелированы.
3. Для допустимости инструмента необходимо, чтобы (1) инструмент был коррелирован с эндогенной переменной и (2) был экзогенным.
4. Для оценки модели с инструментальными переменными необходимо, чтобы число инструментов было не меньше числа эндогенных регрессоров.
5. Оценки двухшаговым методом наименьших квадратов получаются в два шага. На первом шаге оцениваются регрессии используемых эндогенных переменных на экзогенные переменные и инструменты. На втором этапе оценивается регрессия зависимой переменной на экзогенные переменные и предсказанные значения эндогенных переменных, полученные из регрессий, оцененных на первом шаге.
6. Использование слабых инструментов (инструменты, которые практически не коррелированы с используемыми эндогенными переменными) приводит к смещениям в 2МНК-оценках, неточностям в построенных доверительных интервалах и ненадежному тестированию гипотез.
7. Если инструмент не является экзогенным, то 2МНК-оценки не являются состоятельными.

## **Основные понятия**

Модель регрессии с инструментальными переменными (модель IV-регрессии) (с. 439).

Инструментальная переменная (инструмент) (с. 439).

Эндогенная переменная (с. 440).

Экзогенная переменная (с. 441).

Условие релевантности инструмента (с. 441).

Условие экзогенности инструмента (с. 441).

Двухшаговый метод наименьших квадратов (с. 441).

Включенные экзогенные переменные (с. 451).

- Точно определенные (с. 451).  
 Переопределенные (с. 451).  
 Недоопределенные (с. 451).  
 Приведенная форма (с. 453).  
 Оценка регрессии на первом шаге (с. 454).  
 Оценка регрессии на втором шаге (с. 454).  
 Слабые инструменты (с. 459).  
 $F$ -статистика, вычисленная на первом шаге (с. 460).  
 Тест на сверхидентифицирующие ограничения (с. 464).

### **Вопросы для повторения и закрепления основных понятий**

- 12.1. Каким образом (положительно или отрицательно)  $\ln(P_i^{butter})$  коррелирует с  $u_i$  в регрессионной модели спроса, представленной уравнением (12.3)? Если оценка проводится с помощью МНК, то можно ли ожидать, что оценка  $\beta_1$  будет больше или меньше своего истинного значения  $\beta_1$ . Поясните свой ответ.
- 12.2. Предположим, что при изучении спроса на сигареты, рассматриваемого в настоящей главе, в качестве инструмента использовалось бы количество деревьев на душу населения в определенном штате США. Является ли этот инструмент релевантным? Является ли он экзогенным? Является ли инструмент допустимым?
- 12.3. В работе Левитта (Levitt, 1996) изучалось влияние количества заключенных на 100 тыс. человек населения на уровень преступности. Автором в качестве инструментальной переменной рассматривалось количество адвокатов на душу населения. Является ли этот инструмент релевантным? Является ли он экзогенным? Является ли инструмент допустимым?
- 12.4. В работе Макклеллана, Макнейла и Ньюхауса (McClellan, McNeil, Newhouse, 1994) исследовалась эффективность проведения процедуры катетеризации сердца, и в качестве инструментальной переменной рассматривалась разность расстояний от дома пациента до лечебных учреждений, в которых проводится процедура катетеризации сердца, и до «обычных» лечебных учреждений, которые не специализируются на проведении подобной процедуры. Как можно определить, является ли этот инструмент релевантным? Как можно определить, является ли данный инструмент экзогенным?

### **Упражнения**

- 12.1. В данном упражнении рассматриваются результаты оценки регрессий для панельных данных, представленные в таблице 12.1.
- a) Предположим, что правительство рассматривает возможность введения нового налога на сигареты. Его внедрение, по оценкам, может привести к увеличению рыночной цены на 0,50 долл. за пачку сигарет. Если текущая рыночная цена пачки сигарет составляет 7,50 долл., то, используя

результаты оценки регрессии, представленные в столбце (1), оцените величину изменения спроса на сигареты. Постройте 95 %-й доверительный интервал для изменения спроса на сигареты.

- б) Предположим, что экономика США вступает в период рецессии и уровень дохода падает на 2 %. Используя результаты регрессии, представленные в столбце (1), оцените величину изменения спроса.
  - в) Предположим, что рецессия длится более одного года. Можно ли, используя результаты, представленные в столбце (1), получить надежный ответ на вопрос из пункта (б)? Поясните свой ответ.
  - г) Предположим, что  $F$ -статистика, представленная в столбце (1), равна 3,6 вместо 33,6. Можно ли с помощью результатов данной регрессии достоверно ответить на вопрос из пункта (а)? Поясните свой ответ.
- 12.2. Рассмотрим модель регрессии с одним регрессором:  $Y_i = \beta_0 + \beta_1 X_i + u_i$ . Предположим, что все предположения из вставки «Основные понятия 4.3» выполнены.
- а) Покажите, что  $X_i$  является допустимым инструментом. То есть покажите, что выполнено условие  $Z_i = X_i$  из вставки «Основные понятия 12.3».
  - б) Покажите, что при таком выборе  $Z_i$  предположения модели регрессии с инструментальными переменными из вставки «Основные понятия 12.3» выполняются.
  - в) Покажите, что IV-оценки, построенные с помощью данного инструмента, идентичны МНК-оценкам.
- 12.3. Ваш (-а) коллега хочет оценить дисперсию ошибки в уравнении (12.1):
- а) Предположим, что он (она) предполагает использовать оценки, полученные на втором шаге 2МНК:  $\hat{\sigma}_a^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS} \hat{X}_i)^2$ , где  $\hat{X}_i$  – оценка значений  $X_i$  построенная на первом шаге 2МНК. Является ли оценка дисперсии состоятельной? (Для ответа на этот вопрос можно считать, что выборка имеет очень большие размеры, а 2МНК-оценки практически совпадают с  $\beta_0$  и  $\beta_1$ .)
  - б) Является ли  $\hat{\sigma}_b^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS} X_i)^2$  состоятельной оценкой?
- 12.4. Рассмотрим 2МНК-оценку с единственной включенной эндогенной переменной и единственным инструментом. Тогда предсказанное значение, полученное на основе регрессии, оцененной на первом шаге, равно  $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ . Используя определения выборочных дисперсии и ковариации, покажите, что  $s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY}$  и  $s_{\hat{X}}^2 = \hat{\pi}_1^2 s_Z^2$ . Используйте этот результат, чтобы дополнить выкладки в приложении 12.2 для вывода формулы (12.4).
- 12.5. Рассмотрим модель регрессии с инструментальными переменными следующего вида:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i,$$

где  $X_i$  коррелирована с  $u_i$ , а  $Z_i$  – инструмент. Предположим, что три основных предположения из вставки «Основные понятия 12.4» выполнены. Какое из условий инструментальных переменных не выполняется, если:

- a)  $Z_i$  не зависит от  $(Y_i, X_i, W_i)$ ?  
 б)  $Z_i = W_i$ ?  
 в)  $W_i = 1$  для всех  $i$ ?  
 г)  $Z_i = X_i$ ?
- 12.6. В модели регрессии с инструментальными переменными с одним регрессором  $X_i$  и одним инструментом  $Z_i$ , в которой оценивается зависимость  $X_i$  от  $Z_i$ , коэффициент детерминации равен:  $R^2 = 0,05$ , а  $n = 100$ . Является ли  $Z_i$  сильным инструментом? [Подсказка: см. уравнение (7.14).] Как изменится ваш ответ, если  $R^2 = 0,05$ , а  $n = 500$ ?
- 12.7. В модели регрессии с инструментальными переменными с одним регрессором  $X_i$  и двумя инструментами  $Z_{1i}$  и  $Z_{2i}$ , значение J-статистики равно 18,2.  
 а) Означает ли это, что  $E(u_i | Z_{1i}, Z_{2i}) \neq 0$ ? Поясните свой ответ.  
 б) Означает ли это, что  $E(u_i | Z_{1i}) \neq 0$ ? Поясните свой ответ.
- 12.8. Рассмотрим рынок какого-либо продукта, функция предложения для которого имеет вид:  $Q_i^s = \beta_0 + \beta_1 P_i + u_i^s$ , а функция спроса —  $Q_i^d = \gamma_0 + u_i^d$ . Рынок находится в равновесии при условии  $Q_i^s = Q_i^d$ , где  $u_i^s$  и  $u_i^d$  — взаимно независимые одинаково распределенные случайные величины с нулевым математическим ожиданием.  
 а) Покажите, что  $P_i$  и  $u_i^s$  коррелированы.  
 б) Покажите, что МНК-оценка  $\beta_1$  не является состоятельной.  
 в) Каким образом можно оценить  $\beta_0$ ,  $\beta_1$  и  $\gamma_0$ ?
- 12.9. Исследователь изучает влияние военной службы на человеческий капитал. Он выбирает данные из случайной выборки, состоящей из 4000 человек старше 40 лет, и оценивает регрессию с помощью МНК:  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , где  $Y_i$  — годовой заработок человека,  $X_i$  — бинарная переменная, которая равна единице, если работник служил в армии, и равна 0 в противном случае.  
 а) Поясните, почему МНК-оценки могут быть ненадежными. (Подсказка: какие переменные могут быть пропущены в указанном выше уравнении? Коррелированы ли они с фактом военной службы?)  
 б) Во время войны во Вьетнаме происходил призыв на военную службу, где очередность призыва определялась с помощью национальной лотереи. (Даты рождения потенциальных призывников выбирались случайным образом и нумеровались от 1 до 365. Потенциальные призывники, чьи даты рождения имели меньшие по абсолютной величине присвоенные номера, призывались в первую очередь.) Поясните, как лотерея может использоваться для оценки влияния службы в армии на уровень дохода. (Для получения дополнительной информации по данному вопросу можно ознакомиться со следующей работой: Joshua D. Angrist, «Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administration Records», American Economic Review, June 1990: 313–336.)
- 12.10. Рассмотрим модель регрессии с инструментальными переменными  $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$ , где  $Z_i$  — инструментальная переменная. Предположим, что данные по  $W_i$  недоступны, а модель оценивается с исключением из нее переменной  $W_i$ .

- a) Предположим, что  $Z_i$  и  $W_i$  не коррелированы. Являются ли IV-оценки состоятельными?
- б) Предположим, что  $Z_i$  и  $W_i$  коррелированы. Являются ли IV-оценки состоятельными?

## **Компьютерные упражнения**

E12.1. В течение 1880-х годов картель, известный как Объединенный исполнительный комитет [Joint Executive Committee (JEC)], контролировал перевозки зерна железнодорожным транспортом на территориях от Среднего Запада до городов на востоке США. Картель существовал до подписания в 1890 году закона Шермана и поэтому вполне законно поднимал цены на зерно на более высокий уровень, чем в случае свободной конкуренции. Время от времени преступные действия членов картеля приводили к коллапсу договорного механизма установления цен на зерно. В данном упражнении необходимо, используя данные по изменениям предложения зерна, связанные с действиями картеля, оценить эластичность спроса на услуги по перевозке зерна железнодорожным транспортом. На веб-сайте [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) представлен файл JEC, содержащий еженедельные данные по стоимости доставки зерна и ряду других факторов в период с 1880 по 1886 год<sup>1</sup>. Детальное описание данных содержится в файле JEC\_Description, который также доступен на указанном сайте.

Предположим, что кривая спроса на услуги железнодорожного транспорта по перевозке зерна может быть записана в таком виде:

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + \beta_2 Ice_i + \sum_{j=1}^{12} \beta_{2+j} Seas_{j,i} + u_i, \text{ где } Q_i - \text{объем перевезенного зерна (в тоннах) за неделю } i, \text{ и } P_i - \text{стоимость доставки тонны зерна по железной дороге, } Ice_i - \text{бинарная переменная, которая равна 1, если Великие озера являются несудоходными из-за льда, } Seas_j - \text{бинарная переменная, которая отражает сезонные колебания спроса. Переменная } Ice \text{ рассматривается в данной модели, поскольку зерно может быть также перевезено на корабле в те периоды, когда судоходство на Великих озерах возможно.}$$

- a) Оцените уравнение спроса с помощью МНК. Какова оценка эластичности спроса и ее стандартная ошибка?
- б) Объясните, почему взаимодействие спроса и предложения может привести к смещениям в МНК-оценке эластичности.
- в) Рассмотрите использование переменной *cartel* в качестве инструментальной переменной для  $\ln(P_i)$ . Приведите экономические аргументы

---

<sup>1</sup> Данные любезно предоставлены профессором Северо-Западного университета (Northwestern University) Робертом Портером (Robert Porter). Ранее данные использовались в его работе «A Study of Cartel Stability: The Joint Executive Committee, 1880–1886», The Bell Journal of Economics, 1983, 14 (2): 301–314.

- ты в пользу того, что переменная *cartel* с большой вероятностью удовлетворяет условиям допустимости инструмента.
- г) Оцените регрессию первого шага 2МНК. Является ли переменная *cartel* слабым инструментом?
- д) Оцените уравнение спроса с помощью регрессии с инструментальными переменными. Какова оценка эластичности спроса и ее стандартные ошибки?
- е) Говорят ли данные и полученные оценки о том, что картель искусственно завышал цену до ее монопольного (максимизирующего прибыль картеля) уровня? Поясните свой ответ. (*Подсказка:* каковы действия монополиста, если эластичность по цене меньше 1?)
- E12.2. Каким образом рождаемость влияет на предложение труда? То есть в какой степени падает предложение труда среди женщин при рождении еще одного ребенка? В этом упражнении необходимо оценить этот эффект с помощью базы данных за 1980 год U.S. Census (по замужним женщинам)<sup>1</sup>. Данные доступны на веб-сайте [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) в файле Fertility, а их описание представлено в файле Fertility\_Description. База данных содержит информацию о замужних женщинах в возрасте от 21 до 35 лет, имеющих двух или более детей.
- а) Оцените регрессию *weeksworked* на индикаторную переменную *morekids* с помощью МНК. Действительно ли женщины, имеющие более детей, работают меньше, чем женщины с двумя детьми? Насколько меньше?
- б) Объясните, почему оцененная в пункте (а) МНК-регрессия не подходит для оценки влияния рождаемости (*morekids*) на предложение труда (*weeksworked*)?
- в) База данных также содержит переменную *samesex*, которая равна 1, если первые два ребенка имеют одинаковый пол («мальчик-мальчик» или «девочка-девочка»), и равна 0 в противном случае. Действительно ли семейные пары, у которых первые два ребенка одинакового пола, с большей вероятностью имеют третьего ребенка, чем другие семейные пары? Насколько велик данный эффект? Является ли он статистически значимым?
- г) Поясните, почему переменная *samesex* может являться допустимым инструментом для IV-регрессии *weeksworked* на *morekids*?
- д) Является ли переменная *samesex* слабым инструментом?
- е) Оцените регрессию *weeksworked* на *morekids*, используя *samesex* в качестве инструментальной переменной. Какова величина влияния рождаемости на предложение труда?
- ж) Изменятся ли полученные ранее результаты, если в регрессию, описывающую предложение труда, добавить переменные *agem1*, *black*, *hispan*, *othrace* (считая эти переменные экзогенными)? Поясните свой ответ.

<sup>1</sup> Данные были предоставлены Профессором Университета Мэриленд (University of Maryland) Уильямом Эвансом (William Evans). Ранее данные были использованы в его совместной работе с Joshua Angrist «Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size», American Economic Review, 1998, 88 (3): 450–477.

E12.3. (Для выполнения этого упражнения необходимо изучить приложение 12.5.) На веб-сайте [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) представлен файл WeakInstrument, который содержит 200 наблюдений по переменным ( $Y_i, X_i, Z_i$ ) для модели IV-регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$ .

- а) Постройте оценку  $\hat{\beta}_1^{TSLS}$ , ее стандартные ошибки и 95%-й доверительный интервал для  $\beta_1$ .
- б) Вычислите  $F$ -статистику для регрессии  $X_i$  на  $Z_i$ . Свидетельствует ли ее значение о наличии проблемы «слабого инструмента»?
- в) Вычислите 95%-й доверительный интервал для  $\beta_1$ , используя процедуру Андерсона-Рубина. (Для выполнения процедуры предположите, что  $-5 < \beta_1 < 5$ .)
- г) Прокомментируйте различия в доверительных интервалах, построенных в пунктах (а) и (в). Какой из них является более надежным?

## Приложения

### **Приложение 12.1. Панельные данные по потреблению сигарет**

База данных состоит из годовых наблюдений для 48 штатов США в период с 1985 по 1995 год. Объем потребления сигарет измеряется с помощью данных по годовым продажам сигарет (в пачках) на душу населения, полученных из базы данных о собираемости налогов. В качестве переменной, отражающей цену, используется реальная (т.е. с поправкой на инфляцию) средняя розничная стоимость пачки сигарет в течение финансового года с учетом налогов. В качестве переменной, отражающей уровень дохода, используется реальный доход на душу населения. В качестве переменной, отражающей налог с продаж, используется средний налог (в центах за пачку) по широкой корзине, которая применяется к потребительским товарам в разных штатах. Специфический налог на сигареты представляет собой налог, который применяется только к сигаретам (или прочей табачной продукции). Все цены, доходы и налоги, используемые в регрессии в этой главе, скорректированы на индекс потребительских цен и, следовательно, исчисляются в реальном выражении. Авторы благодарны профессору Массачусетского технологического института (MIT) Джонатану Груберу (Jonathan Gruber) за предоставление этих данных.

### **Приложение 12.2. Вывод формулы 2МНК-оценки из уравнения (12.4)**

На первом шаге 2МНК необходимо оценить регрессию  $X_i$  на инструментальную переменную  $Z_i$  и вычислить значения  $\hat{X}_i$ . На втором шаге оценивается регрессия  $Y_i$  на  $\hat{X}_i$  с помощью МНК. Соответственно, формула для 2МНК-оценки, выраженная в терминах оценок значений  $\hat{X}_i$ , представляет собой формулу для МНК-оценки из вставки «Основные понятия 4.2», в которой вместо  $X_i$  используется  $\hat{X}_i$ . То есть  $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y} / s_{\hat{X}}^2$ , где  $s_{\hat{X}}^2$  – выборочная дисперсия  $\hat{X}_i$ , а  $s_{\hat{X}Y}$  – выборочный коэффициент ковариации между  $Y_i$  и  $\hat{X}_i$ . Поскольку  $\hat{X}_i$  – это оценка

значений  $X_i$ , построенная на первом шаге 2МНК, то  $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ . Тогда из определения выборочных дисперсий и ковариаций можно получить:  $s_{\hat{X}Y}^2 = \hat{\pi}_1 s_{ZY}^2$  и  $s_{\hat{X}}^2 = \hat{\pi}_1^2 s_Z^2$  (упражнение 12.4). Таким образом, 2МНК-оценка может быть записана в таком виде:  $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y}^2 / s_{\hat{X}}^2 = s_{ZY}^2 / (\hat{\pi}_1^2 s_Z^2)$ .  $\hat{\pi}_1$  – коэффициент наклона (вычисленный с помощью МНК на первой стадии 2МНК),  $\hat{\pi}_1 = s_{ZX} / s_Z^2$ . Подстановка формулы для  $\hat{\pi}_1$  в выражение для  $\hat{\beta}_1^{TSLS} = s_{ZY}^2 / (\hat{\pi}_1^2 s_Z^2)$  дает формулу для 2МНК-оценки, представленную в уравнении (12.4).

### **Приложение 12.3. Распределение 2МНК-оценок в больших выборках**

В данном приложении изучается распределение 2МНК-оценок в больших выборках в случае, который рассматривался в разделе 12.1, то есть при наличии одной инструментальной переменной, одной эндогенной переменной и в отсутствие экзогенных переменных.

Сначала будет получена формула для 2МНК-оценок в терминах ошибок, что позволит сформировать базис для последующего обсуждения (как было сделано в случае МНК-оценок в уравнении (4.30) приложения 4.3). Из уравнения (12.1) следует, что  $Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$ . Тогда выборочный коэффициент ковариации может быть выражен так:

$$\begin{aligned} s_{ZY} &= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y}) = \\ &= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}) [\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})] = \\ &= \beta_1 s_{ZX} + \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})(u_i - \bar{u}) = \beta_1 s_{ZX} + \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}) u_i, \end{aligned} \quad (12.19)$$

где  $s_{ZX} = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})$  и  $\sum_{i=1}^n (Z_i - \bar{Z}) = 0$ . Подставляя определение для  $s_{ZX}$  и последнее выражение в (12.19) в определение  $\hat{\beta}_1^{TSLS}$ , умножая числитель и знаменатель на  $\frac{n-1}{n}$ , можно получить:

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}) u_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}. \quad (12.20)$$

### **Распределение $\hat{\beta}_1^{TSLS}$ в больших выборках при выполнении предположений IV-регрессии из вставки «Основные понятия 12.4»**

Уравнение (12.20) для 2МНК-оценок похоже на уравнение (4.30) из приложения 4.3 для МНК-оценок с той лишь разницей, что вместо  $X$  в числителе и знаменателе фигурирует  $Z$ , а знаменатель представляет собой коэффициент ковариации между  $Z$  и  $X$ , а не дисперсию  $X$ . Вследствие этого, а также из-за экзогенности  $Z$  аргументы, используемые в приложении 4.3 для доказательства

нормальности распределения МНК-оценок, могут быть расширены и на случай  $\hat{\beta}_1^{TSLS}$ .

В частности, при большом размере выборки  $\bar{Z} \cong \mu_Z$ , таким образом, числитель приблизительно равен  $\bar{q} = \frac{1}{n} \sum_{i=1}^n q_i$ , где  $q_i = (Z_i - \mu_Z) u_i$ . Поскольку инструмент является экзогенным, то  $E(q_i) = 0$ . Согласно предположениям IV-регрессии из вставки «Основные понятия 12.4»,  $q_i$  являются независимыми одинаково распределенными случайными величинами с дисперсией  $\sigma_q^2 = \text{var}[(Z_i - \mu_Z) u_i]$ .

Следовательно,  $\text{var}(\bar{q}) = \sigma_{\bar{q}}^2 = \frac{\sigma_q^2}{n}$  и, согласно центральной предельной теореме,  $\frac{\bar{q}}{\sigma_{\bar{q}}}$  в больших выборках имеет распределение  $N(0; 1)$ .

Поскольку выборочный коэффициент ковариации является состоятельной оценкой теоретического коэффициента ковариации, то  $s_{ZX} \xrightarrow{P} \text{cov}(Z_i, X_i)$  и он не равен нулю, поскольку инструмент является релевантным. Таким образом, согласно уравнению (12.20),  $\hat{\beta}_1^{TSLS} \cong \beta_1 + \bar{q} / \text{cov}(Z_i, X_i)$ , то есть в больших выборках  $\hat{\beta}_1^{TSLS}$  приблизительно имеет распределение  $N(\beta_1, \sigma_{\hat{\beta}_1^{TSLS}}^2)$ , где  $\sigma_{\hat{\beta}_1^{TSLS}}^2 = \sigma_{\bar{q}}^2 / (\text{cov}(Z_i, X_i))^2 = (1/n) \text{var}[(Z_i - \mu_Z) u_i] / [\text{cov}(Z_i, X_i)]^2$ , что представляет собой выражение, заданное уравнением (12.8).

### **Приложение 12.4. Распределение 2МНК-оценок в больших выборках при наличии недопустимых инструментов**

В данном приложении рассматривается распределение 2МНК-оценок в больших выборках в рамках предположений из раздела 12.1 (наличие одного  $X$  и одного  $Z$ ) в тех случаях, когда одно из условий допустимости инструмента не выполняется. Если не выполняется условие релевантности инструмента, то распределение 2МНК-оценок в больших выборках не является нормальным. В действительности оценки имеют распределение отношения двух нормальных случайных величин. Если не выполняется условие экзогенности инструмента, то 2МНК-оценки являются несостоятельными.

#### **Распределение $\hat{\beta}_1^{TSLS}$ при наличии слабых инструментов**

Сначала рассмотрим случай нерелевантности инструмента, то есть  $\text{cov}(Z_i, X_i) = 0$ . Тогда при выводе заключения в приложении 12.3 происходит деление на ноль. Во избежание данной проблемы необходимо более подробно рассмотреть знаменатель в уравнении (12.20) при равенстве нулю выборочного коэффициента ковариации.

Во-первых, перепишем уравнение (12.20) иначе. Поскольку в больших выборках среднее значение представляет собой состоятельную оценку математического ожидания, то  $\bar{Z}$  практически совпадает с  $\mu_Z$ , а  $\bar{X}$  практически совпадает с  $\mu_X$ . Таким образом, знаменатель в выражении (12.20) приблизительно равен:  $\frac{1}{n} \sum_{i=1}^n (Z_i - \mu_Z)(X_i - \mu_X) = \frac{1}{n} \sum_{i=1}^n r_i = \bar{r}$ , где  $r_i = (Z_i - \mu_Z)(X_i - \mu_X)$ . Пусть  $\sigma_r^2 =$

$= \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$ , пусть  $\sigma_r^2 = \sigma_r^2 / n$  и пусть  $\bar{q}$ ,  $\sigma_q^2$  и  $\sigma_{\bar{q}}^2$  определены так же, как в приложении 12.3. Тогда из выражения (12.20) следует, что в больших выборках

$$\hat{\beta}_1^{\text{TSLS}} \cong \beta_1 + \frac{\bar{q}}{\bar{r}} = \beta_1 + \left( \frac{\sigma_q}{\sigma_r} \right) \left( \frac{\bar{q} / \sigma_{\bar{q}}}{\bar{r} / \sigma_{\bar{r}}} \right) = \beta_1 + \left( \frac{\sigma_q}{\sigma_r} \right) \left( \frac{\bar{q} / \sigma_{\bar{q}}}{\bar{r} / \sigma_{\bar{r}}} \right). \quad (12.21)$$

Если инструмент не является релевантным, то  $E(r_i) = \text{cov}(Z_i, X_i) = 0$ . Таким образом,  $\bar{r}$  представляет собой выборочное среднее случайных величин  $r_i$ ,  $i = 1, \dots, n$ , которые являются независимыми одинаково распределенными случайными величинами (согласно второму предположению МНК) и имеют дисперсию  $\sigma_r^2 = \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$  (которая имеет конечное значение согласно третьему предположению модели регрессии с инструментальными переменными). Тогда к  $\bar{r}$  можно применить центральную предельную теорему, с помощью которой можно показать, что  $\bar{r} / \sigma_{\bar{r}}$  имеет распределение  $N(0, 1)$ . Следовательно, итоговый вид выражения (12.21) говорит о том, что  $\hat{\beta}_1^{\text{TSLS}} - \beta_1$  имеет распределение  $aS$ , где  $a = \frac{\sigma_q}{\sigma_r}$  и  $S$  – отношение двух случайных величин, каждая из которых имеет стандартное нормальное распределение (в данном случае эти две случайные величины коррелированы).

Иначе говоря, в тех случаях, когда инструмент не является релевантным, применение ЦПТ к числителю и знаменателю выражения для 2МНК-оценки позволяет показать, что в больших выборках 2МНК-оценка имеет распределение отношения двух нормально распределенных случайных величин. Поскольку  $X_i$  и  $u_i$  коррелированы, то эти нормальные случайные величины тоже коррелированы, поэтому распределение 2МНК-оценки в больших выборках в тех случаях, когда инструмент не является релевантным, имеет достаточно сложный вид. В больших выборках распределение 2МНК-оценки при нерелевантных инструментах сходится по вероятности к МНК-оценке. Таким образом, если инструмент не является релевантным, 2МНК не устраняет смещения, имеющего место в МНК, и, более того, имеет ненормальные распределения даже в больших выборках.

Слабый инструмент представляет собой промежуточный случай между нерелевантным инструментом и случаем с нормальным распределением, рассмотренным в приложении 12.3. Когда инструмент является слабым, но является релевантным, распределение 2МНК-оценки продолжает оставаться ненормальным, поэтому действия в случае нерелевантного инструмента переносятся и на случай слабого инструмента.

### Распределение $\hat{\beta}_1^{\text{TSLS}}$ в больших выборках при наличии эндогенного инструмента

Числитель дроби в итоговом виде выражения (12.20) сходится по вероятности к  $\text{cov}(Z_i, u_i)$ . Если инструмент является экзогенным, то этот коэффициент ковариации равен нулю, а 2МНК-оценка является состоятельной (при наличии инструмента, который не является слабым). Однако если инструмент не является

экзогенным, то если инструмент не является еще и слабым, можно записать  $\hat{\beta}_1^{TSLS} \xrightarrow{P} \beta_1 + \text{cov}(Z_i, u_i) / \text{cov}(Z_i, X_i) \neq \beta_1$ . То есть если инструмент не является экзогенным, то 2МНК-оценка не является состоятельной.

### **Приложение 12.5. Анализ при наличии слабых инструментов**

В данном приложении рассматриваются методы анализа с помощью инструментальных переменных при наличии потенциально слабых инструментов. Основной акцент делается на рассмотрении случая одного эндогенного регрессора (выражения (12.13) и (12.14)).

#### **Тестирование слабых инструментов**

Эмпирическое правило, представленное во вставке «Основные понятия 12.5», говорит о том, что если построенная на первом этапе 2МНК  $F$ -статистика принимает значение, меньшее 10, то инструменты являются слабыми. Одно из обоснований этого эмпирического правила основано на анализе выражения для смещения в 2МНК-оценке. Пусть  $\beta_1^{OLS}$  обозначает предел по вероятности МНК-оценки  $\beta_1$ , а  $(\beta_1^{OLS} - \beta_1)$  обозначает асимптотическое смещение МНК-оценки (если регрессор является эндогенным, то  $\hat{\beta}_1 \xrightarrow{P} \beta_1^{OLS} \neq \beta_1$ ). Можно показать, что при наличии нескольких инструментов смещение 2МНК-оценки приблизительно равно  $E(\hat{\beta}_1^{TSLS}) - \beta_1 \approx (\beta_1^{OLS} - \beta_1) / [E(F) - 1]$ , где  $E(F)$  – математическое ожидание построенной на первом шаге 2МНК  $F$ -статистики. Если  $E(F) = 10$ , то смещение 2МНК-оценки составляет приблизительно  $1/9$  по сравнению со смещением в МНК-оценке (немногим больше 10 %), что составляет достаточно небольшую величину и зачастую лежит в допустимых пределах в рамках различных приложений. Замена  $E(F) > 10$  на  $F > 10$  дает эмпирическое правило, представленное во вставке «Основные понятия 12.5».

В предыдущем параграфе приведена формула для смещения 2МНК-оценки в случае наличия нескольких инструментов. В большинстве практических приложений, как правило, количество инструментов  $m$  мало. В работе Стока и Його (Stock, Yogo, 2005) представлен формальный тест на наличие слабых инструментов, который позволяет избежать предположений о том, что  $m$  принимает большие значения. В teste Стока–Його (Stock–Yogo test) нулевая гипотеза заключается в том, что инструменты являются слабыми, а альтернативная гипотеза заключается в том, что инструменты являются сильными, где сильными являются те инструменты, для которых смещение 2МНК-оценки составляет максимально 10 % процентов от смещения МНК-оценки. Тест заключается в сравнении построенных на первом этапе 2МНК  $F$ -статистик (по техническим причинам рассматривается только случай с гомоскедастичными ошибками) с критическим значением, которое зависит от количества инструментов. Для уровня значимости 5 % это критическое значение лежит в пределах от 9,08 до 11,52, таким образом, эмпирическое правило, заключающееся в сравнении  $F$ -статистики с 10, является хорошим приближением для теста Стока–Його.

### **Тестирование гипотез и доверительные интервалы для $\beta$**

Если инструменты являются слабыми, то 2МНК-оценки смещены и имеют распределение, отличающееся от нормального. Таким образом, построенный на основе  $t$ -статистики, полученной с помощью 2МНК, тест для проверки гипотезы  $\beta_1 = \beta_{1,0}$  является ненадежным, как и доверительные интервалы для  $\beta_1$ . Тем не менее существуют иные тесты для тестирования этой гипотезы, которые являются корректными вне зависимости от того, являются ли инструменты сильными, слабыми или даже нерелевантными. В случае одного эндогенного регрессора наиболее подходящим является тест Морейры (Moreira, 2003), который представляет собой тест отношения условного правдоподобия (Conditional Likelihood Ratio (CLR) Test). Более старый тест, который применим в случае любого количества эндогенных регрессоров, основан на статистике Андерсона–Рубина (Anderson, Rubin, 1949). Поскольку данная статистика концептуально более проста, то она будет описана первой.

Тест Андерсона–Рубина позволяет проверить гипотезу о том, что  $\beta_1 = \beta_{1,0}$ , и состоит из двух шагов. На первом шаге вычисляется новая переменная  $Y_i^* = Y_i - \beta_{1,0}X_i$ . На втором шаге необходимо оценить регрессию  $Y_i^*$  на используемые экзогенные регрессоры ( $W$ ) и инструментальные переменные ( $Z$ ). Статистика Андерсона–Рубина представляет собой  $F$ -статистику для тестирования гипотезы о том, что все коэффициенты при инструментальных переменных равны нулю. Если в рамках нулевой гипотезы о том, что  $\beta_1 = \beta_{1,0}$ , инструменты удовлетворяют условию экзогенности (см. условие 2 во вставке «Основные понятия 12.3»), то они не будут коррелированы с ошибками в данной регрессии, и нулевая гипотеза будет отвергнута на уровне значимости 5 % для всех выборок.

Как обсуждалось в разделах 3.3 и 7.4, доверительный интервал может быть построен как некоторое множество значений параметров, равенство которых не отвергается при тестировании гипотезы. Соответственно, множество значений  $\beta_1$ , которые не отвергаются тестом Андерсона–Рубина на уровне значимости 5 %, составляют 95 %-й доверительный интервал для  $\beta_1$ . В тех случаях, когда  $F$ -статистика Андерсона–Рубина вычисляется в предположении гомоскедастичности ошибок, доверительный интервал может быть построен с помощью решения квадратного уравнения (см. компьютерное упражнение E12.3). В рамках логики, в которой построена статистика Андерсона–Рубина, никогда не предполагается релевантность инструментов, а доверительные интервалы Андерсона–Рубина будут покрывать 95 % вероятности вне зависимости от того, являются ли инструменты сильными, слабыми или нерелевантными.

CLR-статистика также позволяет проверять гипотезу  $\beta_1 = \beta_{1,0}$ . Тест отношения правдоподобия сравнивает значение функции правдоподобия (см. приложение 11.2) в условиях нулевой гипотезы с ее значением в рамках альтернативной гипотезы и отвергает ее, если значение функции правдоподобия в условиях альтернативной гипотезы существенно больше, чем в условиях нулевой гипотезы. Аналогичные тесты, которые рассматривались в данной книге, такие как  $F$ -тест для случая гомоскедастических ошибок для множественной

регрессии, могут быть получены из теста отношения правдоподобия в предложении гомоскедастичности и нормальности остаточного члена. В отличие от любых других тестов, обсуждавшихся в настоящей книге, критические значения CLR-теста зависят от имеющихся данных, в частности – от статистики, которая измеряет силу инструментов. При использовании верного критического значения CLR-тест является корректным даже вне зависимости от того, являются ли инструменты сильными, слабыми или нерелевантными. Основанные на CLR-тесте доверительные интервалы могут быть построены как множество значений  $\beta_1$ , которые не отвергаются CLR-тестом.

CLR-тест эквивалентен 2МНК  $t$ -статистике в тех случаях, когда инструменты являются очень сильными, и имеет очень хорошую мощность в тех случаях, когда инструменты являются слабыми. При наличии подходящего программного обеспечения CLR-тест очень легко использовать. Основным недостатком данного теста является отсутствие простой возможности обобщения на случай нескольких эндогенных регрессоров. В подобных случаях рекомендуется использовать тест Андерсона–Рубина. Тем не менее в тех случаях, когда инструменты являются сильными (т.е. 2МНК можно применять), а коэффициенты определены, тест Андерсона–Рубина является неэффективным в том смысле, что является менее мощным, чем 2МНК  $t$ -статистика.

### Оценка $\beta$

Если инструменты не являются релевантными, то получить несмешенную оценку  $\beta_1$  даже в больших выборках невозможно. Тем не менее при наличии слабых инструментов некоторые IV-оценки в большей степени приближены к истинному значению  $\beta_1$ , чем 2МНК-оценки. Одной из таких оценок является оценка максимального правдоподобия при ограниченной информации (*limited information maximum likelihood (LIML) estimator*). Как можно догадаться из названия, LIML-оценка представляет собой оценку максимального правдоподобия для  $\beta_1$  в системе уравнений (12.13) и (12.14) (для получения дополнительной информации по оценкам максимального правдоподобия можно ознакомиться с приложением 11.2). LIML-оценка также представляет собой значение  $\beta_{1,0}$ , которое минимизирует статистику в teste Андерсона–Рубина (для случая гомоскедастичности). Таким образом, если доверительный интервал Андерсона–Рубина не является пустым множеством, то он будет содержать LIML-оценку. Стоит отметить, что доверительный интервал CLR-теста также содержит LIML-оценку.

Если инструменты являются слабыми, то LIML-оценка в большей степени приближена к истинному значению  $\beta_1$ , чем 2МНК-оценка. Если инструменты являются сильными, то LIML-оценки и 2МНК-оценки асимптотически равны. Основной недостаток LIML-оценки заключается в том, что она может давать существенные выбросы. Доверительные интервалы, построенные с помощью LIML-оценки и соответствующих стандартных ошибок, являются более надежными, нежели доверительные интервалы, построенные с помощью 2МНК, но являются менее надежными, чем доверительные интервалы, построенные с по-

мощью статистик CLR или Андерсона–Рубина, когда инструменты являются слабыми.

Проблемы оценивания, тестирования гипотез и построения доверительных интервалов с помощью модели регрессии с инструментальными переменными в случае слабых инструментов представляют собой перспективное направление для изучения. Для получения дополнительной информации по данной тематике посетите веб-сайт учебника.

### **Приложение 12.6. 2МНК с контрольными переменными**

Во вставке «Основные понятия 12.4» предполагалось, что переменные  $W$  являются экзогенными. В данном приложении рассматривается случай, в рамках которого  $W$  не являются экзогенными, но являются контрольными переменными, которые используются, чтобы  $Z$  были экзогенными. Логика включения контрольных переменных в 2МНК-регрессию схожа с логикой в МНК: если с помощью контрольной переменной можно эффективно контролировать пропущенный фактор, то инструментальная переменная не коррелирует с остаточным членом. Поскольку контрольные переменные коррелированы с остаточным членом, то коэффициенты при контрольных переменных не имеют причинной интерпретации. Математические выкладки в случае контрольных переменных в 2МНК также схожи со случаем МНК, но в них используется ослабление предположения о том, что остаточный член имеет равное нулю условное математическое ожидание при заданных  $Z$  и  $W$ . Вместо этого предполагается, что условное среднее остаточного члена не зависит от  $Z$ . Настоящее приложение опирается на приложение 7.2 (независимость условного среднего) и поэтому будет рассмотрено в первую очередь.

Рассмотрим модель регрессии с инструментальными переменными с одним  $X$  и одним  $W$ , представленную выражением (12.12):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \quad (12.22)$$

Заменим первое предположение модели регрессии с инструментальными переменными из вставки «Основные понятия 12.4», которое говорит о том, что  $E(u_i|W_i) = 0$ , на условие о том, что при заданных значениях  $W_i$  среднее значение  $u_i$  не зависит от  $Z_i$ :

$$E(u_i|W_i, Z_i) = E(u_i|W_i). \quad (12.23)$$

Следуя приложению 7.2, предположим далее, что  $E(u_i|W_i)$  линейно по  $W_i$ , то есть  $E(u_i|W_i) = \gamma_0 + \gamma_2 W_i$ , где  $\gamma_0$  и  $\gamma_2$  – коэффициенты. Положив, что  $\varepsilon_i = u_i - E(u_i|W_i, Z_i)$ , и применив выражение (7.25) к уравнению (12.22), получим:

$$Y_i = \delta_0 + \beta_1 X_i + \delta_2 W_i + \varepsilon_i, \quad (12.24)$$

где  $\delta_0 = \beta_0 + \gamma_0$  и  $\delta_2 = \beta_2 + \gamma_2$ . Тогда  $E(\varepsilon_i|W_i, Z_i) = E[u_i - E(u_i|W_i, Z_i)|W_i, Z_i] = E(u_i|W_i, Z_i) - E(u_i|W_i, Z_i) = 0$ , что говорит о том, что  $\text{corr}(Z_i, \varepsilon_i) = 0$ . Таким образом, первое предположение модели регрессии с инструментальными переменными и условие экзогенности (второе условие из вставки «Основные

понятия 12.3» выполняются для уравнения (12.24) с ошибкой  $\varepsilon_i$ . Следовательно, если первое предположение модели регрессии с инструментальными переменными заменяется условием независимости условного среднего (12.23), то исходные предположения IV-регрессии из вставки «Основные понятия 12.4» применимы к модифицированной регрессии в уравнении (12.24).

Поскольку предположения модели IV-регрессии из вставки «Основные понятия 12.4» выполняются для уравнения (12.24), то все методы построения статистических выводов (как для случая сильных, так и для случая слабых инструментов), которые обсуждались в данной главе, применимы и для уравнения (12.24). В частности, если инструменты являются сильными, то для коэффициентов в уравнении (12.24) могут быть получены состоятельные оценки с помощью 2МНК, а выводы из тестов, проведенных с помощью 2МНК-статистик и доверительных интервалов, будут надежными.

Как и в МНК с контрольными переменными, в общем случае 2МНК-оценка коэффициента при контрольной переменной  $W$  не имеет причинной интерпретации. 2МНК позволяет состоятельно оценить  $\delta_2$  в уравнении (12.24), но  $\delta_2$  представляет собой сумму прямого эффекта влияния  $W$  ( $\beta_2$ ) и коэффициента  $\gamma_2$ , который отражает корреляцию между  $W$  и пропущенными факторами, заключенными в  $u_i$ .

В рамках рассмотренного ранее примера с исследованием потребления сигарет с помощью представленных в таблице 12.1 результатов регрессионного анализа сделана попытка проинтерпретировать коэффициент при 10-летних изменениях логарифма уровня дохода как эластичность спроса по доходу. Если, тем не менее, рост дохода коррелирован с увеличением уровня образования и если дополнительное образование снижает уровень курения, то рост уровня дохода будет являться причиной дополнительного роста дохода, то есть будет иметь место эффект собственной причинности ( $\beta_2$ , эластичность по доходу), а также эффект, возникающий из-за корреляции с уровнем образования ( $\gamma_2$ ). Если последний эффект является негативным ( $\gamma_2 < 0$ ), то представленные в таблице 12.1 оценки коэффициента перед переменной уровня доходов (которые оценивают  $\delta_2 = \beta_2 + \gamma_2$ ) будут смещены вниз (будут «недооценивать» эластичность по доходу). Однако если условие независимости условного среднего в уравнении (12.23) выполняется, то 2МНК-оценка ценовой эластичности является состоятельной.

# **Глава 13. Эксперименты и квазиэксперименты**

Во многих областях, таких как, например, психология и медицина, для анализа причинных взаимосвязей зачастую используются различного рода эксперименты. Например, перед тем как новое лекарство будет одобрено для широкого медицинского применения, оно должно пройти серию тестовых испытаний,ключающихся в том, что отдельным случайно выбранным пациентам назначается данное лекарство, в то время как остальные принимают безвредный заменитель (плацебо). Фирма получает разрешение на запуск лекарства в производство, только если проведенный случайный контролируемый эксперимент статистически подтверждает, что лекарство является эффективным и безопасным.

Существуют три основные причины для изучения контролируемых экспериментов в рамках курса эконометрики. Во-первых, случайный контролируемый эксперимент дает концептуальную возможность получить оценки влияния различного рода причинных взаимосвязей (причинных эффектов, причинного влияния) на основе полученных наблюдений. Во-вторых, результаты случайных контролируемых экспериментов могут быть достаточно важны в тех или иных условиях, поэтому важно понимать ограничения и основные угрозы обоснованности подобных экспериментов, а также их основные сильные стороны. В-третьих, внешние условия и сопутствующие обстоятельства могут вносить «случайность» или случайные искажения в проводимые эксперименты, то есть из-за внешних событий лечение какого-либо индивида назначается как бы «случайным» образом, возможно, при определенных заданных значениях некоторых контрольных переменных. Эта вышеописанная «случайность» дает «квазиэксперимент» или «естественный эксперимент», что позволяет применить множество разработанных методов анализа случайных экспериментов и для случая квазиэкспериментов (с некоторыми модификациями).

В настоящей главе рассматриваются различного рода эксперименты и квазиэксперименты в экономике. Среди статистических методов, применяемых в настоящей главе, можно выделить модель множественной регрессии, регрессионный анализ с помощью моделей панельных данных и моделей с использованием инструментальных переменных. Данная глава отличается от остальных не применяемыми методами анализа, а данными, которые используются при анализе, а также особыми возможностями и целями, которые появляются при анализе экспериментов и квазиэкспериментов.

Методология, представленная в настоящей главе, достаточно часто применяется для оценки социальных и экономических программ. *Оценка программных документов* представляет собой отдельную область, которая изучает вопросы

оценки влияния и последствий принятия программных документов, отдельных видов политики, а также прочих видов «интервенций». Какое влияние оказывает прохождение профессиональной переподготовки на уровень доходов? Каков эффект повышения минимальной заработной платы на уровень занятости низкоквалифицированных работников? Как влияет уровень доступности различного рода кредитов для разных групп студентов на уровень посещаемостиими занятий? В настоящей главе рассматриваются способы, которые позволяют оценить результаты подобных программ и видов политики с помощью экспериментов и квазиэкспериментов.

В разделе 13.1 рассмотрены представленные в главах 1, 3 и 4 методы оценки причинных эффектов, которые могут быть использованы для оценки причинных эффектов на основе случайных контролируемых экспериментов. В действительности эксперименты с участием человека могут быть подвержены влиянию ряда практических проблем, которые несут угрозу их внутренней и внешней обоснованности. Эти проблемы, а также некоторые эконометрические подходы, которые используются для борьбы с ними, представлены в разделе 13.2. В разделе 13.3 анализируется достаточно важный случайный контролируемый эксперимент, в рамках которого учащиеся начальной школы были случайным образом распределены по различным по размеру учебным классам в штате Теннесси в конце 1980-х годов.

В разделе 13.4 рассматриваются оценки причинных эффектов с использованием квазиэкспериментов. Угрозы обоснованности квазиэкспериментов рассматриваются в разделе 13.5. Одна из проблем, которая возникает в экспериментах и квазиэкспериментах, заключается в том, что эффекты могут различаться для разных элементов выборки. Способы интерпретации результатов оценки причинных эффектов при наличии неоднородности в выборке рассматриваются в разделе 13.6.

### **13.1. Потенциальные исходы, причинные эффекты и идеализированные эксперименты**

В данном разделе представлены способы оценки средних индивидуальных причинных эффектов на основе случайных контролируемых экспериментов, а также возможности анализа данных, полученных при помощи подобного рода экспериментов и модели множественной регрессии.

#### ***Потенциальные исходы и средние причинные эффекты***

Предположим, вы принимаете какое-либо лекарство (по медицинским показаниям) и одновременно выбираете между участием в программе повышения квалификации и решением дополнительного набора задач по эконометрике. Разумно было бы задать самому себе вопрос: «Каковы положительные стороны того или иного варианта для меня, учитывая прием лекарств?». Можно представить себе две гипотетические ситуации: в одной из них вы принимаете лекарство, а в другой нет. В рамках каждой из этих ситуаций необходимо иметь

некий результат, который можно было бы измерить (изменение самочувствия, прием на работу после прохождения курсов повышения квалификации, оценки по эконометрике). Различие этих двух потенциальных исходов было бы для вас причинным эффектом в эксперименте.

*Потенциальным исходом* принято называть какой-либо результат для индивида при потенциальном наличии того или иного сопутствующего фактора (в данном случае – лечения). Причинный эффект для такого индивида может рассматриваться как различие в потенциальных исходах между случаями, когда лекарство принимается и когда не принимается. В общем случае причинные эффекты могут быть различными для разных индивидов. Например, результаты применения лекарственных препаратов могут зависеть от вашего возраста, вредных привычек (например курения), а также санитарно-гигиенических условий. Проблема заключается в том, что не существует способов измерения причинных эффектов для одного индивида. Поскольку индивид либо получает лечение, либо нет, то наблюдаемым является только один из исходов, а не оба.

Несмотря на то что различные причинные эффекты не могут быть измерены для отдельных индивидов, в рамках многих приложений достаточно знать среднюю величину причинных эффектов. Например, в рамках оценки результатов программ повышения квалификации могут рассматриваться средние расходы на одного обучаемого или средние показатели успешности обучаемых по поиску работы. Средние значения индивидуальных причинных эффектов в изучаемой выборке называются *средними причинными эффектами* (или *средними эффектами в эксперименте*).

Средние причинные эффекты для рассматриваемой генеральной совокупности могут быть оценены, как минимум, теоретически, на основе идеального случайного контролируемого эксперимента. Сначала предположим, что субъекты выбираются из исследуемой генеральной совокупности случайным образом. Поскольку субъекты выбираются случайным образом, то их потенциальные исходы и, таким образом, причинные эффекты в полученной выборке равны средним причинным эффектам для генеральной совокупности. Далее предположим, что выбранные субъекты случайным образом приписываются к исследуемой (экспериментальной) группе или к контрольной группе. Поскольку принадлежность к группе выбирается случайным образом, то потенциальные исходы для индивида (субъекта) распределены независимо от его принадлежности к той или иной группе. Таким образом, различие ожидаемых значений исходов для тех, кто приписан к экспериментальной группе, и тех, кто приписан к контрольной группе, равна ожидаемому значению причинного эффекта. Когда концепция потенциального исхода объединяется с предположениями о (1) случайному выборе индивидов из генеральной совокупности и о (2) случайному распределении по группам, то ожидаемое значение различия исходов между экспериментальной и контрольной группами равно среднему причинному эффекту в рамках генеральной совокупности. То есть, как предполагалось в разделе 3.5, причинное влияние на  $Y_i$  при наличии лечения ( $X_i = 1$ ) и при его отсутствии ( $X_i = 0$ ) равна разности условных математических ожиданий  $E(Y_i | X_i = 1) - E(Y_i | X_i = 0)$ , где  $E(Y_i | X_i = 1)$  и  $E(Y_i | X_i = 0)$  являются соответственно

ожидаемыми значениями  $Y$  для исследуемой и контрольной групп в рамках идеального случайного эксперимента. В приложении 13.3 представлено математическое обоснование этих утверждений.

В общем случае индивидуальные причинные эффекты могут рассматриваться как зависящие одновременно от наблюдаемых и ненаблюдаемых переменных. Мы уже сталкивались с идеей о том, что причинные эффекты могут зависеть от наблюдаемых переменных. Например, в главе 8 рассматривалась вероятность того, что влияние снижения размера класса на успеваемость может зависеть о того, изучает ли школьник английский язык. В основной части настоящей главы мы фокусируемся на тех случаях, когда причинные эффекты зависят только лишь от наблюдаемых переменных. В разделе 13.6 рассматривается ненаблюдаемая неоднородность в причинных эффектах.

### **Эконометрические методы анализа экспериментальных данных**

Данные случайных контролируемых экспериментов могут быть проанализированы с помощью сравнения разностей средних значений или с помощью оценки регрессий, в которые включены индикаторы, характеризующие эксперимент, и дополнительные контрольные переменные. Последняя спецификация (спецификации в разностях с дополнительными регрессорами) также может быть использована в более сложных случайных схемах, в которых вероятность случайного выбора зависит от конкретных наблюдаемых переменных.

**Оценка разностей.** Оценка разностей представляет собой разность между выборочными средними для экспериментальной и контрольной группами (см. раздел 3.5), которая может быть рассчитана с помощью оценки регрессии переменной исхода  $Y$  на бинарную переменную  $X$  ( $x_i = 1$ , если объект  $i$  получал воздействие):

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots, n. \quad (13.1)$$

Как обсуждалось ранее в разделе 4.4, если  $X$  выбирается случайным образом, то  $E(u_i | X_i) = 0$ , а МНК-оценка  $\hat{\beta}_1$  в уравнении (13.1) представляет собой несмещенную состоятельную оценку причинного эффекта.

**Оценка разностей с дополнительными регрессорами.** Эффективность оценки разностей может быть повышена с помощью включения в уравнение регрессии контрольных переменных  $W$ . Подобные действия позволяют получить оценку разностей с дополнительными регрессорами:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \dots + \beta_{1+r} W_{ri} + u_i, \quad i = 1, \dots, n. \quad (13.2)$$

Если  $W$  помогает объяснить изменения  $Y$ , то включение  $W$  в уравнение регрессии позволяет снизить стандартные ошибки регрессии и  $\hat{\beta}_1$  (в некоторых случаях). Как обсуждалось в разделе 7.5 и приложении 7.2, для того чтобы оценка  $\hat{\beta}_1$  причинного эффекта  $\beta_1$  в уравнении (13.2) была несмещенной, контрольные переменные должны быть такими, чтобы  $u_i$  удовлетворяли условию независимости условного среднего, то есть  $E(u_i | X_i, W_i) = E(u_i | W_i)$ . Это условие удовлетворяется, если  $W_i$  представляет собой индивидуальные характеристики, например пол. Если  $W_i$  – внешняя характеристика,  $X_i$  выбирается случайным образом, то  $X_i$  не зависит от  $u_i$  и  $W_i$ , что означает, что  $E(u_i | X_i, W_i) = E(u_i | W_i)$ .

Регрессоры  $W$  в уравнении (13.2) не должны включать экспериментальные исходы ( $X_i$  не являются случайными при заданных экспериментальных исходах). Как и в других случаях с контрольными переменными, при наличии условной независимости среднего коэффициенты при контрольных переменных не имеют причинной интерпретации.

**Оценка причинных эффектов, зависящих от наблюдаемых переменных.** Как обсуждалось в главе 8, изменение причинных эффектов, зависящее от наблюдаемых переменных, может быть оценено с помощью включения в модель подходящей нелинейной функции от  $X_i$  (или компоненты «взаимодействия»). Например, если  $W_{1i}$  является бинарной переменной, отражающей пол индивида, то различные причинные эффекты для мужчин и женщин могут быть оценены с помощью включения переменной  $W_{1i} \times X_i$  в уравнение регрессии (13.2).

**Случайный выбор, зависящий от наблюдаемых переменных.** Случайный выбор, при котором вероятность попадания индивида в экспериментальную группу зависит от одной или более наблюдаемых переменных  $W$ , называют *случайным выбором, зависящим от наблюдаемых переменных*. Если случайный выбор зависит от наблюдаемых переменных, то в общем случае оценка разностей, полученная на основе уравнения (13.1), будет смещена из-за пропущенных переменных. Например, в приложении 7.2 описан гипотетический эксперимент по оценке причинных влияний обязательных и необязательных домашних заданий в рамках курса эконометрики. В описанном эксперименте студенты-эконометристы ( $W_i = 1$ ) приписывались к экспериментальной группе (с обязательными домашними заданиями,  $X_i = 1$ ) с большей вероятностью, чем прочие студенты ( $W_i = 0$ ). Однако если студенты-эконометристы априори лучше осваивают данный курс, то тогда будет иметь место смещение в оценках, вызванное пропущенными переменными, поскольку попадание в экспериментальную группу будет коррелировано с пропущенной переменной, которая отражает специализацию студента (эконометрика или прочие специализации).

Поскольку  $X_i$  выбирается случайным образом при заданной  $W_i$ , то вызванные наличием пропущенных переменных смещения могут быть устранины с помощью использования оценки разностей с дополнительной контрольной переменной  $W_i$ . Случайный выбор  $X_i$  при заданных  $W_i$  (совместно с предположениями модели линейной регрессии) подразумевает, что  $X_i$  не зависит от  $u_i$  в уравнении (13.2) при заданных  $W_i$ . Эта условная независимость в свою очередь приводит к условной независимости среднего, то есть  $E(u_i | X_i, W_i) = E(u_i | W_i)$ . Таким образом, МНК-оценка  $\hat{\beta}_1$  в уравнении (13.2) представляет собой несмешенную оценку причинного эффекта в случае, когда выбор  $X_i$  случаен при заданных  $W_i$ .

## 13.2. Угрозы обоснованности экспериментов

Во вставке «Основные понятия 9.1» говорилось о том, что статистическое исследование является *внутренне обоснованным*, если полученные на его основе статистические выводы относительно причинных эффектов справедливы для

исследуемой выборки. Статистическое исследование является *внешне обоснованным*, если полученные на его основе статистические выводы и заключения могут быть обобщены с исследуемой выборки на другие выборки и генеральную совокупность в целом. При проведении существенного числа реальных статистических исследований с участием людей возникают проблемы и с внутренней, и с внешней обоснованностью.

### **Угрозы внутренней обоснованности**

Угрозы внутренней обоснованности случайных контролируемых экспериментов могут иметь отношение к проблемам, связанным со случайным отбором, процедурой распределения, истощением выборки, различной спецификой экспериментов и малыми размерами исследуемых выборок.

**Отсутствие случайности отбора.** Если распределение между экспериментальной и контрольной группами происходит не случайно, а частично основывается на характеристиках или предпочтениях субъектов, то исходы экспериментов будут отражать одновременно исследуемые эффекты и эффекты неслучайного распределения. Например, предположим, что участники экспериментальной программы по повышению квалификации приписываются к экспериментальной группе на основании того, с какой буквы начинается их фамилия (из первой половины алфавита или из второй). Из-за возможного наличия этнической специфики фамилий расовый состав может существенно различаться для экспериментальной и контрольной групп. Также можно отметить, что профессиональный опыт, образование и другие характеристики рынка труда могут различаться на основе расовой принадлежности, то есть для экспериментальной и контрольной групп могут возникать систематические отличия в этих пропущенных переменных, влияющих на исходы эксперимента. В целом неслучайное распределение по группам может приводить к корреляции между  $X_i$  и  $u_i$  в уравнениях (13.1) и (13.2), которая в свою очередь приводит к наличию смещений в оценках причинных эффектов.

Стоит отметить, что можно протестировать наличие случайности отбора. Если распределение по группам проводится случайным образом, то  $X_i$  не будут коррелировать с наблюдаемыми индивидуальными характеристиками  $W$ . Таким образом, *тест на случайность распределения между экспериментальной и контрольной группами* или *тест на случайность «назначения воздействия»* представляет собой проверку гипотезы о том, что коэффициенты при  $W_{1i}, \dots, W_{ri}$  в регрессии  $X_i$  на  $W_{1i}, \dots, W_{ri}$  равны нулю. В рассмотренном примере с экспериментальной программой по повышению квалификации оценка регрессии переменной, показывающей, что индивид проходил курс повышения квалификации ( $X_i$ ), на характеристики пола, расовой принадлежности и уровня образования (т.е. переменные  $W$ ) с последующим вычислением  $F$ -статистики для проверки гипотезы о том, что все коэффициенты при  $W$  равны нулю, представляет собой не что иное, как тестирование нулевой гипотезы о том, что распределение между экспериментальной и контрольной группами происходит случайным образом, против альтернативной гипотезы о том, что оно зависит от пола, расо-

вой принадлежности или уровня образования. Если особенности эксперимента подразумевают зависимость процесса случайного выбора от наблюдаемых переменных, то соответствующие переменные должны быть включены в уравнение регрессии. В этом случае с помощью  $F$ -теста необходимо протестировать коэффициенты при оставшихся переменных  $W^1$ .

**Нарушение условий эксперимента.** В реальной жизни люди не всегда следуют наставлениям и советам. Например, в рамках эксперимента с программой по повышению квалификации отдельные участники не посещают занятия. Аналогично субъекты, приписанные к контрольной группе, могут все равно проходить обучение (повышение квалификации) каким-либо другим способом, например по специальному запросу к инструктору или администратору.

Нарушение индивидами условий эксперимента (т.е. случайного влияния рассматриваемого эффекта) называется *частичным соответствием условиям эксперимента*. В некоторых случаях организаторы эксперимента в точности знают, соблюдались ли условия эксперимента (например, посещал ли работник занятия по повышению квалификации), что отражается в переменной  $X_i$ . В случае частичного соответствия условиям эксперимента появляется некий дополнительный элемент выбора, заключающийся в том, как тот или иной индивид соблюдает условия эксперимента и получает ли воздействие. Тогда  $X_i$  будет коррелирован с  $u_i$ , даже в тех случаях, если изначально распределение происходило случайным образом. Таким образом, нарушение условий эксперимента приводит к смещениям в МНК-оценке.

При наличии данных одновременно по участию в эксперименте («получение воздействия»,  $X_i$ ) и исходному случайному распределению между контрольной и исследуемой группами существует возможность оценки рассматриваемых выше эффектов с помощью регрессии с инструментальными переменными. *Оценка эффекта воздействия с помощью инструментальных переменных* включает в себя оценку уравнения (13.1) или уравнения (13.2) при наличии контрольных переменных с использованием данных по исходному случайному распределению ( $Z_i$ ) в качестве инструмента для  $X_i$ . Следует помнить, что для обоснованности используемая инструментальная переменная должна удовлетворять условиям релевантности и экзогенности инструментальных переменных (см. вставку «Основные понятия 12.3»). До тех пор пока условия эксперимента частично выполняются, реальный уровень измеряемых эффектов частично определяется исходным случайнym распределением, то есть  $Z_i$  является релевантной. Если исходное распределение по группам является случайным, то  $Z_i$  распределены независимо от  $u_i$  (условно на  $W_i$ , если случайность выбора зависит от наблюдаемых переменных), то есть инструменты являются экзогенными. Таким образом, в эксперименте со случайным распределением между контрольной и экспериментальной группами и частичным соответствием условиям

<sup>1</sup> В данном примере переменные  $X_i$  являются бинарными. Между тем, как обсуждалось в главе 11, регрессия  $X_i$  на  $W_{i1}, \dots, W_{in}$  представляет собой оценку линейной вероятностной модели, в которой важно использовать устойчивые к наличию гетероскедастичности стандартные ошибки. Еще одним способом протестировать гипотезу о том, что  $E(X_i|W_{i1}, \dots, W_{in})$  не зависит от  $W_{i1}, \dots, W_{in}$  в случае бинарности  $X_i$ , является использование пробит- или логит-модели (см. раздел 11.2).

эксперимента переменная, описывающая исходное случайное распределение по группам, является допустимой инструментальной переменной.

Описанная подобным образом стратегия использования инструментальных переменных требует одновременного наличия данных по предписанному распределению по группам и реальному участию в эксперименте. В некоторых случаях данные по реальному участию в эксперименте могут быть недоступны. Например, если при участии в медицинском эксперименте индивиду назначается некоторый лекарственный препарат, но пациент его просто не принимает, о чем не знают организаторы эксперимента. Тогда наблюдения по «полученному воздействию» в рамках данного эксперимента будут неверны. Неверное измерение реального участия в эксперименте приводит к смещениям оценки разностей.

**Истощение выборки.** Истощение выборки происходит в тех случаях, когда объекты удаляются из выборки уже после того, как они были приписаны к контрольной или экспериментальной группам. Иногда истощение выборки происходит по причинам, не связанным с условиями эксперимента – например, участник программы по повышению квалификации по своим личным, не связанным с экспериментом, причинам вынужден покинуть город. Но если истощение связано непосредственно с условиями эксперимента, то это может приводить к смещениям в МНК-оценках причинных эффектов. Например, предположим, что наиболее способные стажеры уходят из программы по повышению квалификации, поскольку они быстрее получают работу в другом городе, связанную с применением полученных в рамках программы навыков. Таким образом, к концу программы в группах остаются лишь наименее способные участники. Тогда распределение ненаблюдаемых характеристик (способностей) будет различаться для контрольной и экспериментальной групп (участие в эксперименте приводит к тому, что наиболее способные стажеры покидают город). Другими словами, факт участия в эксперименте  $X_i$  будет коррелирован с ошибками  $u_i$  (в которых учитывается наличие способностей) для тех индивидов, которые остаются в выборке до самого конца эксперимента, а получаемая оценка разностей будет смещенной. Поскольку истощение приводит к неслучайному формированию выборки, то по этой причине возникают смещения в оценках, связанные с отбором наблюдений (см. вставку «Основные понятия 9.4»).



### **Эффект Хоторна**

В течение 1920–1930-х годов компания General Electric проводила ряд исследований производительности работников на фабрике в Хоторне. В первой серии экспериментов исследователи варьировали мощность лампочек накаливания, которые использовались для освещения, чтобы увидеть, как освещение повлияло на производительность женщин, работающих на сборке электрических деталей. В других экспериментах варьировалась (увеличивалась или уменьшалась) протяженность периодов отдыха, изменялась рабочая обстановка в мастерской, сокращались рабочие дни. Наиболее впечатляющими были первые результаты

исследования, которые показывали, что производительность труда продолжала расти, несмотря на изменения освещенности, протяженности рабочих дней или рабочих условий. Исследователи пришли к выводу о том, что повышение производительности труда не являлось следствием изменений рабочей обстановки, но из-за того, что участие в эксперименте приводило к тому, что работники ощущали свою ценность и важность для работодателя, они работали все усерднее и усерднее. С течением времени идея о том, что участие в эксперименте влияет на поведение индивида, стала известна как эффект Хоторна (the Hawthorne effect).

Тем не менее в этой истории имеется один интересный факт: тщательное изучение исходных данных по Хоторну показывает отсутствие эффекта Хоторна (см. Gillespie, 1991 и Jones, 1992)! Однако в некоторых экспериментах, особенно в тех, где субъекты заинтересованы в результате, проведение эксперимента может повлиять на результат. Эффект Хоторна и экспериментальные эффекты в целом могут представлять угрозу для внутренней обоснованности, даже если эффект Хоторна не проявляется в исходных данных Хоторна.



**Экспериментальные эффекты.** В экспериментах, в которых задействованы люди, из-за того что они, в частности, могут изменять свое поведение вследствие участия в эксперименте, может возникать феномен, называемый **эффектом Хоторна** (см. вставку «Эффект Хоторна»).

В некоторых экспериментах «двойное слепое» исследование может смягчить влияние эффекта участия в эксперименте: хотя субъекты и организаторы эксперимента одновременно знают, что они участвуют в эксперименте, но не знают, относится ли тот или иной субъект к контрольной или экспериментальной группам. В медицинском эксперименте (с приемом лекарственных препаратов), например, иногда лекарство и плацебо могут выглядеть полностью идентично. Тогда ни медицинские работники, которые дозируют препарат, ни пациенты не знают, является ли принятый препарат лекарственным либо является плацебо. Если эксперимент проводится по «двойной слепой» схеме, то обе группы, экспериментальная группа и контрольная группа, должны испытывать одни и те же экспериментальные эффекты, и тогда разные результаты между двумя группами можно приписать действию препарата.

Проведение экспериментов по «двойной слепой» схеме недостижимо в реальной экономике, поскольку одновременно и участник и организатор эксперимента знают, например, что субъект принимает участие в программе профессиональной подготовки. В плохо сконструированных экспериментах данный экспериментальный эффект может быть значительным. Например, учителя в рамках экспериментальной программы могут прикладывать дополнительные усилия, если считают, что их будущее трудоустройство зависит от результатов эксперимента. Выводы о степени смещения из-за экспериментальных эффектов в полученных результатах эксперимента требуют четкого понимания всех деталей проведения эксперимента.

**Выборки маленьких размеров.** Поскольку проведение экспериментов с участием людей может быть достаточно дорогостоящим, то получаемые в таких экспериментах выборки зачастую не слишком велики по своим размерам. Использование выборок маленьких размеров не приводит к смещению оценок причинных эффектов, но означает, что причинные эффекты оцениваются неточно. Использование выборки небольших размеров также приводит к угрозе достоверности доверительных интервалов и тестовых статистик, которые применяются для тестиования гипотез. Поскольку выводы по результатам эмпирического анализа зачастую основываются на критических значениях для нормального распределения, а устойчивые к наличию гетероскедастичности стандартные ошибки можно использовать только в больших выборках, то экспериментальные данные в малых выборках иногда анализируются в рамках предположений о нормальности распределения ошибок (см. разделы 3.6 и 5.6). Однако предположение о нормальности ошибок достаточно сомнительно как для экспериментальных, так и для реально наблюдаемых данных.

### **Угрозы для внешней обоснованности**

Угрозы внешней обоснованности ставят под вопрос возможность обобщения результатов исследования на другие выборки и эксперименты. Можно выделить два вида подобных угроз, которые возникают, если экспериментальная выборка не является репрезентативной для генеральной совокупности либо условия проводимого эксперимента не являются репрезентативными для более общих случаев.

**Нерепрезентативные выборки.** Изучаемая выборка и потенциально интересующая исследователя выборка могут быть очень похожи, что обосновывает возможность обобщения результатов эксперимента. Если, например, проводится программа по повышению квалификации среди бывших заключенных, результаты подобного эксперимента могут быть обобщены на другие подобные эксперименты (с участием бывших заключенных). Поскольку наличие тюремного заключения довольно сильно влияет на мнение потенциальных работодателей, то результаты подобного эксперимента не могут быть обобщены на обычных людей, которые не имеют судимостей.

Еще один пример нерепрезентативной выборки может возникнуть в случае, если участники эксперимента являются добровольцами. Даже если участники эксперимента случайным образом распределяются между экспериментальной и контрольной группами, добровольцы могут быть более мотивированы, чем другие возможные участники эксперимента (обычные представители генеральной совокупности), из-за чего их усилия будут в большей степени влиять на эксперимент. В общем случае неслучайный характер формирования выборки из генеральной совокупности может приводить к отсутствию возможности обобщения полученных результатов на всю генеральную совокупность.

**Нерепрезентативная программа или политика.** Исходная исследуемая программа или политика должна быть в достаточной степени схожа с той, которая используется в проводимом эксперименте, чтобы результаты эксперимента могли быть обобщены. Еще одним важным аспектом является значительное отличие

осуществляемых в рамках эксперимента действий от того, что используется или будет использоваться в рамках реальной программы. Если реальная программа является широко распространенной, то при ее реализации может не быть того же уровня контроля, что и в рамках эксперимента, либо уровень финансирования в реальной жизни может быть значительно ниже, чем в рамках эксперимента. Каждая из этих ситуаций может приводить к тому, что экспериментальная версия программы будет более эффективной, чем ее оригинальная версия. Еще одно различие между экспериментальной и реальной программой заключается в их различной продолжительности. Экспериментальная программа имеет продолжительность эксперимента, в рамках которого она проводится, в то время как реальная программа может действовать в течение длительных промежутков времени.

**Эффекты общего равновесия.** Аспекты, связанные с размерами и продолжительностью проводимых экспериментов, затрагивают эффекты, которые экономисты часто называют эффектами «общего равновесия». Превращение небольшой временной экспериментальной программы во всеобщую и постоянную программу может привести к изменению общих экономических условий в степени, достаточной для того, чтобы результаты эксперимента не могли быть обобщены. Например, небольшая экспериментальная программа по повышению квалификации сотрудников может стимулировать обучение и рост навыков сотрудников. Однако если подобную программу сделать общедоступной и широко распространенной, то она может вытеснить программы по повышению квалификации, проводимые самими работодателями, что снизит общие выгоды от участия в программе. Аналогично всеобщая реформа образования (например, введение ваучеров на обучение или значительное снижение размеров классов) может увеличивать спрос на услуги преподавателей и изменять внутреннюю структуру группы людей, занимающихся преподавательской деятельностью. Таким образом, общий чистый эффект глобальной реформы будет отражать подобные индуцированные изменения в преподавательском составе. Если переформулировать в эконометрических терминах, то внутренне обоснованный небольшой по размеру эксперимент может позволить корректно измерить причинные эффекты при прочих равных условиях (при неизменных общезэкономических и политических условиях). Однако наличие эффектов общего равновесия говорит о том, что эти «прочие равные условия» зачастую не выполняются в тех случаях, если исследуемая программа в реальности становится широко распространенной.

### 13.3. Эмпирические оценки эффектов уменьшения размера учебного класса

В данном разделе будет рассмотрен вопрос, уже поднимавшийся в части II,— какое влияние оказывает уменьшение размеров классов на результаты тестов. В начале 1980-х годов в Теннесси был проведен масштабный многомиллионный случайный контролируемый эксперимент, призванный ответить на вопрос о том, является ли снижение размера класса эффективным способом улучшения начального образования. Результаты данного эксперимента в значительной степени повлияли на понимание эффектов от снижения размеров классов.

## **Правила эксперимента**

Эксперимент в штате Теннесси по изучению эффектов снижения размеров классов, известный как проект STAR (Student–Teacher Achievement Ratio), представлял собой четырехлетний эксперимент, созданный для изучения влияния малых размеров учебных классов на результаты обучения. Финансирование эксперимента производилось из бюджета штата Теннесси и составило 12 млн долл. США. В рамках эксперимента в школах сравнивались три учебных класса различных размеров: класс обычного размера (22–25 учеников, один преподаватель без дополнительной помощи), малый учебный класс (13–17 учеников, без дополнительной помощи), класс обычного размера (с дополнительной помощью).

В каждой участвующей в эксперименте школе был как минимум один класс каждого типа, а ученики распределялись между ними случайным образом в начале 1985–1986 учебного года. Преподаватели приписывались к одному из классов также случайным образом.

Изначально планировалось, что ученики обучаются в одном и том же классе на протяжении всех четырех лет эксперимента (с «нулевого» по третий класс). Тем не менее по настоянию родителей некоторые ученики, первоначально попавшие в стандартный учебный класс, могли быть случайным образом переведены в другой класс в начале первого класса, а ученики, изначально попавшие в малые учебные классы, могли оставаться только в малых классах. Ученики, поступающие в школу в первый класс («нулевой» класс не является обязательным), на второй год проведения эксперимента также случайным образом распределялись между тремя группами. В течение каждого года ученикам предлагалось проходить стандартизованные тестирования (Stanford Achievement Test) по чтению и математике.

В рамках проекта производилась оплата работы дополнительного числа преподавателей и предоставления сопутствующих услуг, необходимых для достижения требуемых размеров учебных классов. В течение первого года эксперимента приблизительно 6400 учеников проходили обучение в 108 малых учебных классах, 101 стандартном учебном классе и 99 стандартных учебных классах с возможностью дополнительной помощи. За все четыре года эксперимента в его проведении приняли участие порядка 11 600 учеников в 80 школах.

**Отклонения от структуры эксперимента.** Эксперимент предполагал, что ученики не должны иметь возможности перехода в класс другого типа никаким образом, кроме как через проведение повторной процедуры рандомизации в начале первого класса. Тем не менее около 10% учеников сменили свои учебные классы в последующие годы по различным причинам, в том числе из-за несогласованности с другими детьми и проблем с поведением. Эти переходы между классами представляют собой отклонения от процедуры рандомизации и в зависимости от истинной природы подобных переходов могут вносить систематическую ошибку (смещения) в результаты. Переход в другой класс, осуществленный лишь для того, чтобы избежать личностных конфликтов, может быть слабо связан с экспериментом, и поэтому не будет привносить систематическую ошибку. Однако если, например, переход в другой класс происходил из-за того, что родители, переживая за образование своих детей, оказывали давление на школу с целью

перевести ребенка в малый учебный класс, то подобные действия представляют собой несоблюдение правил эксперимента и могут исказить результаты в сторону завышения эффективности малых учебных классов. Еще одним отклонением от правил эксперимента было изменяющееся с течением времени количество учеников в классе, поскольку некоторые ученики переходили из одного класса в другой или переезжали в другие округа с другими школами.

## Анализ данных STAR

Поскольку существуют две экспериментальные группы (малый учебный класс и стандартный учебный класс с дополнительной помощью), то регрессионная версия оценки разностей должна быть модифицирована для случая двух экспериментальных групп и одной контрольной группы. Подобная модификация может быть осуществлена посредством введения двух дополнительных бинарных переменных, одна из которых отражает принадлежность ученика к учебному классу малого размера, а вторая – к стандартному учебному классу с дополнительной помощью, что приводит к новой модели регрессии:

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i, \quad (13.3)$$

где  $Y_i$  – результаты теста, переменная  $SmallClass_i$ , если  $i$ -й ученик приписан к малому учебному классу, и равна нулю в противном случае, а переменная  $RegAid_i = 1$ , если  $i$ -й ученик приписан к стандартному учебному классу с дополнительной помощью, и нулю – в противном случае. Влияние малого размера учебного класса по сравнению с учебным классом стандартного размера равно  $\beta_1$ , а влияние стандартного учебного класса с дополнительной помощью по сравнению с классом стандартного размера равно  $\beta_2$ . Подобные оценки разностей могут быть получены, оценивая  $\beta_1$  и  $\beta_2$  в уравнении (13.3) с помощью МНК.

Таблица 13.1

Проект STAR: оценки влияния размеров классов на успеваемость

Регрессор	Год обучения			
	0	1	2	3
Малый учебный класс	13,90** (2,45)	29,78** (2,83)	19,39** (2,71)	15,59** (2,40)
Стандартный учебный класс с помощью	0,31 (2,27)	11,96** (2,65)	3,48 (2,54)	– 0,29 (2,27)
Свободный член	917,04** (1,63)	1039,39** (1,78)	1157,81** (1,82)	1228,51** (1,68)
Количество наблюдений	5786	6379	6049	5967

Примечание. Регрессии были оценены на основе открытой базы данных проекта STAR (Project STAR Public Access Data Set), описанной в приложении 13.1. Зависимая переменная представляет собой суммарное количество баллов по тестам по математике и чтению в рамках Stanford Achievement Test. Стандартные ошибки представлены в круглых скобках под оценками коэффициентов. \*\* – оценка коэффициента значима на уровне значимости 1% (двухсторонняя статистика).

В таблице 13.1 представлены результаты оценок влияния размера учебного класса на результаты пройденного тестирования. Указанная в таблице 13.1

зависимая переменная  $Y_1$  представляет собой суммарное количество баллов по тестам по математике и чтению в рамках *Stanford Achievement Test*. Оценки, представленные в таблице, показывают, что для учеников нулевого класса пребывание в малых классах повышает успеваемость на 13,9 пункта по сравнению с результатами в стандартных классах, а пребывание в стандартном учебном классе с возможностью помощи повышает успеваемость на 0,31 пункта по сравнению с результатами в стандартных классах. Для каждого года обучения нулевая гипотеза о том, что малые учебные классы не дают каких-либо преимуществ в обучении, отвергается на уровне значимости 1% (двуспиральный тест). Тем не менее нельзя отвергнуть нулевую гипотезу о том, что наличие возможности дополнительной помощи в стандартном классе не приводит к повышению уровня успеваемости для всех годов обучения (кроме первого класса). Величина полученных оценок для малых классов приблизительно одинакова для «нулевого», второго и третьего годов обучения, а оценка для первого класса несколько больше.

Представленные в таблице 13.1 оценки разностей показывают, что снижение размера класса влияет на уровень успеваемости, но появление возможности дополнительной помощи в стандартном классе имеет гораздо меньший эффект, который практически равен нулю. Как отмечалось в разделе 13.1, включение в регрессию в таблице 13.1 дополнительных регрессоров ( $W$ -регрессоры в уравнении (13.2)) может позволить получить более эффективные оценки причинно-следственных эффектов. Кроме того, если попадание в экспериментальную группу происходит неслучайно из-за отклонений от правил проведения эксперимента, то оценки экспериментальных эффектов, полученные с помощью регрессий с дополнительными регрессорами, могут отличаться от оценок разностей, представленных в таблице 13.1. По указанным причинам оценки экспериментальных эффектов, в рамках которых в уравнение (13.3) включены дополнительные регрессоры, приведены для «нулевого» года обучения в таблице 13.2. Результаты, представленные в первом столбце таблицы 13.2, повторяют результаты первого столбца (для «нулевого» класса) из таблицы 13.1, а в остальных трех столбцах представлены результаты с учетом дополнительных регрессоров, которые призваны учесть характеристики преподавателей, школы и учеников.

Основной вывод, который можно получить из таблицы 13.2, заключается в том, что оценки причинно-следственных эффектов, полученные с помощью множественной регрессии в случае двух экспериментальных групп и представленные в трех последних столбцах таблицы 13.2, довольно сильно похожи на оценки разностей, представленными в первом столбце той же таблицы. То что включение дополнительных наблюдаемых регрессоров не приводит к изменению оценок причинно-следственных эффектов, скорее всего, говорит о том, что случайный характер формирования малых учебных классов также не зависит и от ненаблюдаемых переменных. Как и ожидалось, включение этих регрессоров повышает  $R^2$  регрессии, а стандартные ошибки полученных оценок снижаются с 2,45 (в столбце (1)) до 2,16 (в столбце (4)).

Поскольку преподаватели из одной школы распределяются между классами различных типов случайным образом, то эксперимент также предоставляет воз-

можность оценить влияние уровня преподавания (или преподавательского стажа) на успеваемость. В рамках терминологии из раздела 13.1, случайный выбор проводится условно относительно наблюдаемых переменных  $W$ , где  $W$  означает полный набор бинарных переменных, определяющих школу, то есть  $W$  представляет собой полный набор фиксированных эффектов для школы. Таким образом, относительно  $W$  преподаватели с различным стажем распределяются случайным образом, что означает, что  $u_i$  в уравнении (13.2) удовлетворяет предположению об условной независимости среднего, где в качестве переменных  $X$  рассматриваются размер класса, преподавательский стаж, а  $W$  – полный набор фиксированных эффектов, характеризующих школу. Поскольку между школами преподаватели распределяются неслучайным образом, то при отсутствии фиксированных эффектов, характеризующих школу, в оцениваемой регрессии (таблица 13.2, столбец (2)) преподавательский стаж будет в общем случае коррелирован с остаточным членом. Например, в более богатых административных округах могут работать преподаватели с большим стажем. При включении фиксированных эффектов оценка коэффициента при переменной, отражающей преподавательский стаж, снижается почти вдвое, с 1,47 в столбце (2) из таблицы 13.2 до 0,74 в столбце (3). Поскольку преподаватели из одной школы распределяются между классами случайным образом, то в столбце (3) представлены несмешенные оценки влияния дополнительного года опыта на уровень успеваемости. Оценка, равная 0,74, является статистически значимой и относительно большой – десятилетний преподавательский стаж соответствует увеличению уровня успеваемости в тестах на 7,4 пункта.

Велик соблазн попытаться проинтерпретировать некоторые другие коэффициенты в таблице 13.2, но, как коэффициенты при контрольных переменных, они, как правило, не имеют причинной интерпретации. Например, в «нулевом» классе мальчики зачастую имеют более низкие результаты тестов, чем девочки. Однако эти индивидуальные особенности учеников не являются случайными (т.е. пол ученика, который пишет тест, детерминирован, а не задан случайным образом), поэтому дополнительные регрессоры могут быть коррелированы с пропущенными переменными. Аналогичным образом, если расовая принадлежность или наличие права на получение бесплатного обеда коррелирует со снижением возможностей для обучения за пределами школы (которые включены в таблицу 13.2), то оценки их коэффициентов будут отражать эти пропущенные взаимосвязи.

Таблица 13.2

**Проект STAR: разностные оценки с использованием дополнительных регрессоров  
для «нулевого» класса**

Регрессор	(1)	(2)	(3)	(4)
Малый учебный класс	13,90** (2,45)	14,00** (2,45)	15,93** (2,24)	15,89** (2,16)
Стандартный класс с возможностью дополнительной помощи	0,31 (2,27)	-0,60 (2,25)	1,22 (2,04)	1,79 (1,96)
Преподавательский стаж		1,47** (0,17)	0,74** (0,17)	0,66** (0,17)
Бинарная переменная, характеризующая пол ученика (1 – мужской)				-12,09** (1,67)

Окончание таблицы 13.2

Регрессор	(1)	(2)	(3)	(4)
Бинарная переменная, характеризующая право на получение бесплатного обеда				-34,70** (1,99)
Бинарная переменная, характеризующая расовую принадлежность (1 – афроамериканец)				-25,43** (3,50)
Бинарная переменная, характеризующая иную расовую принадлежность (1 – не белый и не черный)				-8,50 (12,52)
Свободный член	918,04** (1,63)	904,72** (2,22)		
Фиксированные эффекты на школу	Нет	Нет	Да	Да
$R^2$	0,01	0,02	0,22	0,28
Количество наблюдений	5786	5766	5766	5748

*Примечание.* Регрессии были оценены на основе открытой базы данных проекта STAR (Project STAR Public Access Data Set), описанной в приложении 13.1. Зависимая переменная представляет собой суммарное количество баллов по тестам по математике и чтению в рамках Stanford Achievement Test. Количество наблюдений для разных регрессий отличается в силу ряда пропущенных значений для отдельных переменных. Стандартные ошибки представлены в круглых скобках под оценками коэффициентов. \*\* – оценка коэффициента значима на уровне значимости 1% (двухсторонняя статистика), \* – оценка коэффициента значима на уровне значимости 5%.

### **Интерпретация оценок влияния размера класса на результаты тестов.**

Какова величина полученных оценок влияния размера класса на результаты обучения, представленных в таблицах 13.1 и 13.2: является ли она большой или маленькой в практическом смысле? Существует два ответа на данный вопрос. Во-первых, можно перевести оценки изменений из единицы шкалы тестов в единицы стандартных отклонений тестов, что позволит сравнивать результаты, представленные в разных столбцах таблицы 13.1 (т.е. для разных годов обучения). Во-вторых, имеет смысл сравнивать оценки влияния размера класса на результаты тестов с оценками других коэффициентов в таблице 13.2.

Поскольку распределение результатов тестов неодинаково для различных годов обучения, то получаемые оценки, представленные в таблице 13.1, не являются в полной мере сопоставимыми для различных лет обучения. Подобная проблема уже возникла в разделе 9.4 при сравнении оценок влияния снижения соотношения учеников и учителей на уровень успеваемости, полученных на основе данных для Калифорнии и для штата Массачусетс. Поскольку тесты различались между собой, то и коэффициенты могли быть несопоставимы напрямую. Решение проблемы из раздела 9.4 заключалось в переводе полученных оценок в единицы стандартных отклонений, и это позволило показать, что снижение соотношения учеников и учителей соответствует снижению оценки доли стандартного отклонения результатов тестов. В текущем разделе также используется подобный подход для сравнения результатов, представленных в таблице 13.1, для различных годов обучения. Например, стандартное отклонение результатов тестов (баллов за тест) для учеников «нулевого» года обучения равно 73,7, то есть влияние обучения в малом учебном классе, основанное на оценках в таблице 13.1, равно  $\frac{13,9}{73,7} = 0,19$  со стандартной ошибкой

$\frac{2,45}{73,7} = 0,03$ . Оценки эффекта влияния размера класса на результаты тестов, полученные в единицах стандартных отклонений, представлены в таблице 13.3.

Выраженные в единицах стандартных отклонений, оценки влияния малого размера класса довольно схожи для «нулевого», второго и третьего классов и составляют приблизительно одну пятую от стандартного отклонения результатов тестов. Аналогично результаты оценок для стандартного учебного класса с возможностью дополнительной помощи практически равны нулю для «нулевого», второго и третьего классов. Полученные оценки экспериментальных эффектов больше по величине для первого класса. Тем не менее полученная оценка разности между малым учебным классом и стандартным классом с возможностью дополнительной помощи равна 0,20 для первого и всех остальных классов. Таким образом, одна из возможных интерпретаций результатов для учеников первого класса в контрольных группах (стандартных классах без возможности дополнительной помощи) может заключаться в том, что плохие результаты тестов получены по какой-то нестандартной причине, возможно, из-за особенностей выборки.

Таблица 13.3

**Оценки влияния размера учебного класса на успеваемость учеников  
в единицах стандартных отклонений**

Экспериментальная группа	Класс (год обучения)			
	«Нулевой»	1	2	3
Малый учебный класс	0,19** (0,03)	0,33** (0,03)	0,23** (0,03)	0,21** (0,03)
Стандартный учебный класс с возможностью дополнительной помощи	0,00 (0,03)	0,13** (0,03)	0,04 (0,03)	0,00 (0,03)
Выборочная стандартная ошибка ( $s_y$ )	73,70	91,30	84,10	73,30

Примечание. Оценки и стандартные ошибки в первых двух строках представляют собой оценки эффектов из таблицы 13.1, разделенные на выборочные стандартные ошибки теста *Stanford Achievement Test* для соответствующего года обучения (последняя строка таблицы), вычисленные на основе данных по ученикам в рамках эксперимента. Стандартные ошибки представлены в круглых скобках под оценками коэффициентов. \*\* – оценка коэффициента является статистически значимой на уровне значимости 1% (двухсторонний тест).

Другим способом оценки величины предполагаемого эффекта влияния малого размера класса является сравнение оценок причинных эффектов с другими коэффициентами в таблице 13.2. Для «нулевого» года обучения оценка влияния малого размера класса составляет 13,9 тестовых баллов (первая строка в таблице 13.2). При заданных значениях расовой принадлежности, преподавательского стажа, возможности получения бесплатных обедов и при заданной экспериментальной группе успеваемость мальчиков по стандартизованным тестам ниже, чем у девочек, приблизительно на 12 пунктов согласно оценкам из столбца (4) таблицы 13.2. Таким образом, оценка влияния попадания в малый учебный класс на результаты обучения в некоторой степени больше, чем разность результатов тестов для мальчиков и девочек. Следует также рассмотреть полученные оценки коэффициента при переменной, отражающей преподавательский стаж учителя. Для модели из столбца (4) эта оценка равна 0,66. То есть результаты тестов у учеников

преподавателя с 20-летним стажем увеличиваются на 13 пунктов. Таким образом, оценка эффекта влияния малого размера класса по величине соответствует эффекту преподавания предметов учителем с 20-летним стажем (по сравнению с учителем без опыта работы). Подобные сравнения показывают, что влияние малого размера класса на успеваемость довольно существенно.

**Дополнительные результаты.** Эконометристы, статистики и специалисты в сфере начального образования изучили результаты рассмотренного эксперимента. Ниже будут кратко описаны основные из сделанных ими выводов. Один из них заключается в том, что эффект «малого размера класса» в большей степени характерен для первых годов обучения, что можно видеть из результатов, представленных в таблице 13.3. За исключением аномальных результатов для первого класса, различие в результатах тестов для обычных и малых учебных классов, представленное в таблице 13.3, имеет практически постоянную величину для всех годов обучения (0,19 единиц стандартного отклонения для «нулевого» учебного года, 0,23 – для второго года и 0,21 – для третьего года обучения). Поскольку дети, которые изначально попадали в малые классы, впоследствии оставались в малых учебных классах, то обучение в рамках малого класса не давало никакого дополнительного улучшения результатов тестирования. С другой стороны, преимущество, получаемое на начальной стадии обучения, сохранялось и в последующие годы, но различия в успеваемости между контрольной и экспериментальной группами не увеличивались. Еще один вывод состоит в том, что, как указано во второй строке таблицы 13.3, эксперимент показал незначительное улучшение успеваемости в стандартных классах с возможностью дополнительной помощи. Одна из возможных интерпретаций этого результата состоит в возможных отклонениях от условий эксперимента для некоторых учеников (некоторые ученики переходили из малых учебных классов в другие). Если исходное распределение в «нулевом» году обучения происходит случайным образом и не оказывает прямого влияния на результаты обучения, то исходное распределение может быть использовано как инструментальная переменная, которая частично, но не полностью, влияет на распределение по классам для более старших классов. Этой стратегии следовал Крюгер в своей работе (Krueger, 1999), в рамках которой использовался двухшаговый МНК (2МНК) для оценки влияния размера класса на результаты тестов с помощью инструментальной переменной распределения по классам. Автор пришел к выводу, что отклонения от условий эксперимента не приводили к существенным смещениям в МНК-оценке<sup>1</sup>.

### **Сравнение экспериментальных оценок эффекта «размера учебного класса»**

Во второй части книги были представлены полученные с помощью модели множественной регрессии оценки влияния размера классов, основанные

<sup>1</sup> Более подробная информация о проекте STAR представлена в следующих работах: Mosteller (1995), Mosteller, Light, Sachs (1996), Krueger (1999). В работе Эренберга, Брюэра, Гаморана и Уиллмса (Ehrenberg, Brewer, Gamoran, and Willms, 2001a; 2001b) проект STAR рассматривается в контексте обсуждения политики относительно размеров классов и сопутствующих исследований по данной тематике. Критика проекта STAR представлена в работе Ханушека (Hanushek, 1999a); в работе Ханушека (Hanushek, 1999b) критика взаимосвязи между размером классов и результатами обучения представлена в более общем виде.

на данных по школьным округам Калифорнии и штата Массачусетс. В этих данных размер класса не устанавливался случайным образом, а определялся руководством школьного учреждения для достижения сбалансированности между образовательными целями и бюджетными возможностями. Каким образом можно провести соответствие между полученными на основе подобных данных оценками и оценками, полученными в рамках проекта STAR?

Для сравнения оценок, полученных по данным Калифорнии и штата Массачусетс, с оценками, представленными в таблице 13.3, необходимо оценить влияние одинакового снижения размера класса и выразить оцениваемый эффект в единицах стандартных отклонений результатов тестов. В течение четырех лет проведения проекта STAR в малых классах в среднем училось на 7,5 учеников меньше, чем в больших классах. Тогда можно использовать полученные оценки для предсказания эффекта влияния снижения числа учеников в классе на 7,5 учеников на результаты обучения. Основываясь на МНК-оценках для линейных спецификаций, представленных в первом столбце таблицы 9.3, оценки по данным Калифорнии показывают увеличение результатов тестов на 5,5 тестовых пунктов при снижении числа учеников в классе на 7,5 и соответствующего изменения соотношения числа учеников и учителей ( $(0,73 \times 7,5 \cong 5,5$  пунктов). Стандартная ошибка результатов тестов в Калифорнии составляет приблизительно 38 пунктов, таким образом, оцененный эффект влияния уменьшения числа учеников в классе на 7,5, измеренный в единицах стандартного отклонения, равен  $\frac{5,5}{38} \cong 0,14$ . Стандартная ошибка углового коэффициента для данных по Калифорнии равна 0,26 (см. табл. 9.3), поэтому стандартная ошибка эффекта снижения числа учеников в классе на 7,5 в единицах стандартного отклонения составляет  $0,26 \times \frac{7,5}{38} \cong 0,05$ . Таким образом, основанная на данных по Калифорнии оценка влияния уменьшения числа учеников в классе на 7,5 человек, выраженная в единицах стандартного отклонения результатов тестов, равна 0,14 стандартных отклонений со стандартной ошибкой 0,05. Эти вычисления и аналогичные вычисления для штата Массачусетс представлены в таблице 13.4 вместе с оценками для «нулевого» года обучения в рамках проекта STAR, которые также были представлены в столбце (1) таблицы 13.2.

Оценки эффектов, полученные на основе данных по Калифорнии и штату Массачусетс, в некоторой степени меньше, чем оценки на основе данных проекта STAR. Одна из возможных причин, по которым эти оценки могут различаться, заключается в изменчивости случайной выборки, поэтому имеет смысл сравнивать доверительные интервалы для полученных оценок. 95%-й доверительный интервал для оценки эффекта влияния обучения в малом классе для «нулевого» года обучения на основе данных проекта STAR (см. последний столбец табл. 13.4) составляет от 0,13 до 0,25. Соответствующий 95%-й доверительный интервал, полученный на основе данных по Калифорнии, составляет от 0,04 до 0,24, а по штату Массачусетс – от 0,02 до 0,22. Таким образом, 95%-е доверительные интервалы, полученные для данных по Калифорнии и штату Массачусетс, содержат в себе значительную часть доверительного интервала,

полученного для оценок по данным проекта STAR. Если рассматривать полученные оценки в подобном ключе, то все они дают очень похожие интервалы для оценок изучаемых эффектов.

Существует множество причин, по которым полученные на разных выборках оценки могут различаться. Одна из них, как обсуждалось в разделе 9.4, следует из имеющихся угроз для внутренней обоснованности исследования. Например, поскольку ученики могут переезжать из одного школьного округа в другой, то соотношение числа учеников и учителей в школьных округах может не отражать реального соотношения числа учеников и учителей, в условиях которого на самом деле обучаются ученики. Поэтому оценки, полученные на основе данных по Калифорнии и штату Массачусетс, могут быть смещены в сторону нуля из-за ошибок измерения переменных. Другие причины связаны с вопросами внешней обоснованности. Среднее по школьному округу соотношение числа учеников и учителей, которое используется для оценок, не в полной мере соответствует переменной, отражающей число учеников в классе в проекте STAR. Стоит также отметить, что проект STAR проводился в 1980-х годах в южном штате, а данные по Калифорнии и штату Массачусетс получены для 1998 года; кроме того, сравниваемые годы обучения также различаются (с «нулевого» по третий класс в проекте STAR, четвертый класс в штате Массачусетс, пятый класс для Калифорнии). В свете всего сказанного можно ожидать существенные различия в оценках эффектов, которые, тем не менее, получились очень похожими. Это может говорить о том, что упомянутые проблемы с внутренней обоснованностью не носят критического характера в рассматриваемых случаях.

Таблица 13.4

**Оценки эффекта снижения соотношения числа учеников и учителей на 7,5 учеников, основанные на данных проекта STAR и данных по Калифорнии и штату Массачусетс**

Данные	$\hat{\beta}_1$	Изменение соотношения учеников и учителей	Стандартное отклонение результатов тестов	Оценка эффекта	95-%-й доверительный интервал
Проект STAR («нулевой» год обучения)	-13,90** (2,45)	Малый учебный класс и стандартный класс	73,8	0,19** (0,03)	(0,13; 0,25)
Калифорния	-0,73** (0,26)	-7,5	38,0	0,14** (0,05)	(0,04; 0,24)
Штат Массачусетс	-0,64* (0,27)	-7,5	39,0	0,12* (0,05)	(0,02; 0,22)

*Примечание.* Оценка коэффициента  $\hat{\beta}_1$  для проекта STAR взята из столбца (1) таблицы 13.2. Оценки коэффициентов по Калифорнии и штату Массачусетс взяты из первого столбца таблицы 9.3. Оценка эффекта представляет собой либо оценку влияния обучения в малом классе по сравнению со стандартным учебным классом (проект STAR), либо оценку эффекта снижения соотношения учеников и учителей на 7,5 (для данных по Калифорнии и штату Массачусетс). 95 %-й доверительный интервал для снижения соотношения учеников и учителей представляет собой оценку эффекта  $\pm 1,96$  стандартной ошибки. Стандартные ошибки представлены в круглых скобках под оценками эффектов. Оценки эффектов являются статистически значимыми на \*5 %-м уровне значимости или \*\*1 %-м уровне значимости (двухсторонний тест).

## 13.4. Квазиэксперименты

Статистические методы и подходы, используемые в рамках случайных контролируемых экспериментов, могут применяться не только в рамках экспериментов. В *квазиэкспериментах* или так называемых *естественных экспериментах* случайность привносится посредством изменения специфических условий, что приводит к тому, что квазиэксперимент проводится *как бы* случайным образом. Такие изменения в специфических обстоятельствах могут возникнуть из-за изменений институциональных условий, местоположения, сроков реализации определенной политики или программы, естественных случайных событий, например, снегопадов, дождей или других факторов, которые не связаны с причинно-следственными эффектами в изучаемых взаимосвязях.

Существует два типа квазиэкспериментов. В рамках первого типа принадлежность индивида (в общем случае – объекта) к экспериментальной группе рассматривается как случайная. В этом случае причинные эффекты могут быть оценены с помощью МНК с использованием переменной  $X_i$ , характеризующей участие в эксперименте, в качестве регрессора. В рамках второго типа квазиэкспериментов «*как бы*» случайное изменение лишь частично определяет участие в эксперименте. В этом случае причинные эффекты оцениваются с помощью регрессий с инструментальными переменными, где «*как бы*» случайный источник изменения также дает инструментальную переменную.

После анализа ряда примеров в настоящем разделе мы расширяем некоторые эконометрические методы, рассмотренные в разделах 13.1 и 13.2, что может быть полезно для анализа данных из квазиэкспериментов.

### Примеры

Проиллюстрируем два типа квазиэкспериментов с помощью примеров. Первый пример представляет собой квазиэксперимент, в котором параметр участия в эксперименте «*как бы*» случайно определен. Второй и третий примеры иллюстрируют квазиэксперименты, в которых «*как бы*» случайные события оказывают влияние, но не полностью определяют параметр участия в эксперименте.

**Пример № 1: влияние иммиграции на рынок труда.** Снижает ли иммиграция уровень заработных плат? В экономической теории предполагается, что при росте предложения труда за счет притока иммигрантов стоимость рабочей силы – заработка плата – должна снижаться. Тем не менее при прочих равных условиях иммигрантов зачастую привлекают города с высоким уровнем спроса на труд, то есть МНК-оценки влияния иммиграции на уровень заработной платы будут смещены. Идеальный случайный контролируемый эксперимент по изучению влияния иммиграции на заработные платы должен проводиться путем случайного распределения различного числа иммигрантов (различные значения параметра участия в эксперименте) по разным рынкам труда («субъекты») и должен измерять последующее влияние на заработные платы («исход»). Проведение подобного эксперимента, тем не менее, довольно сложно из-за многочисленных практических, финансовых и этических проблем.

Один из исследователей рынка труда, Дэвид Кард, в работе (Card, 1990) использовал квазиэксперимент, в рамках которого большое количество кубинских иммигрантов вышло на рынок труда Майами (штат Флорида) в результате массовой эмиграции с Кубы («Mariel boatlift») в 1980 году, которая стала возможной в связи с временным ослаблением ограничений по эмиграции для граждан Кубы. Половина иммигрантов проживала в Майами, поскольку там уже существовала довольно крупная диаспора кубинцев. Кард оценил влияние последствий иммиграции на уровень заработной платы, сравнив изменения в уровне зарплаты для низкоквалифицированных рабочих в Майами и рабочих такого же уровня квалификации в других городах США за соответствующий период. Было установлено, что иммиграция оказывала пренебрежимо малое влияние на уровень заработной платы низкоквалифицированных рабочих.

**Пример № 2: влияние службы в армии на уровень заработной платы гражданских лиц.** Оказывает ли влияние прохождение военной службы на перспективы на рынке труда? Военные проводят различного рода тренинги, в рамках которых утверждается, что будущие работодатели считают прошедших службу в армии работников более привлекательными кандидатами. Тем не менее МНК-оценки влияния факта службы в армии на доходы индивидов могут быть смещены, поскольку прохождение военной службы, по крайней мере частично, предопределется индивидуальными характеристиками и выбором индивида. Например, на военную службу в США принимают лишь тех, кто удовлетворяет минимальным требованиям к физической подготовке, а неудачи на рынке труда могут привести к тому, что человек с большей вероятностью пойдет служить в армию.

В попытке отследить подобные смещения Джошуа Энгрист в работе (Angrist, 1990) использовал квазиэксперимент, в рамках которого он изучил текущее положение и историю рынка труда для тех индивидов, которые прошли службу в армии США во время войны во Вьетнаме. В течение этого периода возможность попадания в армию определялась частично путем национальной лотереи на основании дат рождения – мужчины, датам рождения которых случайным образом приписывались низкие лотерейные номера, призывались в армию, а те, кому приписывались большие номера – не призывались. Кроме того, реальная возможность призыва на военную службу регламентировалась сложным набором правил, который включал в себя требования к физическим характеристикам и некоторые исключения. Некоторые молодые люди хотели пройти военную службу добровольно, поэтому прохождение службы лишь частично определялось тем, призывался ли индивид на службу по результатам лотереи. Поэтому призыв на службу по результатам лотереи может использоваться в качестве инструментальной переменной, которая отчасти определяет прохождение военной службы, но при этом определяется случайным образом. В рассматриваемом случае возможность прохождения военной службы действительно определялась случайным образом при помощи лотереи, но поскольку этот процесс случайного выбора проходил не в рамках эксперимента по оценке эффекта влияния прохождения военной службы, то это был квазиэксперимент. Энгрист показал, что долгосрочный эффект прохождения военной службы снижал доходы белых ветеранов, но не снижал доходы всех остальных ветеранов.

**Пример № 3: эффект процедуры катетеризации сердца.** В разделе 12.5 описана работа Макклеллана, Макнейла и Ньюхаяса (McClellan, McNeil, Newhouse, 1994), в которой в качестве инструментальной переменной для реального проведения процедуры катетеризации сердца использовалась разность между расстояниями от дома пациента до ближайшей больницы, где проводится процедура катетеризации сердца, и от дома пациента до ближайшей больницы. Эта работа представляет собой квазиэксперимент, в котором используется переменная, частично определяющая параметр участия в эксперименте. Непосредственное проведение процедуры катетеризации сердца определяется личными характеристиками пациента, а также решением о проведении операции, которое принимается совместно пациентом и лечащим врачом. Тем не менее на решение о проведении процедуры может влиять также наличие возможности ее проведения в близлежащих медицинских учреждениях. Если местоположение пациента «как бы» случайным образом приписывается к больнице и не имеет какого-либо дополнительного влияния на состояние здоровья жителей (кроме как через возможность проведения процедуры катетеризации сердца), то относительное расстояние до больницы, где может быть проведена процедура катетеризации сердца, представляет собой допустимую инструментальную переменную.

**Другие примеры.** Методика использования квазиэкспериментов в исследованиях применялась также и в других областях. В работе Гарви и Ханка (Garvey, Hanka, 1999) использовались изменения законов США для оценки эффекта влияния законов, направленных против поглощения компаний, на корпоративную финансовую структуру компаний (например объем долга). В работе Мейера, Вискузи и Дарбина (Meyer, Viscusi, Durbin, 1995) использовалось изменение величины страхового возмещения по безработице в штатах Кентукки и Мичиган, которое по-разному влияло на работников с высокими (но не с низкими) заработными платами, при оценке влияния подобных изменений на продолжительность поиска и смены работы. Работы Мейера (Meyer, 1995), Розенцвейга и Уолпина (Rosenzweig, Wolpin, 2000) и Энгриста и Крюгера (Angrist, Krueger, 2001) также представляют собой примеры использования квазиэкспериментов в экономике и социальной политике.

### Оценка «разности разностей»

Если в рамках квазиэксперимента включение в экспериментальную группу происходит «как бы» случайно, условно относительно наблюдаемых переменных  $W$ , то причинное влияние может быть оценено с помощью модели регрессии (13.2). Поскольку исследователь не может каким-либо образом в рамках квазиэксперимента влиять на процесс случайного выбора, то между экспериментальной и контрольной группами могут оставаться некоторые отличия даже после учета  $W$ . Одним из способов учесть упомянутые различия между группами является сравнение изменений исходов до влияния рассматриваемого эффекта и после его влияния (а не сравнение конечных исходов  $Y$ ), то есть с помощью использования разностей значений  $Y$  до начала влияния рассматриваемого

эффекта в двух группах. Поскольку подобная оценка представляет собой различие между изменениями, происходящими в группах во времени, то ее называют оценкой «разности разностей». Например, в работе Карда (Card, 1990), в которой изучался эффект влияния миграции на заработную плату низкоквалифицированных рабочих, оценки «разности разностей» использовались для сравнения изменений заработных плат в Майами с изменениями заработных плат в других городах США. Еще один пример применения оценок «разности разностей» представлен во вставке «Чему равно влияние занятости на минимальный уровень заработных плат?».

**Оценка «разности разностей».** Пусть  $\bar{Y}^{treatment, before}$  – выборочное среднее для исходов  $Y$  в экспериментальной группе до проведения эксперимента, пусть  $\bar{Y}^{treatment, after}$  – выборочное среднее для исходов в экспериментальной группе после проведения эксперимента. Пусть  $\bar{Y}^{control, before}$  – соответствующее выборочное среднее для исходов в контрольной группе до проведения эксперимента, а  $\bar{Y}^{control, after}$  – выборочное среднее для исходов в контрольной группе после проведения эксперимента. Среднее изменение  $Y$  при проведении эксперимента для экспериментальной группы равно  $\bar{Y}^{treatment, after} - \bar{Y}^{treatment, before}$ , а среднее изменение  $Y$  в течение этого периода для контрольной группы равно  $\bar{Y}^{control, after} - \bar{Y}^{control, before}$ . Оценка «разности разностей» равна разности среднего изменения  $Y$  для экспериментальной группы и среднего изменения  $Y$  для контрольной группы:

$$\hat{\beta}_1^{diffs-in-diffs} = (\bar{Y}^{treatment, after} - \bar{Y}^{treatment, before}) - (\bar{Y}^{control, after} - \bar{Y}^{control, before}), \quad (13.4)$$

где  $\Delta\bar{Y}^{treatment}$  – среднее изменение  $Y$  в экспериментальной группе,  $\Delta\bar{Y}^{control}$  – среднее изменение  $Y$  в контрольной группе. Если распределение между контрольной и экспериментальной группами случайно, то  $\hat{\beta}_1^{diffs-in-diffs}$  является несмещенной и состоятельной оценкой исследуемого причинного эффекта.



### **Чему равно влияние занятости на минимальный уровень заработных плат?**

Как сильно рост минимальной заработной платы снижает спрос на труд низкоквалифицированных работников? Экономическая теория говорит о том, что спрос снижается с ростом цены, но насколько сильно – это отдельный вопрос. Поскольку равновесные значения цен и физических объемов определяются спросом и предложением, то МНК-оценки в регрессии уровня безработицы на уровень заработных плат будут иметь смещения, вызванные проблемой эндогенности (см. вставку «Основные понятия 9.6»). Теоретически в рамках случайногоконтролируемого эксперимента можно было бы случайным образом приспать работодателям различные уровни заработных плат, а затем сравнить получившиеся изменения в уровнях занятости (исходы) для экспериментальной и контрольной групп, но возможно ли провести такой случайный эксперимент в реальной жизни?

Экономисты Дэвид Кард и Аллан Крюгер в своей работе (Card, Krueger, 1994) решили провести подобный эксперимент, но предпочли «предоставить природе» –

или, более точно, использовали географический признак – возможность провести процедуру случайного выбора. В 1992 году минимальная заработка плата в Нью-Джерси выросла с 4,25 долл./час до 5,05 долл./час, но минимальная заработка плата в соседней Пенсильвании не изменилась. В этом эксперименте влияние роста минимальной заработной платы, которое происходило в Нью-Джерси вместо Пенсильвании, рассматривается в качестве «как бы» случайного в том смысле, что подобный скачок заработной платы не коррелирует с другими детерминантами заработной платы в рассматриваемый период. Кард и Крюгер собрали базу данных по уровням заработной платы в ресторанах быстрого питания до и после роста заработной платы в двух вышеуказанных штатах США. После того как они вычислили оценку «разности разностей», полученный результат представился довольно интересным: полученные оценки показывали отсутствие роста заработных плат в ресторанах быстрого питания в Нью-Джерси по сравнению с Пенсильванией. Более того, некоторые из полученных оценок показывали, что после роста минимальной заработной платы уровень занятости в ресторанах быстрого питания в Нью-Джерси *вырос* по сравнению с Пенсильванией.

Этот результат противоречит базовой микроэкономической теории и является довольно спорным. Последующий анализ, проведенный с использованием иных источников данных о занятости, показал, что, возможно, в Нью-Джерси после повышения заработной платы могло случиться небольшое падение занятости, но даже в этом случае полученные оценки кривой спроса на труд очень неэластичны (см. Neumark, Wascher, 2000). Несмотря на то что точная эластичность заработной платы в этом квазиэксперименте остается предметом дискуссий, влияние роста минимальной заработной платы на занятость меньше, чем многие экономисты считали ранее.

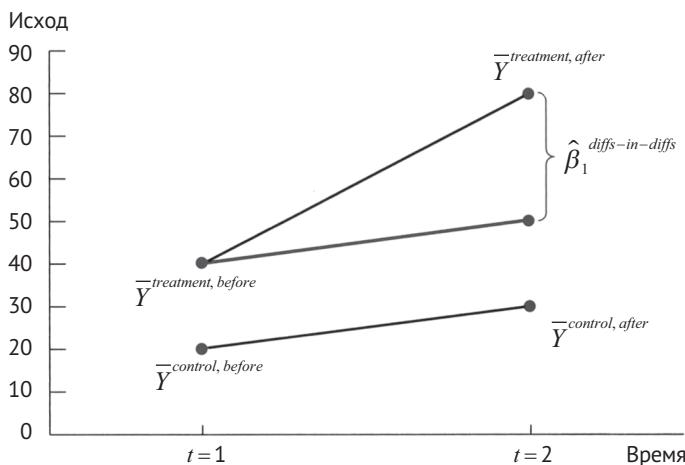


Оценка «разности разностей» может быть записана с использованием обозначений стандартной модели регрессии. Пусть  $\Delta Y_i$  равно разности значений  $Y$  после и до проведения эксперимента. Оценка «разности разностей» может быть получена в виде МНК-оценки  $\beta_1$  в уравнении регрессии:

$$\Delta Y_i = \beta_0 + \beta_1 X_i + u_i. \quad (13.5)$$

Оценка «разности разностей» представлена на рисунке 13.1. Можно видеть, что выборочное среднее  $Y$  для экспериментальной группы до проведения эксперимента равно 40, а аналогичное среднее значение  $Y$  для контрольной группы равно 20. Во время проведения эксперимента выборочное среднее  $Y$  для контрольной группы повышается до 30, а для экспериментальной группы – до 80. Таким образом, средняя разность исходов между экспериментальной и контрольной группами после проведения эксперимента равна  $80 - 30 = 50$ . Стоит отметить, что эта разность возникает из-за того, что экспериментальная и контрольная группы имели разные средние значения – значения для экспериментальной группы изначально были выше, чем для контрольной. Оценка

«разности разностей» отражает прирост рассматриваемого показателя для экспериментальной группы по сравнению с контрольной группой, который в приведенном примере равен:  $(80 - 40) - (30 - 20) = 30$ . Делая акцент на изменении  $Y$  при проведении эксперимента, оценка «разности разностей» устраниет влияние исходных значений  $Y$ , которые отличаются для экспериментальной и контрольной групп.



**Рисунок 13.1. Оценка «разности разностей»**

На представленном графике разность между экспериментальной и контрольной группами после проведения эксперимента равна:  $80 - 30 = 50$ , но она переоценивает эффект проведения эксперимента, поскольку перед экспериментом  $\bar{Y}$  была выше для экспериментальной группы по сравнению с контрольной группой на  $40 - 20 = 20$ . Оценка «разности разностей» равна:  $\hat{\beta}_1^{diffs-in-diffs} = (80 - 30) - (40 - 20) = 50 - 20 = 30$ . Аналогично оценка «разности разностей» равна разности среднего изменения для экспериментальной группы и среднего изменения для контрольной группы, то есть  $\hat{\beta}_1^{diffs-in-diffs} = \Delta\bar{Y}^{treatment} - \Delta\bar{Y}^{control} = (80 - 40) - (30 - 20) = 30$ .

**Оценка «разности разностей» с дополнительными регрессорами.** Оценка «разности разностей» может быть обобщена на случай дополнительных регрессоров  $W_{1i}, \dots, W_{ri}$ , которые отражают дополнительные индивидуальные характеристики участников эксперимента. Эти дополнительные регрессоры могут быть включены в модель множественной регрессии:

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \dots + \beta_{1+r} W_{ri} + u_i, \quad i = 1, \dots, n. \quad (13.6)$$

МНК-оценка коэффициента  $\beta_1$  в уравнении (13.6) представляет собой **оценку «разности разностей» с дополнительными регрессорами**. Если  $X_i$  «как бы» случайно распределены при заданных  $W_{1i}, \dots, W_{ri}$ , то  $u_i$  удовлетворяет условию независимости условного среднего, а МНК-оценка  $\hat{\beta}_1$  в уравнении (13.6) является несмешанной.

Оценка «разности разностей», описанная выше, учитывает два временных периода – до и после проведения эксперимента. Однако в контексте решения некоторых задач возникает необходимость рассматривать панельные данные, в которых фигурирует несколько временных периодов. Оценка «разности разностей» может быть обобщена на случай нескольких временных периодов посредством использования методов оценки регрессионных моделей на панельных данных, которые рассматривались в главе 10.

**Оценка «разности разностей» в случае повторяющихся межобъектных данных.** Повторяющиеся межобъектные данные представляют собой набор отдельных выборок межобъектных данных, каждая из которых соответствуетциальному году. Например, база данных может содержать наблюдения для 400 индивидов в 2004 году, наблюдения для других 500 индивидов в 2005 году, всего 900 различных индивидов. Одними из наиболее часто встречающихся примеров повторяющихся межобъектных данных являются данные опросов по выявлению политических предпочтений, в которых политические предпочтения измеряются при помощи серии опросов отобранных случайным образом потенциальных респондентов, причем все опросы проводятся в разные периоды времени для различных респондентов.

Преимуществом использования повторяющихся межобъектных данных является то, что если индивиды (в общем случае – субъекты) случайным образом выбираются из генеральной совокупности, то индивиды из более ранних опросов могут быть использованы как субъекты (которые заменяют отсутствующих реальных субъектов) в контрольных и экспериментальных группах в более поздних по времени межобъектных выборках.

При наличии лишь двух периодов регрессионную модель для повторяющихся межобъектных данных можно записать в таком виде:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + \dots + \beta_{3+r} W_{rit} + u_{it}, \quad (13.7)$$

где  $X_{it}$  – реальный исследуемый эффект в рамках проведения эксперимента для  $i$ -го индивида (субъекта) в межобъектной выборке в период  $t$  ( $t=1; 2$ ),  $G_i$  – бинарная переменная, отражающая принадлежность индивида к экспериментальной группе (или к заменяющей реальную («суррогатной») экспериментальной группе, если наблюдение относится к периоду, предшествующему проведению эксперимента);  $D_t$  – бинарная переменная, которая равна нулю в первом периоде и равна единице во втором периоде.  $i$ -й индивид подвергается воздействию изучаемого эффекта, если при проведении эксперимента он (или она) находится в экспериментальной группе во втором периоде, тогда в уравнении (13.7)  $X_{it} = G_i \times D_t$ .

Если в квазиэксперименте  $X_{it}$  получается «как бы» случайным образом при заданных  $W$ , то исследуемый в таком квазиэксперимента эффект может быть оценен с помощью МНК-оценки  $\beta_i$  в уравнении (13.7). При наличии более чем двух временных периодов уравнение (13.7) может быть модифицировано с помощью включения  $T-1$  бинарной переменной, характеризующей другие временные периоды (см. раздел 10.4).

### Оценки инструментальных переменных

Если в рамках квазиэксперимента имеется переменная  $Z_i$ , которая влияет на проведение эксперимента, если данные доступны одновременно и для  $Z_i$ , и для переменной, отражающей воздействие, которое было получено на самом деле ( $X_i$ ), и если  $Z_i$  распределена «как бы» случайно (возможно, после учета некоторых дополнительных переменных  $W_i$ ), тогда  $Z_i$  является допустимым

инструментом для  $X$ , а коэффициенты уравнения (13.2) могут быть оценены с помощью двухшагового МНК. Любые контрольные переменные, которые могли появиться в (13.2), также могут присутствовать на первом шаге 2МНК-оценки  $\beta_1$ .

### ***Оценки точек разрыва линии регрессии***

В некоторых ситуациях, в рамках которых проходит квазиэксперимент, его проведение может полностью или частично зависеть от того, проходит или нет наблюдаемая переменная  $W$  через некоторое пороговое значение. Например, предположим, что школьники должны посещать школу летом, если их средний балл по итогам года (GPA) оказывается ниже определенного порогового значения<sup>1</sup>. Тогда одним из способов оценки влияния обязательных летних школ является сравнение исходов (неких показателей) для тех учеников, у которых GPA оказалось чуть ниже границы (и, таким образом, они были обязаны посещать летние школы), и для тех, у кого GPA было чуть выше границы (т.е. тех, кто избежал необходимости посещения летних школ). В качестве таких показателей может использоваться GPA в следующем году или будущий уровень доходов, если речь идет о тех, кто закончил школу. Пока у нас нет какой-либо дополнительной информации относительно описанного выше «порогового» значения кроме того, что оно используется при определении необходимости посещения летней школы; имеет смысл приписывать любые изменения будущих показателей (исходов) к пересечению данного порогового значения. На рисунке 13.2 представлена гипотетическая диаграмма рассеяния для базы данных, в которой эксперимент (летняя школа,  $X$ ) проводится, если GPA ( $W$ ) меньше, чем пороговое значение ( $w_0 = 2,0$ ). На диаграмме отображена зависимость GPA в следующем году ( $Y$ ) от GPA в этом году ( $W$ ) для некоторой гипотетической выборки учеников. Если единственным предназначением порогового значения  $w_0$  является отбор для посещения летней школы, то «скачок» GPA в следующем году в точке  $w_0$  будет представлять собой оценку влияния летней школы на GPA в следующем году.

Из-за подобного «скачка» или «разрыва» линии регрессии в пороговом значении модели, в которых используется подобный разрыв в вероятности получения воздействия в пороговом значении, принято называть моделями разрывных регрессий. Существует два типа моделей *разрывных регрессий*: модели с четким разрывом и модели с нечетким разрывом.

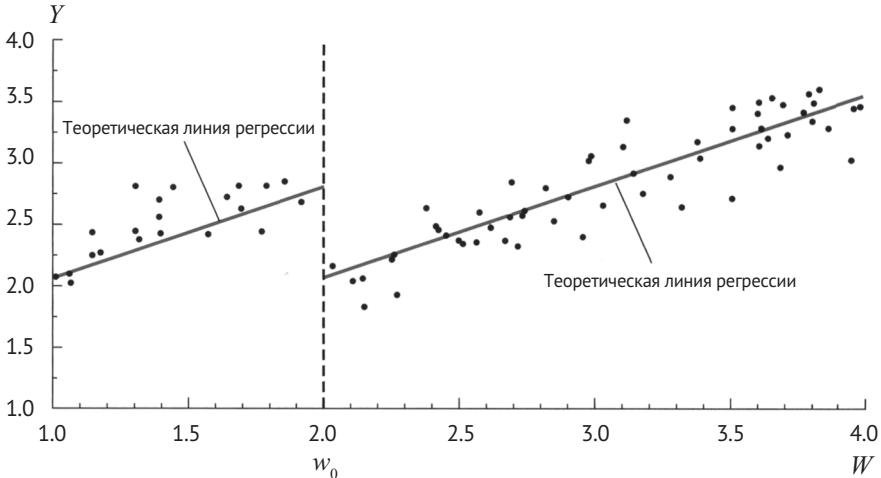
**Модели регрессии с четким разрывом.** В рамках данной модели проведение эксперимента полностью определяется тем, превышает ли  $W$  пороговое значение. Все ученики с  $W < w_0$  посещают летнюю школу, и никто из учеников с  $W \geq w_0$  не посещает. То есть  $X_i = 1$ , если  $W < w_0$ , и  $X_i = 0$ , если  $W \geq w_0$ . В этом случае скачок  $Y$  в пороговом значении отражает влияние проведенного эксперимента для подвыборки с  $W = w_0$ . Если функция регрессии линейна по  $W$ , то эффект проведения эксперимента можно оценить с помощью оценки  $\beta_1$  в регрессии:

---

<sup>1</sup> Этот пример представляет собой упрощенную версию исследования влияния летних школ для учеников начальных и средних школ, проводившегося в работе Мацудайры (Matsudaira, 2008), в рамках которой необходимость посещения летних школ частично определялась на основе баллов академических тестов в конце года.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \quad (13.8)$$

Если же имеет место нелинейная регрессия, то необходимо использовать подходящую нелинейную функцию (см. раздел 8.2).



**Рисунок 13.2. Диаграмма рассеяния с разрывной регрессией**

Предположим, что бинарная переменная  $X$  равна единице, если  $W$  меньше, чем пороговое значение  $w_0 = 2$ . Поскольку единственная роль  $w_0$  заключается в определении того, будет ли проведен эксперимент или нет, то эффект от проведения эксперимента отражается «скачком» или «разрывом» регрессии в точке  $W = 2$ .

**Модели регрессии с нечетким разрывом.** В рамках подобного механизма пересечение порогового значения является определяющим фактором эффекта проведения эксперимента, но не единственным подобным фактором. Например, предположим, что некоторые школьники с GPA ниже порогового значения исключаются из летних школ, а некоторые школьники с GPA выше порогового значения могут посещать летние школы. Подобная ситуация могла бы возникнуть, если «правило» пересечения порогового значения является частью более сложной схемы проведения эксперимента. В моделях с нечетким разрывом регрессии переменная  $X_i$  в уравнении (13.8) будет коррелирована с  $u_i$ . Однако если какой-либо особый эффект пересечения порогового значения влияет только на рост вероятности получения воздействия — то есть отражается только в линейном слагаемом  $W$  — то можно использовать метод инструментальных переменных. В частности, бинарная переменная  $Z_i$ , которая отражает факт пересечения порогового значения (т.е.  $Z_i = 1$ , если  $W < w_0$ , и  $Z_i = 0$ , если  $W \geq w_0$ ), оказывает влияние на факт проведения эксперимента, но не коррелирует с  $u_i$ , то есть является допустимым инструментом для  $X_i$ . Таким образом, в моделях с нечетким разрывом регрессии  $\beta_1$  может быть оценен с помощью уравнения (13.8) и использования бинарной инструментальной переменной, отражающей то, что  $W_i < w_0$ .

## 13.5. Потенциальные проблемы с квазиэкспериментами

Как и во всех эмпирических исследованиях, квазиэксперименты сталкиваются с угрозами для внутренней и внешней обоснованности. Довольно важной

потенциальной угрозой внутренней обоснованности является то, может ли «как бы» случайный выбор рассматриваться как действительно случайный выбор.

## **Угрозы внутренней обоснованности**

Угрозы внутренней обоснованности для действительно случайных экспериментов, которые были рассмотрены в разделе 13.2, с некоторыми изменениями также имеют место и для квазиэкспериментов.

**Отсутствие случайности отбора.** Квазиэксперименты основываются на различиях в индивидуальных обстоятельствах – изменениях законов, неожиданных несвязанных событиях и других ситуациях, которые позволяют привнести «как бы» случайность выбора в эксперимент. Если эта «как бы» случайность выбора не позволяет провести эксперимент, что отражается в случайности переменной  $X$  (или инструментальной переменной  $Z$ ), то в общем случае МНК-оценки являются смещёнными (или оценки на основе инструментальных переменных не являются состоятельными).

Как и в настоящем эксперименте, единственным способом тестирования проблем случайности выбора является проверка систематических отклонений между экспериментальной и контрольной группами, например, с помощью оценки регрессий на  $X$  (или  $Z$ ) и тестирования гипотезы о том, что коэффициенты при  $W$  равны нулю. Если между группами существуют отличия и это не объясняется природой квазиэксперимента, то это является свидетельством того, что квазиэксперимент не обеспечивает настоящей случайности выбора. Даже если взаимосвязи между  $X$  (или  $Z$ ) и  $W$  не обнаруживаются, все равно остается вероятность того, что  $X$  (или  $Z$ ) могут быть связаны с некоторыми ненаблюдаемыми факторами в ошибке  $u$ . Поскольку эти факторы являются ненаблюдаемыми, то наличие подобной взаимосвязи не может быть протестировано, и обоснованность предположения о «как бы» случайности выбора должна быть оценена экспертизой.

**Нарушение условий эксперимента.** Нарушения условий проведения эксперимента могут возникать в тех случаях, когда представители экспериментальной группы не выполняют требуемые согласно условиям проведения эксперимента действия (например не принимают необходимые лекарства), либо представители контрольной группы осуществляют не требующиеся от них действия (например принимают какие-либо посторонние лекарства), либо в обоих случаях одновременно. Как следствие, МНК-оценки эффекта от проведения эксперимента будут смещены. Аналогом нарушения условий проведения эксперимента является ситуация, в которой «как бы» случайность выбора оказывает влияние, но не определяет условия проведения эксперимента. В этом случае оценка на основе инструментальных переменных для квазиэксперимента может быть состоятельной, даже если МНК-оценка таковой не является.

**Истощение выборки.** Проблема истощения выборки в квазиэксперименте схожа с аналогичной проблемой в случае настоящего эксперимента в том смысле, что если она возникает из-за индивидуальных характеристик (или выбора) участников, то приводит к корреляции между параметром, отражающим

влияние эксперимента, и случайным членом. В результате МНК-оценка влияния эксперимента будет смещенной (вследствие наличия смещения из-за отбора наблюдений) и несостоительной.

**Экспериментальные эффекты.** Преимущество квазиэкспериментов заключается в том, что поскольку они не являются в полной мере настоящими экспериментами, то у «участвующих» в них индивидов нет причин думать, что они являются участниками эксперимента. Таким образом, экспериментальные эффекты, такие как эффект Хоторна, как правило, не проявляются в рамках квазиэкспериментов.

**Допустимость инструментов в квазиэкспериментах.** Важным шагом в оценке исследования, в котором используются регрессии с инструментальными переменными, является тщательное рассмотрение допустимости инструментов. Это верно и для квазиэкспериментов, в которых инструмент определен «как бы» случайным образом. Как отмечалось в главе 12, допустимость инструмента требует одновременного выполнения условий релевантности и экзогенности инструмента. Поскольку релевантность инструмента можно проверить с помощью статистических методов, представленных во вставке «Основные понятия 12.5», то далее сконцентрируемся на втором, более оценочном требовании экзогенности инструмента.

Несмотря на то что может показаться, будто случайно выбранные инструментальные переменные обязательно являются экзогенными, это не так. Рассмотрим примеры из раздела 13.4. В работе Энгриста (Angrist, 1990) при использовании выпавших лотерейных чисел (результатов лотереи) в качестве инструментальной переменной при изучении влияния прохождения военной службы на будущий уровень доходов лотерейный номер был действительно случайным. Но как было показано в работе Энгриста (Angrist, 1990), если выпавшее низкое число приводит к стремлению уклониться от службы, а подобное поведение оказывает влияние на последующий уровень доходов после прохождения службы, то маленькое выпавшее лотерейное число ( $Z_i$ ) может быть связано с ненаблюдаемыми факторами, которые определяют будущий уровень доходов после прохождения службы ( $u_i$ ), то есть  $Z_i$  и  $u_i$  будут коррелированы, несмотря на то что  $Z_i$  выбирается случайным образом. В качестве второго примера можно рассмотреть работу Макклеллана, Макнейла и Ньюхауса (McClellan, McNeil, Newhouse, 1994), в которой проводилось изучение влияния случайности распределения относительного расстояния до больницы для пациентов с сердечными приступами, к которым применялась процедура катетеризации сердца. Но, как замечают и впоследствии показывают авторы, если пациенты, которые живут близко к больнице, где может проводиться процедура катетеризации сердца, являются априори более здоровыми, чем те, кто живет на более дальнем расстоянии (возможно, из-за более широкого доступа к медицинской помощи в целом), то относительное расстояние до больницы, где может проводиться процедура катетеризации сердца, может быть коррелирована с пропущенными переменными в остаточном члене уравнения, которое описывает подобную взаимосвязь. Короче говоря, только то, что инструмент определяется случайным образом или «как бы» случайным образом, необязательно

означает его экзогенность в том смысле, что  $\text{corr}(Z_i, u_i) = 0$ . Таким образом, экзогенность инструментов должна подробно анализироваться, даже если инструмент поддерживается в рамках квазиэксперимента.

### **Угрозы внешней обоснованности**

Квазиэкспериментальные исследования используют реальные данные, поэтому угрозы внешней обоснованности исследования, основанного на квазиэксперименте, как правило, похожи на их аналоги, которые обсуждались в разделе 9.1 для исследований с использованием регрессионных оценок, полученных на реальных наблюдениях.

Одним из важных моментов является то, что специальные мероприятия, которые создают «как бы» случайный выбор, лежащий в основе квазиэксперимента, могут привести к ряду особенностей, которые будут угрожать внешней обоснованности. Например, в исследовании Карда (Card, 1990), которое обсуждалось в разделе 13.4, анализировалось влияние иммиграции на рынок труда и использовался «как бы» случайный выбор на основе данных о притоке кубинских иммигрантов во время массовой эмиграции с Кубы в 1980 году. Тем не менее здесь имели место некие особенности кубинских эмигрантов, Майами и его кубинской общины, которые могут затруднить обобщение полученных результатов на случай иммигрантов из других стран или в другие штаты. Аналогично в работе Энгриста (Angrist, 1990) исследование влияния прохождения военной службы во время войны во Вьетнаме на рынок труда, по-видимому, не может быть обобщено на случай прохождения военной службы в мирное время. Как правило, возможность обобщения результатов исследования на какие-то другие конкретные случаи зависит от деталей исследования и должно оцениваться в каждом конкретном случае индивидуально.

## **13.6. Экспериментальные и квазиэкспериментальные оценки в неоднородных выборках**

Как отмечалось в разделе 13.1, причинный эффект может изменяться для различных представителей выборки. В разделе 13.1 описаны оценки причинно-следственных эффектов, которые зависят от такой наблюдаемой переменной, как пол. В этом разделе будут рассмотрены последствия ненаблюдаемых изменений в причинно-следственных эффектах. Будем считать, что ненаблюдаемые изменения причинно-следственных эффектов вызваны неоднородностью выборки. Для упрощения и для того, чтобы сосредоточиться на роли ненаблюдаемой неоднородности, в этом разделе будут опущены контрольные переменные  $W$ . Однако выводы данного раздела могут быть обобщены на случай регрессий с наличием контрольных переменных.

Если выборка неоднородна, то  $i$ -й индивид имеет свой собственный причинно-следственный эффект  $\beta_{1i}$ , который (в терминах раздела 13.1) представляет собой разность потенциальных исходов для  $i$ -го индивида для случаев, если он (или она) подвержен влиянию от проведения эксперимента и если

не подвержен ему (например, в рамках самого простого эксперимента по тестированию действия лекарственного препарата – если принимает новое лекарство или принимает плацебо). Например,  $\beta_{1i}$  может быть равен нулю для программы по обучению написанию резюме, если  $i$ -й индивид уже умеет писать резюме. Тогда, используя такие обозначения, уравнение регрессии можно записать в следующем виде:

$$Y_i = \beta_{0i} + \beta_{1i}X_i + u_i. \quad (13.9)$$

Поскольку  $\beta_{1i}$  может меняться для различных индивидов из генеральной совокупности, а выборка формируется из генеральной совокупности случайным образом, то  $\beta_{1i}$  является случайной переменной, которая так же, как и  $u_i$ , отражает ненаблюдаемую изменчивость между индивидами (например, различие в имеющихся до прохождения программы навыках написания резюме). Средний причинный эффект представляет собой среднее значение причинного эффекта  $E(\beta_{1i})$  в генеральной совокупности. То есть это – ожидаемый причинный эффект для выбранного случайным образом из генеральной совокупности индивида.

Тогда что же показывают оценки в разделах 13.1, 13.2 и 13.4, если в выборке имеет место неоднородность в форме (13.9)? Рассмотрим в первую очередь МНК-оценки в том случае, если  $X_i$  определяются «как бы» случайным образом. В этом случае МНК-оценки представляют собой состоятельные оценки среднего причинного эффекта. Однако в общем случае это не выполняется для оценок, полученных методом инструментальных переменных. Вместо этого, если  $Z_i$  частично влияет на  $X_i$ , то оценки на основе инструмента  $Z$  представляют собой взвешенное среднее причинных эффектов, при этом те индивиды, для которых инструменты имеют наибольшее влияние, получают большие веса.

### **МНК с неоднородными причинными эффектами**

Если в причинных эффектах присутствует неоднородность и если  $X_i$  определяется случайным образом (т.е. индивиды, которым назначается воздействие, выбираются случайно), то оценка разностей является состоятельной оценкой среднего причинного эффекта. Этот результат следует из раздела 13.1 и приложения 13.3 при использовании концепции потенциальных исходов. В настоящем разделе этот результат будет получен без отсылок к концепции потенциальных исходов, а лишь путем применения понятий из разделов 3 и 4 непосредственно к случайным коэффициентам регрессионной модели в уравнении (13.9).

МНК-оценка  $\beta_1$  в уравнении (13.1) равна:  $\widehat{\beta}_1 = s_{xy}/s_x^2$  [уравнение (4.7)]. Если данные являются независимо одинаково распределенными, то выборочные ковариация и дисперсия являются состоятельными оценками соответственно теоретических ковариации и дисперсии, тогда  $\widehat{\beta}_1 \xrightarrow{P} \sigma_{xy}/\sigma_x^2$ . Если  $X_i$  определяется

случайным образом (индивидуды отбираются в экспериментальную группу случайно), то переменная  $X_i$  распределена независимо от других индивидуальных характеристик (наблюдаемых и ненаблюдаемых) и, таким образом, распределена независимо от  $\beta_{0i}$  и  $\beta_{1i}$ . Соответственно, МНК-оценка  $\hat{\beta}_1$  имеет предел:

$$\begin{aligned}\hat{\beta}_1 &= \frac{s_{XY}}{s_X^2} \xrightarrow{p} \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\text{cov}(\beta_{0i} + \beta_{1i}X_i + u_i; X_i)}{\sigma_X^2} = \\ &= \frac{\text{cov}(\beta_{0i} + \beta_{1i}X_i; X_i)}{\sigma_X^2},\end{aligned}\tag{13.10}$$

где третий знак равенства использует свойства ковариации из вставки «Основные понятия 2.3», а также равенство:  $\text{cov}(u_i; X_i) = 0$ , которое следует из  $E(u_i | X_i) = 0$  (уравнение (2.27)); последнее равенство следует из того, что  $\beta_{0i}$  и  $\beta_{1i}$  распределены независимо от  $X_i$ , что получается, если  $X_i$  приписывается индивидам случайным образом (упражнение 13.9). Таким образом, если  $X_i$  определяется для индивидов случайным образом, то  $\hat{\beta}_1$  является состоятельной оценкой среднего причинного эффекта  $E(\beta_{1i})$ .

### *Модель регрессии с инструментальными переменными при неоднородных причинных эффектах*

Предположим, что причинные эффекты оцениваются на основе регрессии с инструментальными переменными  $Y_i$  на  $X_i$  (параметр, характеризующий воздействие, полученное в эксперименте) с использованием  $Z_i$  (первоначальное случайно или «как бы» случайно приписанное воздействие в эксперименте) в качестве инструментальной переменной. Пусть  $Z_i$  является допустимым инструментом (релевантным и экзогенным), а также имеет место неоднородность влияния  $Z_i$  на  $X_i$ . В частности, предположим, что связь  $X_i$  и  $Z_i$  описывается моделью линейной регрессии:

$$X_i = \pi_{0i} + \pi_{1i}Z_i + v_i,\tag{13.11}$$

где коэффициенты  $\pi_{0i}$  и  $\pi_{1i}$  могут быть различны для разных индивидов. Уравнение (13.11) представляет собой уравнение первого шага 2МНК [уравнение (12.2)] с той лишь разницей, что влияние  $Z_i$  на  $X_i$  может быть различным для различных индивидов.

2МНК-оценка имеет вид:  $\hat{\beta}_1^{TSLS} = s_{ZY} / s_{ZX}$  [уравнение (12.4)], то есть она равна отношению выборочной ковариации между  $Z$  и  $Y$  к выборочной ковариации между  $Z$  и  $X$ . Если имеющиеся данные являются i.i.d., то эти выборочные ковариации являются состоятельными оценками истинных (теоретических) ковариаций, тогда  $\hat{\beta}_1^{TSLS} \xrightarrow{p} \sigma_{ZY} / \sigma_{ZX}$ . Предположим, что  $\pi_{0i}$ ,  $\pi_{1i}$ ,  $\beta_{0i}$  и  $\beta_{1i}$  распределены независимо от  $u_i$ ,  $v_i$  и  $Z_i$  и  $E(u_i | Z_i) = E(v_i | Z) = 0$  и  $E(\pi_{1i}) \neq 0$  (релевантность инструмента). В приложении 13.2 будет показано, что в рамках этих предположений

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}. \quad (13.12)$$

Таким образом, 2МНК-оценка сходится по вероятности к отношению ожидаемого значения произведения  $\beta_{1i}$  и  $\pi_{1i}$  к ожидаемому значению  $\pi_{1i}$ .

Последняя дробь в уравнении (13.12) представляет собой взвешенное среднее индивидуальных причинных эффектов  $\beta_{1i}$ . Веса равны  $\pi_{1i}/E(\pi_{1i})$ , что отражает относительную степень влияния инструмента, если  $i$ -й индивид получает воздействие в эксперименте. Таким образом, 2МНК-оценка является состоятельной оценкой средневзвешенного индивидуального причинного эффекта, в котором наибольшие веса получают индивиды, для которых инструмент имеет наибольшую степень влияния. Средневзвешенный причинный эффект, который оценивается с помощью 2МНК, называют *локальным средним эффектом (воздействия) в эксперименте* (LATE<sup>1</sup>). Термин «локальный» отражает то, что наибольшие веса получают те индивиды (в общем случае – субъекты), которые в большей степени подвержены влиянию инструментальной переменной.

Есть три особых случая, в которых локальный средний эффект воздействия в эксперименте совпадает со средним эффектом в эксперименте:

1. Эффект воздействия в эксперименте одинаков для всех индивидов. Этот случай соответствует  $\beta_{1i} = \beta_1$  для всех  $i$ . Тогда последнее выражение в уравнении (13.12) можно упростить до  $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1 E(\pi_{1i})/E(\pi_{1i}) = \beta_1$ .
2. Инструмент влияет на всех индивидов одинаково. Этот случай соответствует  $\pi_{1i} = \pi_1$  для всех  $i$ . Тогда последнее выражение в уравнении (13.12) можно упростить до  $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})\pi_1/\pi_1 = E(\beta_{1i})$ .
3. Неоднородность в эффекте воздействия в эксперименте и неоднородность во влиянии инструмента не коррелированы. Это соответствует случаю, когда  $\beta_{1i}$  и  $\pi_{1i}$  являются случайными величинами, но  $\text{COV}(\pi_{1i}, \beta_{1i}) = 0$ . Поскольку  $E(\beta_{1i}\pi_{1i}) = \text{COV}(\pi_{1i}; \beta_{1i}) + E(\beta_{1i})E(\pi_{1i})$  [уравнение (2.34)] и если  $\text{COV}(\pi_{1i}, \beta_{1i}) = 0$ , то  $E(\beta_{1i}\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})$ . Тогда последнюю часть выражения (13.12) можно упростить до  $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})E(\pi_{1i})/E(\pi_{1i}) = E(\beta_{1i})$ .

В каждом из случаев существует неоднородность генеральной совокупности в эффектах влияния инструментов на индивидов, в эффектах воздействия в эксперименте, или оба этих эффекта, но локальный средний эффект воздействия в эксперименте равен среднему эффекту в эксперименте. То есть во всех трех случаях 2МНК дает состоятельную оценку среднего эффекта воздействия в эксперименте.

В отличие от приведенных выше трех особых случаев, в общем случае локальный средний эффект в эксперименте отличается от среднего причинного эффекта. Например, предположим, что  $Z_i$  не влияет на эксперимент для половины

<sup>1</sup> Local average treatment effect. – Примеч. науч. ред. перевода.

выборки (для них  $\pi_{1i} = 0$ ), но влияет на другую половину выборки (для них  $\pi_{1i}$  – ненулевая константа). Тогда 2МНК дает состоятельную оценку среднего эффекта в эксперименте для половины выборки, для которой имеет место ненулевое влияние инструмента. Более конкретно, предположим, что работникам – кандидатам на прохождение программы по повышению квалификации случайным образом приписывается номер в очереди  $Z$ , который влияет на их вероятность быть принятыми в программу. Половина работников знает, что они выиграют от прохождения программы, и будут участвовать в программе, если будет такая возможность. Для них  $\beta_{1i} = \beta_1^+ > 0$  и  $\pi_{1i} = \pi_1^+ > 0$ . Вторая половина работников знает, что для них прохождение программы будет неэффективным, и они не будут в ней участвовать, даже если им представиться такая возможность. Для них  $\beta_{1i} = 0$  и  $\pi_{1i} = 0$ . Тогда средний эффект в эксперименте равен:  $E(\beta_{1i}) = \frac{1}{2}(\beta_1^+ + 0) = \frac{1}{2}\beta_1^+$ .

А локальный средний эффект в эксперименте равен  $E(\beta_{1i}\pi_{1i})/E(\pi_{1i})$ . Тогда имеем,

что  $E(\pi_{1i}) = \frac{1}{2}\pi_1^+$  и  $E(\beta_{1i}\pi_{1i}) = E[\beta_{1i}E(\pi_{1i}|\beta_{1i})] = \frac{1}{2}(0 + \beta_1^+\pi_1^+) = \frac{1}{2}\beta_1^+\pi_1^+$ . То есть  $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1^+$ . Таким образом, в этом примере локальный средний эффект

воздействия в эксперименте равен среднему эффекту в эксперименте для тех работников, которые, вероятно, будут участвовать в программе, а веса для тех работников, которые не будут участвовать в программе ни при каких обстоятельствах, равны нулю. Поскольку индивиды принимают решение относительно участия в программе на основе их собственных знаний о том, насколько эффективна для них будет программа, в этом примере локальный средний эффект воздействия в эксперименте будет больше среднего эффекта воздействия в эксперименте.

**Выводы.** Если решение индивида об участии в эксперименте зависит от эффективности этого участия, то 2МНК-оценка в общем случае не является состоятельной оценкой среднего причинного эффекта. Вместо этого 2МНК позволяет оценить локальный средний эффект воздействия в эксперименте, в котором причинные эффекты для индивидов, в большей степени подверженных влиянию инструмента, получают большие веса. Этот результат приводит к затруднительной ситуации, в которой два разных исследователя, имея разные инструментальные переменные, каждая из которых является допустимой (экзогенной и релевантной), получат разные оценки причинного эффекта даже в больших выборках. Это различие возникает из-за того, что каждый из исследователей неявно оценивает различные взвешенные средние значения индивидуальных причинных эффектов в генеральной совокупности. Действительно,  $J$ -тест на сверхидентифицирующие ограничения может отвергнуть нулевую гипотезу, если обе инструментальные переменные оценивают различные локальные средние эффекты воздействия в эксперименте, даже если обе этих инструментальные переменные являются допустимыми. Несмотря на то что обе описанные оценки предоставляют некоторое описание распределения причинных эф-

фектов через их соответствующие взвешенные средние в форме, записанной в уравнении (13.12), в общем случае ни одна из этих оценок не является состоятельной оценкой среднего причинного эффекта<sup>1</sup>.

**Пример: изучение последствий катетеризации сердца.** В разделах 12.5 и 13.4 рассматривалась работа Макклеллана, Макнейла и Ньюхауза (McClellan, McNeil, Newhouse, 1991), в которой проводилось изучение влияния на продолжительность жизни проведения катетеризации сердца для пациентов с сердечными приступами. Авторы использовали модель регрессии с инструментальными переменными, в которой присутствовала переменная, отражающая относительное расстояние до лечебных учреждений, где может осуществляться процедура катетеризации сердца, по сравнению с расстоянием до обычной больницы. Полученные 2МНК-оценки показали, что возможность проведения процедуры катетеризации сердца не оказывает практически никакого влияния на продолжительность жизни. Это довольно странный результат, поскольку медицинские процедуры, подобные катетеризации сердца, подвергаются очень тщательному клиническому тестированию, прежде чем быть допущенными к широкому применению. Кроме того, катетеризация сердца позволяет хирургам проводить такие хирургические процедуры, которые за десять лет до этого требовали серьезного операционного вмешательства, что делает подобные процедуры безопаснее и, возможно, лучше для здоровья пациента в долгосрочной перспективе. Почему же тогда эконометрическое исследование не обнаружило положительного влияния процедуры катетеризации сердца?

Одним из возможных ответов является наличие неоднородности в эффектах воздействия, получаемых от катетеризации сердца. Для некоторых пациентов эта процедура представляется эффективной, но для других, которые, возможно, более здоровы, менее эффективна или даже (при наличии некоторых рисков при проведении любого рода хирургического вмешательства) неэффективна. Таким образом, средний причинный эффект в генеральной совокупности пациентов с сердечным приступом, вероятно, может быть положительным (и, по-видимому, является положительным). Оценка с помощью инструментальных переменных отражает предельный, а не средний эффект, где предельный эффект представляет собой эффект влияния от проведения данной процедуры для тех пациентов, для которых расстояние до больницы является важным фактором проведения этой процедуры. Но эти пациенты могут быть просто относительно более здоровыми по сравнению с теми пациентами (в пределе), для которых процедура катетеризации относительно менее эффективна. Если

<sup>1</sup> Существует несколько хороших работ (продвинутого уровня), в которых проводится обсуждение влияния неоднородности генеральной совокупности на оценку результативности программ. Среди них исследование Хекмана, Лалонда и Смита (Heckman, LaLonde, Smith, 1999, Section 7) и нобелевская лекция Джеймса Хекмана (James Heckman), прочитанная им на церемонии вручения Нобелевской премии по экономике (Heckman, 2001, Section 7). В упомянутой лекции и в работе Энгриста, Грэди и Имбенса (Angrist, Graddy, Imbens, 2000) представлено более детальное обсуждение модели со случайными эффектами (что может приводить к изменчивости  $\beta_{ij}$  между индивидами), а также более общий результат в уравнении (13.12). Понятие локального среднего эффекта в эксперименте было впервые введено в работе Энгриста и Имбенса (Angrist, Imbens, 1994), где было показано, что в общем случае он не совпадает со средним эффектом в эксперименте.

это имеет место, то 2МНК-оценка, полученная в работе Макклеллана, Макнейла и Ньюхауза (McClellan, McNeil, Newhouse, 1991), измеряет эффект от проведения процедуры для «предельного» пациента (для которого процедура относительно менее эффективна), а не «среднего» пациента (для которого процедура может быть эффективна).

### 13.7. Заключение

В главе 1 было введено определение причинного эффекта в терминах ожидаемого исхода идеального случайного контролируемого эксперимента. Если случайный контролируемый эксперимент может быть проведен, с его помощью можно получить убедительные свидетельства исследуемых причинных эффектов, даже если случайные контролируемые эксперименты сталкиваются с угрозами внешней и внутренней обоснованности.

Несмотря на свои преимущества, случайные контролируемые эксперименты в экономике сталкиваются со значительными препятствиями, в том числе с этическими проблемами. Однако часть экспериментальных методов может быть применена и к квазиэкспериментам, в которых особые обстоятельства могут приводить к тому, что процесс случайного выбора «как бы» имел место. В квазиэкспериментах причинные эффекты могут быть оценены с помощью оценки «разности разностей» (в которой могут использоваться дополнительные регрессоры). Если «как бы» случайный выбор лишь частично влияет на проведение эксперимента, то для получения оценок можно использовать модель регрессии с инструментальными переменными. Важное преимущество квазиэкспериментов заключается в том, что источник «как бы» случайности в данных обычно является «прозрачным» и, таким образом, может быть использован при получении оценок явным образом. Важной угрозой, с которой сталкиваются квазиэксперименты, является то, что иногда «как бы» случайность выбора на самом деле не является полностью случайной, поэтому параметр, отражающий результат проведения эксперимента (или инструментальная переменная), коррелирует с пропущенными переменными, что приводит к смещению получаемых оценок причинных эффектов.

Квазиэксперименты, являются связующим звеном между базами данных, полученными путем наблюдений за субъектами, и истинными случайными контролируемыми экспериментами. Эконометрические методы, используемые в этой главе для анализа квазиэкспериментов, были представлены в различных контекстах в предыдущих главах: МНК, методы оценки панельных данных и регрессии с инструментальными переменными. Что отличает квазиэксперименты от методов, рассмотренных в части II и в первых главах части III, так это способ интерпретации полученных результатов, а также базы данных, к которым могут применяться эти методы. Квазиэксперименты дают экономистам способ изучения новых наборов данных, применения к ним инструментальных переменных, а также способ оценки правдоподобности

предположений об экзогенности, лежащих в основе МНК и метода инструментальных переменных<sup>1</sup>.

## **Выходы**

1. Средний причинный эффект в изучаемой генеральной совокупности представляет собой ожидаемую разность средних исходов в экспериментальной и контрольной группах в идеальном случайном контролируемом эксперименте. Реальные эксперименты с людьми часто отличаются от идеального эксперимента из-за множества практических причин, среди которых может быть, например, несоблюдение правил проведения эксперимента его участниками.
2. Если фактическое воздействие  $X_i$  является случайной величиной, то эффект воздействия может быть оценен оценкой регрессии исхода на переменную, характеризующую наличие воздействия –  $X_i$ . Если назначенное воздействие  $Z_i$  является случайной величиной, а фактическое участие в эксперименте –  $X_i$  – частично определяется личными характеристиками индивидов, принимающих участие в эксперименте, то причинный эффект может быть оценен с помощью модели регрессии с инструментальными переменными, включающей  $Z_i$  в качестве инструмента. Если воздействие (или назначенное воздействие) является случайным условно относительно некоторых переменных  $W$ , то эти контрольные переменные должны быть включены в модель регрессии.
3. В рамках квазиэкспериментов изменения в законах, обстоятельствах или происшествиях рассматриваются в качестве «как бы» случайных, то есть вызывающих случайное распределение субъектов между экспериментальной и контрольной группами. Если фактическое воздействие является «как бы» случайным, то причинный эффект может быть оценен с помощью модели регрессии (возможно, с использованием дополнительных регрессоров, отражающих различные параметры, влияющие на квазиэксперимент). Если назначенное воздействие проводится «как бы» случайным образом, то причинный эффект может быть оценен с помощью модели регрессии с инструментальными переменными.
4. Ключевую угрозу внутренней обоснованности квази-экспериментов можно сформулировать как вопрос: следует ли из «как бы» случайного выбора экзогенность инструментов. Из-за поведенческих реакций только лишь то, что инструмент формируется «как бы» случайно не означает, что он обязательно является экзогенным в том смысле,

---

<sup>1</sup> В работе Шадиса, Кука и Кэмпбелла (Shadish, Cook, Campbell, 2002) представлен подробный обзор экспериментов и квазиэкспериментов в социальных науках и психологии. Важным направлением исследований развивающихся экономик является проведение экспериментальных оценок образовательных программ и программ по здравоохранению в развивающихся странах. См., например, работу Кремера, Мигеля и Торнтона (Kremer, Miguel, Thornton, 2009) и веб-сайт MIT Poverty Action Laboratory (<http://www.povertyactionlab.org>)

- который рассматривается в определении допустимости инструментальных переменных.
5. Когда эффект воздействия в эксперименте различен для разных индивидов, МНК-оценка является состоятельной оценкой среднего причинного эффекта, если фактическое воздействие назначается случайно или «как бы» случайно. Однако оценка на основе метода инструментальных переменных представляет собой средневзвешенное значение индивидуальных эффектов воздействия, причем индивиды, для которых инструмент оказывает наибольшее влияние, получают наибольший вес.

## **Основные понятия**

- Оценка программных документов (с. 493).  
Потенциальный исход (с. 495).  
Средний причинный эффект (с. 495).  
Средний эффект в эксперименте (с. 495).  
Оценка разностей (с. 496).  
Оценка разностей с дополнительными регрессорами (с. 496).  
Случайный выбор, зависящий от наблюдаемых переменных (с. 497).  
Тест на случайность распределения между экспериментальной и контрольной группами (с. 498).  
Частичное соответствие (с. 499).  
Оценка эффекта воздействия с помощью инструментальных переменных (с. 499).  
Истощение выборки (с. 500).  
Эффект Хоторна (с. 501).  
Квазиэксперимент (с. 513).  
Естественный эксперимент (с. 513).  
Оценка «разности разностей» (с. 516).  
Оценка «разности разностей» с дополнительными регрессорами (с. 518).  
Повторяющиеся межобъектные данные (с. 519).  
Модель разрывной регрессии (с. 520).  
Локальный средний эффект воздействия в эксперименте (с. 527).

## **Вопросы для повторения и закрепления основных понятий**

- 13.1. Исследователь изучает влияние новых удобрений на урожайность сельскохозяйственных культур и планирует провести эксперимент, в котором различные количества удобрений применяются для 100 различных земельных участков площадью в 1 акр. Исследователь предполагает использовать четыре различные схемы добавления удобрений. Первая схема предполагает отсутствие удобрений, вторая – добавление 50 % от рекомендованного производителем количества удобрений, третья – добавление 100 %, и четвертая – 150 %. Исследователь планирует применить первую схему к первым 25 земельным

участкам, вторую – ко вторым 25 участкам и так далее. Можете ли вы предложить лучшую схему поведения эксперимента? Почему ваше предложение лучше, чем метод, используемый исследователем?

- 13.2. Пусть проводится клиническое тестирование нового лекарства, которое должно понижать уровень холестерина в крови. Случайным образом выбираются 500 пациентов, которые принимают настоящее лекарство, и еще 500 пациентов принимают плацебо. Как вы оценили бы эффект от приема лекарственного препарата? Предположим, что у вас есть данные о весе, возрасте и поле каждого пациента. Можете ли вы использовать эти данные, чтобы улучшить свои оценки? Объясните. Предположим, что у вас есть данные об уровне холестерина в крови каждого пациента до того момента, как он или она начали участвовать в эксперименте. Можете ли вы использовать эти данные, чтобы улучшить свои оценки? Объясните.
- 13.3. Исследователи, изучающие данные STAR, приводят забавные случаи того, что директора школ подвергались давлению со стороны некоторых родителей, которые хотели отдать своих детей в малые учебные классы. Предположим, что некоторые директора поддались этому давлению и перевели нескольких детей в малые учебные классы. Как подобные переводы учеников повлияли бы на внутреннюю обоснованность исследования? Предположим, что у вас есть данные об исходном случайному распределении каждого школьника по классам до вмешательства директора. Как вы могли бы использовать эту информацию, чтобы восстановить внутреннюю обоснованность исследования?
- 13.4. Поясните, важны ли эффекты проведения эксперимента (например, подобные эффекту Хоторна) в каждом из сформулированных выше трех вопросов?
- 13.5. В разделе 12.1 представлен гипотетический пример, в котором некоторые школы были повреждены во время землетрясения. Поясните, почему он является примером квазиэксперимента? Как вы могли бы использовать вызванные этими событиями изменения в размерах классов для оценки влияния размера класса на успеваемость школьников?

## Упражнения

- 13.1. Используя результаты, представленные в таблице 13.1, вычислите для каждого из годов обучения оценку эффекта влияния малого размера учебного класса на успеваемость по сравнению с обычным учебным классом, ее стандартную ошибку и 95 %-й доверительный интервал. (В данном вопросе не учитывайте результаты, полученные для обычных классов с наличием дополнительной помощи.)

13.2. Для вычислений в этом упражнении используйте столбец (4) таблицы 13.2. Рассмотрите два учебных класса – А и Б с одинаковыми значениями регрессоров в столбце (4) таблицы 13.2, предположим, что:

- Класс А – малый учебный класс, а класс Б – стандартный учебный класс. Постройте 95 %-й доверительный интервал для ожидаемой разности средних результатов тестов в этих классах.
- В классе А преподает учитель с пятилетним стажем, а в классе Б – учитель с 10-летним стажем. Постройте 95 %-й доверительный интервал для ожидаемой разности средних результатов тестов в этих классах.
- Класс А – малый учебный класс, в котором преподает учитель с пятилетним стажем, а в класс Б – стандартный учебный класс, в котором преподает учитель с 10-летним стажем. Постройте 95 %-й доверительный интервал для ожидаемой разности средних результатов тестов в этих классах. (Подсказка: в рамках STAR учителя случайным образом распределялись по учебным классам).
- Почему в столбце (4) нет оценки свободного члена?

13.3. Предположим, что после проведения случайного контролируемого эксперимента по оценке влияния подготовительных курсов к тесту SAT (Scholastic Assessment Test) на итоговые результаты по этому тесту получены следующие результаты:

	Экспериментальная группа	Контрольная группа
Среднее количество баллов по SAT ( $\bar{X}$ )	1241	1201
Стандартное отклонение количества баллов по тесту ( $s_x$ )	93,2	97,1
Количество участников мужского пола	55	45
Количество участников женского пола	45	55

- Оцените средний эффект воздействия в эксперименте на набранное количество баллов.
  - Есть ли какие-то свидетельства того, что эксперимент проводился неслучайным образом? Объясните.
- 13.4. Прочитайте вставку «Чему равно влияние занятости на минимальный уровень заработных плат?» в разделе 13.4. Для большей ясности предположим, что Кард и Крюгер собирали данные в 1991 году (перед изменением минимальной заработной платы в штате Нью-Джерси) и в 1993 году (после изменения минимальной заработной платы в штате Нью-Джерси). Рассмотрим уравнение (13.7) с включенными в него регрессорами  $W$ .
- Каковы значения  $X_i$ ,  $G_i$  и  $D_i$  для:
    - ресторана в штате Нью-Джерси в 1991 году?

- (ii) ресторана в штате Нью-Джерси в 1993 году?  
 (iii) ресторана в штате Пенсильвания в 1991 году?  
 (iv) ресторана в штате Пенсильвания в 1993 году?
- б) Поясните в терминах коэффициентов  $\beta_0, \beta_1, \beta_2, \beta_3$ , каково ожидаемое количество работников для:  
 (i) ресторана в штате Нью-Джерси в 1991 году?  
 (ii) ресторана в штате Нью-Джерси в 1993 году?  
 (iii) ресторана в штате Пенсильвания в 1991 году?  
 (iv) ресторана в штате Пенсильвания в 1993 году?
- в) Каков в терминах коэффициентов  $\beta_0, \beta_1, \beta_2, \beta_3$  средний причинный эффект влияния изменения минимальной заработной платы на занятость?
- г) Поясните, почему Кард и Крюгер использовали оценку «разности разностей» причинного эффекта вместо оценки разностей (Нью-Джерси «до» – Нью-Джерси «после») или оценки разностей (Пенсильвания «до» – Пенсильвания «после»)?
- 13.5. Рассмотрим исследование, в рамках которого оценивается влияние наличия доступа в интернет в комнатах студентов в общежитии на их уровень успеваемости. В больших общежитиях половина комнат оснащена высокоскоростным доступом в интернет (экспериментальная группа), а итоговые оценки по учебным курсам собираются для всех проживающих в общежитии. Что из перечисленного ниже может представлять угрозы внутренней обоснованности и почему:
- а) В середине года все спортсмены-мужчины переезжают в другое общежитие и выпадают из исследования (их итоговые оценки не наблюдаются).
- б) Студенты технических специальностей, приписанные к контрольной группе, создают локальную сеть, в результате чего они могут совместно использовать подключение одного пользователя к интернету, за которое они платят совместно.
- в) Студенты гуманитарных специальностей в экспериментальной группе не знают, как получить доступ к своим интернет-аккаунтам.
- г) Студенты экономических специальностей в экспериментальной группе предоставляют доступ в интернет представителям контрольной группы за определенную плату.
- 13.6. Предположим, что имеются панельные данные с  $T=2$  временными периодами для случайного контролируемого эксперимента, где первое наблюдение ( $t=1$ ) получено до проведения эксперимента, а второе наблюдение ( $t=2$ ) – после проведения эксперимента. Предположим, что воздействие в эксперименте носит бинарный характер, то есть  $X_{it}=1$ , если  $i$ -й индивид приписан к экспериментальной группе и  $t=2$ , и  $X_{it}=0$  в противном случае. Далее предположим, что эффект влияния эксперимента может описываться с помощью модели:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it},$$

где  $\alpha_i$  – индивидуальные эффекты (см. уравнение (13.11)) с нулевым математическим ожиданием и дисперсией  $\sigma_{\alpha}^2$ ,  $u_{ii}$  – гомоскедастичный остаточный член,  $\text{cov}(u_{ii}, u_{ij}) = 0$ , а  $\text{cov}(u_{ii}, \alpha_i) = 0$  для всех  $i$ . Пусть  $\hat{\beta}_1^{differences}$  – оценка разностей коэффициента, то есть МНК-оценка коэффициента в регрессии  $Y_{i2}$  на  $X_{i2}$  со свободным членом. Пусть  $\hat{\beta}_1^{diffs-in-diffs}$  – оценка «разности разностей», то есть оценка коэффициента  $\beta_1$ , основанная на МНК регрессии  $\Delta Y_i = Y_{i2} - Y_{i1}$  на  $\Delta X_i = X_{i2} - X_{i1}$  со свободным членом.

a) Покажите, что  $n \text{var}(\hat{\beta}_1^{differences}) \rightarrow (\sigma_u^2 + \sigma_{\alpha}^2) / \text{var}(X_{i2})$ . (Подсказка: используйте

зуйте применимые в случае гомоскедастичности ошибок формулы для дисперсии МНК-оценок, представленные в приложении 5.1.)

б) Покажите, что  $n \text{var}(\hat{\beta}_1^{diffs-in-diffs}) \rightarrow 2\sigma_u^2 / \text{var}(X_{i2})$ . (Подсказка: используйте

равенство  $X_i = X_{i2}$ . Почему оно выполняется?)

в) Основываясь на результатах, полученных в пунктах (a) и (б), какую из оценок вы использовали бы в будущем в подобных ситуациях, исходя только лишь из соображений эффективности?

- 13.7. Предположим, что имеются панельные данные с  $T=2$  временными периодами (т.е.  $t = 1, 2$ ). Рассмотрим оцениваемую на панельных данных регрессию с фиксированными индивидуальными и временными эффектами и индивидуальными характеристиками  $W_i$ , которые не изменяются с течением времени (например пол человека). Пусть эффект влияния эксперимента носит бинарный характер, то есть  $X_{it} = 1$  для  $t = 2$  для индивидов в экспериментальной группе,  $X_{it} = 0$  в противном случае. Рассмотрим теоретическую модель регрессии:

$$Y = a_i + \beta_1 X_{it} + \beta_2 (D_t \times W_i) + \beta_0 D_t + v_{it},$$

где  $a_i$  – индивидуальные фиксированные эффекты,  $D_t$  – бинарная переменная, которая равна 1, если  $t = 2$ , и равна 0, если  $t = 1$ ,  $D_t \times W_i$  – произведение  $D_t$  и  $W_i$ ,  $a$  и  $\beta$  – неизвестные коэффициенты. Пусть  $\Delta Y_i = Y_{i2} - Y_{i1}$ . Выведите уравнение (13.6) (в случае одного регрессора  $W$ , т.е. при  $r = 1$ ) на основе представленного выше уравнения регрессионной модели.

- 13.8. Предположим, что у вас имеются те же данные, что и в упражнении 13.7 (панельные данные для двух временных периодов,  $n$  наблюдений), кроме регрессора  $W$ . Рассмотрите альтернативную модель регрессии:

$$Y_{it} = a_i + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + u_{it},$$

где  $G_i = 1$ , если индивид приписан к экспериментальной группе, и где  $G_i = 0$ , если индивид приписан к контрольной группе. Покажите, что МНК-оценка  $\beta_1$  в рамках этой модели представляет собой оценку «разности разностей» в уравнении (13.4). (Подсказка: см. раздел 8.3.)

13.9. Выведите последнее равенство в выражении (13.10). (Подсказка: используйте определение ковариации и то, что поскольку  $X_i$  является случайной переменной, то  $\beta_{1i}$  и  $X_i$  распределены независимо.)

13.10. Рассмотрите модель регрессии с неоднородными коэффициентами:

$$Y_i = \beta_{0i} + \beta_{1i}X_i + v_i,$$

где  $(v_i, X_i, \beta_{0i}, \beta_{1i})$  являются независимыми одинаково распределенными случайными величинами с  $\beta_0 = E(\beta_{0i})$  и  $\beta_1 = E(\beta_{1i})$ .

а) Покажите, что модель может быть записана в виде  $Y_i = \beta_0 + \beta_1 X_i + u_i$ ,

$$\text{где } u_i = (\beta_{0i} - \beta_0) + (\beta_{1i} - \beta_1)X_i + v_i.$$

б) Предположим, что  $E[\beta_{0i}|X_i] = \beta_0$ ,  $E[\beta_{1i}|X_i] = \beta_1$  и  $E[v_i|X_i] = 0$ . Покажите, что  $E[u_i|X_i] = 0$ .

в) Покажите, что предположения 1 и 2 из вставки «Основные понятия 4.3» выполняются.

г) Предположим, что выбросы в имеющихся данных редки, то есть  $(u_i, X_i)$  имеют конечные моменты с первого по четвертый. Можно ли использовать МНК и методы, рассмотренные в главах 4 и 5, для оценки средних значений  $\beta_{0i}$  и  $\beta_{1i}$  и соответствующих выводов?

д) Предположим, что  $\beta_{1i}$  и  $X_i$  положительно коррелированы, то есть наблюдения со значениями большими, чем среднее значение  $X_i$ , чаще имеют большие значения, чем среднее значение  $\beta_{1i}$ . Выполняются ли в подобной ситуации предположения из вставки «Основные понятия 4.3»? Если не выполняются, то какие из них нарушены? Возможно ли использовать МНК и методы, представленные в главах 4 и 5, для оценки средних значений  $\beta_{0i}$  и  $\beta_{1i}$  и соответствующих выводов?

13.11. В главе 12 для оценки эластичности спроса на сигареты по цене использовались панельные данные по штатам США, где в качестве инструментальной переменной использовался налог с продаж. Рассмотрите регрессию (1) в таблице 12.1. Как, по вашему мнению, в этом случае локальный средний эффект воздействия в эксперименте отличается от среднего эффекта воздействия в эксперименте? Объясните.

## Компьютерные упражнения

E13.1. Потенциальный работодатель получает два резюме: резюме от белого кандидата и аналогичное резюме от афроамериканца. Действительно ли работодатель с большей вероятностью пригласит белого кандидата на собеседование, нежели чернокожего? Марианна Берtrand (Marianne Bertrand) и Сэндхил Маллэннатан (Sendhil Mullainathan) провели случайный контролируемый эксперимент, чтобы ответить на этот вопрос. Поскольку расовая принадлежность обычно не включается в резюме, они разделили резюме на основе имен, которые чаще

встречаются среди европеоидной расы (например, Эмили Уолш или Грегори Бейкер), и имен, которые чаще встречаются у афроамериканцев (например, Лакиша Вашингтон или Джамал Джонс). В итоге была составлена большая база данных «дополненных» расовой принадлежностью резюме (на основе описанного выше принципа). Эти резюме были отправлены потенциальным работодателям для определения того, какие из резюме вызовут обратную реакцию со стороны работодателей (приглашение на собеседование). Данные, полученные в ходе этого эксперимента, и их подробное описание представлены на сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) в файлах Names и Names\_Description<sup>1</sup>.

- a) Вычислите «коэффициент реакции» работодателя как долю тех резюме, которые вызвали реакцию работодателя в виде приглашения работника на собеседование, в общем количестве резюме. Каков был коэффициент реакции для белых? А для афроамериканцев? Постройте 95 %-й доверительный интервал для разности коэффициентов реакции этих групп. Является ли эта разность статистически значимой? Является ли она большой в житейском смысле?
  - б) Является ли разность коэффициентов реакции для белых и афроамериканцев различной для мужчин и женщин?
  - в) Каково различие в коэффициентах реакции для высококвалифицированных и низкоквалифицированных работников? Существует ли подобное различие для белых? Существует ли подобное различие для афроамериканцев? Является ли подобное различие статистически значимым?
  - г) Авторы работы утверждают, что расовая принадлежность приписывалась тому или иному резюме случайным образом. Обнаружили ли вы какое-либо подтверждение тому, что это не так?
- E13.2. Потребителю предоставляется возможность купить бейсбольную карточку за 1 долл., но он отказывается от предложения. Если же потребителю сейчас дать бейсбольную карточку, будет ли он готов продать ее за 1 долл.? Стандартная теория потребителя предполагает положительный ответ на этот вопрос, но поведенческая экономика утверждает, что наличие права «собственности» увеличивает ценность товаров для потребителей. То есть потребитель может захотеть продать товар несколько дороже, чем за 1 долл. (например за 1,20 долл.), несмотря на то что он был готов заплатить только менее 1 долл. (например 0,88 долл.) при ее покупке. Поведенческая экономика называет это явление «эффектом владения». Джон Лист (John List) исследовал этот

---

<sup>1</sup> Эти представленные профессором университета Чикаго (University of Chicago) Марианной Берtrand (Marianne Bertrand) данные использовались в ее совместной с Сэндхилом Маллннатаном (Sendhil Mullainathan) работе «Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination», American Economic Review, 2004, 94 (4): 991–1013.

эффект с помощью случайного эксперимента с участием коллекционеров спортивных сувениров во время выставки спортивных карточек. Торговцам случайным образом вручался один из двух коллекционных спортивных сувениров, например сувениры *A* и *B*, имеющие приблизительно одинаковую рыночную стоимость<sup>1</sup>. Коллекционеры, получившие сувенир *A*, имели возможность обменять его на сувенир *B* с человеком, проводившим эксперимент. Коллекционеры, получившие сувенир *B*, имели возможность обменять его на сувенир *A* с организатором эксперимента. Данные, полученные в эксперименте, и их подробное описание представлены на веб-сайте учебника [http://www.pearsonhighered.com/stock\\_watson/](http://www.pearsonhighered.com/stock_watson/) в файлах *Sportscards* и *Sportscards\_Description*<sup>2</sup>.

- а) i. Предположим, что в отсутствие эффекта владения все индивиды предпочитают сувенир *A* сувениру *B*. Чему равна доля коллекционеров, участвующих в эксперименте, которые будут менять свой сувенир на другой? (*Подсказка:* случайное распределение сувениров означает, что приблизительно 50 % коллекционеров получили сувенир *A*, а другие 50 % – сувенир *B*.)
  - ii. Предположим, что в отсутствие эффекта владения 50 % коллекционеров предпочитают сувенир *A* сувениру *B*, а вторые 50 % предпочитают сувенир *B* сувениру *A*. Чему равна доля коллекционеров, участвующих в эксперименте, которые будут менять свой сувенир на другой?
  - iii. Предположим, что в отсутствие эффекта владения  $X\%$  коллекционеров предпочитают сувенир *A* сувениру *B*, а другие  $(100-X)\%$  предпочитают сувенир *B* сувениру *A*. Покажите, что можно ожидать, что 50 % коллекционеров будут обменивать свой сувенир на другой.
- б) На основе имеющихся данных оцените, какая доля коллекционеров обменивала свой сувенир на другой? Сильно ли эта доля отличается от 50 %? Какая доля коллекционеров, получивших сувенир *A*, обменивала его на сувенир *B*? Какая доля коллекционеров, получивших сувенир *B*, обменивала его на сувенир *A*? Можно ли говорить о наличии свидетельств эффекта владения?
- в) Обсудим вопрос о том, что эффект владения имеет место, но постепенно исчезает по мере того, как торговцы приобретают дополнительный опыт в торговле. Пусть половина коллекционеров, участвующих в эксперименте, являлись профессиональными торговцами (про-

<sup>1</sup> Сувенир *A* представлял собой корешок билета от игры, в которой Кэл Рипкен-младший (Cal Ripken, Jr.) установил рекорд по количеству последовательно сыгранных матчей, а сувенир *B* был сувениром 300-го выигрышного матча Нолана Райан (Nolan Ryan).

<sup>2</sup> Данные, предоставленные профессором Университета Чикаго (University of Chicago) Джоном Листом (John List), использовались в его работе «Does Market Experience Eliminate Market Anomalies», Quarterly Journal of Economics, 2003, 118 (1): 41–71.

- фессионалы), а вторая половина ими не являлась (любители). Существенно ли различалось их поведение? Повторите пункт (б) для профессионалов и любителей. Согласуются ли полученные результаты с мнением о том, что эффект владения исчезает по мере приобретения опыта?
- г) База данных содержит два дополнительных показателя, характеризующих опыт коллекционеров: количество торговых сделок в течение месяца и количество лет ведения торговой деятельности. Имеют ли место свидетельства того, что для любителей эффект владения снижается по мере приобретения опыта торговли?

## Приложения

### *Приложение 13.1. База данных проекта STAR*

Открытая база данных проекта STAR содержит данные по результатам учебных тестов, экспериментальным группам, характеристикам учеников и преподавателей для всех четырех лет проведения эксперимента, начиная с 1985–1986 учебного года, заканчивая 1988–1989 учебным годом. Анализируемые в настоящей главе данные по результатам тестов представляют собой сумму результатов тестов по математике и чтению в рамках теста SAT (Stanford Achievement Test). Бинарная переменная, характеризующая пол ученика в таблице 13.2, равна единице, если ученик является мальчиком, и равна нулю в случае, если ученик – девочка. Бинарные переменные, характеризующие расовую принадлежность (афроамериканец) и иную расовую принадлежность (не белый и не черный), отражают расовую принадлежность ученика. Бинарная переменная, характеризующая право на получение бесплатного обеда, отражает возможность получения школьником бесплатного обеда в течение соответствующего года обучения в школе. Преподавательский стаж учителя представляет собой суммарное количество лет преподавания того учителя, который преподавал в тот год, когда ученики писали тест. Данные также отражают, какую школу посещал ребенок в том или ином году, что дает возможность создания дополнительных бинарных переменных для отдельных школ.

### *Приложение 13.2. Оценка метода инструментальных переменных при наличии различных причинных эффектов у индивидов*

В данном приложении будут выведены значения предела по вероятности 2МНК-оценки (13.12) при наличии неоднородности в выборке относительно эффекта влияния в эксперименте и влияния инструментальных переменных

на участие в эксперименте. В частности, предполагается, что выполняются основные предположения модели регрессии с инструментальными переменными, представленные во вставке «Основные понятия 12.4», за исключением того, что равенства (13.9) и (13.11) выполняются при неоднородных эффектах. Далее предположим, что  $\pi_{0i}$ ,  $\pi_{1i}$ ,  $\beta_{0i}$  и  $\beta_{1i}$  распределены независимо от  $u_i$ ,  $v_i$  и  $Z_i$ , а также то, что  $E(u_i|Z_i)=E(v_i|Z_i)=0$  и  $E(\pi_{1i})\neq 0$ .

Поскольку  $(X_i, Y_i, Z_i)$ ,  $i=1, \dots, n$  являются независимыми одинаково распределенными с конечными первыми четырьмя моментами, то по закону больших чисел из вставки «Основные понятия 2.6» получаем:

$$\hat{\beta}_1^{TSLS} = \frac{S_{ZY}}{S_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}} \quad (13.13)$$

(см. приложение 3.3 и упражнение 17.2). Таким образом, основная задача заключается в получении выражений для  $\sigma_{ZY}$  и  $\sigma_{ZX}$  в терминах моментов  $\pi_{1i}$  и  $\beta_{1i}$ . Можно записать так:  $\sigma_{ZX}=E[(Z_i-\mu_Z)(X_i-\mu_X)]=E[(Z_i-\mu_Z)X_i]$ . Подставляя выражение (13.11) в выражение для  $\sigma_{ZX}$ , получаем:

$$\begin{aligned} \sigma_{ZX} &= E[(Z_i - \mu_Z)(\pi_{0i} + \pi_{1i}Z_i + v_i)] = \\ &= E[\pi_{0i}] \times 0 + E[\pi_{1i}Z_i(Z_i - \mu_Z)] + \text{cov}(Z_i, v_i) = \sigma_z^2 E(\pi_{1i}), \end{aligned} \quad (13.14)$$

где второе равенство следует из равенства  $\text{cov}(Z_i, v_i)=0$  [которое следует из предположения о том, что  $E(v_i|Z_i)=0$ ; см. уравнение (2.27)], поскольку  $E[(Z_i - \mu_Z)\pi_{0i}] = E\{E[(Z_i - \mu_Z)\pi_{0i}|Z_i]\} = E[(Z_i - \mu_Z)E(\pi_{0i}|Z_i)] = E(Z_i - \mu_Z)E(\pi_{0i})$  (используется закон повторного математического ожидания и предположение о том, что  $\pi_{0i}$  не зависит от  $Z_i$ ) и поскольку  $E[\pi_{1i}Z_i(Z_i - \mu_Z)] = E\{E[\pi_{1i}Z_i(Z_i - \mu_Z)|Z_i]\} = E(\pi_{1i})E[Z_i(Z_i - \mu_Z)] = \sigma_z^2 E(\pi_{1i})$  (используется закон повторного математического ожидания и предположение о том, что  $\pi_{1i}$  и  $Z_i$  не зависят друг от друга).

Далее рассмотрим  $\sigma_{ZY}$ . Подстановка выражения (13.11) в выражение (13.9) приводит к  $Y_i = \beta_{0i} + \beta_{1i}(\pi_{0i} + \pi_{1i}Z_i + v_i) + u_i$ , поэтому

$$\begin{aligned} \sigma_{ZY} &= E[(Z_i - \mu_Z)Y_i] = \\ &= E[(Z_i - \mu_Z)(\beta_{0i} + \beta_{1i}\pi_{0i} + \beta_{1i}\pi_{1i}Z_i + \beta_{1i}v_i + u_i)] = \\ &= E(\beta_{0i}) \times 0 + \text{cov}(Z_i, \beta_{1i}\pi_{0i}) + E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] + \\ &\quad + E[\beta_{1i}v_i(Z_i - \mu_Z)] + \text{cov}(Z_i, u_i). \end{aligned} \quad (13.15)$$

Поскольку  $(\beta_{1i}\pi_{0i})$  и  $Z_i$  независимо распределены, то  $\text{cov}(Z_i, \beta_{1i}\pi_{0i})=0$ ; поскольку  $\beta_{1i}$  распределен независимо от  $v_i$  и  $Z_i$ , а  $E(v_i|Z_i)=0$ , то  $E[\beta_{1i}v_i(Z_i - \mu_Z)] = E[\beta_{1i}]E[v_i(Z_i - \mu_Z)] = 0$ ; так как  $E(u_i|Z_i)=0$ ,  $\text{cov}(u_i, Z_i)=0$ , и так как  $\beta_{1i}$  и  $\pi_{1i}$  распределены независимо от  $Z_i$ , получаем:  $E[\beta_{1i}\pi_{1i}Z_i(Z_i - \mu_Z)] = \sigma_z^2 E(\beta_{1i}\pi_{1i})$ . Таким образом, финальное выражение в уравнении (13.15) дает

$$\sigma_{ZY} = \sigma_z^2 E(\beta_{1i}\pi_{1i}). \quad (13.16)$$

Подстановка выражений (13.14) и (13.16) в выражение (13.13) дает

$$\hat{\beta}_1^{TSLS} \xrightarrow{p} \frac{\sigma_Z^2 E(\beta_{1i}\pi_{1i})}{\sigma_Z^2 E(\pi_{1i})} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})},$$

что и является конечным результатом, представленным в выражении (13.12).

### **Приложение 13.3. Методология потенциальных исходов для анализа экспериментальных данных**

В данном приложении представлен математический обзор методологии потенциальных исходов, которая обсуждалась в разделе 13.1. Методология потенциальных исходов совместно с постоянным эффектом влияния в эксперименте приводит к модели регрессии (13.1). Если процедура распределения субъектов между экспериментальной и контрольной группами является случайной и зависит от наблюдаемых переменных, то методология потенциальных исходов приводит к модели (13.2) и предположению об условной независимости среднего. Рассмотрим бинарную переменную, отражающую эффект воздействия в эксперименте, которая равна:  $X_i = 1$  для экспериментальной группы.

Пусть  $Y_i(1)$  означает  $i$ -й индивидуальный потенциальный исход, если индивид находится в экспериментальной группе (эксперимент оказывает влияние),  $Y_i(0)$  – потенциальный исход, если индивид находится в контрольной группе. Тогда эффект воздействия эксперимента можно вычислить как  $Y_i(1) - Y_i(0)$ . Поскольку индивид либо оказывается под влиянием эксперимента, либо нет, то только один из описанных выше потенциальных исходов является наблюдаемым. Наблюдаемый исход  $Y_i$  связан с потенциальными исходами следующим образом:

$$Y_i = Y_i(1)X_i + Y_i(0)(1 - X_i). \quad (13.17)$$

Если некоторые индивиды получают экспериментальное воздействие, а некоторые – нет, то ожидаемая разность в наблюдаемых исходах между двумя группами равна  $E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0]$ . Это равенство выполняется вне зависимости от того, что собой представляет влияние эксперимента, и просто говорит о том, что ожидаемая разность равна разности средних исходов в эксперименте для тех, кто получил воздействие в эксперименте, и средних исходов в эксперименте для тех, кто не получил воздействие. Если, кроме того, индивиды случайным образом приписываются к экспериментальной и контрольной группам, то  $X_i$  распределены независимо от всех характеристик индивидов, то есть независимо от  $[Y_i(1), Y_i(0)]$ . При случайному распределении между группами средняя разность между контрольной и экспериментальной группами равна:

$$E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = E[Y_i(1)X_i = 1] - E[Y_i(0)|X_i = 0] = E[Y_i(1) - Y_i(0)], \quad (13.18)$$

где во втором равенстве используется тот факт, что  $[Y_i(1), Y_i(0)]$  распределены независимо от  $X_i$  при случайному распределении, а также линейность математического ожидания [уравнение (2.28)]. Таким образом, если  $X_i$  распределяется между группами случайным образом, то разность средних экспериментальных исходов между двумя группами представляет собой средний эффект в эксперименте в рассматриваемой генеральной совокупности, из которой были выбраны субъекты.

Методология потенциальных исходов может быть непосредственным образом представлена в обозначениях, используемых в регрессионном анализе в учебнике. Пусть  $u_i = Y_i(0) - E[Y_i(0)]$ , обозначим  $E[Y_i(0)] = \beta_0$ . Также обозначим  $Y_i(1) - Y_i(0) = \beta_{1i}$ , тогда  $\beta_{1i}$  представляет собой эффект влияния в эксперименте для индивида  $i$ . Используя выражение (13.17), можно показать:

$$\begin{aligned} Y_i &= Y_i(1)X_i + Y_i(0)(1-X_i) = Y_i(0) + [Y_i(1) - Y_i(0)]X_i = \\ &= E[Y_i(0)] + [Y_i(1) - Y_i(0)]X_i + \{Y_i(0) - E[Y_i(0)]\} = \\ &= \beta_0 + \beta_{1i}X_i + u_i. \end{aligned} \quad (13.19)$$

Таким образом, начав со взаимосвязи между наблюдаемыми и ненаблюдаемыми исходами с помощью простой смены обозначений, можно получить модель регрессии со случайными коэффициентами, представленную в (13.9). [В уравнении (13.9)  $\beta_0$  различен для различных индивидов, но это эквивалентно уравнению (13.19), поскольку  $u_i$  также различна для различных индивидов.] Если  $X_i$  назначается случайным образом, то  $X_i$  распределены независимо от  $[Y_i(1), Y_i(0)]$  и, следовательно, независимо от  $\beta_{1i}$  и  $u_i$ . Если эффект влияния в эксперименте постоянен, то  $\beta_{1i} = \beta_1$ , а уравнение (13.9) превращается в уравнение (13.1).

Как обсуждалось в приложении 7.2 и в разделах 13.1 и 13.3, в некоторых случаях  $X_i$  назначается случайным образом на основе третьей переменной  $W_i$ . Если  $W_i$  и потенциальные исходы не являются независимыми, то в общем случае средняя разность между группами не равна среднему эффекту в эксперименте, то есть равенство (13.18) не выполняется. Тем не менее случайное распределение  $X_i$  на основе  $W_i$  подразумевает, что  $X_i$  и  $[Y_i(1), Y_i(0)]$  не зависят друг от друга условно относительно  $W_i$ . Это условие – независимость  $X_i$  и  $[Y_i(1), Y_i(0)]$  при заданных  $W_i$  – часто называется несмешиваемостью.

Если эффект влияния эксперимента не меняется между индивидами и если  $E[Y_i | X_i, W_i]$  является линейной функцией, то несмешиваемость предполагает независимость условного среднего ошибки регрессии в уравнении (13.2). Для того чтобы это продемонстрировать, допустим, что  $Y_i(0) = \beta_0 + \gamma W_i + u_i$ , где  $\gamma$  – причинное влияние (если оно есть)  $W_i$  на  $Y_i(0)$ , и пусть  $Y_i(1) - Y_i(0) = \beta_1$  (постоянный эффект влияния в эксперименте). Согласно логике, которая привела к уравнению (13.19), получим:  $Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + u_i$ , что эквивалентно уравнению (13.2). Тогда

$E(u_i | X_i, W_i) = E[Y_i(0) - \beta_0 - \gamma W_i | X_i, W_i] = E[Y_i(0) - \beta_0 - \gamma W_i | W_i] = E(u_i | W_i)$ , где второй знак равенства следует из условия несмешиваемости (если  $[Y_i(1), Y_i(0)]$  не зависит от  $X_i$  при заданном  $W_i$ , то  $E[Y_i(0)|X_i, W_i] = E[Y_i(0)|W_i]$ ). Таким образом, несмешиваемость позволяет показать, что  $E[u_i | X_i, W_i] = E[u_i | W_i]$  в (13.2). Аргументы, используемые в приложении 7.2, означают, что если  $E[u_i | W_i]$  – линейная функция от  $W_i$ , то МНК-оценка  $\beta_1$  в уравнении (13.2) является несмещенной, несмотря на то что в общем случае МНК-оценка  $\gamma$  является смещенной, поскольку  $E(u_i | W_i) \neq 0$ .

Часть IV

РЕГРЕССИОННЫЙ  
АНАЛИЗ  
ЭКОНОМИЧЕСКИХ  
ВРЕМЕННЫХ РЯДОВ



# **Глава 14. Введение в модели временных рядов и прогнозирование**

Временные ряды представляют собой данные, собранные для одного показателя в различные моменты времени; они могут использоваться, чтобы ответить на количественные вопросы, для которых межобъектные выборки не являются адекватными. Одним из таких вопросов может быть вопрос о том, какое влияние на интересующую нас переменную  $Y$  оказывают изменения во времени, происходящие с другой переменной  $X$ ? Иными словами, какой динамический причинный эффект оказывают на  $Y$  изменения в  $X$ ? Например, какое влияние имеет на показатель смертности в автокатастрофах закон, требующий от пассажиров пристегиваться ремнями безопасности во время езды на автомобиле, как сразу после принятия закона, так и впоследствии? Другим таким вопросом является вопрос о наилучшем прогнозе будущих значений некоторой переменной. К примеру, какой из наших прогнозов показателя инфляции, ставки процента или цен акций в следующем месяце является лучшим? Ответы на оба вопроса (первый – о причинных динамических эффектах, второй – об экономических прогнозах) могут быть получены при использовании данных, имеющих структуру временных рядов. Но такие данные предполагают наличие некоторых специальных проблем, преодоление которых требует знания некоторых новых методов.

В главах 14–16 вводится технический аппарат для эконометрического анализа временных рядов и рассматриваются приложения этого аппарата для решения проблем прогнозирования и оценки динамических причинных эффектов. В главе 14 описываются базовые понятия и инструменты регрессионного анализа временных рядов и их приложения к экономическому прогнозированию. В главе 15 понятия и инструменты, рассмотренные в предыдущей главе, применяются для решения проблем оценки динамических причинных эффектов при использовании временных рядов. В главе 16 обсуждаются некоторые более сложные вопросы, касающиеся анализа временных рядов, включая прогнозирование многомерных временных рядов и моделирование изменений волатильности во времени.

К эмпирическим проблемам, изучаемым в данной главе, можно отнести прогнозирование инфляции, то есть процентного роста уровня цен. Несмотря на то что прогнозирование является всего лишь приложением регрессионного анализа, оно довольно сильно отличается от оценки причинных эффектов,

на которых мы концентрировали свое внимание до сих пор. Как обсуждается в разделе 14.1, модели, являющиеся полезными для целей прогнозирования, не обязательно имеют причинную интерпретацию: если мы видим пешеходов, несущих зонты, мы можем предположить (спрогнозировать), что будет дождь, даже несмотря на то что зонт не является причиной дождя. В разделе 14.2 вводятся некоторые базовые понятия анализа временных рядов и приводятся некоторые примеры экономических временных рядов. В разделе 14.3 представлены регрессионные модели временных рядов, в которых в качестве объясняющих переменных рассматриваются прошлые значения зависимой переменной; такие «авторегрессионные» модели используют исторические значения инфляции для прогнозирования ее будущих значений. Часто прогнозы, базирующиеся на авторегрессиях, могут быть улучшены, если добавить в них в качестве regressоров дополнительные переменные и их прошлые значения или «лаги» («запаздывания»). Подобные так называемые авторегрессионные модели с распределенными лагами рассматриваются в разделе 14.4. Например, мы находим, что при прогнозе инфляции можно использовать запаздывающие значения показателя уровня безработицы в дополнение к запаздывающим значениям инфляции, то есть мы рассматриваем прогнозы, базирующиеся на эмпирической кривой Филлипса, и такие модели дают лучшие прогнозы по сравнению с авторегрессионными моделями. На практике мы должны решить, сколько запаздываний нужно включать в авторегрессии и авторегрессионные модели с распределенными лагами, а в разделе 14.5 описываются методы, позволяющие принять такое решение.

Предположение о том, что будущие значения будут похожи на прошлые, является настолько важным в регрессионном анализе временных рядов, что получило собственное название – «стационарность». Временные ряды могут оказаться нестационарными по двум причинам: (1) временные ряды могут иметь устойчивую долгосрочную динамику, то есть тренды; (2) теоретическая регрессия может быть неустойчивой во времени, то есть в ней могут быть сдвиги. Такие отклонения от стационарности подвергают опасности прогнозы и гипотезы, лежащие в основе регрессионного анализа временных рядов. К счастью, существуют статистические процедуры для определения трендов и сдвигов, позволяющие скорректировать спецификацию модели. Такие процедуры описаны в разделах 14.6 и 14.7.

## **14.1. Использование регрессионных моделей для прогнозирования**

В главах 4–9 в качестве примера эмпирического приложения эконометрических методов мы рассматриваем оценку причинного влияния соотношения учеников и учителей (*STR*) на результаты тестов (*TestScore*). Простейшая регрессионная модель, оцененная в главе 4, показывает, что результаты тестов следующим образом зависят от соотношения учеников и учителей:

$$\widehat{\text{TestScore}} = 989,9 - 2,28 \times \text{STR}. \quad (14.1)$$

Как обсуждалось в шестой главе, окружной школьный инспектор, планирующая нанять больше учителей для того, чтобы уменьшить размеры классов, не сочла бы это уравнение очень полезным. Оцененный коэффициент наклона в уравнении (14.1) не позволяет получить полезные оценки причинного эффекта влияния соотношения учителей и учеников на результаты тестов, потому что возможно наличие смещения оценок, вызванного вероятным пропуском переменных, являющихся характеристиками школы и учеников и, с одной стороны, влияющими на результаты тестов, а с другой — коррелированными с соотношением учеников и учителей.

В противоположность этому, как обсуждалось в главе 9, родители, рассматривающие переезд в другой школьный округ, могут найти уравнение (14.1) более полезным. Хотя коэффициент наклона не имеет причинной интерпретации, регрессия может помочь родителям получить прогноз результатов тестов в районе, для которого они не являются доступными. Вообще говоря, регрессионная модель может быть полезна для прогнозирования, даже если ни один из коэффициентов не имеет причинной интерпретации. Исходя из перспектив прогнозирования, самым важным является то, чтобы модель давала настолько точные прогнозы, насколько это возможно. И хотя не существует такой вещи, как совершенный прогноз, регрессионные модели могут, тем не менее, давать прогнозы, являющиеся аккуратными и заслуживающими доверия.

Приложения в данной главе отличаются от проверки предсказаний относительно зависимости результатов тестов от размеров классов, потому что в ней данные, имеющие структуру временных рядов, используются для прогнозирования будущих событий. Например, родителю действительно интересно знать, какие результаты будут в следующем году после того, как его или ее ребенок поступит в школу. Конечно, эти тесты еще не сданы, так что родители должны прогнозировать их результаты, используя доступную в настоящий момент информацию. Если доступны результаты тестов за последние несколько лет, то это хорошая стартовая точка для прогнозирования будущих результатов тестов на основе их текущих и прошлых значений. Такие доводы приводят нас непосредственно к авторегрессионным моделям, представленным в разделе 14.3, в которых прошлые значения переменной используются в линейных регрессиях для прогнозирования будущих значений временного ряда. Следующим шагом, рассмотренным в разделе 14.4, является расширение этой модели посредством включения в нее дополнительной объясняющей переменной — размера класса. Так же, как и уравнение (14.1), такая регрессионная модель может позволить получить аккуратные и заслуживающие доверия прогнозы, даже если ее коэффициенты не имеют причинной интерпретации. В главе 15 мы возвращаемся к проблемам, похожим на те, с которыми сталкивается окружной школьный инспектор, и обсудим оценку причинных эффектов во временных рядах.

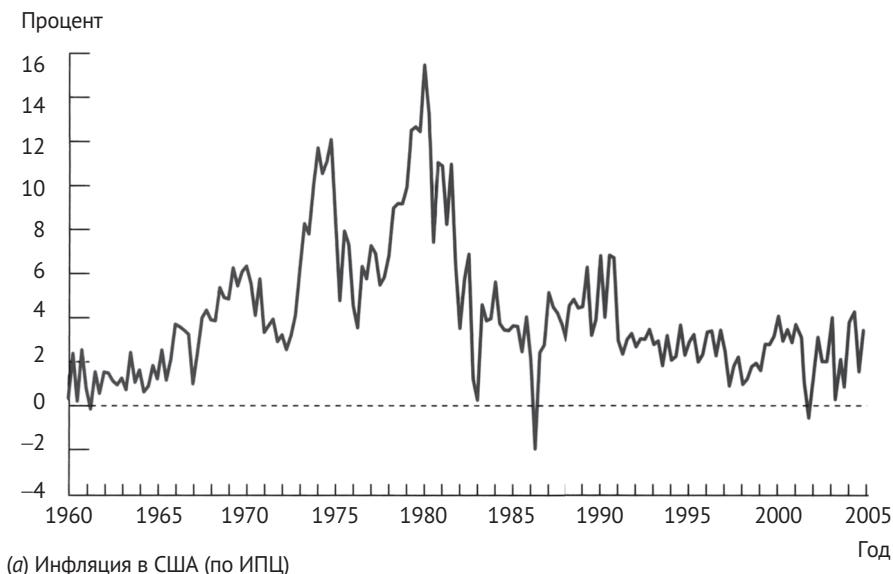
## 14.2. Введение во временные ряды и серийную корреляцию

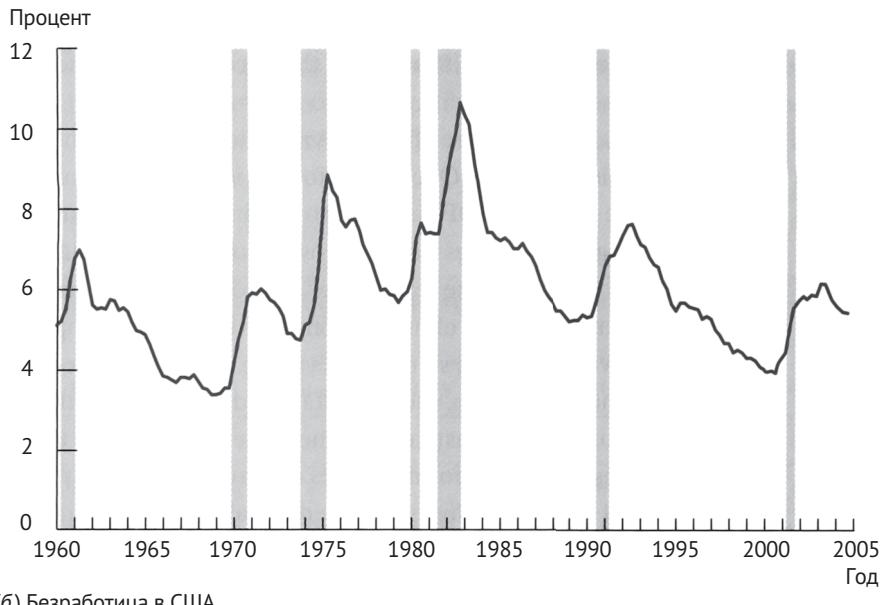
В данном разделе вводятся некоторые базовые понятия и определения, возникающие в эконометрике временных рядов. Хорошой стартовой точкой для анализа временных рядов является рассмотрение графиков имеющихся данных, с чего мы и начнем данный раздел.

### *Инфляция и безработица в Соединенных Штатах*

На рисунке 14.1а изображен график американской инфляции – годовое процентное изменение уровня цен в США, рассчитанное как индекс потребительских цен (ИПЦ), – с 1 квартала 1960 года по 4 квартал 2004 года (данные описаны в приложении 14.1). Инфляция была низкой в 1960-х годах, выросла в течение 1970-х до уровня послевоенного пика в 15,5 % в первом квартале 1980 года (т.е. в январе, феврале и марте 1980 года) и затем упала менее чем до 3 % к концу 1990-х годов. Как можно видеть на рисунке 14.1а, инфляция может изменяться на величину, достигающую 1 % или более, в течение одного квартала.

Уровень американской безработицы – доля неработающих американцев, взятая из Текущего обследования населения США (Current Population Survey, см. приложение 3.1), – представлен на рисунке 14.1б. Изменения уровня безработицы в основном связаны с бизнес-циклами в США. Например, уровень безработицы возрастал в течение периодов рецессии в 1960–1961, 1970, 1974–1975 годах, двойной рецессии в 1980 и 1981–1982 годах и в течение рецессий 1990–1991 и 2001 годов, то есть в течение периодов, обозначенных тенью на рисунке 14.1б.





(б) Безработица в США

Рисунок 14.1. Инфляция и безработица в США, 1960–2004 годы

Инфляция потребительских цен в США (рис. 14.1а) росла с 1960 по 1980 год и затем падала в начале 1980-х годов.  
Уровень безработицы в США (рис. 14.1б) рос во время рецессий (выделенные тенью периоды)  
и падал в периоды роста.

### *Запаздывания, первые разности, логарифмы и темпы прироста*

Наблюдение временного ряда  $Y$ , сделанное в момент времени  $t$ , обозначают  $Y_t$ , а общее число наблюдений обозначают  $T$ . Интервал между наблюдениями, то есть период времени между  $t$  и  $t+1$ , является некоторой единицей времени, такой как неделя, месяц, квартал или год. Например, данные по американской инфляции, изучаемые в этой главе, являются квартальными, так что единица времени («период») есть квартал.

Мы используем специальную терминологию и обозначения, чтобы обозначить будущие и прошлые значения  $Y$ . Значение  $Y$  в предыдущий период называется его первым запаздывающим значением или, более просто, его *первым запаздыванием (лагом)* и обозначается  $Y_{t-1}$ . Его  $j$ -м запаздывающим значением (или просто – его  $j$ -м лагом) является его значение  $j$  периодов назад, то есть  $Y_{t-j}$ . Аналогично,  $Y_{t+1}$  обозначает значение  $Y$  на один период вперед.

Изменением значения  $Y$  между периодами  $t-1$  и  $t$  является величина  $Y_t - Y_{t-1}$ ; это изменение называется *первой разностью* переменной  $Y_t$ . В данных, имеющих структуру временных рядов, символ « $\Delta$ » используется для обозначения первых разностей, так что  $\Delta Y_t = Y_t - Y_{t-1}$ .

Экономические временные ряды часто анализируются после логарифмирования или после вычисления изменений их логарифмированных значений. Одной из причин этого является то, что многие экономические временные ряды, такие как валовой внутренний продукт (ВВП), имеют приблизительно

экспоненциальный рост, то есть в долгосрочном периоде временной ряд растет в среднем на определенный процент в год; таким образом, логарифм временного ряда растет почти линейно. Другой причиной является то, что стандартное отклонение многих экономических временных рядов почти пропорционально их уровням, то есть стандартное отклонение довольно точно выражается как процент уровня временного ряда; таким образом, стандартное отклонение логарифма временного ряда приблизительно равно константе. В любом случае, полезно преобразовывать временные ряды так, чтобы изменения в преобразованных рядах были пропорциональными изменениями (или процентами) исходных временных рядов, а это можно получить, прологарифмировав ряды<sup>1</sup>.

## ОСНОВНЫЕ ПОНЯТИЯ

### 14.1

#### Запаздывания, первые разности, логарифмы и темпы прироста

Первое запаздывание временного ряда  $Y_t$  – это  $Y_{t-1}$ ; его  $j$ -е запаздывание –  $Y_{t-j}$ .

Первая разность временного ряда,  $\Delta Y_t$ , это его изменение между периодами  $t-1$  и  $t$ , то есть  $\Delta Y_t = Y_t - Y_{t-1}$ .

Первая разность логарифма  $Y_t$  есть  $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$ .

Процентное изменение временного ряда  $Y_t$  между периодами  $t-1$  и  $t$  приблизительно равно  $100 \Delta \ln(Y_t)$ , причем это приближение более точно, когда процентное изменение мало.

Определения запаздываний, первых разностей и темпов прироста представлены во вставке «Основные понятия 14.1».

Понятия запаздываний, первых разностей и темпов прироста проиллюстрированы на примере временного ряда инфляции в США в таблице 14.1. В первом столбце таблицы показана дата (или период), где первый квартал 2004 года обозначен 2004: I, второй квартал 2004 года – 2004: II, и так далее. Второй столбец содержит данные об индексе потребительских цен в США в рассматриваемом квартале; третий столбец показывает, чему был равен показатель инфляции. Например, во втором квартале 2004 года ИПЦ вырос до 188,60% по сравнению со значением 186,57% в первом квартале 2004 года. Процентный рост составил тогда  $100 \times (188,60 - 186,57) / 186,57 = 1,09\%$ . Это число есть процентный рост во втором квартале по сравнению с первым. Тогда довольно удобно представлять инфляцию (и другие экономические показатели) в годо-

<sup>1</sup> Изменение логарифма переменной приблизительно равно пропорциональному изменению этой переменной, то есть  $\ln(X+a) - \ln(X) \cong a/X$ , где приближение работает тем лучше, чем  $a/X$  меньше [см. уравнение 8.16 и сопровождающее его обсуждение]. Теперь заменим  $X$  на  $Y_{t-1}$  и  $a$  на  $\Delta Y_t$  и заметим, что  $Y_t = Y_{t-1} + \Delta Y_t$ . Это означает, что пропорциональное изменение временного ряда  $Y_t$  между периодами  $t-1$  и  $t$  приблизительно равно  $\ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_{t-1} + \Delta Y_t) - \ln(Y_{t-1}) \cong \Delta Y_t / Y_{t-1}$ . Выражение  $\ln(Y_t) - \ln(Y_{t-1})$  есть первая разность  $\ln(Y_t)$ , то есть  $\Delta \ln(Y_t)$ . Таким образом,  $\Delta \ln(Y_t) \cong \Delta Y_t / Y_{t-1}$ . Процентное изменение равно 100-кратному пропорциональному изменению временного ряда  $Y_t$  и приблизительно равно  $100 \Delta \ln(Y_t)$ .

вом исчислении, считая, что инфляция (в процентах) будет расти тем же самым темпом, что и в текущем квартале. Поскольку в году четыре квартала, инфляция в годовом измерении во втором квартале 2004 года составит  $1,04 \times 4 = 4,36$ , или 4,4 % в год.

Таблица 14.1

## Инфляция в США в 2004 году и первом квартале 2005 года

Квартал	ИПЦ	Уровень инфляции в годовом исчислении ( $Inf_t$ )	Первое запаздывание ( $Inf_{t-1}$ )	Изменение уровня инфляции ( $\Delta Inf_t$ )
2004: I	186,57	3,8	0,9	2,9
2004: II	188,60	4,4	3,8	0,6
2004: III	189,37	1,6	4,4	-2,8
2004: IV	191,03	3,5	1,6	1,9
2005: I	192,17	2,4	3,5	-1,1

Примечание. Инфляция в годовом исчислении представляет собой процентное изменение ИПЦ за один квартал (по сравнению с предыдущим кварталом), умноженное на четыре. Первое запаздывание инфляции – это ее значение в предыдущем квартале; изменение инфляции – это разность между текущим значением инфляции и ее значением в предыдущий квартал. Все значения в 3–5 столбцах округлены до десятых.

Это процентное изменение также может быть вычислено приближенно, используя первые разности логарифмов (см. вставку «Основные понятия 14.1»). Разность логарифмов ИПЦ в первом-втором кварталах 2004 года составит  $\ln 188,60 - \ln 186,57 = 0,0108$ , а соответствующее приближенное значение квартального прироста тогда равно  $100 \times 0,0108 = 1,08\%$ . Следовательно, в годовом исчислении получаем  $1,08 \times 4 = 4,32$ , или 4,3 %, что практически равно полученному выше годовому приросту. Эти вычисления могут быть записаны в таком виде:

$$\begin{aligned} \text{Инфляция в годовом исчислении} &= Inf_t \cong 400[\ln(IPI_t) - \\ &\quad - \ln(IPI_{t-1})] = \\ &= 400\Delta\ln(IPI_t), \end{aligned} \quad (14.2)$$

где  $IPI_t$  – значение индекса потребительских цен в период  $t$ . Множитель, равный 400, переводит разность логарифмов из долей в проценты (умножаем на 100) и из квартального изменения в годовое (умножаем на 4).

В двух последних колонках таблицы 14.1 приведены значения первых запаздываний инфляции и ее первых разностей. Первое запаздывание инфляции во втором квартале 2004 года равно 3,8 % и совпадает со значением инфляции в первом квартале 2004 г. Изменение инфляции во втором квартале 2004 года по сравнению с первым кварталом составило  $4,4\% - 3,8\% = 0,6\%$ .

**Автокорреляция**

Во временных рядах значение  $Y$  в один период обычно коррелирует со своими значениями в другие периоды. Корреляцию временного ряда со своими собственными запаздывающими значениями называют *автокорреляцией*, или *серийной корреляцией*. Первая автокорреляция (или *коэффициент автокорреляции*, или *автокорреляция первого порядка*) – это корреляция между  $Y_t$  и  $Y_{t-1}$ , то есть корреляция между двумя соседними значениями  $Y$ . Автокорреляция

второго порядка равна корреляции между  $Y_t$  и  $Y_{t-2}$ , и автокорреляция  $j$ -го порядка есть автокорреляция между  $Y_t$  и  $Y_{t-j}$ . Аналогично, автоковариацией порядка  $j$  называется автоковариация между  $Y_t$  и  $Y_{t-j}$ . Формальные определения автокорреляции и автоковариации приведены во вставке «Основные понятия 14.2».

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**14.2**

**Автокорреляция (серийная корреляция)  
и автоковариация**

Автоковариацией порядка  $j$  временного ряда  $Y_t$  называется автоковариация между  $Y_t$  и его  $j$ -м лагом<sup>1</sup>, а коэффициентом автокорреляции  $j$ -го порядка называется коэффициент корреляции между  $Y_t$  и  $Y_{t-j}$ . То есть:

$$\text{Автоковариация порядка } j = \text{cov}(Y_t, Y_{t-j}). \quad (14.3)$$

$$\begin{aligned} \text{Автокорреляция}^2 \text{ порядка } j &= \rho_j = \text{corr}(Y_t, Y_{t-j}) = \\ &= \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t) \text{ var}(Y_{t-j})}}. \end{aligned} \quad (14.4)$$

Коэффициент автокорреляции  $j$ -го порядка иногда называют коэффициентом серииной корреляции  $j$ -го порядка.

Теоретические значения автоковариаций и автокорреляций  $j$ -го порядка, рассмотренные во вставке «Основные понятия 14.2», могут быть оценены с использованием выборочных автоковариаций и автокорреляций  $j$ -го порядка, то есть  $\widehat{\text{cov}}(Y_t, Y_{t-j})$  и  $\widehat{\rho}_j$ :

$$\widehat{\text{cov}}(Y_t, Y_{t-j}) = \frac{1}{T} \sum_{t=j+1}^T (Y_t - \bar{Y}_{j+1, T})(Y_{t-j} - \bar{Y}_{1, T-j}), \quad (14.5)$$

$$\widehat{\rho}_j = \frac{\widehat{\text{cov}}(Y_t, Y_{t-j})}{\widehat{\text{var}}(Y_t)}, \quad (14.6)$$

где  $\bar{Y}_{j+1, T}$  обозначает выборочное среднее значение  $Y_t$ , вычисленное по наблюдениям в моменты времени  $t = j+1, \dots, T$ , а  $\widehat{\text{var}}(Y_t)$  есть выборочная дисперсия  $Y$ <sup>1</sup>.

Первые четыре значения выборочной автокорреляции показателя инфляции и ее первых разностей представлены в таблицы 14.2. Эти значения показывают, что инфляция сильно положительно автокоррелирована: ее первая

<sup>1</sup> Автоковариацию  $j$ -го порядка часто обозначают  $\gamma_j$  и называют автоковариационной функцией. – Примеч. науч. ред. перевода.

<sup>2</sup>  $\rho_j$  также часто называют автокорреляционной функцией. – Примеч. науч. ред. перевода.

<sup>1</sup> Сумма в уравнении (14.5) делится на  $T$  несмотря на то, что в обычной формуле для выборочной ковариации [см. уравнение (3.24)] сумма делится на число наблюдений, скорректированное на число степеней свободы. Формула (14.5) является стандартной для вычисления автоковариаций. В уравнении (14.6) используется предположение о том, что  $\text{var}(Y_t)$  и  $\text{var}(Y_{t-j})$  совпадают, – следствие предположения о стационарности  $Y$ , которое обсуждается в разделе 14.4.

автокорреляция равна 0,84. Выборочные автокорреляции убывают по мере увеличения лага (или глубины запаздывания), но продолжают оставаться довольно большими даже для запаздывания на четыре квартала. Изменение инфляции (первые разности) отрицательно автокоррелированы: увеличение темпа инфляции в одном квартале влечет за собой его снижение в следующем.

На первый взгляд может показаться, что эти два обстоятельства (положительная автокорреляция уровней инфляции и отрицательная для их изменений) противоречат друг другу. Однако эти две автокорреляции отражают разные свойства. Сильная положительная автокорреляция в уровнях инфляции показывает долгосрочные тенденции в поведении инфляции (см. рисунок 14.1): инфляция была низкой в первом квартале 1965 года, а затем и во втором; но за высоким значением в первом квартале 1981 года следовало высокое значение во втором квартале. Напротив, отрицательная автокорреляция первых разностей инфляции означает, что в среднем за ростом инфляции в одном квартале последует ее снижение в следующем.

Таблица 14.2

**Первые четыре значения выборочной автокорреляции уровня инфляции в США и ее изменений, I квартал 1960 года – IV квартал 2004 года**

Автокорреляция:		
Лаг	Уровень инфляции ( $Inf_t$ )	Изменение уровня инфляции ( $\Delta Inf_t$ )
1	0,84	-0,26
2	0,76	-0,25
3	0,76	0,29
4	0,67	-0,06

### *Другие примеры экономических временных рядов*

Экономические временные ряды очень разнообразны. На рисунке 14.2 представлено четыре примера экономических временных рядов: ставка процента межбанковского кредитования США (ставка процента по федеральным фондам); обменный курс между долларом США и британским фунтом; логарифм ВВП Японии и ежедневные доходности индекса S&P500.

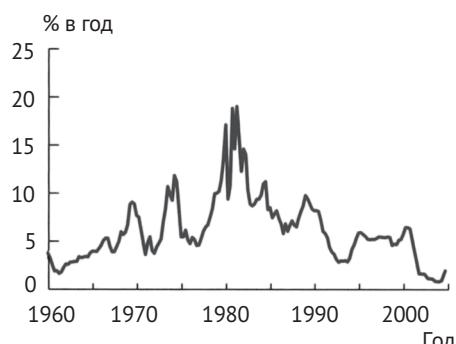
Ставка процента по федеральным фондам США (рис. 14.2а) представляет собой ставку процента, которую банки платят друг другу по займам овернайт. Эта ставка является важной, потому что она контролируется Федеральным резервным банком США и является основным инструментом его монетарной политики. Если вы сравните графики межбанковской ставки процента и уровней безработицы и инфляции, изображенных на рисунке 14.1, вы увидите, что резкий рост межбанковской ставки процента часто сопровождается последующей рецессией.

Обменный курс британского фунта к доллару США (рис. 14.2б) – это цена британского фунта (£), измеренная в американских долларах. До 1972 года развитые экономики придерживались политики фиксированных обменных курсов – так

называемая «Бреттон-Вудская» система, когда правительства государств-участников поддерживали валютные курсы неизменными. В 1972 году из-за сильной инфляции эта система была отменена, вследствие чего курсы основных валют стали «плавающими», то есть их значения определялись спросом и предложением валют на валютном рынке. До 1972 года обменный курс был практически неизменным, за исключением единственной девальвации 1968 года, когда официальная стоимость фунта по отношению к доллару упала до 2,40 долл. С 1972 года обменный курс менялся довольно сильно на рассматриваемом периоде времени.

Квартальный ВВП Японии (рис. 14.2в) представляет собой общую стоимость товаров и услуг, произведенных в Японии в течение рассматриваемого квартала. Показатель ВВП является широко распространенной мерой общей экономической активности. Логарифм этого временного ряда представлен на рисунке 14.2в, и изменения в нем могут быть интерпретированы как темп прироста. В течение 1960-х годов и в первой половине 1970-х годов японский ВВП рос быстро, но этот рост замедлился во второй половине 1970–1980-х годов. Еще более медленный рост продолжился в 1990-х: среднегодовой рост ВВП составил 1,2 % в течение периода с 1990 по 2004 год.

Сводный фондовый индекс Нью-Йоркской фондовой биржи является общим индексом, характеризующим стоимость компаний, акции которых торгуются на Нью-Йоркской фондовой бирже. Рисунок 14.2г показывает ежедневные процентные изменения этого индекса за торговые дни с 2 января 1990 по 11 ноября 2005 года (всего 4003 наблюдения). В отличие от других временных рядов, изображенных на рисунке 14.2, здесь мы видим очень низкую серийную корреляцию между дневными процентными изменениями: если бы корреляция присутствовала, мы могли бы предсказывать изменения, используя значения доходностей в предыдущие дни, и зарабатывать деньги, покупая акции, если бы мы ожидали рост на рынке, и продавая их в противном случае. Хотя доходности являются по существу непредсказуемыми, на графике 14.2г видна модель их волатильности. Например, стандартное отклонение дневных процентных изменений было относительно большим в 1990–1991 и в 1998–2003 годах и относительно маленьким в 1995 и 2005 годах. Такая «кластеризованная волатильность» обнаруживается у многих финансовых временных рядов, и эконометрические методы для моделирования этого специального типа гетероскедастичности мы рассмотрим в разделе 16.5.



(а) Межбанковская ставка процента в США



(б) Обменный курс британского фунта к американскому доллару

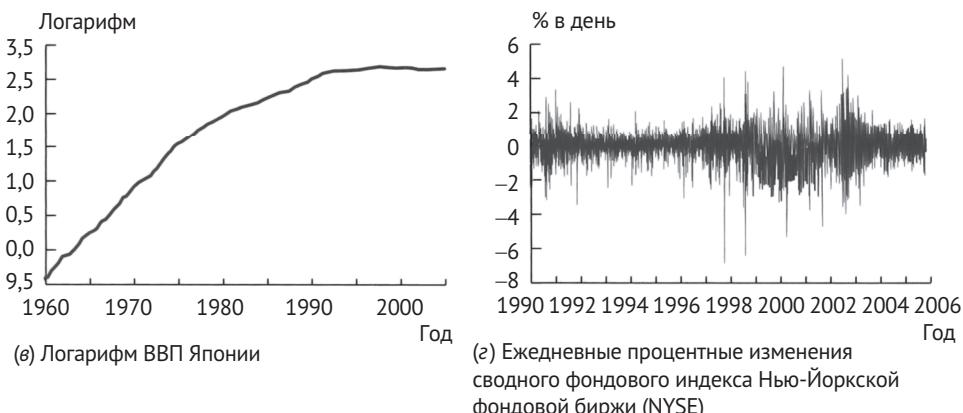


Рисунок 14.2. Четыре экономических временных ряда

Четыре временных ряда заметно различаются. График ставки процента по федеральным фондам (рис. 14.2а) похож на график ценовой инфляции. Обменный курс между долларом США и британским фунтом (рис. 14.2б) показывает дискретные изменения после раз渲ала Бреттон-Вудской системы фиксированных обменных курсов в 1972 году. Логарифм ВВП Японии (рис. 14.2в) демонстрирует относительно гладкий рост, хотя его темп и замедляется в 1970-х годах, а потом снова в 1990-х. Дневные изменения фондового индекса Нью-Йоркской фондовой биржи (рис. 14.2г) являются по существу непредсказуемыми, но их дисперсия меняется: этот временной ряд демонстрирует так называемую кластеризованную волатильность.

### 14.3. Авторегрессии

Каким будет уровень ценовой инфляции – процентный прирост уровня цен в следующем году? Инвесторы с Уолл-стрит доверяют прогнозам инфляции, когда принимают решения о том, сколько облигаций (и за какую цену) покупать. Экономисты центральных банков, например Федерального резервного банка США, используют прогнозы инфляции, когда принимают решения о том, какую монетарную политику проводить. Фирмы используют прогнозы инфляции при прогнозировании объемов продаж своей продукции, а местные органы власти учитывают прогнозы инфляции при составлении своих бюджетов на следующий год. В этом разделе мы рассмотрим прогнозы, которые можно получить, используя модель авторегрессии – регрессионной модели, в которой переменная связывается со своими прошлыми значениями.

#### *Модель авторегрессии первого порядка*

Если вы хотите предсказать будущее значение временного ряда, то хорошей отправной точкой может служить ближайшее прошлое. Например, если вы хотите получить прогноз изменения инфляции в следующем квартале, вы можете посмотреть, как росла и падала инфляция в последнем квартале. Один из способов предсказать изменение инфляции,  $\Delta Inf_t$ , используя изменение в предыдущем квартале,  $\Delta Inf_{t-1}$ , – просто оценить методом наименьших квадратов регрессию  $\Delta Inf_t$  на  $\Delta Inf_{t-1}$ . Используя данные с 1962 по 2004 год, получаем:

$$\widehat{\Delta Inf}_t = 0,017 - 0,238 \Delta Inf_{t-1}, \quad (14.7)$$

где, как обычно, в скобках под оценками коэффициентов приведены их стандартные ошибки, и  $\widehat{\Delta Inf}_t$  – это предсказанное значение  $\Delta Inf_t$ , полученное на основе оцененной регрессии. Модель из уравнения (14.7) называется моделью авторегрессии первого порядка: авторегрессии, потому что эта регрессия представляет зависимость только от своего запаздывания,  $\Delta Inf_{t-1}$ , а первого порядка, потому что в ней используется только первое запаздывание. Коэффициент в уравнении (14.7) отрицательный, поэтому за ростом инфляции в одном квартале следует ее снижение в следующем. Авторегрессия первого порядка обычно обозначается как AR(1), где «1» обозначает ее порядок (первый). Теоретическая модель авторегрессии первого порядка AR(1) для временного ряда  $Y_t$  имеет вид:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t, \quad (14.8)$$

где  $u_t$  – случайная ошибка.

**Прогнозы и ошибки прогнозирования.** Предположим, что у вас есть исторические данные некоторого временного ряда  $Y$  и вы хотите предсказать его будущие значения. Если  $Y$  является авторегрессией первого порядка AR(1), описываемой уравнением (14.8), и  $\beta_0$  и  $\beta_1$  известны, то прогноз  $Y_{T+1}$ , основанный на  $Y_T$ , равен  $\beta_0 + \beta_1 Y_T$ .

На практике коэффициенты  $\beta_0$  и  $\beta_1$  неизвестны, поэтому прогнозы основываются на их оценках. Мы будем использовать МНК-оценки  $\hat{\beta}_0$  и  $\hat{\beta}_1$ , которые можно получить, используя исторические данные. В общем случае  $\hat{Y}_{T+1|T}$  будет обозначать прогноз значения  $Y_{T+1}$ , основанный на информации до периода  $T$  включительно и полученный исходя из оценок регрессии по имеющимся на момент  $T$  данным. Следовательно, прогноз по модели авторегрессии первого порядка AR(1) из уравнения (14.8) есть

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T, \quad (14.9)$$

где  $\hat{\beta}_0$  и  $\hat{\beta}_1$  – оценки коэффициентов, полученные на основе исторических данных до момента  $T$  включительно.

**Ошибка прогнозирования** – это разность между фактическим значением  $Y_{T+1}$  и его прогнозом, полученным с использованием  $Y_T$ :

$$\text{Ошибка прогнозирования} = Y_{T+1} - \hat{Y}_{T+1|T}. \quad (14.10)$$

**Прогнозы и предсказанные значения.** Прогноз не является предсказанным на основе метода наименьших квадратов значением  $Y$ , а ошибка прогнозирования не является остатком МНК-регрессии. Предсказанные по МНК значения вычисляются для наблюдений внутри выборки с использованием оценки регрессии. Напротив, прогноз строится для некоторой даты, находящейся за пределами имеющихся данных, используемых для оценки регрессии. Таким образом, прогнозируемые значения зависимой переменной не принадлежат выборке, по которой оценивается регрессия. Аналогично, МНК-остатки – это разность между фактическим значением  $Y$  и его предсказанным по МНК-регрессии значением. В то же время ошибка прогнозирования – это разность между будущим значением  $Y$ , которое не включено в исходную выборку, и прогнозом этого будущего значения. Говоря иначе, прогнозы и ошибки прогнозирования относятся к вневыборочным наблюдениям, а предсказанные значения и остатки – к внутривыборочным наблюдениям.

**Квадратный корень из среднеквадратичной ошибки прогнозирования.** Квадратный корень из среднеквадратичной ошибки прогнозирования (RMSFE) – это характеристика размера ошибки прогнозирования, то есть величина типичной ошибки, получаемой по прогнозной модели. RMSFE – корень квадратный из среднего значения величины, равной сумме квадратов ошибок прогнозирования:

$$\text{RMSFE} = \sqrt{E[(Y_{T+1} - \hat{Y}_{T+1|T})^2]} . \quad (14.11)$$

У RMSFE есть два источника ошибки: это ошибка, возникающая из-за того, что будущие значения случайного остатка  $u_t$  неизвестны, и ошибки оценок коэффициентов  $\beta_0$  и  $\beta_1$ . Если первый источник ошибки больше, чем второй, что может быть в случае наличия большой выборки данных, то RMSFE хорошо приближается теоретическим значением стандартного отклонения авторегрессии случайной ошибки  $u_t$ , то есть  $\sqrt{\text{var}(u_t)}$  [см. уравнение (14.8)]. Стандартное отклонение  $u_t$  может быть оценено как стандартная ошибка регрессии (см. раздел 4.3). Таким образом, если неопределенность относительно коэффициентов регрессии достаточно мала, чтобы это обстоятельство можно было проигнорировать, RMSFE может быть оценена с использованием стандартной ошибки регрессии. Оценка RMSFE, включающая оба источника неопределенности ошибки прогнозирования, будет рассмотрена в разделе 14.4.

**Пример: инфляция в США.** Чему равен прогноз инфляции на I квартал 2005 года (2005: I), который прогнозист сделал бы в IV квартале 2004 года, основываясь на модели авторегрессии первого порядка AR(1), оценки которой представлены уравнением (14.7), которое было оценено на основе массива данных до IV квартала 2004 года включительно? Из таблицы 14.1 следует, что инфляция в IV квартале 2004 года составила 3,5% (т.е.  $\text{Inf}_{2004:\text{IV}} = 3,5\%$ ) и выросла на 1,9% по сравнению с III кварталом 2004 года (т.е.  $\Delta\text{Inf}_{2004:\text{IV}} = 1,9$ ). Подставляя эти значения в уравнение (14.7), получаем, что прогноз изменения уровня инфляции в I квартале 2005 года по сравнению с IV кварталом 2004 года равен:  $\widehat{\Delta\text{Inf}}_{2005:\text{I}|2004:\text{IV}} = 0,017 - 0,238 \times \Delta\text{Inf}_{2004:\text{IV}} = 0,017 - 0,238 \times 1,9 = -0,43 \cong -0,4$ . Тогда прогноз инфляции есть сумма последнего значения инфляции и ее прогнозируемого изменения:

$$\widehat{\text{Inf}}_{T+1|T} = \text{Inf}_T + \widehat{\Delta\text{Inf}}_{T+1|T} . \quad (14.12)$$

Так как  $\text{Inf}_{2004:\text{IV}} = 3,5\%$  и ее прогнозируемое изменение в следующем квартале равно  $-0,4$ , то прогнозируемая инфляция в первом квартале 2005 года равна:  $\widehat{\text{Inf}}_{2005:\text{I}|2004:\text{IV}} = \text{Inf}_{2004:\text{IV}} + \widehat{\Delta\text{Inf}}_{2005:\text{I}|2004:\text{IV}} = 3,5\% - 0,4\% = 3,1\%$ . Таким образом, авторегрессионная модель первого порядка AR(1) прогнозирует снижение инфляции с 3,5% в IV квартале 2004 года до 3,1% в I квартале 2005 года.

Насколько хорош такой прогноз? Фактическое значение инфляции в I квартале 2005 года составило 2,4% (см. таблицу 14.1), следовательно, AR(1)-прогноз завышен на 0,7%, то есть ошибка прогнозирования равна  $-0,7$ .  $\bar{R}^2$  в AR(1)-модели из уравнения (14.7) – равен лишь 0,05, следовательно, запаздывающее значение инфляции объясняет очень малую долю дисперсии инфляции в выборке, используемой для оценки авторегрессии. Такой низкий  $\bar{R}^2$  вполне согласуется с плохим прогнозом инфляции в I квартале 2005 года, полученным по уравнению (14.7). Говоря общими словами, низкий  $\bar{R}^2$  предполагает, что эта AR(1)-модель будет прогнозировать очень небольшую долю дисперсии изменения инфляции.

Стандартная ошибка регрессии (14.7) равна 1,65. Игнорируя неопределенность, возникающую из-за ошибок в оценках коэффициентов, оценка *RMSFE* для прогнозов, полученных по уравнению (14.7), равна, таким образом, 1,65 %.

### **Модель авторегрессии порядка $p$**

В модели AR(1) для прогнозирования  $Y_t$  используется  $Y_{t-1}$  и при этом игнорируется полезная информация из более отдаленного прошлого. Одним из способов учесть такую информацию являются модели авторегрессии порядка  $p$ , AR( $p$ ), которые включают в себя в качестве регрессов более глубокие лаги, чем первый.

**Модель авторегрессии порядка  $p$  [AR( $p$ )-модель]** представляет  $Y_t$  как линейную функцию от своих  $p$ -запаздываний. То есть в AR( $p$ )-модели в качестве регрессоров присутствуют  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$  плюс свободный член. Число запаздываний  $p$ , включенных в AR( $p$ )-модель, называют порядком (или глубиной запаздывания) авторегрессии.

Например, модель AR(4) изменения инфляции включает четыре запаздывания изменения инфляции в качестве регрессоров. Оценка МНК-регрессии на интервале 1962–2004 годов дает следующие результаты:

$$\widehat{\Delta Inf}_t = 0,02 - 0,26 \Delta Inf_{t-1} - 0,32 \Delta Inf_{t-2} + 0,16 \Delta Inf_{t-3} - 0,03 \Delta Inf_{t-4}. \quad (14.13)$$

Коэффициенты при трех последних лагах в уравнении (14.13) значимо отличаются от нуля на 5 %-м уровне значимости:  $F$ -статистика равна 6,91 ( $p$ -значение < 0,01). Этот факт также подтверждается увеличением  $\bar{R}^2$  с 0,05 в AR(1)-модели до 0,18 в модели AR(4). Аналогично, SER в AR(4)-модели в уравнении (14.13) равна 1,52, что лучше, чем SER в модели AR(1), которая равна 1,65.

Определение авторегрессии порядка  $p$  приведено во вставке «Основные понятия 14.3».

### **ОСНОВНЫЕ ПОНЯТИЯ 14.3**

#### **Авторегрессии**

Модель авторегрессии порядка  $p$  [AR( $p$ )-модель] представляет  $Y_t$  как линейную функцию от  $p$  своих запаздываний:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + u_t, \quad (14.14)$$

где  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$ . Число запаздываний  $p$  называется порядком или глубиной запаздывания авторегрессии.

**Свойства прогнозов и случайных ошибок в модели AR( $p$ ).** Предположение о том, что условное математическое ожидание случайной ошибки  $u_t$  относительно прошлых значений  $Y_t$  равно нулю [т.е.  $E(u_t | Y_{t-1}, \dots, Y_{t-2}, \dots) = 0$ ], имеет два важных следствия.

Первое следствие заключается в том, что лучший прогноз  $\hat{Y}_{T+1}$ , основанный на своей собственной истории, зависит только от  $p$  своих прошлых значений. Более конкретно, пусть  $\hat{Y}_{T+1|T} = E(Y_{T+1}|Y_T, Y_{T-1}, \dots)$  обозначает условное математическое ожидание  $Y_{T+1}$  относительно своей собственной истории. Тогда  $\hat{Y}_{T+1|T}$  имеет наименьшую RMSFE среди всех прогнозов, базирующихся на информации об  $Y$  (упражнение 14.5). Если  $Y_t$  описывается моделью AR( $p$ ), то лучшим прогнозом  $\hat{Y}_{T+1}$ , основанном на  $Y_T, Y_{T-1}, \dots$ , является:

$$\hat{Y}_{T+1|T} = \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \dots + \beta_p Y_{T-p+1}, \quad (14.15)$$

что следует из модели авторегрессии порядка  $p$  (см. уравнение 14.14) и предположения о том, что  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$ .

Вторым следствием является серийная некоррелированность ошибок  $u_t$ , что следует из уравнения (2.27) (упражнение 14.5).

**Пример: инфляция в США.** Чему равен прогноз инфляции на I квартал 2005 года (2005: I), полученный по AR(4)-модели инфляции, представленной уравнением (14.13)? Чтобы вычислить этот прогноз, подставим численные значения изменений инфляции в каждом из четырех кварталов 2004 года в уравнение (14.13):  $\widehat{\Delta Inf}_{2005:1|2004:IV} = 0,02 - 0,26\Delta Inf_{2004:IV} - 0,32\Delta Inf_{2004:III} + 0,16\Delta Inf_{2004:II} - 0,03\Delta Inf_{2004:I} = 0,02 - 0,26 \times 1,9 - 0,32 \times (-2,8) + 0,16 \times 0,6 - 0,03 \times 2,9 \cong 0,4$ , где значения изменений инфляции в 2004 году взяты из последнего столбца таблицы 14.1.

Соответствующий прогноз инфляции в I квартале 2005 года – это значение инфляции в четвертом квартале 2004 года плюс прогнозируемое изменение, то есть  $3,5\% + 0,4\% = 3,9\%$ . Ошибка прогнозирования – это разность фактического значения инфляции ( $2,4\%$ ) и ее прогноза ( $3,9\%$ ):  $2,4\% - 3,9\% = -1,5\%$ , что больше по абсолютному значению, чем ошибка прогнозирования по AR(1)-модели, которая была равна  $-0,7$  процентных пунктов.



### *Можем ли мы перехитрить рынок? Часть 1*

Мечтали ли вы когда-нибудь быстро разбогатеть, удачно сыграв на фондовой бирже? Если вы думаете, что рынок будет расти, то вам следует покупать акции сегодня и продавать их завтра до того момента, когда они начнут дешеветь. Если вы умеете хорошо прогнозировать колебания цен на фондовой бирже, активная торговая стратегия принесет большую прибыль, чем пассивная стратегия «покупать и держать», когда вы покупаете акции и держите их на руках. Хитрость, конечно, заключается в наличии надежного прогноза будущих доходностей на фондовом рынке.

Прогнозы, основанные на прошлых значениях доходностей, иногда называются «инерционными»: если значение цены выросло в этом месяце, то, возможно, оно инерционно и также будет расти в следующем месяце. Если так, то доходности будут автокоррелированными и авторегрессионные модели дадут хорошие прогнозы. Вы можете использовать такую инерционную стратегию для конкретных акций или для фондового индекса, который является некоторой мерой стоимости активом на рынке.

Таблица 14.3

## Авторегрессионные модели избыточной доходности, 1960:1–2002:12

Зависимая переменная: избыточная доходность взвешенного индекса CRSP			
	(1)	(2)	(3)
Спецификация	AR(1)	AR(2)	AR(3)
Регрессоры			
• избыточная доходность <sub>t-1</sub>	0,050 (0,051)	0,053 (0,051)	0,054 (0,051)
• избыточная доходность <sub>t-2</sub>		-0,053 (0,048)	-0,054 (0,048)
• избыточная доходность <sub>t-3</sub>			0,009 (0,050)
• избыточная доходность <sub>t-4</sub>			-0,016 (0,047)
Константа	0,312 (0,197)	0,328 (0,199)	0,331 (0,202)
F-статистика для лагов избыточной доходности ( <i>p</i> -значение)	0,968 (0,325)	1,342 (0,261)	0,707 (0,587)
<i>R</i> <sup>2</sup>	0,0006	0,0014	-0,0022

*Примечание.* Показатель избыточной доходности измеряется в процентах в месяц. Данные описаны в приложении 14.1. Все регрессии оценены на интервале 1960:1–2002:12 ( $T = 516$  наблюдений) с более ранними значениями, используемыми как начальные значения лаговых переменных. Значения в строках – это оценки коэффициентов со стандартными ошибками в скобках. В двух последних строках даны значения F-статистики для проверки гипотезы о равенстве нулю коэффициентов при лаговых значениях избыточной доходности с ее *p*-значением и значение скорректированного коэффициента детерминации.

В таблице 14.3 представлены оценки авторегрессионных моделей избыточной доходности сводного взвешенного индекса CRPS на интервале с января 1960 по декабрь 2002 года. Ежемесячная избыточная доходность – это то, что вы зарабатываете (в процентах), покупая акции в конце предыдущего месяца и продавая их в конце рассматриваемого месяца, минус то, что вы могли бы заработать, если бы вложили эти деньги в казначейские векселя. Показатель доходности акций включает доходы (потери) от изменений цен плюс любые дивиденды, которые вы получаете в течение месяца. Данные описаны в приложении 14.1.

Грустно видеть, насколько плохи результаты, представленные в таблице 14.3. Коэффициент при первом запаздывании доходности в AR(1)-модели статистически незначим, и мы не можем отвергнуть нулевую гипотезу о том, что все коэффициенты при лагах доходности равны нулю в моделях AR(2) и AR(4). На самом деле скорректированный  $R^2$  одной из этих моделей отрицателен, а в двух других положителен, но практически не отличается от нуля, что предполагает не очень большую пользу от оцененных моделей с точки зрения прогнозирования.

Эти плохие результаты согласуются с гипотезой эффективности рынков, которая утверждает, что избыточную доходность нельзя предсказать, потому что цены акций уже включают всю доступную на текущий момент информацию о них. Причина проста: если участники рынка думают, что цены акций будут иметь положительную избыточную доходность в следующем месяце, то они будут покупать соответствующие акции сейчас. Но совершая это, участники рынка стимулируют рост цен на акции в точности до того уровня, когда они перестанут приносить избыточную доходность. В результате

мы не можем прогнозировать будущую избыточную доходность, используя только доступную информацию, и не можем сделать это, используя регрессии из таблицы 14.3.



## 14.4. Модели временных рядов с дополнительными переменными и авторегрессионные модели с распределенными лагами

Экономическая теория часто предполагает, что для прогнозирования конкретной переменной могут быть полезны и другие экономические показатели. Эти другие переменные могут быть добавлены в авторегрессию. Таким образом, мы получим модель временных рядов с несколькими регрессорами (предикторами). Когда другие переменные и их запаздывания добавляются в авторегрессию, мы получаем авторегрессионную модель с распределенными лагами.

### *Изменения прогноза уровня инфляции при использовании прошлых значений уровня безработицы*

Высокий уровень безработицы влечет за собой снижение уровня инфляции в будущем. Эта отрицательная связь, известная как краткосрочная кривая Филлипса, очевидна в представленном на рисунке 14.3 графике, на котором изображены годовые изменения уровня ценовой инфляции в зависимости от уровня безработицы в предыдущем году. Например, в 1982 году уровень инфляции составил 9,7%, а в следующем году снизился на 2,9%. В целом коэффициент корреляции между показателями, представленными на рисунке 14.3, равен  $-0,36$ .



**Рисунок 14.3. Диаграмма рассеяния изменений инфляции в году  $t+1$  по сравнению с годом  $t$  и уровня безработицы в году  $t$ , 1961–2004 годы**

В 1982 году уровень безработицы в США составил 9,7%, а уровень инфляции в 1983 году упал на 2,9% (большая точка). В целом за высокими значениями уровня безработицы в году  $t$  следует снижение уровня инфляции в следующем году  $t+1$  с коэффициентом корреляции, равным  $-0,36$ .

Диаграмма рассеяния, представленная на рисунке 14.3, предполагает, что прошлые значения уровня безработицы могут содержать информацию о будущем поведении цен, которой нет в прошлых изменениях инфляции. Догадка легко проверяется включением в модель AR(4) из уравнения (14.13) первого запаздывания уровня безработицы:

$$\widehat{\Delta Inf_t} = 1,28 - 0,31 \Delta Inf_{t-1} - 0,39 \Delta Inf_{t-2} + 0,09 \Delta Inf_{t-3} - \\ - 0,08 \Delta Inf_{t-4} - 0,21 Unemp_{t-1}. \quad (14.16)$$

$t$ -статистика коэффициента при первом запаздывании уровня безработицы (переменная  $Unemp_{t-1}$ ) равна  $-2,23$ , таким образом, этот показатель значим на 5%-м уровне значимости.  $R^2$  этой регрессии равен  $0,21$  и превышает  $R^2$  модели AR (4), равный  $0,18$ .

Прогноз изменения инфляции в первом квартале 2005 года получается подстановкой в уравнение (14.16) значений изменений инфляции в 2004 году и значения уровня безработицы в четвертом квартале 2004 года (который равен  $5,4\%$ ). Итоговый прогноз равен:  $\widehat{\Delta Inf}_{2005:I|2004:IV} = 0,4$ . Таким образом, прогноз инфляции на I квартал 2005 года равен:  $3,5\% + 0,4\% = 3,9\%$  и ошибка прогнозирования равна  $-1,5\%$ .

Если один лаг уровня безработицы улучшает прогноз инфляции, то несколько запаздываний могут оказаться еще более полезными; добавляя еще три запаздывания уровня безработицы, получаем:

$$\widehat{\Delta Inf_t} = 1,30 - 0,42 \Delta Inf_{t-1} - 0,37 \Delta Inf_{t-2} + 0,06 \Delta Inf_{t-3} - \\ - 0,04 \Delta Inf_{t-4} - 2,64 Unemp_{t-1} + 3,04 Unemp_{t-2} - \\ - 0,38 Unemp_{t-3} - 0,25 Unemp_{t-4}. \quad (14.17)$$

Значение  $F$ -статистики, при помощи которой проверяется гипотеза о равенстве нулю второго, третьего и четвертого запаздываний уровня безработицы, равно  $10,76$  ( $p$ -значение  $< 0,001$ ), следовательно, они совместно значимы.  $R^2$  регрессии (14.17) равен  $0,34$ , что является серьезным улучшением значения  $0,21$  для регрессии (14.16). Значение  $F$ -статистики для проверки гипотезы о значимости всех запаздываний уровня безработицы равно  $8,91$  ( $p$ -значение  $< 0,001$ ), что говорит о значимом улучшении модели AR (4), рассмотренной в разделе 14.3 [уравнение (14.13)]. Стандартная ошибка регрессии (14.17) равна  $1,36$ , что существенно лучше  $SER$  в модели AR(4), которая равна  $1,52$ .

Прогнозируемое изменение инфляции в I квартале 2005 года по сравнению с IV кварталом 2004 года может быть вычислено подстановкой значений переменных в уравнение (14.17). Уровень безработицы в I квартале 2004 года был равен  $5,7\%$ , во втором —  $5,6\%$ , в третьем и четвертом — по  $5,4\%$ . Прогноз изменения инфляции в I квартале 2005 года по сравнению с IV кварталом 2004 года в таком случае равен:

$$\widehat{\Delta Inf}_{2005:I|2004:IV} = 1,30 - 0,42 \times 1,9 - 0,37 \times (-2,8) + 0,06 \times 0,6 - \\ - 0,04 \times 2,9 - 2,66 \times 5,4 + 0,34 \times 5,4 - \\ - 0,38 \times 5,6 - 0,25 \times 5,7 = 0,1 . \quad (14.18)$$

Таким образом, прогноз инфляции в I квартале 2005 года составляет  $3,5\% + 0,1\% = 3,6\%$ . Ошибка прогнозирования в этом случае равна  $-1,2\%$ .

**Авторегрессионная модель с распределенными лагами.** Модели из уравнений (14.16) и (14.17) являются *авторегрессионными моделями с распределенными лагами (ADL)*: «авторегрессионными» потому, что в них в качестве объясняющих переменных включены запаздывания объясняющей переменной, и «с распределенными лагами», поскольку регрессия включает несколько запаздываний («распределенных лагов») дополнительного регрессора. В общем случае авторегрессионная модель с распределенными лагами с  $p$ -запаздываниями зависимой переменной  $Y_t$  и  $q$ -запаздываниями дополнительного регрессора называется моделью  $ADL(p, q)$ . В этих обозначениях модель (14.16) является моделью  $ADL(4,1)$ , а модель (14.17) – моделью  $ADL(4, 4)$ .

### Авторегрессионная модель с распределенными лагами

Авторегрессионная модель с распределенными лагами с  $p$ -запаздываниями переменной  $Y_t$  и  $q$ -запаздываниями переменной  $X_t$ , обозначаемая  $ADL(p, q)$ , это:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \\ + \delta_2 X_{t-2} + \dots + \delta_q X_{t-q} + u_t \quad (14.19)$$

где  $\beta_0, \beta_1, \dots, \beta_p, \delta_1, \dots, \delta_q$  – неизвестные коэффициенты и  $u_t$  – случайная ошибка с  $E(u_t | Y_{t-1}, Y_{t-2}, \dots, X_{t-1}, X_{t-2}, \dots) = 0$ .

## ОСНОВНЫЕ ПОНЯТИЯ

### 14.4

Определение авторегрессионной модели с распределенными лагами приведено во вставке «Основные понятия 14.4». Обозначение, используемое в уравнении (14.19), довольно громоздкое, поэтому в приложении 14.3 приводится альтернативная запись, в которой используется так называемый оператор запаздывания (лаговый оператор).

Предположение о том, что условное математическое ожидание случайной ошибки в ADL-модели относительно всех прошлых значений  $Y$  и  $X$  равно нулю, то есть что  $E(u_t | Y_{t-1}, Y_{t-2}, \dots, X_{t-1}, X_{t-2}, \dots) = 0$ , подразумевает, что в данную модель больше не нужно включать никаких дополнительных запаздываний. Другими словами, это означает, что длина лагов  $p$  и  $q$  является истинной длиной лагов и коэффициенты при дополнительных запаздываниях равны нулю.

ADL-модель включает запаздывания зависимой (авторегрессионная составляющая) и распределенные лаги единственной дополнительной

переменной  $X$ . Однако в общем случае прогнозы могут быть улучшены, если включить в регрессию дополнительные объясняющие переменные. Прежде чем перейти к обсуждению общего понятия множественной регрессии временных рядов, мы введем концепцию стационарности, которая будет использована в дальнейшем.

## Стационарность

Регрессионный анализ временных рядов использует прошлые данные, чтобы моделировать исторические взаимосвязи. Если будущее похоже на прошлое, то эти взаимоотношения могут быть использованы для прогнозирования будущего. Но если будущее серьезно отличается от прошлого, то взаимосвязи, оцененные на основе исторических данных, не очень полезны при прогнозировании будущего.

В контексте регрессионного анализа временных рядов идея о том, что исторические взаимосвязи могут быть обобщены на случай будущего, формализуется при помощи концепции *стационарности*. Определение стационарности приводится во вставке «Основные понятия 14.5» и фактически означает, что функция плотности временного ряда не изменяется во времени.

### ОСНОВНЫЕ ПОНЯТИЯ 14.5

#### Стационарность

Временной ряд  $Y_t$  называется *стационарным*, если его функция плотности не изменяется во времени, то есть если совместное распределение случайных величин  $(Y_{s+1}, Y_{s+2}, \dots, Y_{s+T})$  не зависит от  $s$  для любых  $T$ . В противном случае он называется *нестационарным*. Пара временных рядов  $X_t$  и  $Y_t$  называется *совместно стационарными*, если их совместное распределение  $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \dots, X_{s+T}, Y_{s+T})$  не зависит от  $s$  для любых  $T$ . Чтобы временной ряд был стационарным, необходимо, чтобы будущее было похоже на прошлое, по крайней мере в вероятностном смысле.

## Множественная регрессия временных рядов

Общая регрессионная модель временных рядов с несколькими объясняющими переменными является расширением авторегрессионной модели с распределенными лагами на случай включения в число регрессоров запаздываний нескольких переменных. Общий вид модели и предпосылки, лежащие в ее основе, приведены во вставке «Основные понятия 14.6». Присутствие в модели нескольких объясняющих переменных и их запаздываний приводит к необходимости использования двойных индексов в обозначениях регрессионных коэффициентов и регрессоров.

**Предположения регрессионной модели временных рядов.** Предпосылки, перечисленные во вставке «Основные понятия 14.6», являются модификациями

четырех предположений метода наименьших квадратов для множественной регрессии межобъектных данных (вставка «Основные понятия 6.4») для временных рядов.

Первое предположение говорит о равенстве нулю условного среднего ошибки регрессии  $u_t$ , относительно информационного множества, включающего все объясняющие переменные и дополнительные запаздывания всех регрессов, не включенные в модель. Эта предпосылка расширяет предположение, используемое в AR- и ADL-моделях, и предполагает, что лучший прогноз  $\hat{Y}_t$  может быть получен, используя только те запаздывания  $Y$  и  $X$ -ов, которые включены в выражение (14.20).

### Множественная регрессия временных рядов

В общем случае регрессионная модель временных рядов расширяется на случай  $k$  дополнительных регрессов и включает  $q_1$ -запаздывания первой объясняющей переменной,  $q_1$ -запаздывания второй объясняющей переменной и так далее, то есть:

$$\begin{aligned} Y_t = & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \delta_{11} X_{1t-1} + \\ & + \delta_{12} X_{1t-2} + \dots + \delta_{1q_1} X_{1t-q_1} + \dots + \delta_{k1} X_{kt-1} + \\ & + \delta_{k2} X_{kt-2} + \dots + \delta_{kq_k} X_{kt-q_k} + u_t, \end{aligned} \quad (14.20)$$

где

1.  $E(u_t | Y_{t-1}, Y_{t-2}, \dots, X_{1t-1}, X_{1t-2}, \dots, X_{kt-1}, X_{kt-2}, \dots) = 0$ .
2. (а) Случайные величины  $(Y_t, X_{1t}, \dots, X_{kt})$  имеют стационарное распределение;  
(б)  $(Y_t, X_{1t}, \dots, X_{kt})$  и  $(Y_{t-j}, X_{1t-j}, \dots, X_{kt-j})$  становятся независимыми при увеличении  $j$ .
3. Большие выбросы маловероятны: то есть  $X_{1t}, \dots, X_{kt}$  и  $Y_t$  имеют ненулевые, конечные четвертые моменты.
4. Отсутствует совершенная мультиколлинеарность.

### ОСНОВНЫЕ ПОНЯТИЯ

14.6

Второе предположение метода наименьших квадратов для межобъектных выборок (вставка «Основные понятия 6.4») говорит, что  $(X_{1i}, \dots, X_{ki}, Y_i)$ ,  $i = 1, \dots, n$  являются независимыми одинаково распределенными случайными величинами (i.i.d.). Во втором предположении для регрессии временных рядов предпосылка об i.i.d. заменяется на более подходящую и состоит из двух частей. Часть (а) – это предпосылка о том, что случайная выборка сделана из стационарного распределения, распределение данных сегодня не отличается от их распределения вчера. Это предположение является аналогом i.i.d. предпосылки для случая временных рядов: для межобъектных данных требование о том, чтобы каждая выборка была одинаково распределена, заменяется требованием о том, чтобы совместное распределение всех переменных, включая запаздывания, не менялось во времени. На практике очень многие временные ряды

не выглядят стационарными, что означает, что это предположение может не выполняться в приложениях. Если временной ряд нестационарен, то возникает много проблем: прогнозы могут быть смещены, неэффективны (могут существовать прогнозы, основанные на тех же данных, но с меньшей дисперсией) или обычновенные, основанные на МНК-оценках статистики (например тестирование гипотез), используя МНК  $t$ -статистики и сравнивая их с  $\pm 1,96$ , могут приводить к неверным выводам. Какая именно из этих проблем возникает и как ее исправить, зависит от источника (типа) нестационарности. В разделах 14.6 и 14.7 мы познакомимся с двумя основными типами нестационарности экономических временных рядов — трендами и сдвигами — и с соответствующими проблемами, тестированием этих проблем и методами их решения.

В части (б) второго предположения требуется, чтобы случайные величины становились независимо распределенными по мере удаления от конкретного момента времени на большое расстояние. Это требование заменяет предпосылку для межобъектных данных о том, что переменные являются независимыми друг от друга, на требование, что они были независимо распределены, когда они сильно отдалены друг от друга во времени. На это предположение иногда ссылаются как на *слабую зависимость* и оно гарантирует, что в больших выборках данные являются достаточно случайными, чтобы мы могли считать, что выполняются закон больших чисел и центральная предельная теорема. Мы не будем формулировать точное математическое утверждение, касающееся условий слабой зависимости: читатель может найти их в книге Хаяши (Hayashi, 2000. Ch. 2).

Третье предположение является аналогичным третьей предпосылке метода наименьших квадратов для межобъектных данных и говорит о том, что большие выбросы являются маловероятными и математически формализуются как предположение, что все переменные имеют ненулевые и конечные четвертые моменты.

И наконец, четвертое предположение также аналогично случаю межобъектных данных и говорит об отсутствии совершенной мультиколлинеарности.

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**14.7**

**Тестирование причинности по Грейндджеру  
(тестирование предсказательной способности)**

Статистика, используемая для тестирования причинности по Грейндджеру, является  $F$ -статистикой для проверки гипотезы о том, что все значения одной из объясняющих переменных в регрессионной модели (4.20) (например коэффициенты при  $X_{1t-1}, X_{1t-2}, \dots, X_{1t-q_1}$ ) равны нулю. Эта нулевая гипотеза предполагает, что эти регрессоры не имеют предсказательной силы для  $Y_t$ , в то время как у других регрессов она есть. Тестирование такой нулевой гипотезы называется тестированием причинности по Грейндджеру.

**Тестирование статистических гипотез и тест Грейнджера на причинность.**  
Если выполняются предположения из вставки «Основные понятия 14.6», то те-

стирование гипотез о коэффициентах регрессии с использованием МНК-оценок происходит точно так же, как и для случая межобъектных выборок.

Одним из полезных приложений  $F$ -статистики при прогнозировании временных рядов является проверка того, являются ли запаздывания одного из включенных регрессоров полезными с точки зрения прогнозирования объясняющей переменной вне зависимости от других объясняющих переменных, включенных в модель. Утверждение о том, что переменная не имеет прогнозной силы, равносильно нулевой гипотезе о равенстве нулю коэффициентов при всех запаздывающих значениях этой переменной.  $F$ -статистика, проверяющая эту гипотезу, называется *статистикой Грейнджера для проверки причинности*, а соответствующий тест – *тестом Грейнджера на причинность* (Granger, 1969). Основные моменты, связанные с этим тестом, представлены во вставке «Основные понятия 14.7».

Причинность по Грейнджеру не отражает смысла, который мы вкладываем в понятие причинности, используемое в этом учебнике. В первой главе мы определили причинность в терминах идеального случайного управляемого эксперимента, в котором различные значения переменной  $X$  получаются экспериментально и мы имеем возможность наблюдать реакцию переменной  $Y$  на это. Напротив, причинность по Грейнджеру означает, что если переменная  $X$  является по Грейнджеру причиной переменной  $Y$ , то  $X$  является полезным с точки зрения предсказания  $Y$  при фиксированных других переменных в регрессии. «Предсказуемость по Грейнджеру» является более аккуратным термином по сравнению с термином «причинность по Грейнджеру», который стал частью эконометрического жаргона.

В качестве примера рассмотрим отношение между изменением показателя инфляции и его прошлыми значениями и прошлыми значениями показателя безработицы. Основываясь на МНК-оценках регрессии (14.17), получаем, что  $F$ -статистика, используемая для тестирования нулевой гипотезы о том, что коэффициенты при всех четырех лагах показателя безработицы равны нулю, равна 8,91 ( $p$ -значение <0,001): используя жargon из вставки «Основные понятия 14.7», мы можем сделать вывод (на уровне значимости 1%) о том, что безработица является по Грейнджеру причиной изменений в инфляции. Это не обязательно означает, что изменение показателя безработицы будут вызывать – в терминах главы 1 – соответствующее изменение показателя безработицы. Это означает, что прошлые значения показателя безработицы представляются содержащими информацию, которая является полезной для прогнозирования изменений инфляции и выходит за рамки информации, содержащейся в прошлых значениях инфляции.

### ***Неопределенность прогнозирования и интервальные прогнозы***

Всякий раз, когда используется какой-либо метод оценивания, полезно приводить характеристики, измеряющие неопределенность получаемых оценок, и прогнозирование не является исключением. Одной из мер неопределенности прогнозов является квадратный корень из среднеквадратичной ошибки

прогнозирования. При дополнительном предположении о том, что случайные ошибки  $u_t$  являются нормально распределенными, RMSFE может быть использован для построения интервальных прогнозов, то есть интервала, который будущие значения переменной спрогнозирует с определенной вероятностью.

**Неопределенность прогнозирования.** Ошибка прогнозирования состоит из двух частей: неопределенности, возникающей при оценке коэффициентов прогнозирования, и неопределенности, связанной с неизвестными будущими значениями  $u_{t+1}$ . Для регрессий с небольшим числом коэффициентов и большим числом наблюдений неопределенность из-за будущих значений  $u_{t+1}$  может сильно превышать неопределенность, возникающую при оценках параметров. Однако в общем случае оба источника неопределенности являются важными, поэтому сейчас мы получим разложение RMSFE так, чтобы учесть оба источника неопределенности.

Сохраняя простые обозначения, рассмотрим прогнозы  $\hat{Y}_{T+1}$ , полученные по модели ADL(1, 1) с одним регрессором, то есть модель  $Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + u_t$ , и предположим, что  $u_t$  гомоскедастична. Тогда прогноз равен:  $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\delta}_1 X_T$ , а ошибка прогнозирования имеет вид:

$$Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - \left[ (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) X_T \right]. \quad (14.21)$$

Поскольку  $u_{T+1}$  имеет нулевое условное среднее и является гомоскедастичной,  $u_{T+1}$  имеет дисперсию, равную  $\sigma_u^2$ , и является некоррелированной с выражением в квадратных скобках в (14.21). Таким образом, среднеквадратичная ошибка прогнозирования (MSFE) равна:

$$\begin{aligned} \text{MSFE} &= E \left[ \left( Y_{T+1} - \hat{Y}_{T+1|T} \right)^2 \right] = \\ &= \sigma_u^2 + \text{var} \left[ (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) X_T \right] \end{aligned} \quad (14.22)$$

и RMSFE является корнем квадратным из MSFE.

Оценка MSFE включает в себя два слагаемых из уравнения (14.22). Как уже обсуждалось в разделе 14.3, первое слагаемое –  $\sigma_u^2$  – может быть оценено как квадрат стандартной ошибки регрессии. Чтобы оценить второе слагаемое, необходимо оценить дисперсию взвешенного среднего коэффициентов регрессии, и методы решения этой проблемы мы обсуждали в разделе 8.1 [см. обсуждение, следующее за уравнением 8.7].

В качестве альтернативного метода для оценки MSFE можно использовать псевдовневыборочные прогнозы, процедура расчета которых обсуждается в разделе 14.7.

**Интервальные прогнозы.** Интервальные прогнозы похожи на доверительные интервалы, за исключением того что они относятся к прогнозам. То есть 95 %-й интервальный прогноз представляет собой интервал, содержащий будущие значения временного ряда в 95 % повторяющихся выборок.

Важным отличием интервальных прогнозов от доверительных интервалов является то, что стандартная формула для 95 %-го доверительного интервала (оцен-

ка коэффициента  $\pm 1,96$  его стандартной ошибки) подтверждается центральной предельной теоремой и, следовательно, верна для широкого класса распределений остаточного члена. Напротив, из-за того что ошибка прогнозирования, заданная уравнением (14.21), включает будущие значения ошибки  $u_{T+1}$ , вычисление интервального прогноза требует либо оценивать распределение остаточного члена, либо делать некоторые предположения об этом распределении.

На практике удобно предполагать, что  $u_{T+1}$  является нормально распределенным. Если это так, то уравнение (14.21) и центральная предельная теорема, примененные к  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  и  $\hat{\delta}_1$ , означают, что ошибка прогнозирования является суммой двух независимых нормально распределенных слагаемых, так что ошибка прогнозирования сама является нормально распределенной с дисперсией, равной MSFE. Из этого следует, что 95 %-й доверительный интервал, рассчитываемый как  $\hat{Y}_{T+1|T} \pm 1,96SE(Y_{T+1} - \hat{Y}_{T+1|T})$ , где  $SE(Y_{T+1} - \hat{Y}_{T+1|T})$ , представляет собой оценку RMSFE.

Вышесказанное относилось к ситуации гомоскедастичности остаточного члена  $u_{T+1}$ . Если теперь  $u_{T+1}$  гетероскедастичен, то нам необходимо расширить модель на случай гетероскедастичности, так чтобы можно было оценить  $\sigma_u^2$  из уравнения (14.22) при известных прошлых значениях  $Y$  и  $X$ . Соответствующие методы моделирования условной гетероскедастичности рассматриваются в разделе 16.5.

Из-за неопределенности относительно будущих событий, то есть неопределенности относительно  $u_{T+1}$ , 95 %-й интервальный прогноз может быть настолько широк, что будет иметь ограниченную ценность при принятии решений. Поэтому профессиональные прогнозисты часто публикуют более узкие доверительные интервалы, чем 95 %-й, например используя одно стандартное отклонение (что является 68 %-м интервальным прогнозом, если ошибки распределены нормально). В качестве альтернативы некоторые прогнозисты публикуют несколько интервальных прогнозов, как, например, делают экономисты Банка Англии при публикации прогнозов инфляции (см. вставку «Река крови»).



### ***Река крови***

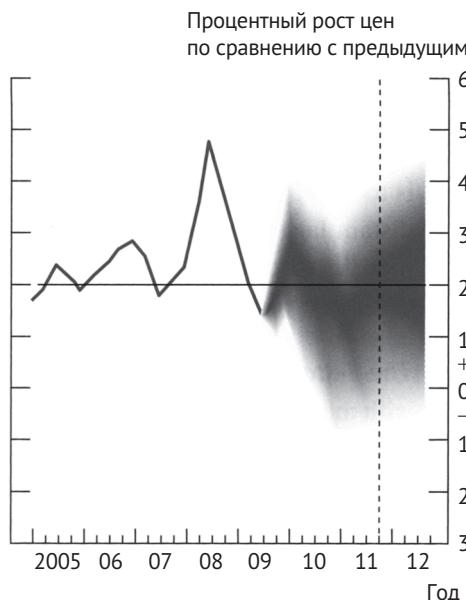
Банк Англии регулярно публикует прогнозы инфляции в Великобритании, что является частью его обязанностей по информированию населения о проводимой им монетарной политике. Эти прогнозы являются комбинацией прогнозов, полученных различными профессиональными прогнозистами, работающими в банке, на основе эконометрических моделей, и корректируются на экспертное мнение высокопоставленных сотрудников банка и членов Комитета по монетарной политике. Прогнозы, представленные множеством интервальных прогнозов, отражают взгляд этих экономистов на множество возможных траекторий, по которым может изменяться инфляция. В своем Докладе об инфляции банк печатает эти траектории красным цветом, выделяя темно-красным центральное направление. Несмотря на то

что в банке этот график прозаически называют «веерной диаграммой», пресса называет эти разбросанные тени красного цвета «рекой крови».

Река крови для ноября 2009 года показана на рисунке 14.4 (на этом графике кровь серая, а не красная, поэтому вам придется использовать ваше воображение). Эта диаграмма показывает, что в ноябре 2009 года экономисты банка ожидали резкий рост инфляции до 3 % в начале 2010 года и его падение до 1 % к концу 2010 года, а затем устойчивый рост до 2 % к 2012 году. Несмотря на это, экономисты выражали сомнение относительно прогноза. Они приводили рост налога с продаж в качестве важного фактора роста инфляции в краткосрочной перспективе и обсуждали неопределенность, связанную с реакцией инфляции на ослабление экономики, и необходимое для восстановления экономики время и его интенсивность, как источники неопределенности инфляции. Как оказалось, ближайшие прогнозы были очень близки к истинному значению инфляции: инфляция во втором квартале 2010 года была равна 3,5 %.

Банк Англии был пионером с точки зрения движения по направлению открытости центральных банков, и другие центральные банки сейчас также публикуют прогнозы инфляции. Решения, принимаемые влияющими на экономическую политику экономистами, сложны и влияют на жизнь (и кошельки) множества их сограждан. В демократический информационный век, по мнению экономистов Банка Англии, особенно важно понимать взгляд банка на экономику и причины, по которым принимаются эти сложные решения.

Чтобы увидеть реку крови в ее оригинальном красном цвете, посетите веб-сайт Банка Англии [www.bankofengland.co.uk](http://www.bankofengland.co.uk). Чтобы узнать больше о том, как прогнозируют инфляцию в Банке Англии, см. Clements (2004).



**Рисунок 14.4. Река крови**

На веерной диаграмме показана область прогнозов инфляции Банка Англии в ноябре 2009 года. Пунктирная линия показывает второй квартал 2011 года – два года спустя после публикации прогноза.

## 14.5. Информационные критерии и выбор глубины запаздывания

Регрессии, оцененные для инфляции в разделах 14.3 и 14.4, имеют либо одно, либо четыре запаздывания объясняющих переменных. Включение одного запаздывания имеет смысл, но зачем включать четыре? Другими словами, сколько запаздываний нужно включать в регрессию? В данном разделе мы обсуждаем статистические методы, используемые для выбора количества лагов сначала в авторегрессии, а затем в регрессии, включающей множество объясняющих переменных.

### Определение порядка авторегрессии

На практике выбор порядка  $p$  авторегрессии требует сбалансированности между предельной пользой от включения дополнительных запаздываний и предельных издержек от дополнительной неопределенности оценок. С одной стороны, если порядок оцененной авторегрессии слишком мал, вы пропускаете потенциально значимую информацию, содержащуюся в значениях более глубоких запаздываний. С другой стороны, если он слишком велик, вы оцениваете больше коэффициентов, чем необходимо, что приводит к появлению дополнительной ошибки в ваших прогнозах.

**Подход на основе F-статистики.** Один из подходов к выбору  $p$  заключается в оценке модели с достаточно большим числом запаздываний с последующим тестированием гипотез для выбора итогового лага. Например, вы можете начать с оценки модели AR(6) и проверить гипотезу о значимости коэффициента при шестом запаздывании на 5 %-м уровне значимости; если он незначим, исключите его и оценивайте AR(5), тестируйте значимость коэффициента при пятом запаздывании и так далее. Недостатком этого метода является то, что он будет выбирать слишком большую модель хотя бы иногда: даже если истинный порядок авторегрессии равен 5, так что гипотеза о равенстве коэффициента при шестом запаздывании нулю при использовании 5 %-го уровня значимости для тестиирования будет некорректно отвергнута случайным образом в 5 % случаев. Таким образом, когда истинное значение  $p$  равно 5, этот метод будет оценивать глубину запаздывания как 6 в 5 % случаев.

**BIC.** Одним из способов решения этой проблемы является оценка  $p$  при помощи минимизации «информационного критерия». Одним из таких информационных критериев является *байесовский информационный критерий (BIC<sup>1</sup>)*, также называемый *информационным критерием Шварца (SIC<sup>2</sup>)*, вычисляемый по формуле:

$$BIC(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{\ln(T)}{T}, \quad (14.23)$$

<sup>1</sup> Bayes information criterion. – Примеч. науч. ред. перевода.

<sup>2</sup> Schwartz information criterion. – Примеч. науч. ред. перевода.

где  $SSR(p)$  является суммой квадратов остатков оцененной модели AR( $p$ ).  $BIC$ -оценка  $p$ ,  $\hat{p}$  – это значение, которое минимизирует  $BIC(p)$  среди всех возможных значений  $p = 0, 1, \dots, p_{max}$ , где  $p_{max}$  является наибольшим из рассматриваемых значений  $p$ , а  $p = 0$  относится к модели, содержащей только константу.

Формула для расчета  $BIC$ , на первый взгляд, может выглядеть немного загадочной, но за ней стоит интуитивное объяснение. Рассмотрим первое слагаемое в выражении (14.23). Так как коэффициенты регрессии оцениваются при помощи МНК, сумма квадратов остатков обязательно уменьшается (или, по крайней мере, не увеличивается) при добавлении запаздываний. В отличие от этого, второе слагаемое представляет собой число оцениваемых коэффициентов регрессии (количество лагов,  $p$  плюс константа), умноженное на  $In(T)/T$ . Это второе слагаемое растет при добавлении запаздываний.  $BIC$  зависит от этих двух сил, так что количество лагов, которое минимизирует  $BIC$ , является состоятельной оценкой истинного значения глубины запаздывания. Математический вывод этого утверждения приведен в приложении 14.5.

В качестве примера рассмотрим оценку порядка AR для авторегрессии изменений инфляции. Различные этапы расчета  $BIC$  приведены в таблице 14.4 для авторегрессий с максимальным порядком, то есть  $p_{max} = 6$ . Например, для модели AR(1) в уравнении (14.7)  $SSR(1)/T = 2,737$ , поэтому  $In[SSR(1)/T] = 1,007$ . Так как  $T = 172$  (43 года, четыре квартала в год),  $In(T)/T = 0,030$  и  $(p+1)In(T)/T = 2 \times 0,030 = 0,060$ . Таким образом,  $BIC(1) = 1,007 + 0,060 = 1,067$ .

$BIC$  принимает самое маленькое значение, когда  $p = 2$  в таблице 14.4. Таким образом,  $BIC$ -оценка глубины запаздывания равна 2. Как можно видеть в таблице 14.4, при росте числа запаздываний увеличивается  $R^2$  и уменьшается  $SSR$ . Увеличение  $R^2$  больше при переходе от модели с одним лагом к модели с двумя лагами, меньше при переходе от двух лагов к трем и довольно небольшое при переходе от трех к четырем.  $BIC$  помогает решить, насколько большим должно быть увеличение  $R^2$ , чтобы оправдать включение дополнительных запаздываний.

Таблица 14.4

**Байесовский информационный критерий (BIC) и  $R^2$   
для авторегрессионных моделей инфляции в США, 1962–2004 годы**

<b>p</b>	<b><math>SSR(p)/T</math></b>	<b><math>In[SSR(p)/T]</math></b>	<b><math>(p+1)In(T)/T</math></b>	<b><math>BIC(p)</math></b>	<b><math>R^2</math></b>
0	2,900	1,065	0,030	1,095	0,000
1	2,737	1,007	0,060	1,067	0,056
2	2,375	0,865	0,090	0,955	0,181
3	2,311	0,838	0,120	0,957	0,203
4	2,309	0,837	0,150	0,986	0,204
5	2,308	0,836	0,180	1,016	0,204
6	2,308	0,836	0,209	1,046	0,204

**AIC.**  $BIC$  не является единственным информационным критерием; информационный критерий Акаике ( $AIC^1$ ):

<sup>1</sup> Akaike information criterion. – Примеч. науч. ред. перевода.

$$AIC(p) = In\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T}. \quad (14.24)$$

Различие между AIC и BIC заключается в том, что множитель  $In(T)$  в BIC заменяется на 2 в AIC, так что второе слагаемое в AIC меньше. Например, для 172 наблюдений, используемых для оценки авторегрессионной модели инфляции,  $In(T) = In(172) = 5,15$ , так что второй член в BIC более чем вдвое больше, чем второй член в AIC. Таким образом, меньшее снижение  $SSR$  необходимо в AIC, чтобы оправдать включение дополнительного лага. Как следует из теории, второе слагаемое в AIC не является достаточно большим для того, чтобы был выбран лаг правильной длины, даже в больших выборках, поэтому AIC-оценка  $p$  не является состоятельной. Как обсуждается в приложении 14.5, в больших выборках AIC переоценивает  $p$  с ненулевой вероятностью.

Несмотря на такой теоретический недостаток, AIC широко используется на практике. Если вас не устраивает то, что BIC может выбрать модель со слишком малым числом запаздываний, AIC предоставляет вполне разумную альтернативу.

**Замечание о расчете информационных критериев.** Насколько хорошо две оцененные регрессии соответствуют данным, лучше всего пытаться понять, когда они оцениваются с использованием одной и той же базы данных. Поскольку BIC и AIC являются формальными способами, используемыми для такого сравнения, рассматриваемые авторегрессии должны оцениваться с использованием одних и тех же наблюдений. Например, в таблице 14.4 все регрессии оцениваются с использованием данных с I квартала 1962 года по IV квартал 2004 года, в общей сложности – 172 наблюдения. Поскольку авторегрессии включают лаги изменения инфляции, это означает, что более ранние значения изменения инфляции (значения до I квартала 1962 года) были использованы для оценки моделей. Иначе говоря, каждая из регрессий, рассмотренных в таблице 14.4, включает наблюдения переменных  $\Delta Inf_t, \Delta Inf_{t-1}, \dots, \Delta Inf_{t-p}$  для  $t = 1962:I, \dots, 2004:IV$ , соответствующие 172 наблюдениям зависимой переменной и регрессоров, поэтому  $T = 172$  в уравнениях (14.23) и (14.24).

## Выбор длины лага в регрессии временных рядов со множественными регрессорами

Компромисс, связанный с выбором длины запаздывания в общей модели регрессии временных рядов с несколькими регрессорами [уравнение (14.20)], аналогичен случаю авторегрессии: использование слишком малого числа лагов может снижать точность прогноза, потому что будет потеряна ценная информация, но включение дополнительных лагов увеличивает неопределенность оценки. При выборе лагов мы должны балансировать между выигрышем от использования дополнительной информации и издержками от оценки дополнительных коэффициентов.

**Подход на основе F-статистики.** Как и в авторегрессии, один из способов определения количества лагов заключается в использовании F-статистики для проверки совместных гипотез о равенстве нулю нескольких коэффициентов.

Например, обсуждая уравнение (14.17), мы проверили гипотезу о том, что коэффициенты при втором, третьем и четвертом лагах безработицы равны нулю против альтернативы о том, что они отличны от нуля; эта гипотеза была отвергнута на 1%-м уровне значимости, что послужило аргументом в пользу выбора спецификации с большим числом запаздываний. Если число сравниваемых моделей невелико, то метод, основанный на  $F$ -статистике, прост в использовании. Однако в целом метод, основанный на  $F$ -статистике, может приводить к выбору моделей, которые слишком велики, в том смысле, что истинная глубина запаздывания переоценена.

**Информационные критерии.** Как и в авторегрессии, BIC и AIC могут быть использованы для оценки числа лагов и переменных в регрессионной модели временных рядов с несколькими регрессорами. Если в регрессионной модели оценивается  $K$  коэффициентов (в том числе константа), BIC равен:

$$BIC(K) = \ln\left[\frac{SSR(K)}{T}\right] + K \frac{\ln(T)}{T}. \quad (14.25)$$

AIC определяется аналогичным образом, заменяя  $\ln(T)$  на 2 в выражении (14.25). Для каждой из возможных моделей может быть оценен BIC (или AIC), а модель с самым низким значением BIC (или AIC) является предпочтительной моделью, выбранной на основе информационного критерия.

Существуют два важных практических соображения при использовании информационных критериев для оценки глубины запаздывания. Во-первых, как и в случае авторегрессии, все предполагаемые модели должны быть оценены на одном и том же массиве данных, то есть в обозначениях выражения (14.25); количество наблюдений, используемых для оценки модели, равное  $T$ , должно быть одинаковым для всех моделей. Во-вторых, при наличии нескольких объясняющих переменных этот подход требует довольно многочисленных вычислений, поскольку нужно оценить множество различных моделей (с большим количеством комбинаций лагов). На практике удобно требовать, чтобы все регрессоры имели одинаковое число запаздываний, то есть чтобы  $p = q_1 = \dots = q_k$ , и таким образом нужно было бы сравнивать только  $p_{max} + 1$  моделей (т.е. необходимо сравнить модели с  $p = 0, 1, \dots, p_{max}$  лагами).

## 14.6. Нестационарность I: тренды

Во вставке «Основные понятия 14.6» предполагается, что зависимая переменная и регрессоры являются стационарными. Если это не так, то есть если зависимая переменная и/или регрессоры являются нестационарными, то обычные методы тестирования гипотез, построения доверительных интервалов и прогнозов могут быть недостаточными. Точный характер проблемы, возникающей из-за нестационарности, и методы ее решения зависят от типа нестационарности.

В данном и следующем разделах мы рассмотрим два наиболее важных типа нестационарности в экономических временных рядах: тренды и структурные сдвиги. В каждом разделе мы сначала опишем характер рассматриваемого

типа нестационарности, а затем обсудим последствия наличия нестационарности такого типа в регрессиях временных рядов, если нестационарность игнорируется. Затем мы рассмотрим тесты для проверки наличия нестационарности во временных рядах и обсудим методы решения проблем, вызванных нестационарностью определенного типа. Мы начнем с обсуждения типов трендов во временных рядах.

## **Что такое тренд?**

Трендом называется постоянное долгосрочное изменение переменной во времени. Временные ряды колеблются вокруг своих трендов.

Из рисунка 14.1а следует, что инфляция в США имеет тренд, который отражает тенденцию к росту до 1982 года, а затем имеет тенденцию к снижению. Временные ряды на рисунках 14.2а, б, в также имеют тренды, но они довольно сильно различаются. Тренд в межбанковской процентной ставке США похож на тренд в показателе инфляции США. Обменный курс британского фунта к доллару США явно имел длительную тенденцию к снижению после распада системы фиксированных обменных курсов в 1972 году. Логарифм ВВП Японии имеет сложный тренд: быстрый рост в начале, затем умеренный рост и, наконец, медленный рост.

**Детерминированный и стохастический тренды.** Существует два основных типа трендов, которые могут присутствовать во временных рядах: детерминированный и стохастический. Детерминированный тренд является неслучайной функцией от времени. Например, детерминированный тренд может быть линейным по времени; если инфляция имела детерминированный линейный тренд такой, что она увеличивалась на 0,1% в квартал, то этот тренд может быть записан в виде  $0,1t$ , где  $t$  измеряется в кварталах. В противоположность этому стохастический тренд является случайным и изменяется с течением времени. Например, стохастический тренд в инфляция может проявляться в наличии длительного периода роста с последующим длительным периодом снижения, аналогично тренду инфляции на рисунке 14.1.

Как и многие эконометристы, мы считаем, что для моделирования экономических временных рядов больше подходят модели стохастических, а не детерминированных трендов. Экономика – очень сложная вещь. Трудно поверить в предсказуемость, которая подразумевается в модели детерминированного тренда, со всеми проблемами и неожиданностями, с которыми сталкиваются из года в год работники, предприятия и правительства. Например, несмотря на то что инфляция в США росла в 1970-е, ей не было суждено расти вечно и упасть снова. Скорее, сейчас считается, что медленный рост инфляции в настоящее время происходит из-за невезения и ошибок в денежно-кредитной политике, и ее укрощение было во многом следствием жестких решений, принятых Советом управляющих Федеральной резервной системы. Аналогично, обменный курс британского фунта к доллару США падал с 1972 по 1985 год, а затем дрейфовал, но эти движения также были следствием сложной экономической политики; из-за того что такие усилия изменяются непредсказуемо, подобные

тренды полезно рассматривать как имеющие большую непредсказуемую или случайную компоненту.

По описанным причинам, наше внимание к трендам в экономических временных рядах фокусируется на стохастических, а не детерминированных трендах, и когда мы говорим о «тренде» во временных рядах, мы имеем в виду стохастический тренд, если мы явно не говорим об ином. В данном разделе представлена простейшая модель стохастического тренда – модель случайного блуждания; другие модели трендов обсуждаются в разделе 16.3.

**Модель случайного блуждания.** Простейшей моделью стохастического тренда является случайное блуждание. Говорят, что временной ряд является *случайным блужданием*, если его изменения являются i.i.d., то есть если

$$Y_t = Y_{t-1} + u_t, \quad (14.26)$$

где  $u_t$  является i.i.d. Мы, однако, использует термин «случайное блуждание» в более общем смысле для обозначения временных рядов, которые описываются уравнением (14.26), где  $u_t$  имеет нулевое условное среднее, то есть  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$ .

Основная идея случайного блуждания состоит в том, что значение ряда завтра равно его значению сегодня плюс непредсказуемое изменение: из-за того что путь, пройденный  $Y_t$ , состоит из случайных «шагов»  $u_t$ , этот путь является «случайным блужданием». Условное среднее  $Y_t$  относительно данных до момента  $t-1$  включительно равно  $Y_{t-1}$ , то есть из-за того что  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$ , получаем, что  $E(Y_t | Y_{t-1}, Y_{t-2}, \dots) = Y_{t-1}$ . Другими словами, если  $Y_t$  является случайным блужданием, то лучшим прогнозом значения завтра является его значение сегодня.

Некоторые временные ряды, такие как логарифм ВВП Японии, изображенный на рисунке 14.2c, имеют очевидные тенденции к росту, и в этом случае лучший прогноз временного ряда должен включать поправки на эту его тенденцию к росту. Такая корректировка приводит к расширению модели случайного блуждания, включающему эту тенденцию к росту (или падению) или «дрейф» («снос») в одном или в другом направлении. Это расширение называют *моделью случайного блуждания с дрейфом или со сносом*:

$$Y_t = \beta_0 + Y_{t-1} + u_t, \quad (14.27)$$

где  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$  и  $\beta_0$  является «дрейфом» в случайном блуждании. Если  $\beta_0$  положительно, то  $Y_t$  увеличивается в среднем. В модели случайного блуждания с дрейфом лучший прогноз временного ряда на завтра равен его значению сегодня плюс дрейф.

Модель случайного блуждания (со сносом в соответствующих случаях) проста, но универсальна, и является основной моделью для трендов, которая используется в данной книге.

**Случайное блуждание является нестационарным процессом.** Если временной ряд  $Y_t$  является случайным блужданием, то он не является стационарным: дисперсия процесса случайного блуждания увеличивается со временем, поэтому распределение  $Y_t$  также изменяется во времени. Один из способов понять

это заключается в том, чтобы признать, что, так как  $u_t$  не коррелирован с  $Y_{t-1}$  в уравнении (14.26), то  $\text{var}(Y_t) = \text{var}(Y_{t-1}) + \text{var}(u_t)$ ; а для того чтобы  $Y_t$  был стационарным,  $\text{var}(Y_t)$  не должен зависеть от времени, так что, в частности, должно выполняться равенство  $\text{var}(Y_t) = \text{var}(Y_{t-1})$ , но это может произойти, только если  $\text{var}(u_t) = 0$ . Чтобы убедиться в этом иным способом, представьте себе, что начальное значение  $Y_t$  равно нулю, то есть  $Y_0 = 0$ . Тогда  $Y_1 = u_1$ ,  $Y_2 = u_1 + u_2$  и так далее, так что  $Y_t = u_1 + u_2 + \dots + u_t$ . Так как  $u_t$  серийно не коррелированы,  $\text{var}(Y_t) = \text{var}(u_1 + u_2 + \dots + u_t) = t\sigma^2$ . Таким образом, дисперсия  $Y_t$  зависит от  $t$  и действительно растет при увеличении  $t$ . Поскольку дисперсия  $Y_t$  зависит от  $t$ , его распределение зависит от  $t$ , то есть он нестационарен.

Поскольку дисперсия процесса случайного блуждания увеличивается неограниченно, его теоретические автокорреляции (т.е. автокорреляции в генеральной совокупности) не определены (первая автоковариация и дисперсия бесконечны, а их отношение не определено). Тем не менее особенностью процесса случайного блуждания является то, что его выборочная автокорреляция, как правило, очень близка к 1; в действительности  $j$ -я выборочная автокорреляция процесса случайного блуждания сходится к 1 по вероятности.

**Стохастические тренды, авторегрессионные модели и единичные корни.** Модель случайного блуждания является частным случаем модели AR(1) [уравнение (14.8)], в которой  $\beta_1 = 1$ . Другими словами, если  $Y_t$  является процессом AR(1) с  $\beta_1 = 1$ , то он содержит стохастический тренд и является нестационарным. Однако, если  $|\beta_1| < 1$  и  $u_t$  стационарна, то совместное распределение  $Y_t$  и его лагов не зависит от  $t$  (результат выводится в приложении 14.2), и тогда  $Y_t$  стационарен.

Аналогичное условие стационарности выполняется для AR( $p$ ), но оно является более сложным, чем условие  $|\beta_1| < 1$  для AR(1). Его точная формулировка включает в себя условия для корней многочлена  $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \dots - \beta_p z^p$ . (Корни этого многочлена являются значениями  $z$ , которые удовлетворяют уравнению  $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \dots - \beta_p z^p = 0$ .) Для того чтобы процесс AR( $p$ ) был стационарным, все корни этого многочлена должны быть больше единицы по абсолютной величине. В частном случае процесса AR(1) корень этого многочлена равен значению  $z$ , которое является решением уравнения  $1 - \beta_1 z = 0$ , так что его корень равен:  $z = 1/\beta_1$ . Таким образом, утверждение о том, что корень больше единицы по абсолютной величине, эквивалентно утверждению  $|\beta_1| < 1$ .

Если AR( $p$ ) имеет корень, равный единице, считается, что временной ряд имеет *авторегрессионный единичный корень* или, проще говоря, *единичный корень*. Если  $Y_t$  содержит единичный корень, то он содержит стохастический тренд. Если  $Y_t$  является стационарным (и, следовательно, не содержит единичного корня), он не содержит стохастического тренда. По этой причине мы будем использовать термины «стохастический тренд» и «единичный корень» как взаимозаменяемые.

### Проблемы, возникающие из-за стохастических трендов

Если регрессор содержит стохастический тренд (имеет единичный корень), то МНК-оценки его коэффициента и соответствующая  $t$ -статистика могут иметь

нестандартные (т.е. ненормальные) распределения, даже в больших выборках. Обсудим три конкретных аспекта этой проблемы: (1) смещение к нулю оценки коэффициента при первом запаздывании в AR(1), если его истинное значение равно 1; (2)  $t$ -статистикой коэффициента при регрессоре, характеризующем стохастический тренд, может быть ненормальное распределение, даже в больших выборках, и (3) как крайний пример ситуации, связанной с наличием стохастического тренда в данных, является ложная (кажущаяся, мнимая) регрессия, возникающая в ситуации, когда два независимых временных ряда с высокой вероятностью покажут наличие обманчивой регрессионной зависимости, если они оба содержат стохастический тренд.

**Проблема № 1: смещенные к нулю авторегрессионные коэффициенты.** Предположим, что  $Y_t$  является процессом случайного блуждания (14.26), но это неизвестно, и эконометрист вместо модели случайного блуждания оценивает модель AR(1), задаваемую уравнением (14.8). Из-за того что  $Y_t$  является нестационарным, предположения метода наименьших квадратов для регрессии временных рядов из вставки «Основные понятия 14.6» не выполняются, так как в общем случае мы не можем считать, что получаемые оценки и тестовые статистики имеют обычное асимптотическое нормальное распределение. В самом деле, в этом примере МНК-оценка авторегрессионного коэффициента  $\beta_1$  является состоятельной, но имеет ненормальное распределение, даже в больших выборках: асимптотическое распределение  $\hat{\beta}_1$  смещается в сторону нуля. Ожидаемое значение  $\hat{\beta}_1$  равно примерно  $E(\hat{\beta}_1) = 1 - 5,3/T$ . Этот результат приводит к большему смещению в выборках большего размера и обычно встречается в экономических приложениях. Например, квартальные данные за 20 лет содержат 80 наблюдений, в этом случае ожидаемое значение  $\hat{\beta}_1$  равно:  $E(\hat{\beta}_1) = 1 - 5,3/80 = 0,934$ . Более того, рассматриваемое распределение имеет длинный левый хвост: 5-й процентиль  $\hat{\beta}_1$  равен приблизительно  $1 - 14,1/T$  и для  $T = 80$  соответствует значению 0,824, так что в 5 % случаев  $\hat{\beta}_1 < 0,824$ .

Одним из следствий этого смещения к нулю является то, что если  $Y_t$  является процессом случайного блуждания, то прогнозы, основанные на модели AR(1), могут оказаться значительно хуже, чем те, которые основаны на модели случайного блуждания, в которой используется истинное значение  $\beta_1 = 1$ . Этот вывод также относится к авторегрессиям более высокого порядка, прогноз по которым можно улучшить, если учесть наличие единичного корня (т.е. оценить авторегрессию в первых разностях, а не в уровнях), когда ряд на самом деле содержит единичный корень.

**Проблема № 2: распределения  $t$ -статистик не являются нормальными.** Если регрессор имеет стохастический тренд, то его обычная МНК  $t$ -статистика может иметь ненормальное распределение в условиях нулевой гипотезы, даже в больших выборках. Это не нормальное распределение означает, что обычные доверительные интервалы не являются корректными и проверка гипотез не может проводиться стандартным образом. В общем случае распределение этой  $t$ -статистики нелегко заключить в таблицу, так как это распределение зависит от соотношения между регрессором, о котором идет речь, и другими регрессорами. Важным примером того, как эта проблема возникает в регрессии,

является попытка прогнозирования доходностей акций с использованием регрессоров, которые могут содержать стохастический тренд (см. вставку в разделе 14.7 «Можем ли мы перехитрить рынок? Часть II»).

Одним из важных случаев, когда распределение  $t$ -статистики можно представить в виде таблицы, является ситуация, когда регрессор содержит стохастический тренд в контексте авторегрессии с единичным корнем. Мы вернемся к этому частному случаю, когда приступим к рассмотрению задачи тестирования наличия стохастического тренда во временном ряде.

**Проблема № 3: ложная регрессия.** Стохастические тренды, присутствующие в двух временных рядах, могут привести к появлению регрессионной связи между ними, когда на самом деле ее нет, и эта проблема называется *ложной (минимой, кажущейся) регрессией*.

Например, инфляция в США неуклонно росла с середины 1960-х годов до начала 1980-х, и в то же время ВВП Японии (приведен график логарифмов на рис. 14.2в) неуклонно рос. Эти два тренда приводят к регрессии, которая кажется «значимой», если использовать традиционные меры оценки значимости. Оцененная с помощью МНК с использованием данных с 1965 по 1981 год, эта регрессия имеет вид

$$\widehat{U.S.Inflation}_t = -37,78 + 3,83 \times \ln(\text{Japanese GDP}_t), \bar{R}^2 = 0,56. \quad (14.28)$$

$t$ -статастика коэффициента наклона превышает 10, что указывает на сильную положительную связь между двумя временными рядами по обычным критериям, и  $\bar{R}^2$  также высок. Однако оценка этой регрессии с использованием данных с 1982 по 2004 год дает следующие результаты:

$$\widehat{U.S.Inflation}_t = 31,20 - 2,17 \times \ln(\text{Japanese GDP}_t), \bar{R}^2 = 0,08. \quad (14.29)$$

Регрессии в уравнениях (14.28) и (14.29) вряд ли могли бы быть более разными. В буквальном смысле уравнение (14.28) указывает на сильную положительную связь, в то время как уравнение (14.29) указывает на слабую, но, видимо, статистически значимую отрицательную связь.

Источником этих противоречивых результатов является то, что оба временных ряда содержат стохастические тренды. Эти тренды присутствуют с 1965 по 1981 год, но их нет с 1982 по 2004 год. Но, по сути, нет никаких убедительных экономических или политических причин считать, что тренды, присутствующие в этих двух временных рядах, связаны между собой. Короче говоря, эти регрессии являются кажущимися.

Регрессии (14.28) и (14.29) эмпирически иллюстрируют теоретическое положение о том, что МНК может вводить в заблуждение, если временные ряды содержат стохастический тренд (см. упражнение 14.6, в котором компьютерные симуляции иллюстрируют этот результат). Есть один частный случай, в котором используемые регрессионные методы являются надежными: такая ситуация возникает, когда компоненты трендов двух временных рядов одинаковы, то есть когда временные ряды содержат *общий* стохастический тренд; в таком случае говорят, что временные ряды *коинтегрированы*. Эконометрические методы для

обнаружения и анализа коинтегрированных экономических временных рядов обсуждаются в разделе 16.4.

## **Обнаружение стохастических трендов: тестирование на авторегрессионный единичный корень**

Тренды во временных рядах могут быть обнаружены как при помощи неформальных, так и формальных методов. Неформальные методы включают анализ графиков временных рядов и вычисление коэффициентов автокорреляции, как это было сделано в разделе 14.2. Поскольку, по крайней мере в больших выборках, коэффициент автокорреляции первого порядка приблизительно равен единице, если ряд содержит стохастический тренд, небольшой коэффициент автокорреляции первого порядка в сочетании с графиком временного ряда, не имеющим очевидного тренда, предполагает, что временной ряд не содержит тренда. Однако если вы сомневаетесь в ваших выводах, существуют формальные статистические процедуры, которые можно использовать для проверки гипотезы о том, что существует стохастический тренд во временном ряде против альтернативы о том, что стохастического тренда нет.

В этом разделе мы будем использовать тест Дики–Фуллера (названный в честь его авторов Дэвида Дики и Уэйна Фуллера (David Dickey, Wayne Fuller, 1979)), чтобы проверить наличие стохастического тренда в данных. Несмотря на то что тест Дики–Фуллера – не единственный тест на наличие стохастического тренда (еще один тест описан в разделе 16.3), на практике он наиболее часто используется и является одним из самых надежных.

**Тест Дики–Фуллера для модели AR(1).** Отправной точкой для *теста Дики–Фуллера* является модель авторегрессии. Как обсуждалось ранее, модель случайного блуждания (14.27) является частным случаем модели AR(1) с  $\beta_1 = 1$ . Если  $\beta_1 = 1$ ,  $Y_t$  является нестационарным и содержит (стохастический) тренд. Таким образом, в AR(1)-модели гипотеза о том, что  $Y_t$  содержит тренд, может быть проверена путем тестирования:

$$H_0 : \beta_1 = 1 \text{ против } H_1 : \beta_1 < 1 \text{ в регрессии } Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t. \quad (14.30)$$

Если  $\beta_1 = 1$ , AR(1) имеет авторегрессионный корень, равный единице, так что нулевая гипотеза из (14.30) состоит в том, что AR(1) имеет единичный корень, а альтернативная гипотеза говорит о том, что временной ряд стационарен.

Этот тест наиболее просто реализовать, оценивая модифицированную версию уравнения (14.30), полученную вычитанием из обеих его частей  $Y_{t-1}$ . Пусть  $\delta = \beta_1 - 1$ , тогда уравнение (14.30) принимает вид:

$$H_0 : \delta = 0 \text{ против } H_1 : \delta < 0 \text{ в регрессии } \Delta Y_t = \beta_0 + \delta Y_{t-1} + u_t. \quad (14.31)$$

МНК *t*-статистика для проверки гипотезы  $\delta = 0$  в регрессии (14.31) называется *статистикой Дики–Фуллера*. Формулировка в уравнении (14.31) удобна тем, что эконометрические программные пакеты автоматически выводят на печать *t*-статистики для тестирования гипотезы  $\delta = 0$ . Обратите внимание на то, что тест Дики–Фуллера является односторонним, поскольку соответствующая

альтернативная гипотеза заключается в том, что  $Y_t$  стационарен, или, что эквивалентно,  $\delta < 0$ . Статистика Дики–Фуллера вычисляется с использованием «неустойчивых» стандартных ошибок, то есть рассчитываются оценки стандартных ошибок в предположении гомоскедастичности ошибок регрессии, формулы для вычисления которых приведены в приложении 5.1 [уравнение (5.29) для случая парной регрессии и выражения в разделе 18.4 для модели множественной регрессии]<sup>1</sup>.

**Тест Дики–Фуллера для модели AR( $p$ ).** Статистика Дики–Фуллера, представленная в контексте уравнения (14.31), применима только к модели AR(1). Как отмечалось в разделе 14.3, для некоторых временных рядов модели AR(1) не учитывают все серийные корреляции временного ряда  $Y_t$ , и в этом случае использование авторегрессии более высокого порядка является более целесообразным.

Расширение теста Дики–Фуллера на случай модели AR( $p$ ) приведено во вставке «Основные понятия 14.8». В условиях нулевой гипотезы  $\delta = 0$  и  $Y_t$  является стационарным процессом AR( $p$ ). В соответствии с альтернативной гипотезой  $\delta < 0$ , так что  $Y_t$  стационарен. Поскольку регрессия, используемая для вычисления этой версии статистики Дики–Фуллера, дополнена лагами  $\Delta Y_t$ , получающуюся  $t$ -статистику называют *расширенной статистикой Дики–Фуллера (ADF)*.

В общем случае глубина запаздывания  $p$  неизвестна, но она может быть оценена с использованием информационных критериев, применяемых к регрессии в форме уравнения (14.32) для различных значений  $p$ . Исследование ADF-статистики предполагает, что лучше иметь слишком много лагов, чем слишком мало, поэтому для оценки  $p$  при расчете ADF-статистики рекомендуется использовать AIC вместо BIC<sup>2</sup>.

**Тестирование против альтернативной гипотезы о стационарности вокруг детерминированного линейного временного тренда.** До сих пор мы рассматривали нулевую гипотезу о том, что временному ряду имеет единичный корень, и альтернативную гипотезу — о том, что он является стационарным. Эта альтернативная гипотеза о стационарности подходит для таких временных рядов, как уровень инфляции, не демонстрирующих долгосрочный рост. Но другие экономические временные ряды, такие как ВВП Японии (рис. 14.2e), показывают наличие долгосрочного роста, и для таких временных рядов альтернативная гипотеза о стационарности без тренда является неуместной. Вместо этого обычно используется альтернативная гипотеза о том, что временному ряду является стационарным около детерминированного временного тренда, то есть тренда, являющегося детерминированной функцией времени.

<sup>1</sup> В условиях нулевой гипотезы о наличии единичного корня обычные «неустойчивые» стандартные ошибки дают  $t$ -статистику, которая на самом деле устойчива к гетероскедастичности, что является удивительным и очень специфическим результатом.

<sup>2</sup> См. работы Стоука (Stock, 1994) и Халдрупа и Янссона (Haldrup, Jansson, 2006) для изучения исследований свойств теста Дики–Фуллера в конечных выборках и ознакомления с другими тестами на единичные корни.

## ОСНОВНЫЕ ПОНЯТИЯ

### 14.8

#### Расширенный тест Дики–Фуллера на авторегрессионный единичный корень

Тест Дики–Фуллера (ADF) на авторегрессионный единичный корень проверяет нулевую гипотезу  $H_0 : \delta = 0$  против односторонней альтернативы  $H_1 : \delta < 0$  в регрессии:

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \dots + \gamma_p \Delta Y_{t-p} + u_t. \quad (14.32)$$

В условиях нулевой гипотезы  $Y_t$  имеет стохастический тренд; в условиях альтернативной гипотезы  $Y_t$  – стационарен. ADF-статистика является МНК  $t$ -статистикой, используемой для тестирования гипотезы  $\delta = 0$  в уравнении (14.32).

Если вместо рассматриваемой альтернативной гипотезы использовать гипотезу о том, что  $Y_t$  является стационарным вокруг детерминированного линейного временного тренда, то этот тренд « $t$ » (номер наблюдения) должен быть добавлен в качестве дополнительного регрессора, и в этом случае регрессия Дики–Фуллера принимает вид:

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \dots + \gamma_p \Delta Y_{t-p} + u_t, \quad (14.33)$$

где  $\alpha$  – неизвестный коэффициент и ADF-статистика является МНК  $t$ -статистикой для тестирования гипотезы  $\delta = 0$  в уравнении (14.33).

Глубину запаздывания  $p$  можно оценить с помощью ВIC или AIC. Когда  $p = 0$ , лаги не включены в качестве регрессоров в уравнения (14.32) и (14.33), и ADF-тест упрощается до случая теста Дики–Фуллера в модели AR(1). ADF-статистика не имеет нормального распределения даже в больших выборках. Критические значения для одностороннего ADF-теста зависят от того, основан ли он на уравнении (14.32) или (14.33), и приведены в таблице 14.5.

Точная формулировка этой альтернативной гипотезы заключается в том, что временной тренд является линейным, то есть тренд является линейной функцией  $t$ , таким образом, нулевая гипотеза говорит о том, что временной ряд имеет единичный корень, а альтернативная – о том, что он не содержит единичного корня, но имеет детерминированный временной тренд. Регрессия Дики – Фуллера должна быть изменена для проверки нулевой гипотезы о наличии единичного корня против альтернативы о том, что временной ряд является стационарным вокруг линейного тренда. Как показано в уравнении (14.33) из вставки «Основные понятия 14.8», это достигается при помощи включения временного тренда (регрессора  $X_t = t$ ) в регрессию.

Линейный тренд не является единственным способом смоделировать детерминированный временной тренд, например, детерминированный временной тренд может быть квадратичным, или линейным, но со структурным сдвигом (т.е. быть линейным с угловыми коэффициентами, которые отличаются в двух частях выборки). Использование альтернативной гипотезы, предполагающей наличие такого нелинейного детерминированного тренда, должно быть обосновано экономической теорией. Для ознакомления с тестами на единичный

корень против альтернативной гипотезы о стационарности вокруг нелинейных детерминированных трендов см. Maddala, Kim (1998. Ch. 13).

**Критические значения для ADF-статистики.** В условиях нулевой гипотезы о наличии единичного корня ADF-статистика не является нормально распределенной, даже в больших выборках.

Из-за того что ее распределение не является стандартным, обычные критические значения нормального распределения не могут быть использованы в случае использования ADF-статистики для тестирования гипотезы о единичном корне; мы должны использовать специальные критические значения, основанные на распределении ADF-статистики в условиях нулевой гипотезы.

Критические значения для ADF-теста приведены в таблице 14.5. Поскольку альтернативная гипотеза о стационарности предполагает, что в уравнениях (14.32) и (14.33)  $\delta < 0$ , ADF-тест является односторонним. Например, если регрессия не включает в себя временной тренд, то гипотеза о наличии единичного корня отвергается на уровне значимости 5 %, если ADF-статистика меньше -2,86. Если в регрессию включен временной тренд, критическое значение становится равным -3,41.

Критические значения в таблице 14.5 существенно превышают по абсолютному значению (являются более отрицательными) по сравнению с односторонними критическими значениями, равными -1,28 (на уровне значимости 10 %) и -1,645 (на уровне значимости 5 %), стандартного нормального распределения. Нестандартное распределение ADF-статистики является примером того, как МНК  $t$ -статистики для коэффициентов при regressorах со стохастическим трендом могут иметь ненормальное распределение. Причины, по которым асимптотическое распределение ADF-статистики является нестандартным, обсуждаются в разделе 16.3.

**Есть ли стохастический тренд в американской инфляции?** Нулевая гипотеза о том, что инфляция имеет стохастический тренд, против альтернативы о том, что ряд стационарен, может быть протестирована при помощи ADF-теста о наличии авторегрессионного единичного корня. ADF-регрессия с четырьмя лагами  $\Delta \text{Infl}_t$  имеет вид:

$$\widehat{\Delta \text{Infl}_t} = 0,51 \underset{(0,21)}{-0,11} \text{Infl}_{t-1} - 0,19 \underset{(0,04)}{\Delta \text{Infl}_{t-1}} - 0,26 \underset{(0,08)}{\Delta \text{Infl}_{t-2}} + \\ + 0,20 \underset{(0,08)}{\Delta \text{Infl}_{t-3}} + 0,01 \underset{(0,08)}{\Delta \text{Infl}_{t-4}}. \quad (14.34)$$

ADF  $t$ -статистика равна  $t$ -статистике для проверки гипотезы о том, что коэффициент при  $\text{Infl}_{t-1}$  равен нулю и  $t = -2,69$ . Из таблицы 14.5 следует, что 5 %-е критическое значение равно -2,86. Поскольку ADF-статистика, равная -2,69, лежит правее, чем -2,86 (т.е. менее отрицательная), тест не отвергает нулевую гипотезу на уровне значимости 5 %. Поэтому на основе регрессии (14.34) мы не можем отвергнуть (на 5 %-м уровне значимости) нулевую гипотезу о том, что инфляция имеет единичный авторегрессионными корень, то есть гипотезу о том, что инфляция содержит стохастический тренд, против альтернативы о его стационарности.

Таблица 14.5

## Асимптотические критические значения расширенного теста Дики–Фуллера

Детерминированные переменные	10 %	5 %	1 %
Константа	-2,57	-2,86	-3,43
Константа и тренд	-3,12	-3,41	-3,96

ADF-регрессия (14.34) включает в себя четыре лага для вычисления ADF-статистики. Однако если оценивать количество лагов при помощи AIC, где  $0 < p < 5$ , AIC-оценкой длины лага будет 3. Если использовать три лага (т.е. если включать  $\Delta Inf_{t-1}$ ,  $\Delta Inf_{t-2}$  и  $\Delta Inf_{t-3}$  в качестве регрессоров), ADF-статистика будет равна -2,72, что опять находится правее на числовой оси, чем -2,86. Таким образом, когда количество лагов в ADF-регрессии выбирается при помощи AIC, гипотеза о том, что инфляция содержит стохастический тренд, не отвергается на уровне значимости 5 %.

Эти тесты проводились на уровне значимости 5 %. Однако на 10 %-м уровне значимости тесты будут отвергать нулевую гипотезу о наличии единичного корня: ADF-статистики (четыре лага) и (три лага) лежат левее (более отрицательные), чем 10 %-е критическое значение. Таким образом, ADF-статистика дает весьма неоднозначную картину, и прогнозист должен принять обоснованное решение о том, включать ли в модель инфляции стохастический тренд или нет. Очевидно, что на графике инфляции на рисунке 14.1а видны долгосрочные колебания, соответствующие модели стохастического тренда. На практике многие прогнозисты считают, что ряд инфляции в США содержит стохастический тренд, и мы также придерживаемся этого мнения здесь.

### Как избежать проблем, возникающих из-за стохастических трендов

Наиболее надежный способ избавиться от тренда во временном ряде заключается в его последовательном преобразовании, так чтобы тренд исчез. Если временной ряд имеет стохастический тренд, то есть если он содержит единичный корень, то его первая разность не имеет тренда. Например, если  $Y_t$  является случайным блужданием, так что  $Y_t = \beta_0 + Y_{t-1} + u_t$ , то его первая разность  $\Delta Y_t = \beta_0 + u_t$  будет стационарной. Таким образом, используя первые разности, мы исключаем стохастические тренды из временного ряда.

На практике вы редко можете утверждать с уверенностью, содержит ли временной ряд стохастический тренд. Напомним, что в общем случае невозможность отвергнуть нулевую гипотезу необязательно означает, что нулевая гипотеза верна, скорее всего, это означает, что у вас недостаточно данных, чтобы сделать вывод о том, что она ложна. Таким образом, то что мы не можем отвергнуть нулевую гипотезу о наличии единичного корня с помощью ADF-теста, не означает, что на самом деле временной ряд содержит единичный корень. Например, в модели AR(1) истинный коэффициент  $\beta_1$  может быть очень близок

к 1, скажем 0,98, в этом случае мощность ADF-теста была бы мала для отвержения нулевой гипотезы, то есть вероятность правильного отвержения нулевой гипотезы о наличии единичного корня во временном ряде инфляции была бы невысокой в выборках имеющейся у нас размера. Даже если мы неспособны отвергнуть нулевую гипотезу о наличии единичного корня, это не означает, что временной ряд имеет единичный корень, но все равно можно считать, что истинный авторегрессионный корень равен 1, и, следовательно, использовать первые разности временного ряда, а не его уровни<sup>1</sup>.

## 14.7. Нестационарность II: структурные сдвиги

Второй тип нестационарности возникает тогда, когда теоретическая функция регрессии изменяется в имеющейся выборке. В экономике это может произойти по целому ряду причин, таких как наличие изменений в экономической политике и в структуре экономики или появление изобретения, которое изменит конкретную отрасль. Если такие изменения, или «переломы», происходят, то модель регрессии, в которой такие изменения не учитываются, может не быть надежной основой для статистических выводов и прогнозирования.

В этом разделе представлены два метода проверки наличия структурных сдвигов в функции регрессии временных рядов во времени. Первый метод рассматривает потенциальные структурные сдвиги с точки зрения проверки гипотез и основан на тестировании гипотез об изменении коэффициентов регрессии с использованием  $F$ -статистики. Второй метод рассматривает потенциальные структурные сдвиги с точки зрения прогнозирования: вы считаете, что ваша выборка заканчивается раньше, чем есть на самом деле, и стремитесь оценить прогнозы для оставшейся части выборки. Структурные сдвиги обнаруживаются, если качество прогнозов значительно хуже, чем ожидалось.

### *Что такое структурный сдвиг?*

Структурные сдвиги могут возникать либо дискретно, представляя собой изменение коэффициентов в теоретической регрессии в разные даты, либо как постепенное изменение коэффициентов в течение более длительного периода времени.

Одним из источников дискретных структурных сдвигов в макроэкономических данных является изменение в макроэкономической политике. Например, распад Бреттон-Вудской системы фиксированных валютных курсов в 1972 году привел к структурному сдвигу в поведении временного ряда обменного курса британского фунта к доллару США, что видно на рисунке 14.26. До 1972 года обменный курс был практически постоянным, за исключением одной девальвации в 1968 году, во время которой официальная стоимость фунта по отношению к доллару снизилась. Напротив, с 1972 года обменный курс колебался в очень широком диапазоне.

<sup>1</sup> Для того чтобы ознакомиться с дополнительными вопросами, касающимися стохастических трендов в экономических временных рядах и проблем, возникающих из-за этого в регрессионном анализе, см. Stock, Watson (1988).

Структурные сдвиги могут происходить и более медленно, так как теоретическая регрессия изменяется во времени. Например, такие изменения могут возникнуть из-за медленного изменения экономической политики и происходящих изменений в структуре экономики. Методы, используемые для обнаружения структурных сдвигов и описанные в этом разделе, позволяют обнаружить оба типа структурных сдвигов: и резкие дискретные изменения, и медленную эволюцию.

**Проблемы, возникающие из-за структурных сдвигов.** Если структурный сдвиг происходит в теоретической функции регрессии в течение рассматриваемого временного интервала, то МНК-оценки регрессии на всем исследуемом периоде будут характеризовать отношения, которые имеют место «в среднем» в том смысле, что оценка сочетает в себе свойства двух различных периодов. В зависимости от расположения и размеров структурного сдвига «средняя» функция регрессии может существенно отличаться от истинной функции регрессии на подвыборках, что приведет к плохим прогнозам.

### **Тестирование структурных сдвигов**

Одним из способов проверки наличия структурных сдвигов является тестирование дискретных изменений или структурных сдвигов в коэффициентах регрессии. Как это сделать, зависит от того, известна ли дата возможного структурного сдвига (*дата или момент структурного сдвига*) или нет.

**Тестирование структурного сдвига, произошедшего в известный момент времени.** В некоторых случаях вы можете подозревать, что структурный сдвиг произошел в известный момент времени. Например, если вы изучаете международные торговые отношения с использованием данных с 1970-х годов, то можно предположить, что существует структурный сдвиг в интересующей вас теоретической функции регрессии в 1972 году, когда была отменена Бреттон-Вудская система фиксированных валютных курсов и произошел переход к плавающим валютным курсам.

Если дата гипотетического структурного сдвига в коэффициентах известна, то нулевая гипотеза об отсутствии структурного сдвига может быть проверена с помощью модели регрессии с бинарной компонентой взаимодействия, которая обсуждается в главе 8 (вставка «Основные понятия 8.4»). Для простоты рассмотрим модель ADL (1, 1), включающую константу, один лаг  $Y_t$  и один лаг  $X_t$ . Обозначим через  $\tau$  предполагаемую дату структурного сдвига, и пусть  $D_t(\tau)$  является бинарной переменной, равной нулю до даты структурного сдвига и единице – после, так что  $D_t(\tau) = 0$ , если  $t \leq \tau$ , и  $D_t(\tau) = 1$ , если  $t > \tau$ . Тогда регрессия, включающая бинарную переменную, характеризующую структурный сдвиг, и все компоненты взаимодействия, имеет вид:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) \times Y_{t-1}] + \\ + \gamma_2 [D_t(\tau) \times X_{t-1}] + u_t. \quad (14.35)$$

Если структурного сдвига нет, то теоретическая функция регрессии одинакова на обоих подпериодах, так что слагаемые с бинарной переменной, характе-

ризующей структурный сдвиг, не входят в уравнение (14.35). То есть в условиях нулевой гипотезы об отсутствии структурного сдвига  $\gamma_0 = \gamma_1 = \gamma_2 = 0$ . Если верна альтернативная гипотеза о том, что структурный сдвиг есть, то теоретическая функция регрессии различна для подпериодов до и после даты структурного сдвига  $\tau$ , и в этом случае по крайней мере один из коэффициентов,  $\gamma$ , – не нулевой. Таким образом, гипотезу о наличии структурного сдвига можно тестировать с помощью  $F$ -статистики, которая проверяет гипотезу  $\gamma_0 = \gamma_1 = \gamma_2 = 0$  против гипотезы о том, что по крайней мере один из коэффициентов,  $\gamma$ , не равен нулю. Этот тест часто называют тестом Чоу на структурный сдвиг при известной дате структурного сдвига, названным по имени его автора Грегори Чоу (Chow, 1960).

При наличии нескольких регрессоров или большего числа запаздываний этот тест может быть расширен путем включения компонент взаимодействия с бинарной переменной для всех регрессоров и проверкой гипотезы о том, что все коэффициенты при переменных, в которых присутствует  $D_i(\tau)$ , равны нулю.

Этот подход может быть изменен для того, чтобы проверить наличие структурных сдвигов в некотором подмножестве коэффициентов, включая компоненты взаимодействия с бинарной переменной только для этого интересующего нас подмножества регрессоров.

**Тестирование структурного сдвига, произошедшего в неизвестный момент времени.** Часто дата возможного структурного сдвига неизвестна или известна только в пределах некоторого диапазона. Предположим, что вы подозреваете, что структурный сдвиг произошел между двумя датами  $\tau_0$  и  $\tau_1$ . Тест Чоу может быть изменен для того, чтобы проверить наличие структурных сдвигов во все возможные моменты времени  $\tau$  между  $\tau_0$  и  $\tau_1$ , а затем с помощью наибольшей полученной  $F$ -статистики проверить наличие структурного сдвига в неопределененный момент времени. Этот модифицированный тест Чоу называют *стatisтикой отношения правдоподобия Куандта* (*QLR1*) (Quandt, 1960) (термин, который мы будем использовать), или, обобщенно, *супремум-стatisтикой Вальда*<sup>2</sup>.

Поскольку QLR-стatisтика является наибольшей из множества  $F$ -стatisтик, ее распределение не является таким же, как у отдельных  $F$ -стatisтик. Поэтому критические значения для QLR-стatisтики должны быть получены из специального распределения. Как и  $F$ -стatisтика, это распределение зависит от числа тестируемых ограничений  $q$ , то есть числа коэффициентов (включая константу), в которых предполагаются структурные сдвиги (или изменения), в альтернативной гипотезе. Распределение QLR-стatisтики также зависит от  $\tau_0/T$  и  $\tau_1/T$ , то есть от значений концов  $\tau_0$  и  $\tau_1$  подвыборки, для которой вычисляется  $F$ -стatisтика, выраженных в виде доли от общего размера выборки.

Для того чтобы QLR-стatisтика хорошо приближалась к своему асимптотическому распределению, концы подвыборки  $\tau_0$  и  $\tau_1$  не должны быть слишком близки к началу или концу самой выборки. По этой причине на практике QLR-стatisтики вычисляются над «усеченной» выборкой или подмножеством имеющейся

<sup>1</sup> Quandt likelihood ratio statistic. – Примеч. науч. ред. перевода.

<sup>2</sup> Sup-Wald statistic. – Примеч. науч. ред. перевода.

выборки. Распространенным является использование 15 %-го усечения, то есть выбора  $\tau_0 = 0,15T$  и  $\tau_1 = 0,85T$  (с округлением до целого числа). С 15 %-м усечением  $F$ -статистика для даты структурного сдвига вычисляется в центральных 70 % выборки.

Критические значения для QLR-статистики, рассчитанные с 15 %-м усечением, приведены в таблице 14.6. Сравнение этих критических значений со значениями распределения  $F_{q, \infty}$  (табл. 4 приложения) показывает, что критические значения для QLR-статистики больше. Это отражает тот факт, что QLR-статистика строится как наибольшая из множества отдельных  $F$ -статистик. Изучив  $F$ -статистики для всех возможных дат структурного сдвига, QLR-статистика может довольно часто отвергать нулевую гипотезу, что приводит к критическим значениям QLR-статистики, большим, чем критические значения отдельных  $F$ -статистик.

Таблица 14.6

## Критические значения QLR-статистики с 15 %-м отсечением выборки

	10 %	5 %	1 %
1	7,12	8,68	12,16
2	5,00	5,86	7,78
3	4,09	4,71	6,02
4	3,59	4,09	5,12
5	3,26	3,66	4,53
6	3,02	3,37	4,12
7	2,84	3,15	3,82
8	2,69	2,98	3,57
9	2,58	2,84	3,38
10	2,48	2,71	3,23
11	2,40	2,62	3,09
12	2,33	2,54	2,97
13	2,27	2,46	2,87
14	2,21	2,40	2,78
15	2,16	2,34	2,71
16	2,12	2,29	2,64
17	2,08	2,25	2,58
18	2,05	2,20	2,53
19	2,01	2,17	2,48
20	1,99	2,13	2,43

*Примечание.* Критические значения получены для случая, когда  $\tau_0 = 0,15T$  и  $\tau_1 = 0,85T$  (округленными до ближайшего целого числа), так что  $F$ -статистика вычислялась для всех возможных дат структурных сдвигов в центральных 70 % выборки. Число ограничений  $q$  равно числу ограничений, тестируемых при помощи  $F$ -статистики в каждом конкретном случае. Критические значения для других процентов отсечения выборки приведены в работе Эндрюса (Andrews, 2003).

Аналогично тесту Чоу, QLR-тест может быть использован, чтобы проверить на возможность наличия структурных сдвигов только некоторые коэффициенты регрессии. Для того чтобы сделать это, нужно сначала провести тесты Чоу для различных дат структурных сдвигов, используя бинарную компоненту взаимо-

действия только для переменных, коэффициенты которых подозреваются в наличии структурного сдвига, а затем вычислить наибольшее значение среди этих тестов Чоу для дат потенциальных структурных сдвигов, находящихся в интервале  $\tau_0 \leq \tau \leq \tau_1$ . Критические значения для этой спецификации QLR-теста также можно взять из таблицы 14.6, где число ограничений ( $q$ ) соответствует количеству ограничений, тестируемых при помощи  $F$ -статистик.

### QLR-тест на стабильность коэффициентов

Пусть  $F(\tau)$  обозначает  $F$ -статистику для проверки гипотезы об отсутствии структурного сдвига в коэффициентах регрессии в момент времени  $\tau$ , например, в регрессии (14.35) это  $F$ -статистика для тестирования нулевой гипотезы  $\gamma_0 = \gamma_1 = \gamma_2 = 0$ . QLR-статистика (или супремум-статистика Вальда) представляет собой наибольшее значение статистики среди всех значений статистик, рассчитанных для моментов структурных сдвигов, находящихся в диапазоне  $\tau_0 \leq \tau \leq \tau_1$ :

$$QLR = \max[F(\tau_0), F(\tau_0 + 1), \dots, F(\tau_1)]. \quad (14.36)$$

Как и  $F$ -статистика, QLR-статистика может быть использована для проверки гипотезы об отсутствии структурного сдвига во всех или только в некоторых коэффициентах регрессии.

В больших выборках распределение QLR-статистики в условиях нулевой гипотезы зависит от числа ограничений  $q$  и от значений концов интервала  $\tau_0$  и  $\tau_1$  в виде их доли от  $T$ . Критические значения приведены в таблице 14.6 для 15%-го отсечения ( $\tau_0 = 0,15T$  и  $\tau_1 = 0,85T$ , округленные до ближайшего целого числа).

QLR-тест может обнаружить один дискретный структурный сдвиг, несколько структурных сдвигов и/или медленное изменение функции регрессии.

Если существует четкий структурный сдвиг в функции регрессии, то оценкой даты структурного сдвига является момент, в который статистика теста Чоу принимает наибольшее значение.

## ОСНОВНЫЕ ПОНЯТИЯ

### 14.9

Если структурный сдвиг происходит в момент времени, находящийся в тестируемом диапазоне, то в больших выборках QLR-статистика с высокой вероятностью будет отвергать нулевую гипотезу. Кроме того, момент времени  $\hat{\tau}$ , в который  $F$ -статистика принимает максимальное значение, является оценкой даты структурного сдвига  $\tau$ . Эта оценка хороша в том смысле, что при выполнении определенных условий  $\hat{\tau}/T \xrightarrow{P} \tau/T$ , то есть оценка даты структурного сдвига в долях длины выборки является состоятельной.

QLR-статистика также с высокой вероятностью отвергает нулевую гипотезу в больших выборках, если есть несколько дискретных структурных сдвигов

или если структурный сдвиг имеет вид медленно изменяющейся функции регрессии. Это означает, что QLR-статистика обнаруживает формы нестабильности, отличные от единичного дискретного структурного сдвига. В результате, если QLR-статистика отвергает нулевую гипотезу, это может означать, что имеется один дискретный структурный сдвиг, что существует множество дискретных структурных сдвигов или что функция регрессии медленно изменяется.

Все основные моменты, связанные с использованием QLR-статистики, приведены во вставке «Основные понятия 14.9».

**Внимание: очень вероятно, что вы не знаете даты структурного сдвига, даже если вы думаете, что знаете.** Иногда эксперт может считать, что он или она знает дату возможного структурного сдвига, поэтому использует тест Чоу вместо QLR-теста. Но если эта уверенность основана на мнении эксперта об исследуемом временном ряде, то на самом деле дата структурного сдвига была рассчитана с использованием данных, хотя и таким неформальным способом. Предварительная оценка даты структурного сдвига означает, что критические значения обычной  $F$ -статистики не могут быть использованы для теста Чоу на структурный сдвиг, произошедший в этот момент времени. Таким образом, и в этой ситуации целесообразно использовать QLR-статистику.

**Приложение: стабильна ли кривая Филлипса?** QLR-тест дает возможность проверить, является ли кривая Филлипса стабильной с 1962 по 2004 год. Более того, мы попытаемся понять, происходили ли изменения в коэффициентах при запаздывающих значениях уровня безработицы и в константе в модели ADL (4,4) в спецификации (14.17), содержащей по четыре лага  $\Delta Inf_t$  и  $Unemp_t$ .

$F$ -статистика теста Чоу для проверки гипотезы о том, что константа и коэффициенты при  $Unemp_{t-1}$ , ...,  $Unemp_{t-4}$  в уравнении (14.17) постоянны против альтернативы о том, что они различны для двух подвыборок относительно определенной даты для структурного сдвига, находящейся в центральных 70 % выборки, представлены на рисунке 14.5. Например,  $F$ -статистика для тестирования гипотезы о наличии структурного сдвига в первом квартале 1980 года равна 2,85, и это значение соответствует значению для этой даты на графике. Каждая  $F$ -статистика тестирует наличие пяти ограничений (отсутствие изменений в константе и в четырех коэффициентах при лагах уровня безработицы), так что  $q = 5$ . Самая большая  $F$ -статистика равна 5,16 и имеет место в IV квартале 1981 года; это и есть QLR-статистика. Сравнение 5,16 с критическим значением из таблицы 14.6 указывает на то, что гипотеза о том, что эти коэффициенты являются стабильными, отвергнута на 1 %-м уровне значимости (критическое значение равно 4,53). Таким образом, существуют доказательства того, что по крайней мере один из пяти рассматриваемых коэффициентов менялся в выборке.

### **Псевдовневыборочное прогнозирование**

Основным критерием качества прогнозирования по модели являются свойства ее вневыборочных прогнозов, то есть качество прогнозов в «реальном

времени» после того, как модель оценена. *Псевдовневыборочное прогнозирование* является методом, позволяющим оценить в реальном времени качество прогнозов по модели, используемой для целей прогнозирования. Идея псевдовневыборочного прогнозирования проста: выберите дату, близкую к концу выборки, оцените модель, по которой вы потом будете прогнозировать, используя данные до этой даты, а затем постройте прогноз по оцененной модели. Повторив оценки для нескольких дат, близких к концу вашей выборки, вы получите ряд псевдопрогнозов и, как следствие, ошибки псевдопрогноза. Ошибки псевдопрогноза могут быть рассмотрены, чтобы понять, являются ли прогнозы такими, какими вы могли бы ожидать, предполагая устойчивость прогнозов.



**Рисунок 14.5. F-статистики теста на отсутствие структурного сдвига в уравнении (14.17) в различные моменты времени**

При данной дате структурного сдвига изображенная здесь  $F$ -статистика проверяет нулевую гипотезу о наличии структурного сдвига по крайней мере в одном из коэффициентов при  $Unemp_{t-1}$ ,  $Unemp_{t-2}$ ,  $Unemp_{t-3}$ ,  $Unemp_{t-4}$  или константе в уравнении (14.17). Например,  $F$ -статистика для тестирования гипотезы о наличии структурного сдвига в I квартале 1980 года равна 2,85. QLR-статистика является наибольшей из всех этих  $F$ -статистик и равна 5,16, что превышает 1%-е критическое значение, равное 4,53.

Причиной, по которой выбрано название «псевдовневыборочное прогнозирование», является то, что оно не является на самом деле вневыборочным прогнозированием. Вневыборочное прогнозирование имеет место в реальном времени, то есть, когда вы делаете свой прогноз, не зная будущих значений ряда. При псевдовневыборочном прогнозировании вы делаете вид, что прогнозируете в реальном времени используя свою модель, но у вас есть «будущие» данные, для которых вы строите рассчитанные или псевдопрогнозы. Псевдовневыборочное прогнозирование имитирует процесс прогнозирования, который

происходил бы в реальном времени, но нам не нужно дожидаться появления новых данных.

Псевдовневыборочное прогнозирование дает прогнозисту возможность понять, насколько хорошо прогнозирует модель в конце выборки. Оно может дать ценную информацию, которая либо укрепит доверие относительно того, что модель прогнозирует хорошо, либо, наоборот, покажет, что модель прогнозирует плохо. Методология псевдовневыборочного прогнозирования приведена во вставке «Основные понятия 14.10».

**Другие приложения псевдовневыборочных прогнозов.** Вторым применением псевдовневыборочных прогнозов является возможность оценки RMSFE. Поскольку псевдовневыборочные прогнозы вычисляются с использованием данных только до прогнозируемой даты, ошибки псевдовневыборочных прогнозов отражают как неопределенность, связанную с будущими значениями остаточного члена регрессии, так и неопределенность, возникающую из-за возможных ошибок в оценках коэффициентов регрессии, то есть ошибки псевдовневыборочных прогнозов включают в себя оба источника ошибок из уравнения (14.21). Таким образом, выборочное стандартное отклонение ошибок псевдовневыборочных прогнозов является оценкой RMSFE. Как обсуждалось в разделе 14.4, эта оценка RMSFE может быть использована как количественная мера неопределенности прогноза и для построения интервальных прогнозов.

## ОСНОВНЫЕ ПОНЯТИЯ

### 14.10

#### Псевдовневыборочные прогнозы

Псевдовневыборочные прогнозы вычисляются с использованием следующих шагов:

1. Выберите количество наблюдений  $P$ , для которых вы будете строить псевдовневыборочные прогнозы, например,  $P$  может быть равно 10 или 15 % от всей выборки. Пусть  $s = T - P$ .
2. Оцените регрессию, на основе которой вы будете строить прогнозы, используя сокращенную выборку наблюдений с  $t = 1, \dots, s$ .
3. Вычислите прогноз для значения под номером  $s + 1$ ; назовем его  $\tilde{Y}_{s+1|s}$ .
4. Вычислите ошибку прогноза  $\tilde{u}_{s+1} = Y_{s+1} - \tilde{Y}_{s+1|s}$ .
5. Повторите шаги со 2 по 4 для оставшихся дат с  $s = T - P + 1$  до  $T - 1$  (переоцените регрессии для каждой даты). Множество псевдовневыборочных прогнозов имеет вид:  $\{\tilde{Y}_{s+1|s}, s = T - P, \dots, T - 1\}$ , а ошибок псевдовневыборочных прогнозов:  $\{\tilde{u}_{s+1}, s = T - P, \dots, T - 1\}$ .

Третьим применением псевдовневыборочного прогнозирования является возможность сравнения двух или более моделей, используемых для целей про-

гнозирования. Две модели, которые кажутся одинаково хорошо приближающими данные, могут совершенно по-разному вести себя с точки зрения псевдовневыборочного прогнозирования. Например, если в модели включаются различные регрессоры, псевдовневыборочное прогнозирование дает возможность сравнить этих две модели, сфокусировавшись на сравнении их способности давать достоверные прогнозы.

**Приложение: изменилась ли кривая Филлипса в 1990-х годах?** Используя QLR-статистику, мы отвергли на 1%-м уровне значимости нулевую гипотезу о том, что кривая Филлипса была стабильной, в пользу альтернативной гипотезы о наличии структурного сдвига (см. рис. 14.5). F-статистика приняла максимальное значение в IV квартале 1981 года, указывая на то, что структурный сдвиг произошел в начале 1980-х годов. Все это предполагает, что если прогнозист использует запаздывание показателя безработицы для прогнозирования инфляции, то следует использовать оценки регрессии на выборке после структурного сдвига, произошедшего в IV квартале 1981 года. Тем не менее остается вопрос: является ли кривая Филлипса стабильной моделью для прогнозирования после структурного сдвига в IV квартале 1981 года?



### **Можем ли мы перехитрить рынок? Часть II**

Возможно, вы слышали совет, следуя которому вы должны покупать акции, когда доходы по ним высоки по сравнению с их ценой. Покупка акций является, в сущности, покупкой потока будущих дивидендов, выплачиваемых этой компанией из своих доходов. Если поток дивидендов необычно велик по отношению к цене акций компании, то компания может считаться недооцененной. Если текущие дивиденды характеризуют будущие дивиденды, то доходность по дивидендам – отношение текущих дивидендов к цене акции – может служить прогнозом будущей избыточной доходности акций. Если доходность по дивидендам высока, то акции недооценены и будет прогнозироваться растущая доходность.

Такие рассуждения предлагают изучить авторегрессионные модели с распределенными лагами для избыточной доходности, в которых регрессором будет доходность по дивидендам. Но здесь возникает трудность: доходность по дивидендам отличается высокой устойчивостью и даже может содержать стохастический тренд. Используя ежемесячные данные логарифма отношения стоимости дивидендов к цене акции для индекса CRSP, взвешенного по стоимости (данные приведены в приложении 14.1), с января 1960 года по декабрь 2002 года, при помощи теста Дики–Фуллера на единичный корень (с включением константы) получаем, что мы не можем отклонить нулевую гипотезу о наличии единичного корня даже на 10%-м уровне значимости. Как всегда, эта неспособность отвергнуть нулевую гипотезу не означает, что нулевая гипотеза верна, но она подчеркивает, что доходность по дивидендам является высоко устойчивым регрессором. Следуя логике раздела 14.6, этот результат показывает, что мы должны использовать первую

разность логарифма доходности по дивидендам в качестве регрессора, а не уровень логарифма доходности по дивидендам.

В таблице 14.7 представлены оценки ADL-модели зависимости избыточной доходности от взвешенного по стоимости индекса CRSP. В столбцах (1) и (2) показатель доходности по дивидендам рассматривается в первых разностях, а индивидуальные *t*-статистики и совместная *F*-статистика не отвергают нулевые гипотезы об отсутствии предсказуемости. Но несмотря на то что эти спецификации соответствуют рекомендациям по моделированию из раздела 14.6, они противоречат экономическим аргументам, приведенным во вступительном абзаце, согласно которым избыточная доходность зависит от уровня доходности по дивидендам. В этой связи в столбце (3) таблицы 14.7 приводятся оценки ADL(1, 1)-модели для зависимости избыточной доходности от логарифма доходности по дивидендам, оцененной на интервале до декабря 1992 года, *t*-статистика коэффициента при логарифме первого лага дивидендной доходности равна 2,25, что превышает обычное 5%-е критическое значение, равное 1,96. Однако из-за того что регрессор отличается высокой устойчивостью, распределение этой *t*-статистики подозрительно, и критическое значение, равное 1,96, может не подходить для использования в этом случае. (*F*-статистика для этой регрессии не сообщается, потому что она не обязательно имеет распределение хи-квадрат, даже в больших выборках, из-за высокой устойчивости регрессора.)

Одним из способов оценки явной предсказуемости, следующей из оценок (3) таблицы 14.7, является расчет псевдовневыборочных прогнозов. Рассчитывая псевдовневыборочные прогнозы для периода 1993:1–2002:12, получаем, что выборочная среднеквадратическая ошибка прогнозирования равна 4,08%. В противоположность этому выборочная RMSFE для прогноза, заключающегося в том, что избыточная доходность равна нулю, составляет 4,00%, а выборочная RMSFE «постоянного прогноза» (в котором рекурсивно оцененная модель, используемая для прогнозирования, включает только константу) равна 3,98%. Псевдовневыборочный прогноз на основе ADL(1, 1)-модели с доходностью по дивидендам оказывается хуже прогнозов, в которых нет регрессоров!

Это отсутствие предсказуемости согласуется с гипотезой об эффективности рынка в сильной форме, которая утверждает, что вся общедоступная информация включена в цены на акции, так что их доходность не должна быть предсказуемой при использовании только этой общедоступной информации (гипотеза в слабой форме предполагает, что прогнозы основываются только на информации о прошлых доходностях). Основная идея о том, что избыточную доходность не так легко предсказать, имеет смысл: если бы это было возможно, цены на акции выравнивались бы до уровня, когда не существовало бы ожидаемой избыточной доходности.

Интерпретация результатов, аналогичных результатам из таблицы 14.7, является предметом жарких споров среди финансовых экономистов. Некоторые из них рассматривают отсутствие у регрессии предсказательной способности как доказательство гипотезы об эффективности рынков (см., напр., Goyal, Welch, 2003). Другие говорят, что регрессии, оцененные на больших выборках и для больших горизонтов, при анализе с помощью методов, специально предназначенных для работы с высоко устойчивыми регрессорами, демонстрируют предсказуемость

(см. Campbell, Yogo, 2006). Такая предсказуемость может возникать в результате рационального экономического поведения, при котором отношение инвесторов к риску меняется в течение экономического цикла (Campbell, 2003) или это может отражать «иррациональное изобилие» (Shiller, 2005).

Результаты из таблицы 14.7 получены для месячных доходностей, но некоторые финансовые эконометристы используют более высокочастотные данные. Теория «микроструктуры рынка», рассматривающая ежеминутные изменения фондового рынка, предполагает, что могут существовать короткие периоды предсказуемости и что умные и ловкие могут заработать на этом деньги. Но такая работа требует нервов и больших вычислительных расчетов, а также наличия штата талантливых эконометристов.

Таблица 14.7

**Авторегрессионные модели с распределенными лагами  
для месячного показателя избыточной доходности**

Зависимая переменная: избыточные доходности взвешенного по стоимости индекса CRSP			
	(1)	(2)	(3)
Спецификация модели	ADL (1,1)	ADL (2,2)	ADL (1,1)
Период оценки	1960:1 – 2002:12	1960:1 – 2002:12	1960:1 – 1992:12
Объясняющие переменные			
$excess\ return_{t-1}$	0,059 (0,158)	0,042 (0,162)	0,078 (0,057)
$excess\ return_{t-2}$		-0,213 (0,193)	
$\Delta \ln(divident\ yield_{t-1})$	0,009 (0,157)	-0,012 (0,163)	
$\Delta \ln(divident\ yield_{t-2})$		-0,161 (0,185)	
$\ln(divident\ yield_{t-1})$			0,026 <sup>a</sup> (0,012)
Константа	0,003 1 (0,002 0)	0,003 7 (0,002 1)	0,090 <sup>a</sup> (0,039)
<i>F</i> -статистика для проверки значимости регрессии ( <i>p</i> -значение)	0,501 (0,606)	0,843 (0,497)	
$\bar{R}^2$	-0,001 4	-0,000 8	0,013 4

*Примечания.* Данные описаны в приложении 14.1. В строках с названиями регрессоров приведены значения коэффициентов при них с *p*-значениями в скобках. В последних двух строках приведены *F*-статистики для проверки гипотезы о том, что все коэффициенты регрессии равны нулю и их

*p*-значения в скобках, а также скорректированный  $\bar{R}^2$ .

<sup>a</sup>  $|t| > 1,96$ .



Если коэффициенты кривой Филлипса изменились несколько раз в период с I квартала 1982 по I квартал 2004 года, то псевдовневыборочные прогнозы,

рассчитываемые с использованием данных, начиная с I квартала 1982 года должны ухудшиться. Псевдовневыборочные прогнозы инфляции на I квартал 1999 – IV квартал 2004 года, вычисленные по кривой Филлипса, включающей четыре запаздывания и оцененной на интервале с I квартала 1982 года, представлены на рисунке 14.6 вместе с фактическими значениями инфляции. Например, прогноз инфляции на I квартал 1999 года вычислен следующим образом. Сначала была оценена регрессия зависимости  $\Delta\text{Inf}_t$  от  $\Delta\text{Inf}_{t-1}, \dots, \Delta\text{Inf}_{t-4}, \text{Unemp}_{t-1}, \dots, \text{Unemp}_{t-4}$  и константы с использованием данных по IV квартал 1998 года. А затем с использованием оценок коэффициентов этой регрессии и имеющихся данных до IV квартала 1998 года был рассчитан прогноз  $\widehat{\Delta\text{Inf}}_{1999:\text{I}|1998:\text{IV}}$ . Прогноз инфляции на I квартал 1999 года тогда равен  $\text{Inf}_{1999:\text{I}|1998:\text{IV}} = \text{Inf}_{1998:\text{IV}} + \widehat{\Delta\text{Inf}}_{1999:\text{I}|1998:\text{IV}}$ . Мы повторили эту процедуру с использованием данных до I квартала 1999 года, чтобы вычислить прогноз  $\widehat{\Delta\text{Inf}}_{1999:\text{II}|1999:\text{I}}$ . Повторяя эту процедуру для всех 24 кварталов с I квартала 1999 по IV квартала 2004 года, мы получили 24 псевдовневыборочных прогноза, которые графически представлены на рисунке 14.6. Ошибки псевдовневыборочных прогнозов равны разности между фактической инфляцией и ее псевдовневыборочным прогнозом, то есть разности между двумя линиями на рисунке 14.6. Например, в IV квартале 2000 года уровень инфляции снизился на 0,8 %, но псевдовневыборочный прогноз  $\Delta\text{Inf}_{2000:\text{IV}}$  был равен 0,3 %, так что ошибка псевдовневыборочного прогноза составила  $\Delta\text{Inf}_{2000:\text{IV}} - \widehat{\Delta\text{Inf}}_{2000:\text{IV}|2000:\text{III}} = -0,8 - 0,3 = -1,1$  процентных пункта. Другими словами, прогнозист при помощи ADL(4, 4) – модели кривой Филлипса, оцененной по данным до III квартала 2000 года, спрогнозировал бы, что инфляция увеличится на 0,3 % в IV квартале 2000 года, тогда как в действительности она снизилась на 0,8 %.

Как среднее значение и стандартное отклонение ошибок псевдовневыборочных прогнозов сравнить с внутривыборочным качеством подгонки модели? Стандартная ошибка регрессии кривой Филлипса с четырьмя запаздываниями, оцененная на интервале с I квартала 1982 по IV квартал 1998 года, равна 1,30, так что, основываясь на таком внутривыборочном качестве подгонки, мы ожидаем, что ошибка вневыборочного прогноза будет иметь нулевое среднее, а квадратный корень из среднеквадратической ошибки прогноза составит 1,30. На самом деле на интервале с I квартала 1999 по IV квартал 2004 года средняя ошибка псевдовневыборочного прогноза равна 0,11, и  $t$ -статистика для гипотезы о том, что средняя ошибка прогноза равна нулю, составляет 0,41, поэтому предположение о равенстве нулю среднего прогноза не отвергается. Кроме того, RMSFE псевдовневыборочных прогнозов для этого периода составляет 1,32, что очень близко к значению 1,30 для стандартной ошибки регрессии, оцененной на интервале с I квартала 1999 по IV квартал 2004 года. Более того, на графике, изображающем прогнозы и их ошибки на рисунке 14.6, не видно наличия больших выбросов или необычных расхождений.

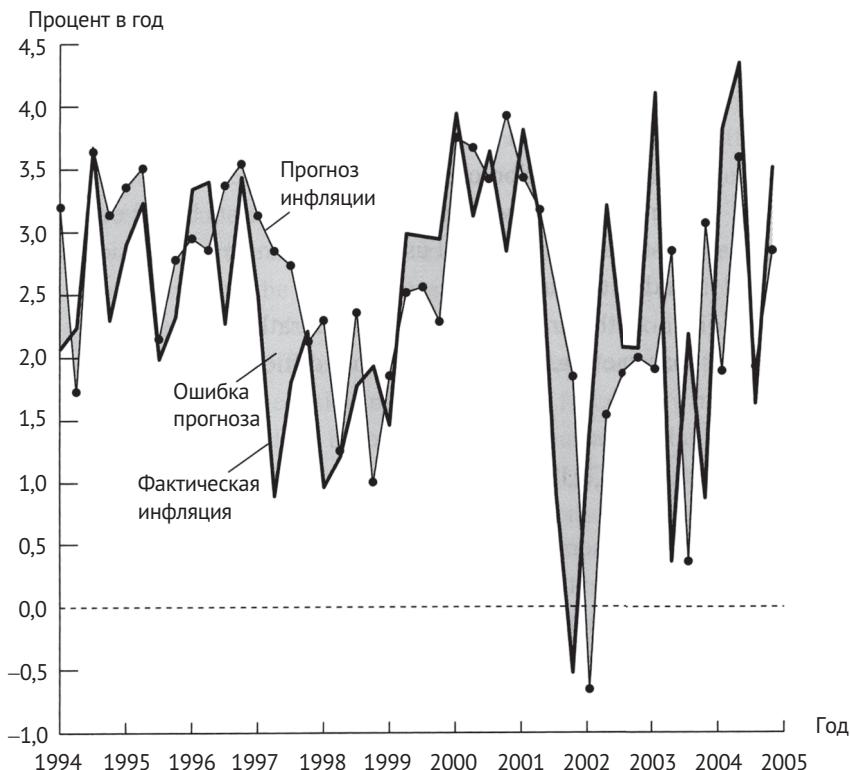


Рисунок 14.6. Инфляция в США и псевдовневыборочные прогнозы

Псевдовневыборочные прогнозы, полученные при помощи кривой Филлипса с четырьмя запаздываниями, описываемой уравнением (14.17), в целом согласуются как с фактической инфляцией, так и с устойчивой кривой Филлипса, оцененной на интервале с 1982 года.

В соответствии с полученным псевдовневыборочным прогнозом, качество кривой Филлипса как модели, используемой для прогнозирования инфляции в течение псевдовневыборочного периода с I квартала 1999 по IV квартал 2004 года, сравнимо с ее качеством во внутривыборочный период с I квартала 1982 по IV квартал 1998 года. Таким образом, несмотря на то что QLR-тест указывает на нестабильность кривой Филлипса в начале 1980-х годов, этот псевдовневыборочный анализ показывает, что после структурного сдвига, произошедшего в начале 1980-х годов, кривая Филлипса, используемая для прогнозирования, была стабильной.

### **Как избежать проблем, возникающих из-за структурных сдвигов**

Лучшим способом учета структурного сдвига в теоретической функции регрессии является построение зависимости от источника этого структурного сдвига. Если структурный сдвиг происходит в конкретный момент времени, то он будет с высокой вероятностью обнаружен при помощи QLR-статистики,

и тогда дата структурного сдвига может быть оценена. Таким образом, функция регрессии может быть оценена с использованием бинарной переменной, которая является индикатором двух различных подвыборок, связанных с этим структурным сдвигом, и связанная с другими регрессорами, если в этом есть необходимость. Если все коэффициенты имеют структурный сдвиг, то такая регрессия принимает форму уравнения (14.35), где  $\tau$  заменяется на оценку даты структурного сдвига  $\hat{\tau}$ , если же только некоторые из коэффициентов имеют структурный сдвиг, то в регрессии появляются только необходимые регрессоры, характеризующие эти изменения. Если имеет место явно выраженный отдельный структурный сдвиг, то все выводы о коэффициентах регрессии могут быть сделаны как обычно, используя, например, критические значения обычного нормального распределения для проверки гипотез на основе  $t$ -статистики. Кроме того, оцененная функция регрессии может быть использована для построения прогнозов.

Если структурный сдвиг нечетко выражен и является, скорее, результатом медленного, постепенного изменения параметров, то его идентификация является более сложным процессом и выходит за рамки этой книги<sup>1</sup>.

## 14.8. Заключение

Данные, имеющие структуру временных рядов, обычно коррелированы во времени. Следствием этой корреляции является то, что оцененная линейная регрессия может быть использована для прогнозирования будущих значений временного ряда, основываясь на его текущем и прошлых значениях. Отправной точкой для регрессии временных рядов является модель авторегрессии, в которой регрессорами являются запаздывающие значения зависимой переменной. Если есть дополнительные объясняющие переменные, то их запаздывания могут быть включены в регрессию.

В данной главе рассматривается несколько технических вопросов, возникающих при оценке и использовании регрессии, оцененной для данных, имеющих структуру временных рядов. Одним из таких вопросов является определение количества лагов, которые надо включать в регрессию. Как отмечалось в разделе 14.5, если число запаздываний выбирается минимизацией критерия ВIC, то оцененное число запаздываний является состоятельной оценкой истинного числа запаздываний.

Другой из этих вопросов связан с проблемой стационарности исследуемых временных рядов. Если временные ряды являются стационарными, то могут быть использованы обычные методы статистического анализа (например, сравнение  $t$ -статистик с критическими значениями нормального распределения), и так как теоретическая функция регрессии устойчива во времени, регрессия, оцененная на основе исторических данных, может быть использована для прогнозирова-

---

<sup>1</sup> Дополнительное обсуждение методов оценки и тестирования при наличии дискретных структурных сдвигов см.: Hansen (2001). Для обсуждения более продвинутых методов оценки и прогнозирования при наличии медленных изменений в коэффициентах см.: Hamilton (1994. Ch. 13).

ния. Однако если временные ряды не являются стационарными, то ситуация осложняется, и характер осложнений зависит от природы нестационарности. Например, если ряд является нестационарным из-за присутствия стохастического тренда, то МНК-оценки и  $t$ -статистика могут иметь нестандартные (ненормальные) распределения даже в больших выборках, и качество прогнозов может быть улучшено, если оценить регрессию в первых разностях. Тест для обнаружения этого типа нестационарности, расширенный тест Дики–Фуллера на единичный корень, был рассмотрен в разделе 14.6. В качестве альтернативы, если теоретическая функция регрессии имеет структурный сдвиг, то пренебрежение этим структурным сдвигом приводит к оценке усредненной теоретической функции регрессии, что, в свою очередь, может привести к смешанным и / или неточным прогнозам. Процедуры идентификации наличия структурных сдвигов в теоретической функции регрессии были рассмотрены в разделе 14.7.

В данной главе методы регрессионного анализа временных рядов были применены к экономическому прогнозированию, а коэффициенты в этих прогнозных моделях не имели причинной интерпретации. Для прогнозирования нам не нужно присутствие причинных связей, и игнорирование таких причинных интерпретаций дает нам определенную свободу с точки зрения получения хороших прогнозов. Однако в некоторых случаях стоящая перед нами задача заключается не в получении прогнозов, а в оценке причинно-следственных связей между временными рядами, то есть нам нужно оценить динамический причинный эффект влияния изменений переменной  $X$  во времени на  $Y$ . При определенных условиях методы, рассматриваемые в данной главе, или тесно связанные с ними методы могут быть использованы для оценки динамических причинных эффектов, что подробно обсуждается в следующей главе.

## Выводы

1. Регрессионные модели, используемые для прогнозирования, могут не иметь причинной интерпретации.
2. Временные ряды обычно коррелированы во времени, то есть серийно коррелированы.
3. Авторегрессии порядка  $p$  являются моделью множественной линейной регрессии, в которой регрессоры – первые  $p$  лагов зависимой переменной. Коэффициенты AR( $p$ ) можно оценить с помощью МНК, а оцененная функция регрессии может быть использована для прогнозирования. Глубина запаздывания  $p$  может быть оценена с использованием информационных критериев, таких как критерий BIC.
4. Включение в регрессию других переменных и их запаздываний может улучшить качество прогнозирования. При выполнении предположений метода наименьших квадратов для регрессии временных рядов (вставка «Основные понятия 14.6») МНК-оценки асимптотически нормально распределены, и статистические выводы можно делать так же, как и для случая межобъектных данных.

5. Интервальные прогнозы являются одним из способов количественной оценки неопределенности прогноза. Если ошибки регрессии распределены нормально, то приближенный 68 %-й интервальный прогноз может быть построен как значение прогноза плюс или минус одно значение оценки корня квадратного из среднеквадратической ошибки прогноза.
6. Временной ряд, содержащий стохастический тренд, является нестационарным, что нарушает второе предположение метода наименьших квадратов из вставки «Основные понятия 14.6».  $t$ -статистика для коэффициента при стохастическом тренде может иметь нестандартное распределение, что может привести к смещению МНК-оценок, неэффективным прогнозам и некорректным выводам. ADF-статистика может быть использована для тестирования гипотезы о наличии стохастического тренда. Стохастический тренд (случайное блуждание) может быть устранен, если рассмотреть первую разность временного ряда.
7. Если теоретическая функция регрессии изменяется с течением времени, то МНК-оценки, не учитывающие эту неустойчивость, не являются надежной основой для статистических выводов или прогнозов. QLR-статистика может быть использована для проверки отсутствия структурных сдвигов, и в случае обнаружения дискретного структурного сдвига функция регрессии может быть переоценена таким образом, чтобы учесть этот структурный сдвиг.
8. Псевдовневыборочные прогнозы могут быть использованы для оценки стабильности модели в конце выборки, для оценки квадратного корня из среднеквадратической ошибки прогноза и для сравнения различных моделей, использующихся для прогнозирования.

## **Основные понятия**

- Первое запаздывание (лаг) (с. 551).  
 $j$ -й лаг (с. 551).  
Первая разность (с. 551).  
Автокорреляция (с. 553).  
Серийная корреляция (с. 553).  
Коэффициент автокорреляции (с. 553).  
Автокорреляция  $j$ -го порядка (с. 554).  
Авторегрессия (с. 557).  
Ошибка прогнозирования (с. 558).  
Квадратный корень из среднеквадратичной ошибки прогнозирования (RMSFE) (с. 559).  
Модель авторегрессии порядка  $p$  [AR( $p$ )] (с. 560).  
Авторегрессионная модель с распределенными лагами (ADL) (с. 565).  
 $ADL(p, q)$  (с. 565).  
Стационарность (с. 566).  
Слабая зависимость (с. 568).  
Статистика Грейнджа для проверки причинности (с. 569).

Тест Грейнджа на причинность (с. 569).  
Интервальный прогноз (с. 570).  
Байесовский информационный критерий (BIC) (с. 573).  
Информационный критерий Акаике (AIC) (с. 574).  
Тренд (с. 577).  
Детерминированный тренд (с. 577).  
Стохастический тренд (с. 577).  
Случайное блуждание (с. 578).  
Случайное блуждание с дрейфом (сносом) (с. 578).  
Единичный корень (с. 579).  
Ложная (мнимая, кажущаяся) регрессия (с. 581).  
Тест Дики—Фуллера (с. 582).  
Статистика Дики—Фуллера (с. 582).  
Расширенная статистика Дики—Фуллера (ADF) (с. 583).  
Дата (момент) структурного сдвига (с. 588).  
Статистика отношения правдоподобия Куандта (QLR) (с. 589).  
Псевдовневыборочное прогнозирование (с. 593).  
Оператор запаздывания (с. 611).  
Лаговый полином (с. 611).  
Модель авторегрессии с ошибками в форме скользящего среднего (ARMA) (с. 612).

### ***Вопросы для повторения и закрепления основных понятий***

- 14.1. Посмотрите на график логарифма ВВП Японии на рисунке 14.2в. Выглядит ли этот временной ряд стационарным? Объясните. Предположим, что вы вычислили первую разность этого временного ряда. Будет ли этот временной ряд стационарным? Объясните.
- 14.2. Многие финансовые экономисты считают, что модель случайного блуждания хорошо описывает логарифмы цен на акции. Это означает, что процентное изменение цены акции непрогнозируемо. Финансовый аналитик утверждает, что у него есть новая модель, которая позволяет получать лучшие прогнозы, чем по модели случайного блуждания. Объясните, как вы проверили бы утверждение аналитика о том, что прогнозы по его модели лучше наивного прогноза (по модели случайного блуждания).
- 14.3. Исследователь оценивает модель AR(1) с константой и обнаруживает, что МНК-оценка коэффициента  $\beta_1$  равна 0,95 со стандартной ошибкой, равной 0,02. Будет ли 95 %-й доверительный интервал включать  $\beta_1 = 1$ ? Объясните.
- 14.4. Предположим, что вы подозреваете, что константа в уравнении (14.17) изменилась в I квартале 1992 года. Как вы изменили бы уравнение, чтобы учесть это изменение? Как вы проверили бы наличие такого изменения в константе? Как вы проверили бы наличие изменения в константе, если бы не знали его даты?

## Упражнения

- 14.1. Рассмотрим модель AR(1)  $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$ . Предположим, что процесс является стационарным.
- Покажите, что  $E(Y_t) = E(Y_{t-1})$ . (Подсказка: прочитайте вставку «Основные понятия 14.5».)
  - Покажите, что  $E(Y_t) = \beta_0 / (1 - \beta_1)$ .
- 14.2. Месячный индекс промышленного производства ( $IP$ ) является временными рядом, характеризующим количество промышленных товаров, произведенных в данном месяце. В задаче используются данные по этому показателю для США. Все регрессии оцениваются на интервале с января 1960 по декабрь 2000 года. Пусть  $Y_t = 1200 \times \ln(IP_t / IP_{t-1})$ .
- Прогнозист утверждает, что  $Y_t$  показывает ежемесячное процентное изменение в  $IP$ , измеренное в процентных пунктах в год. Правильно ли это? Почему?
  - Предположим, что прогнозист оценивает следующую AR(4)-модель для  $Y_t$ :

$$\hat{Y}_t = 1,377 - 0,318Y_{t-1} - 0,123Y_{t-2} + 0,068Y_{t-3} + 0,001Y_{t-4},$$

(0,062)      (0,078)      (0,055)      (0,068)      (0,056)

Используйте эту AR(4)-модель для прогнозирования значения  $Y_t$  в январе 2001 года с использованием следующих значений  $IP$  с августа 2000 по декабрь 2000 года:

Дата	2000:7	2000:8	2000:9	2000:10	2000:11	2000:12
$IP$	147,595	148,650	148,973	148,660	148,206	147,300

- Предполагая возможное наличие сезонных колебаний в производстве, прогнозист добавляет  $Y_{t-12}$  в авторегрессию. Оценка соответствующего коэффициента равна  $-0,054$  со стандартной ошибкой, равной  $0,053$ . Является ли этот коэффициент статистически значимым?
- Предполагая возможность наличия структурного сдвига, она вычисляет QLR-тест (с 15 %-м отсечением) для константы и AR-коэффициентов в AR(4)-модели. Полученная QLR-статистика равна 3,45. Является ли это доказательством наличия структурного сдвига? Объясните.
- Предполагая, что она могла включить в модель слишком мало или слишком много лагов, прогнозист оценивает AR( $p$ )-модели для  $p=1, \dots, 6$  на том же временном периоде. Суммы квадратов остатков каждой из оцененных моделей приведены в таблице. Используйте BIC, чтобы оценить количество лагов, которые должны быть включены в авторегрессию. Будут ли результаты другими, если вы используете AIC?

Порядок AR	1	2	3	4	5	6
SSR	29,175	28,538	28,393	28,391	28,378	28,317

- 14.3. Используя данные из упражнения 14.2, исследователь тестирует наличие стохастического тренда в  $\ln(IP_t)$  с помощью регрессии:

$$\widehat{\Delta \ln(IP_t)} = 0,061 + 0,00004t - 0,018 \ln(IP_{t-1}) + 0,333 \Delta \ln(IP_{t-1}) + 0,162 \Delta \ln(IP_{t-2}),$$

(0,024)                    (0,00001)                    (0,007)                    (0,075)                    (0,055)

где стандартные ошибки в скобках вычисляются в предположении гомоскедастичности, а « $t$ » – линейный тренд.

- а) Используйте ADF-статистику для проверки гипотезы о наличии стохастического тренда (единичного корня) в  $\ln(IP_t)$ .
- б) Согласуются ли эти результаты со спецификацией, оцененной в упражнении 14.2? Объясните.
- 14.4. Прогнозист из упражнения 14.2 расширяет AR(4)-модель для темпа роста  $IP$ , включив в нее четыре запаздывающих значения  $\Delta R_t$ , где  $R_t$  – процентная ставка по трехмесячным казначейским векселям США (измеряется в процентных пунктах в годовом исчислении).
- а)  $F$ -статистика теста Грейндженера на причинность для случая четырех лагов равна 2,35. Помогает ли процентная ставка предсказать темп роста  $IP$ ? Объясните.
- б) Исследователь также оценивает регрессию  $\Delta R_t$  на константу, четыре лага и четыре лага темпа роста  $IP$ . Результирующая  $F$ -статистика теста Грейндженера на причинность для четырех лагов темпа роста  $IP$  равна 2,87. Помогает ли темп роста  $IP$  предсказать процентные ставки? Объясните.
- 14.5. Докажите следующие утверждения об условных средних, прогнозах и ошибках прогноза:
- а) Пусть  $W$  – случайная величина со средним  $\mu_W$  и дисперсией  $\sigma_W^2$  и пусть  $c$  – некоторая константа. Покажите, что  $E[(W - c)^2] = \sigma_W^2 + (\mu_W - c)^2$ .
- б) Рассмотрим задачу прогнозирования  $Y_t$ , используя  $Y_{t-1}, Y_{t-2}, \dots$ . Обозначим через  $f_{t-1}$  некоторый прогноз  $Y_t$ , где индекс  $t-1$  у  $f_{t-1}$  указывает на то, что прогноз построен на основе данных до момента  $t-1$  включительно. Пусть  $E[(Y_t - f_{t-1})^2 | Y_{t-1}, Y_{t-2}, \dots]$  – условная среднеквадратичная ошибка прогноза  $f_{t-1}$  относительно  $Y_t$ , наблюдаемых до момента  $t-1$ . Покажите, что условная среднеквадратичная ошибка прогноза минимальна, когда  $f_{t-1} = Y_{t|t-1}$ , где  $Y_{t|t-1} = E(Y_t | Y_{t-1}, Y_{t-2}, \dots)$ . (Подсказка: вспомните упражнение 2.27.)
- в) Пусть  $u_t$  обозначает ошибку в уравнении (14.14). Покажите, что  $\text{cov}(u_t, u_{t-j}) = 0$  для  $j \neq 0$ . [Подсказка: используйте уравнение (2.27).]
- 14.6. В этом упражнении вы будете проводить эксперимент Монте-Карло, чтобы познакомиться с ложной регрессией, обсуждаемой в разделе 14.6. В методе Монте-Карло искусственные данные создаются при помощи компьютера, а затем эти искусственные данные используются для расчета изучаемых статистик. Все это позволяет вычислить распределения статистик в известных моделях в ситуации, когда математические выражения для этих распределений сложны (как здесь) или даже неизвестны. В упражнении вы будете генерировать данные, так чтобы два временных ряда  $Y_t$  и  $X_t$  были независимо распределенными случайными блужданиями. Для этого необходимо сделать следующие шаги:

- (i) Используя компьютер, сгенерируйте последовательность, состоящую из  $T = 100$  независимых одинаково распределенных стандартных нормальных случайных величин. Назовем эти случайные величины  $e_1, e_2, \dots, e_{100}$ . Пусть  $Y_1 = e_1$  и  $Y_t = Y_{t-1} + e_t$  для  $t = 2, 3, \dots, 100$ .
- (ii) Используя компьютер, создайте новую последовательность  $a_1, a_2, \dots, a_{100}$ , состоящую из  $T = 100$  i.i.d. стандартных нормальных случайных величин. Пусть  $X_1 = a_1$  и  $X_t = X_{t-1} + a_t$  для  $t = 2, 3, \dots, 100$ .
- (iii) Оцените регрессию  $Y_t$  на константу и  $X_t$ . Вычислите МНК-оценки, коэффициент детерминации  $R^2$  и  $t$ -статистику (только для случая гомоскедастичности) для проверки нулевой гипотезы о том, что  $\beta_1$  (коэффициент при  $X_t$ ) равен нулю.

Используя этот алгоритм, ответьте на следующие вопросы:

- a) Пройдите шаги с (i) до (iii) один раз. С помощью  $t$ -статистики из (iii) проверьте нулевую гипотезу о том, что  $\beta_1 = 0$  с помощью обычного 5 %-го критического значения, равного 1,96. Чему равен  $R^2$  вашей регрессии?
  - б) Повторите пункт (a) 1000 раз, сохраняя каждое полученное значение  $R^2$  и  $t$ -статистики. Постройте гистограммы  $R^2$  и  $t$ -статистики. Чему равны 5, 50 и 95-е процентили распределений  $R^2$  и  $t$ -статистики? Для какой части из 1000 смоделированных вами рядов  $t$ -статистика превышает 1,96 по абсолютной величине?
  - в) Повторите пункт (б) для различного числа наблюдений выборки, например, для  $T = 50$  и  $T = 200$ . Можно ли сказать, что при увеличении размера выборки доля случаев отверждения нулевой гипотезы приближается к 5 %, как и положено, потому что вы сгенерировали  $Y$  и  $X$  независимо распределенными? Приближается ли эта доля к другой границе при увеличении  $T$ ? Чему она равна?
- 14.7. Предположим, что  $Y_t$  описывается стационарной AR(1)-моделью  $Y_t = 2,5 + 0,7Y_{t-1} + u_t$ , в которой  $u_t$  – i.i.d. с  $E(u_t) = 0$  и  $var(u_t) = 9$ .
- а) Вычислите математическое ожидание и дисперсию  $Y_t$ . (Подсказка: см. упражнение 14.1.)
  - б) Вычислите первые две автоковариации  $Y_t$ . (Подсказка: см. приложение 14.2)
  - в) Вычислите первые два значения коэффициента автокорреляции  $Y_t$ .
  - г) Предположим, что  $Y_T = 102,3$ . Вычислите  $Y_{T+1|T} = E(Y_{T+1} | Y_T, Y_{T-1}, \dots)$ .
- 14.8. Предположим, что  $Y_t$  представляет ежемесячные значения числа новых жилых проектов в Соединенных Штатах, находящихся на начальной стадии строительства. Из-за особенностей погоды  $Y_t$  имеет ярко выраженный сезонный характер, например, строительство нового жилья редко начинается в январе и часто в июне. Обозначим через  $\mu_{Jan}$  среднее значение числа начинающихся строек нового частного жилья в январе, а  $\mu_{Feb}, \mu_{Mar}, \dots, \mu_{Dec}$  – средние значения в другие месяцы. Покажите, что значения  $\mu_{Jan}, \mu_{Feb}, \dots, \mu_{Dec}$  могут быть оценены при помощи МНК-регрессии  $Y_t = \beta_0 + \beta_1 Feb_t + \beta_2 Mar_t + \dots + \beta_{11} Dec_t + u_t$ , где  $Feb_t$  – бинарная переменная, равная единице, если  $t$  является февралем;  $Mar_t$  – бинар-

ная переменная, равная единице, если  $t$  является мартом, и так далее. Покажите, что  $\beta_0 + \beta_2 = \mu_{Mar}$  и так далее.

14.9. Модель скользящего среднего порядка  $q$  имеет вид:

$$Y_t = \beta_0 + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \dots + b_q e_{t-q},$$

где  $e_t$  является серией некоррелированной случайной величиной со средним, равным нулю, и дисперсией  $\sigma_e^2$ .

- a) Покажите, что  $E(Y_t) = \beta_0$ .
- б) Покажите, что дисперсия  $Y_t$  равна  $var(Y_t) = \sigma_e^2(1 + b_1^2 + b_2^2 + \dots + b_q^2)$ .
- в) Покажите, что  $\rho_j = 0$  для  $j > q$ .
- г) Предположим, что  $q = 1$ . Выведите автоковариационную функцию для  $Y_t$ .

14.10. Исследователь проводит QLR-тест с 25 %-м отсечением выборки при наличии  $q = 5$  ограничений. Используя значения из таблицы 14.6 («Критические значения QLR-статистики с 15 %-м отсечением выборки») и из таблицы приложения 4 («Критические значения распределения  $F_{m, \infty}$ »), ответьте на следующие вопросы:

- а) Пусть  $F$ -статистика QLR-теста равна 4,2. Может ли исследователь отвергнуть нулевую гипотезу на уровне значимости 5 %?
- б) Пусть  $F$ -статистика QLR-теста равна 2,1. Может ли исследователь отвергнуть нулевую гипотезу на уровне значимости 5 %?
- в) Пусть  $F$ -статистика QLR-теста равна 3,5. Может ли исследователь отвергнуть нулевую гипотезу на уровне значимости 5 %?

14.11. Предположим, что  $\Delta Y_t$  описывается AR(1)-моделью  $\Delta Y_t = \beta_0 + \beta_1 \Delta Y_{t-1} + u_t$ .

- а) Покажите, что  $Y_t$  описывается AR(2)-моделью.
- б) Выведите выражения для коэффициентов модели AR(2) как функций от  $\beta_0$  и  $\beta_1$ .

## Компьютерные упражнения

На официальном сайте учебника [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson), вы найдете базу данных USMacro\_Quarterly, содержащую квартальные данные по ряду макроэкономических временных рядов США; данные описаны в файле USMacro\_Description. Вычислите  $\ln(GDP_t)$  – логарифм реального ВВП и  $\Delta Y_t$  – квартальный темп роста ВВП. В заданиях Е14.1–Е14.6 используйте выборку с I квартала 1955 по IV квартал 2009 года, где данные до 1955 года могут быть использованы при необходимости, в качестве начальных значений для запаздывающих значений регрессоров.

- Е14.1. а) Оцените среднее значение  $\Delta Y_t$ .
- б) Выразите средний темп роста в процентных пунктах в годовом исчислении. [Подсказка: умножьте выборочное среднее из (а) на 400.]
- в) Оцените стандартное отклонение  $\Delta Y_t$ . Выразите ответ в процентах в годовом исчислении.
- г) Оцените первые четыре значения автокорреляционной функции  $\Delta Y_t$ . Каковы единицы измерения автокорреляций (ежеквартальные темпы

роста, процентные пункты в годовом исчислении или это безразмерная величина)?

- E14.2. a) Оцените AR(1)-модель для  $\Delta Y_t$ . Чему равен коэффициент при первом запаздывании? Отличается ли этот коэффициент статистически значимо от нуля? Постройте 95 %-й доверительный интервал для теоретического значения коэффициента AR(1)-модели.
- б) Оцените AR(2)-модель для  $\Delta Y_t$ . Является ли коэффициент при втором запаздывании статистически значимым? Является ли эта модель предпочтительнее AR(1)-модели?
- в) Оцените модели AR(3) и AR(4).
- (i) Используя оценки моделей AR(1) – AR(4), выберите количество лагов в модели AR при помощи BIC.
- (ii) Сколько запаздываний будет выбрано при помощи AIC?
- E14.3. Используйте статистики расширенного теста Дики–Фуллера для проверки гипотезы о наличии авторегрессионного единичного корня в AR-модели для  $Y_t$ . В качестве альтернативы предположите, что  $Y_t$  стационарен около детерминированного тренда.
- E14.4. Проверьте гипотезу о наличии структурного сдвига в параметрах AR(1)-модели для  $\Delta Y_t$ , используя QLR-тест.
- E14.5. а) Пусть  $R_t$  обозначает процентную ставку по трехмесячным казначейским векселям. Оцените модель ADL(1, 4) для  $\Delta Y_t$ , используя лаги  $\Delta R_t$ , в качестве дополнительных regressоров. Насколько сильно изменился  $\bar{R}^2$  в модели ADL(1,4) по сравнению с моделью AR(1)?
- б) Является ли значимой  $F$ -статистика теста Грейндженера на причинность?
- в) Проверьте наличие структурного сдвига в коэффициентах для константы и коэффициента при запаздывающих значениях  $\Delta R_t$ , при помощи QLR-теста. Существуют ли доказательства наличия структурного сдвига?
- E14.6. а) Постройте псевдовневыборочные прогнозы по AR(1)-модели, начиная с IV квартала 1989 года и до конца выборки. (То есть вычислите  $\hat{\Delta Y}_{1990:1|1989:4}$ ,  $\hat{\Delta Y}_{1990:2|1990:1}$  и так далее.)
- б) Постройте псевдовневыборочные прогнозы, используя модель ADL(1,4).
- в) Постройте псевдовневыборочные прогнозы, используя следующую «наивную» модель:
- $$\Delta Y_{t+1|t} = (\Delta Y_t + \Delta Y_{t-1} + \Delta Y_{t-2} + \Delta Y_{t-3})/4.$$
- г) Вычислите ошибки псевдовневыборочных прогнозов для каждой модели. Есть ли среди прогнозов смещенные? Какая модель имеет наименьший квадратный корень из среднеквадратической ошибки прогноза (RMSFE)? Насколько велика RMSFE (выраженная в процентных пунктах в годовом исчислении) в лучшей модели?
- E14.7. Прочитайте вставки «Можем ли мы перехитрить рынок? Часть I» и «Можем ли мы перехитрить рынок? Часть II» в этой главе. Затем перейдите на веб-сайт, где вы найдете расширенную версию базы

данных, которая описана во вставках; данные находятся в файле Stock\_Returns\_1931\_2002 и описаны в файле Stock\_Returns\_1931\_2002\_Description.

- а) Повторите расчеты, представленные в таблице 14.3, используя регрессию, оцененную на интервале с января 1932 по декабрь 2002 года.
- б) Повторите расчеты, представленные в таблице 14.7, используя регрессию, оцененную на интервале с января 1932 по декабрь 2002 года.
- в) Является ли переменная  $\ln(\text{dividend yield})$  высоко устойчивой? Объясните.
- г) Постройте псевдовневыборочный прогноз избыточной доходности на период с января 1983 по декабрь 2002 года, используя регрессию, оцененную на интервале с января 1932 года.
- д) Предполагают ли результаты, оцененные в пунктах (а)–(г), какие-либо серьезные изменения в выводах, сделанных во вставках? Объясните.

## Приложения

### Приложение 14.1. Данные, используемые в главе 14

В Соединенных Штатах Америки макроэкономические временные ряды собираются и публикуются различными государственными службами. Индекс потребительских цен США строится на основе ежемесячных обследований и собирается в Бюро статистики труда (BLS<sup>1</sup>). Уровень безработицы вычисляется на основе Текущего обследования населения, также проводимого BLS (см. приложение 3.1). Квартальные данные, используемые здесь, вычислялись путем усреднения месячных значений. Межбанковская ставка процента представляет собой взвешенные данные по ежедневным ставкам процента, публикуемые Федеральной резервной системой США, месячные данные по обменному курсу британского фунта к доллару США также представляют собой средние значения ежедневных данных; и оба ряда рассматриваются для последнего месяца квартала. Данные по ВВП Японии взяты из базы данных ОЭСР. Ежедневные процентные изменения сводного фондового индекса Нью-Йоркской фондовой биржи (NYSE<sup>2</sup>) были рассчитаны по формуле  $100\Delta\ln(\text{NYSE}_t)$ , где  $\text{NYSE}_t$  представляет собой значение индекса на момент закрытия Нью-Йоркской фондовой биржи; и поскольку биржа не работает по выходным и праздничным дням, рассматриваются данные только за рабочие дни. Все эти данные, как и тысячи других экономических временных рядов, находятся в свободном доступе на веб-сайтах различных агентств, занимающихся сбором данных.

---

<sup>1</sup> The Bureau of Labor Statistics (BLS). – Примеч. науч. ред. перевода.

<sup>2</sup> The New York Stock Exchange (NYSE). – Примеч. науч. ред. перевода.

Для оценки регрессий из таблиц 14.3 и 14.7 использовались ежемесячные финансовые данные по Соединенным Штатам Америки. Цены акций ( $P_t$ ) изменяются как взвешенный по стоимости широкий индекс цен акций, рассчитываемый Центром по исследованию рынка ценных бумаг (CRSP<sup>1</sup>). Ежемесячная избыточная доходность рассчитывается как  $100 \times \{ \ln[(P_t + Div_t)/P_{t-1}] - \ln(TBill_t) \}$ , где  $Div_t$  – дивиденды, выплачиваемые по акциям, включенным в индекс CRSP, и  $TBill_t$  – валовая доходность (1 плюс процентная ставка) по 30-дневным казначейским векселям в течение месяца  $t$ . Отношение дивидендов к цене строится как сумма дивидендов за последние 12 месяцев, деленная на цену в текущем месяце. Мы благодарим Motohiro Yogo (Motohiro Yogo) за помощь и предоставление этих данных.

### Приложение 14.2. Стационарность в моделях AR(1)

В данном приложении мы показываем, что если  $|\beta_1| < 1$  и ошибка  $u_t$  стационарна, то  $Y_t$  стационарен. Вспомним из вставки «Основные понятия 14.5», что временной ряд  $Y_t$  является стационарным, если совместное распределение  $(Y_{s+1}, \dots, Y_{s+T})$  не зависит от  $s$  для любого  $T$ . Чтобы упростить вывод, мы покажем это формально для  $T = 2$  и при наличии упрощающих выкладки предположений о том, что  $\beta_0 = 0$  и  $\{u_t\}$  i.i.d.  $N(0, \sigma_u^2)$ .

На первом шаге перепишем выражение для  $Y_t$  в терминах  $u_t$ . Поскольку  $\beta_0 = 0$ , из уравнения (14.8) следует, что  $Y_t = \beta_1 Y_{t-1} + u_t$ . Подставляя  $Y_{t-1} = \beta_1 Y_{t-2} + u_{t-1}$  в это выражение, получаем:  $Y_{t-1} = \beta_1(\beta_1 Y_{t-2} + u_{t-1}) + u_t = \beta_1^2 Y_{t-2} + \beta_1 u_{t-1} + u_t$ . Продолжая аналогично, на следующем шаге получаем:  $Y_t = \beta_1^3 Y_{t-3} + \beta_1^2 u_{t-2} + \beta_1 u_{t-1} + u_t$ , и продолжая подстановку до бесконечности, находим:

$$Y_t = u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \beta_1^3 u_{t-3} + \dots = \sum_{i=0}^{\infty} \beta_1^i u_{t-i}. \quad (14.37)$$

Таким образом,  $Y_t$  представляет собой средневзвешенное текущего и прошлых значений ошибки  $u_t$ . Поскольку  $u_t$  распределены нормально, а средневзвешенное от нормальных случайных величин также является нормально распределенной случайной величиной (раздел 2.4),  $Y_{s+1}$  и  $Y_{s+2}$  имеют двухмерное нормальное распределение. Вспомним из раздела 2.4, что двухмерное нормальное распределение полностью определяется средними значениями двух переменных, их дисперсиями и их ковариацией. Таким образом, чтобы показать, что  $Y_t$  является стационарным, мы должны показать, что средние, дисперсии и ковариация  $(Y_{s+1}, Y_{s+2})$  не зависят от  $s$ . Приведенное ниже доказательство может быть расширено, чтобы показать, что распределение  $(Y_{s+1}, Y_{s+2}, \dots, Y_{s+T})$  не зависит от  $s$ .

Средние и дисперсии случайных величин  $Y_{s+1}$  и  $Y_{s+2}$  могут быть вычислены, используя уравнения (14.37) и заменяя индекс  $t$  на  $s+1$  или  $s+2$ . Во-первых, поскольку  $E(u_t) = 0$  для всех  $t$ ,  $E(Y_t) = E\left(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}\right) = \sum_{i=0}^{\infty} \beta_1^i E(u_{t-i}) = 0$ , так что

<sup>1</sup> The Center for Research in Security Prices (CRSP). – Примеч. науч. ред. перевода.

средние значения и  $Y_{s+1}$  и  $Y_{s+2}$  равны нулю и, в частности, не зависят от  $s$ . Во-вторых,  $\text{var}(Y_t) = \text{var}\left(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}\right) = \sum_{i=0}^{\infty} (\beta_1^i)^2 \text{var}(u_{t-i}) = \sigma_u^2 \sum_{i=0}^{\infty} (\beta_1^i)^2 = \sigma_u^2 / (1 - \beta_1^2)$ , где последнее равенство вытекает из того, что если  $|a| < 1$ , то  $\sum_{i=0}^{\infty} a^i = 1/(1-a)$ ; и таким образом,  $\text{var}(Y_{s+1}) = \text{var}(Y_{s+2}) = \sigma_u^2 / (1 - \beta_1^2)$  и не зависит от  $s$ , если  $|\beta_1| < 1$ . И наконец, поскольку  $Y_{s+2} = \beta_1 Y_{s+1} + u_{s+2}$ ,  $\text{cov}(Y_{s+1}, Y_{s+2}) = E[Y_{s+1} Y_{s+2}] = E[Y_{s+1} (\beta_1 Y_{s+1} + u_{s+2})] = \beta_1 \text{var}(Y_{s+1}) + \text{cov}(Y_{s+1}, u_{s+2}) = \beta_1 \text{var}(Y_{s+1}) = \beta_1 \sigma_u^2 / (1 - \beta_1^2)$ . Ковариация не зависит от  $s$ , поэтому совместное распределение  $Y_{s+1}$  и  $Y_{s+2}$  также не зависит от  $s$ , то есть их совместное распределение является стационарным. Если  $|\beta_1| \geq 1$ , то такой вывод сделать нельзя, потому что бесконечная сумма в уравнении (14.37) не сходится и дисперсия  $Y_t$  бесконечна. Таким образом,  $Y_t$  является стационарным, если  $|\beta_1| < 1$ , но не является таковым, если  $|\beta_1| \geq 1$ .

Предыдущее доказательство было получено в предположении, что  $\beta_0 = 0$  и  $u_t$  нормально распределена. Если  $\beta_0 \neq 0$ , то весь вывод сохранится, за исключением того, что средние значения  $Y_{s+1}$  и  $Y_{s+2}$  равны  $\beta_0 / (1 - \beta_1)$  и уравнение (14.37) должно быть изменено для случая этого ненулевого среднего. Предположение о том, что это  $u_t$  является нормальной одинаково независимо распределенной случайной величиной, может быть заменено предположением о стационарности  $u_t$  с конечной дисперсией, так как по уравнению (14.37),  $Y_t$  все еще может быть выражена в виде функции от текущего и прошлых значений  $u_t$ , следовательно, распределение  $Y_t$  стационарно до тех пор, пока стационарно распределение  $u_t$  и бесконечная сумма в уравнении (14.37) существует в смысле сходимости, для чего необходимо, чтобы  $|\beta_1| < 1$ .

### Приложение 14.3. Оператор запаздывания

Обозначения, используемые в этой и следующих двух главах, значительно упрощаются, если ввести в рассмотрение так называемый оператор запаздывания. Пусть  $L$  обозначает *оператор запаздывания*, который ставит в соответствие переменной ее лаг. То есть оператор запаздывания  $L$  обладает свойством, которое математически выражается следующим образом:  $LY_t = Y_{t-1}$ . Применяя оператор запаздывания дважды, получаем второй лаг:  $L^2 Y_t = L(LY_t) = LY_{t-1} = Y_{t-2}$ . И наконец, применяя оператор запаздывания  $j$  раз, получаем  $j$ -й лаг. Таким образом, оператор запаздывания обладает свойством, выражющимся в следующем:

$$LY_t = Y_{t-1}, L^2 Y_t = Y_{t-2}, L^j Y_t = Y_{t-j}. \quad (14.38)$$

Обозначение, использующее оператор запаздывания, позволяет определить *лаговый полином* (многочлен), являющийся полиномом оператора запаздывания:

$$a(L) = a_0 + a_1 L + a_2 L^2 + \dots + a_p L_p = \sum_{j=0}^p a_j L^j, \quad (14.39)$$

где  $a_0, \dots, a_p$  – коэффициенты лагового многочлена и  $L^0 = 1$ . Степень лагового полинома  $a(L)$  в уравнении (14.39) равна  $p$ . Умножая  $Y_t$  на  $a(L)$ , получаем:

$$a(L)Y_t = \left( \sum_{j=0}^p a_j L^j \right) Y_t = \sum_{j=0}^p a_j (L^j Y_t) = \\ = \sum a_j Y_{t-j} = a_0 Y_t + a_1 Y_{t-1} + \dots + a_p Y_{t-p}. \quad (14.40)$$

Из выражения (14.40) следует, что модель AR( $p$ ), задаваемая уравнением (14.14), может быть записана более компактно:

$$a(L)Y_t = \beta_0 + u_t, \quad (14.41)$$

где  $a_0 = 1$  и  $a_j = -\beta_j$  для  $j = 1, \dots, p$ . Кроме того, модель ADL( $p, q$ ) может быть записана в таком виде:

$$a(L)Y_t = \beta_0 + c(L)X_{t-1} + u_t, \quad (14.42)$$

где  $a(L)$  является лаговым многочленом степени  $p$  (с  $a_0 = 1$ ) и  $c(L)$  является лаговым многочленом степени  $q-1$ .

#### Приложение 14.4. ARMA-модели

**Модель авторегрессии с ошибками в форме скользящего среднего (ARMA<sup>1</sup>)** является расширением модели авторегрессии, в котором ошибка  $u_t$ , предполагается серийно коррелированной или, более точно, моделируется как распределенный лаг («скользящее среднее») другой ненаблюдаемой ошибки. В обозначениях оператора запаздывания из приложения 14.3 пусть  $u_t = b(L)e_t$ , где  $b(L)$  является лаговым многочленом степени  $q$  с  $b_0 = 1$ , а  $e_t$  — серийно некоррелированная ненаблюдаемая случайная величина. Тогда ARMA( $p, q$ ) модель имеет вид:

$$a(L)Y_t = \beta_0 + b(L)e_t, \quad (14.43)$$

где  $a(L)$  — лаговый многочлен степени  $p$  с  $a_0 = 1$ .

Обе модели, AR и ARMA, можно рассматривать как способ аппроксимировать автоковариации  $Y_t$ . Причина этого заключается в том, что любой стационарный временной ряд  $Y_t$  с конечной дисперсией можно записать либо в виде AR-модели, либо в виде MA-модели с серийно некоррелированным остаточным членом, хотя эти AR- или MA-модели могут иметь бесконечный порядок. Второй из этих результатов, то есть вывод о том, что стационарный процесс можно записать в форме скользящего среднего, известен как теорема Вольда (Wold) о декомпозиции и является одним из основных результатов, лежащих в основе методов анализа стационарных временных рядов.

С теоретической точки зрения семейства моделей AR, MA и ARMA обладают похожими характеристиками при достаточно высоких степенях лаговых полиномов. Тем не менее в некоторых случаях может быть лучше аппроксимировать автоковариации, используя ARMA( $p, q$ )-модель с небольшими  $p$  и  $q$ , чем чистую модель AR с несколькими лагами. С практической точки зрения, однако, оценить ARMA-

<sup>1</sup> The autoregressive-moving average (ARMA) model. В российских учебниках часто можно встретить название «модель авторегрессии скользящего среднего». — Примеч. науч. ред. перевода.

модели гораздо сложнее, чем AR-модели, и ARMA-модели тяжелее расширить с точки зрения включения в них дополнительных регрессоров, чем AR-модели.

### **Приложение 14.5. Состоительность критерия BIC при оценке глубины запаздывания**

В настоящем приложении приведено доказательство того, что информационный критерий Шварца (BIC) оценивает длину лага  $p$  в авторегрессии корректно в больших выборках, то есть  $\Pr(\hat{p} = p) \rightarrow 1$ . Это утверждение неверно для критерия AIC, который может переоценить  $p$  даже в больших выборках.

#### **BIC**

Рассмотрим, во-первых, специальный случай, когда при помощи BIC мы выбираем модель среди авторегрессий с числом запаздываний, равным 0, 1 и 2, зная, что истинный лаг равен единице. Ниже будет показано, что (i)  $\Pr(\hat{p} = 0) \rightarrow 0$  и (ii)  $\Pr(\hat{p} = 2) \rightarrow 0$ , из чего следует, что  $\Pr(\hat{p} = 1) \rightarrow 1$ . Расширение этого доказательства на общий случай  $0 \leq p \leq p_{max}$  — перебор по всем таким  $p$  и доказательство того, что  $\Pr(\hat{p} < p) \rightarrow 0$  и  $\Pr(\hat{p} > p) \rightarrow 0$ ; стратегия доказательства этого аналогична доказательству пунктов (i) и (ii) ниже.

#### **Доказательство (i) и (ii)**

**Доказательство (i).** Для случая  $\hat{p} = 0$  должно выполняться, что  $\text{BIC}(0) < \text{BIC}(1)$ ; то есть  $\text{BIC}(0) - \text{BIC}(1) < 0$ . Имеем  $\text{BIC}(0) - \text{BIC}(1) = [\ln(\text{SSR}(0)/T) + (\ln T)/T] - [\ln(\text{SSR}(1)/T) + 2(\ln T)/T] = \ln(\text{SSR}(0)/T) - \ln(\text{SSR}(1)/T) - (\ln T)/T$ . Далее,  $\text{SSR}(0)/T = [(T-1)/T]s_y^2 \xrightarrow{p} \sigma_y^2$ ,  $\text{SSR}(1)/T \xrightarrow{p} \sigma_u^2$  и  $(\ln T)/T \rightarrow 0$ ; подставляя все это в выражение выше, получаем  $\text{BIC}(0) - \text{BIC}(1) \xrightarrow{p} \ln \sigma_y^2 - \ln \sigma_u^2 > 0$ , так как  $\sigma_y^2 > \sigma_u^2$ . Из этого следует, что  $\Pr[\text{BIC}(0) < \text{BIC}(1)] \rightarrow 0$ , так что  $\Pr(\hat{p} = 0) \rightarrow 0$ .

**Доказательство (ii).** Для случая  $\hat{p} = 2$  должно выполняться, что  $\text{BIC}(2) < \text{BIC}(1)$  или  $\text{BIC}(2) - \text{BIC}(1) < 0$ . Имеем  $T[\text{BIC}(2) - \text{BIC}(1) < 0] = T\{\ln(\text{SSR}(2)/T) + 3(\ln T)/T\} - [\ln(\text{SSR}(1)/T) + 2(\ln T)/T\} = T\ln[\text{SSR}(2)/\text{SSR}(1)] + (\ln T) = -T\ln[1 + F/(T-2)] + \ln T$ , где  $F = [\text{SSR}(1) - \text{SSR}(2)]/[\text{SSR}(2)/(T-2)]$  является  $F$ -статистикой, рассчитанной в предположении гомоскедастичности (уравнение 7.13) и тестирующей гипотезу  $\beta_2 = 0$  в AR(2). Если  $u$ , гомоскедастична, то  $F$  распределена асимптотически по  $\chi_1^2$ ; если нет, то согласно какому-то другому распределению. Таким образом,  $\Pr[\text{BIC}(2) - \text{BIC}(1) < 0] = \Pr\{T[\text{BIC}(2) - \text{BIC}(1) < 0]\} = \Pr\{-T\ln[1 + F/(T-2)] + (\ln T) < 0\} = \Pr\{T\ln[1 + F/(T-2)] > \ln T\}$ . Так как  $T$  возрастает,  $T\ln[1 + F/(T-2)] - F \xrightarrow{p} 0$  [это является следствием логарифмического

приближения  $\ln(1+a) \approx a$ , которое становится точным равенством в пределе при  $a \rightarrow 0$ . Таким образом,  $\Pr[BIC(2) - BIC(1) < 0] \rightarrow \Pr(F > \ln T) \rightarrow 0$ , так что  $\Pr(\hat{p} > 2) \rightarrow 0$ .

### AIC

Для случая AR(1)-модели при оценке моделей с нулем, одним или двумя запаздываниями в пункте (i) применительно к AIC нужно заменить  $\ln T$  на 2, поэтому  $\Pr(\hat{p} = 0) \rightarrow 0$ . Все шаги в доказательстве (ii) для BIC также повторяются для AIC, заменяя  $\ln T$  на 2; поэтому  $\Pr[AIC(2) - AIC(1) < 0] \rightarrow \Pr(F > 2) > 0$ . Если  $u_t$  гомоскедастична, то  $\Pr(F > 2) \rightarrow \Pr(\chi^2_1 > 2) = 0,16$ , следовательно,  $\Pr(\hat{p} = 2) \rightarrow 0,16$ . В общем случае, если  $\hat{p}$  выбирается, используя AIC, то  $\Pr(\hat{p} < p) \rightarrow 0$ , но  $\Pr(\hat{p} > p)$  стремится к положительному числу, поэтому  $\Pr(\hat{p} = p)$  не стремится к единице.

## Глава 15. Оценка динамического причинного влияния

Герои фильма «Поменяться местами»<sup>1</sup>, которых играют Дэн Эйкройд (Dan Aykroyd) и Эдди Мерфи (Eddie Murphy), используют инсайдерскую информацию о том, насколько хорошо обстоят дела зимой у флоридских фермеров, производящих апельсины, чтобы заработать миллионы на фьючерсном рынке концентратов апельсинового сока, на рынке контрактов на покупку или продажу больших объемов концентратов апельсинового сока по определенной цене в определенный день в будущем. В реальной жизни трейдеры, продающие фьючерсы на апельсиновый сок, должны обращать пристальное внимание на погоду во Флориде: заморозки губят урожай апельсинов, которые являются источником почти всех замороженных концентратов апельсинового сока, производимых в Соединенных Штатах, так что если их предложение снижается, цена поднимается. Но насколько именно растет цена при ухудшении погоды во Флориде? Происходит ли рост цен сразу или через какое-то время, и если да, то через какое? На все эти вопросы необходимо отвечать трейдерам, торгующим фьючерсами на апельсиновый сок, если они хотят добиться успеха.

В данной главе мы концентрируемся на проблеме оценки влияния изменений  $X$  на  $Y$ , как сейчас, так и в будущем, то есть на *динамическом причинном влиянии* изменений  $X$  на  $Y$ . Например, как проявляется эффект влияния заморозков во Флориде на динамику цен на апельсиновый сок во времени? Отправной точкой для моделирования и оценки динамического причинного влияния является так называемая модель регрессии с распределенными лагами, в которой  $Y$ , выражается в виде функции от текущих и прошлых значений  $X$ . В разделе 15.1 вводится понятие модели с распределенными лагами в контексте оценки влияния холодной погоды во Флориде на цену концентрата апельсинового сока во времени. В разделе 15.2 подробно обсуждается вопрос о том, что именно понимается под динамическим причинным влиянием.

Одним из способов оценки динамического причинного влияния является оценка коэффициентов модели регрессии с распределенными лагами с использованием МНК. Как отмечается в разделе 15.3, эта оценка является состоятельной, если ошибки регрессии имеют нулевое условное среднее

<sup>1</sup> Английское название фильма *Trading Places*. Фильм снят в 1983 году. – Примеч. науч. ред. перевода.

относительно текущего и прошлых значений  $X$  – условие, которое (как определено в главе 12) называют экзогенностью. Из-за пропуска факторов, определяющих  $Y$ , и являющихся коррелированными во времени, то есть серийно коррелированных, остаточный член в модели с распределенными лагами может быть серийно коррелирован. Это, в свою очередь, требует оценивать стандартные ошибки как «состоятельные при наличии гетероскедастичности и автокорреляции» (HAC<sup>1</sup>), которые рассматриваются в разделе 15.4.

Второй способ оценки динамического причинного влияния, обсуждаемый в разделе 15.5, связан с оценкой модели серийной корреляции, присущей в ошибках как авторегрессии, с тем чтобы затем использовать эту модель авторегрессии для вывода авторегрессионной модели с распределенными лагами (ADL). Кроме того, коэффициенты исходной модели с распределенными лагами могут быть оценены при помощи обобщенного метода наименьших квадратов (ОМНК, GLS<sup>2</sup>). Оба метода, и ADL, и GLS, требуют, однако, выполнения более сильной версии экзогенности, чем мы использовали до сих пор: строгой экзогенности, согласно которой ошибки регрессии должны иметь нулевое условное среднее относительно прошлых, настоящего и будущих значений  $X$ .

В разделе 15.6 рассматривается более полный анализ зависимости между ценами на апельсиновый сок и погодой. В этом примере погода не может контролироваться человеком и, следовательно, является экзогенной (хотя, как обсуждается в разделе 15.6, экономическая теория предполагает, что это не означает строгую экзогенность). Из-за того что экзогенность является необходимым условием для оценки динамического причинного влияния, это предположение анализируется в разделе 15.7 на примере нескольких приложений, взятых из макроэкономики и финансов.

Данная глава строится на материале из разделов 14.1–14.4 и, за исключением подраздела (который может быть пропущен), касающегося эмпирического анализа, в разделе 15.6, не требует изучения материала разделов 14.5–14.7.

## 15.1. Знакомство с данными по апельсиновому соку

Орландо, исторический центр выращивающего апельсины региона во Флориде, как правило, солнечный и теплый. Но время от времени там случаются кратковременные заморозки, и если отрицательная температура держится слишком долго, то апельсины опадают с деревьев. Если похолодание будет слишком длительным, деревья замерзнут. После заморозков предложение концентрата апельсинового сока падает и его цена возрастает.

---

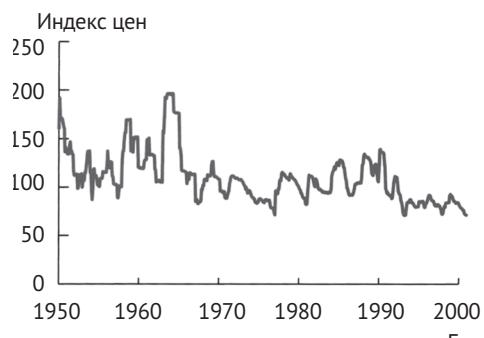
<sup>1</sup> Heteroskedasticity – and autocorrelation-consistent (HAC). – Примеч. науч. ред. перевода.

<sup>2</sup> General least square (GLS). – Прим. научн. ред. перевода.

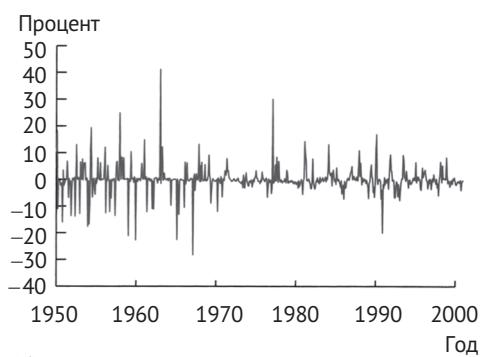
ет. Однако предсказать момент повышения цен довольно сложно. Концентрат апельсинового сока является «товаром длительного пользования» или непортящимся товаром, то есть он может храниться в замороженном состоянии, хотя и за счет некоторых издержек (которые необходимы для работы морозильной камеры). При этом цена концентрата апельсинового сока зависит не только от предложения, но и от ожиданий относительно будущих поставок. Заморозка концентрата сегодня означает, что будущие поставки концентрата будут низкими, а поскольку в настоящее время концентрат отправляется на склад вместо того, чтобы использоваться для удовлетворения текущего или будущего спроса, цена на концентрат растет уже сегодня. Но насколько точно вырастет цена концентрата, если он замораживается? Ответ на этот вопрос представляет интерес не только для трейдеров, продающих апельсиновый сок, но и для всех экономистов, изучающих механизмы, которые действуют на современных товарных рынках. Чтобы узнать, как цена апельсинового сока изменяется в ответ на погодные условия, мы должны проанализировать данные о ценах на апельсиновый сок и о погоде.

Ежемесячные данные о ценах замороженного концентрата апельсинового сока, их ежемесячные изменения в процентах и температура в выращивающем апельсины регионе Флориды с января 1950 до декабрь 2000 года приведены на рисунке 15.1. Цены на рисунке 15.1а представляют собой реальные средние оптовые цены на замороженный концентрат апельсинового сока. Цена дефлировалась с помощью сводного индекса цен производителей на произведенную продукцию, чтобы скорректировать последствия общего роста цен. Процентное изменение цены приведено на рисунке 15.1б и представляет собой процентное изменение цены за месяц. Данные по температуре воздуха приведены на рисунке 15.1в и отражают число «морозных дней» в аэропорту г. Орландо штата Флорида, рассчитанные как сумма градусов по Фаренгейту, в которые минимальная температура падала ниже нуля (по Цельсию) в данный день за все дни месяца – например, в ноябре 1950 года температура в аэропорту падала ниже нуля два раза: 25-го ( $31^{\circ}\text{F}$ ) и 29-го ( $29^{\circ}\text{F}$ ), поэтому суммарное число отрицательных (морозных) температур равно  $4$  [т.е.  $(32 - 31) + (32 - 29) = 4$ ]<sup>1</sup>. (Данные описаны более подробно в приложении 15.1). Как можно видеть, сравнивая графики на рисунке 15.1, цена концентрата апельсинового сока сильно менялась в рассматриваемый период времени, и некоторые из этих колебаний кажутся связанными с холодной погодой во Флориде.

<sup>1</sup> В английском тексте используется название «the number of freezing degree days during the month» или, как более короткий вариант, «freezing degree days». Применительно к рассматриваемому примеру авторы говорят о четырех морозных днях (хотя на самом деле это суммарное количество отрицательных температур (градусов)). Мы далее также будем говорить о «морозных днях» (а не о температурах), время от времени напоминая, что в действительности имеется в виду температура воздуха. – Прим. научн. ред. перевода.



(а) Индекс цен на замороженный концентрат апельсинового сока



(б) Процентное изменение цены на замороженный концентрат апельсинового сока



(в) Ежемесячное суммарное количество отрицательных температур в Орландо, Флорида

#### **Рисунок 15.1. Цены на апельсиновый сок и погода во Флориде, 1950–2000 годы**

На графиках видно, что имели место большие ежемесячные изменения в ценах замороженных концентратов апельсинового сока. Многие из этих изменений совпадают с заморозками в Орландо, где выращивается большая часть апельсинов в США.

Мы начинаем количественный анализ взаимосвязи между ценами на апельсиновый сок и погодой с оценки регрессии процентного изменения цены апельсинового сока от заморозков. Зависимой переменной является процентное изменение цены за текущий месяц  $\widehat{\%ChgP}_t$ , где  $\widehat{\%ChgP}_t = 100 \times \Delta \ln(P_t^{OJ})$  и  $P_t^{OJ}$  – реальная цена апельсинового сока]. Регрессором является суммарное число отрицательных температур в течение этого месяца ( $FDD_t$ ). Регрессия оценивается на выборке с января 1950 по декабрь 2000 года (как и все регрессии в этой главе), в общей сложности  $T=612$  наблюдения:

$$\widehat{\%ChgP}_t = -0,40 + 0,47 FDD_t. \quad (15.1)$$

Стандартные ошибки коэффициентов регрессий, приводимых в этом разделе, не являются обычными стандартными МНК-ошибками, а являются стандартными ошибками состоятельных при наличии гетероскедастичности и автокорреляции (НАС-стандартные ошибки), которые более корректно использовать при наличии автокорреляции между остаточным членом и регрессорами. НАС-стандартные ошибки обсуждаются в разделе 15.4, и на данный момент они будут использоваться без подробного объяснения.

Согласно оцененной регрессии, дополнительный морозный день в месяце повышает цену концентрата апельсинового сока в этом месяце на 0,47%. Если в месяце было четыре морозных дня, например как в ноябре 1950 года, цена концентрата апельсинового сока согласно оценкам, увеличилась на 1,88% по сравнению с месяцем без морозных дней.

Поскольку регрессия (15.1) включает в себя только характеристику погоды в тот же месяц, она не отражает каких-либо долгосрочных последствий влияния похолодания на цену апельсинового сока в последующие месяцы. Для того чтобы учесть это влияние, мы должны рассмотреть, как влияют на цену и одновременное, и запаздывающие значения  $FDD$ , что, в свою очередь, может быть сделано при помощи расширения регрессии (15.1), например, включением в нее запаздывающих значений  $FDD$  за последние 6 месяцев:

$$\begin{aligned} \widehat{\%ChgP}_t = & -0,65 + 0,47 FDD_t + 0,14 FDD_{t-1} + \\ & + 0,06 FDD_{t-2} + 0,07 FDD_{t-3} + 0,03 FDD_{t-4} + \\ & + 0,05 FDD_{t-5} + 0,05 FDD_{t-6}. \end{aligned} \quad (15.2)$$

Уравнение (15.2) представляет собой модель регрессии с распределенными лагами. Коэффициент при  $FDD_t$  в уравнении (15.2) оценивает процентный рост цен в течение месяца, в котором были заморозки; дополнительный день заморозков, по оценкам, повышает цену в этом месяце на 0,47%. Коэффициент при первом лаге  $FDD_t$ , то есть при  $FDD_{t-1}$ , оценивает процентный рост цен в результате заморозков в предыдущем месяце, коэффициент при втором лаге – оценку влияния заморозков два месяца назад, и так далее. Или равносильно,

коэффициент при первом лаге  $FDD$  является оценкой влияния единичного увеличения  $FDD$  через один месяц после заморозков. Таким образом, оценки коэффициентов в уравнении (15.2) являются оценками эффекта влияния единичного роста  $FDD_t$  на текущие и будущие значения  $\%ChgP_t$ , то есть они являются оценками динамического воздействия  $FDD_t$  на  $\%ChgP_t$ . Например, четыре морозных дня в ноябре 1950 года, по оценкам, увеличили цену апельсинового сока на 1,88 % в ноябре 1950 года, на дополнительные 0,56 % ( $=4 \times 0,14$ ) – в декабре 1950 года, еще на 0,24 ( $=4 \times 0,06$ ) – в январе 1951 года, и так далее.

## 15.2. Динамическое причинное влияние

Прежде чем узнать больше о методах оценки динамического причинного влияния, мы должны понять, что именно понимается под динамическим причинным влиянием. Имея четкое представление о том, что такое динамическое причинное влияние, мы сможем более четко понимать условия, при которых оно может быть оценено.

### ***Причинное влияние и временные ряды***

В разделе 1.2 определена причинно-следственная связь как результат идеального случайного контролируемого эксперимента: случайным образом садовод использует удобрения лишь для части ростков помидоров, а затем измеряет урожайность помидоров и ожидаемую разность в урожайности между удобрившимися и неудобрившимися кустами помидоров, которая характеризует влияние удобрений на урожайность помидоров. Однако такая концепция эксперимента подходит для случая, когда в распоряжении исследователя имеется несколько объектов (несколько кустов помидоров или несколько человек), так что данные являются либо межобъектными (урожай помидоров, собранный в конце эксперимента) или панельными (индивидуальные доходы до и после экспериментальной программы по обучению персонала). При наличии нескольких объектов можно разбить выборку на исследуемую и контрольную группы и тем самым оценить причинное влияние эксперимента.

Во временных рядах такое определение причинно-следственного влияния с точки зрения идеального случайного управляемого эксперимента должно быть модифицировано. Чтобы быть более конкретными, рассмотрим важную макроэкономическую проблему оценки влияния непредвиденного изменения краткосрочной процентной ставки на текущее и будущее экономическое положение данной страны, измеряемое ВВП. Буквальное применение случайного контролируемого эксперимента из раздела 1.2 требует разделить страны случайным образом на исследуемую и контрольную группы. Центральные банки стран из исследуемой группы должны случайным образом менять процентные ставки, в то время как в контрольной группе процентные ставки должны оставаться неизменными; и показатель экономического положения (например, ВВП) должен измеряться в течение ближайших нескольких лет после этого для

обеих групп. Но что будет, если мы заинтересованы в оценке такого эффекта для конкретной страны, скажем, для Соединенных Штатов? Для проведения такого эксперимента необходимо «клонировать» Соединенные Штаты, чтобы разделить эти «клоны» на исследуемую и экспериментальную группу. Очевидно, что такой эксперимент с «параллельными вселенными» является невозможным.

Вместо этого для временных рядов полезно думать о случайному контролируемом эксперименте как состоящем из одних и тех же объектов (например, экономика США), получающих различное воздействие (случайно выбранные изменения процентных ставок) в различные моменты времени (в 1970-е годы, 1980-е годы и т.д.). В такой постановке один объект в различные моменты времени играет роль и исследуемой и контрольной групп: ФРС иногда изменяет процентные ставки, но в другие моменты времени этого не происходит. Поскольку данные собираются во времени, можно оценить динамическое причинное влияние, то есть понять, как изменение ставки процента влияет на выпуск и как это влияние меняется во времени. Например, неожиданное увеличение краткосрочной процентной ставки на два процентных пункта, произошедшее в конкретном квартале, поначалу может оказывать незначительное влияние на объем производства; через два квартала темп роста ВВП может замедлиться с наибольшим замедлением в течение полутора лет; а затем в ближайшие два года темп роста ВВП может вернуться к обычному. Такое количественное изменение во времени причинно-следственных связей является динамическим причинным влиянием неожиданного изменения процентной ставки на темп роста ВВП.

В качестве второго примера рассмотрим причинное влияние на цену апельсинового сока изменения количества заморозков. Можно представить различные гипотетические эксперименты, в каждом из которых исследуются различные причинно-следственные связи. В одном из экспериментов можно было бы изменить погоду на апельсиновых фермах Флориды, сохраняя постоянную погоду в других местах, например, на грейпфрутовых фермах Техаса и в других регионах, выращивающих цитрусовые. Такой эксперимент будет измерять частный эффект в условиях постоянства погоды в других регионах. Во втором эксперименте можно изменить погодные условия в тех регионах, где «исследование» заключается в применении общих моделей влияния изменения погодных условий. Если погодные условия в регионах, производящих различные культуры, коррелируют между собой, то динамические причинные влияния различны. В данной главе мы рассмотрим причинное влияние в последнем эксперименте, то есть причинное влияние при применении общих моделей. Все это подразумевает измерение динамического влияния изменений в погодных условиях во Флориде на цены без фиксации погодных условий в других сельскохозяйственных регионах.

**Динамическое влияние и модель с распределенными лагами.** Из-за того что динамическое влияние – это процесс, происходящий во времени, эконометрические модели, используемые для оценки динамического причинного влияния, должны включать в себя лаги. Для того чтобы сделать это, можно выразить  $Y_t$  как распределенный лаг текущего и  $r$  прошлых значений  $X_t$ :

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \dots + \beta_{r+1} X_{t-r} + u_t, \quad (15.3)$$

где  $u_t$  компонента ошибки, включающая ошибку измерения  $Y_t$  и эффект от пропущенных переменных, влияющих на  $Y_t$ . Модель (15.3) называется моделью регрессии с распределенными лагами, описывающей влияние  $X_t$  и  $r$  ее запаздываний на  $Y_t$ .

В качестве иллюстрации интерпретации уравнения (15.3) рассмотрим модифицированный эксперимент с помидорами и удобрениями: из-за того что удобрения, применяемые сегодня, могут остаться в земле в будущем, садовод хочет определить влияние удобрений на урожай помидоров в течение некоторого времени после применения удобрений. Соответственно, она также планирует трехлетний эксперимент и случайным образом делит свои участки с саженцами помидоров на четыре группы: первый участок удобряется только в первый год, второй – только во второй год, третий – только в третий год и четвертый – контрольная группа – никогда не удобряется. Помидоры выращивают на всех участках ежегодно, а урожай, собранный в третий год, взвешивают. Для первых трех групп вводятся бинарные переменные  $X_{t-2}, X_{t-1}$  и  $X_t$ , где  $t$  – третий год (год, когда урожай взвешивали),  $X_{t-2}=1$ , если участок находится в первой группе (удобренной двумя годами ранее),  $X_{t-1}=1$ , если участок был удобрен во второй год, и  $X_t=1$ , если участок удобрялся в последний год. В контексте уравнения (15.3) (которое применяется к одному участку) эффект от влияния удобрения, внесенного в последний год, равен  $\beta_1$ , эффект от влияния удобрения, внесенного годом ранее, равен  $\beta_2$ , и эффект от влияния удобрения, внесенного за два года до окончания эксперимента, равен  $\beta_3$ . Если влияние удобрений, внесенных в год окончания эксперимента, больше, чем для предыдущих лет, то  $\beta_1$  будет больше, чем  $\beta_2$  и  $\beta_3$ .

В более общем случае коэффициент при одновременном значении  $X_t$ ,  $\beta_1$ , характеризует одновременное или немедленное влияние единичного изменения  $X_t$  на  $Y_t$ . Коэффициент при  $X_{t-1}$ ,  $\beta_2$ , является характеристикой эффекта влияния единичного изменения  $X_{t-1}$  на  $Y_t$ , или, эквивалентно, влияние единичного изменения  $X_t$  на  $Y_t$ , то есть  $\beta_2$  является эффектом от единичного изменения  $X$  на  $Y$  на период позже. В целом коэффициент при  $X_{t-h}$  характеризует эффект влияния единичного изменения  $X$  на  $Y$  через  $h$  периодов. Динамическое причинное влияние характеризует эффект влияния изменения  $X_t$  на  $Y_t, Y_{t-1}, Y_{t-2}$  и так далее, то есть это последовательность причинных эффектов влияния на текущие и будущие значения  $Y$ . Таким образом, в рамках модели с распределенными лагами (15.3) динамическое причинное влияние представляет собой последовательность коэффициентов  $\beta_1, \beta_2, \dots, \beta_{r+1}$ .

**Приложение к эмпирическому анализу временных рядов.** Такое понимание динамического причинного влияния во временных рядах как ожидаемого результата эксперимента, в котором различные уровни воздействия неоднократно применяются к одному и тому же объекту, имеет два следствия для попытки измерить эмпирически динамическое причинное влияние во временных рядах. Первое следствие заключается в том, что динамическое причинное влияние не должно меняться в выборке данных, которая есть у нас. В свою очередь это

означает, что данные должны быть совместно стационарны (вставка «Основные понятия 14.5»). Как отмечалось в разделе 14.7, гипотеза о том, что теоретическая функция регрессии стабильна во времени, может быть проверена с помощью QLR-теста на структурный сдвиг, и возможно оценить динамическое причинное влияние в различных подвыборках. Вторым следствием является то, что переменная  $X$  должна быть не коррелирована с остаточным членом, и к обсуждению именно этого вывода мы и переходим.

## Два вида экзогенности

В разделе 12.1 мы определили «экзогенную» переменную как переменную, которая не коррелирует с ошибкой регрессии, и «эндогенную» переменную как переменную, которая коррелирует с ошибкой. Эта терминология распространяется на модели с несколькими уравнениями, в которых «эндогенные» переменные определяются как переменные, которые моделируются в системе, в то время как «экзогенные» переменные определяются вне ее. Грубо говоря, если мы хотим оценить динамическое причинное влияние с помощью модели с распределенными лагами из уравнения (15.3), регрессоры ( $X$ -ы) не должны быть коррелированы с ошибкой. Таким образом, переменная  $X$  должна быть экзогенной. Однако из-за того что мы работаем с временными рядами, нужно усовершенствовать определение экзогенности. На самом деле мы будем использовать два разных понятия экзогенности.

Первая концепция экзогенности заключается в том, что ошибка имеет нулевое условное среднее относительно текущего и всех прошлых значений  $X$ , то есть что  $E(u_t | X_t, X_{t-1}, X_{t-2}, \dots) = 0$ . Это предположение изменяет стандартное предположение об условном среднем для множественной регрессии для межобъектных данных (предположение № 1 из вставки «Основные понятия 6.4»), в котором требовалось, чтобы  $u_t$  имела нулевое условное среднее только относительно включенных регрессоров, то есть чтобы  $E(u_t | X_t, X_{t-1}, \dots, X_{t-r}) = 0$ . Включая все запаздывающие значения  $X$ , в условное математическое ожидание, мы подразумеваем, что все более отдаленные причинные эффекты – все причинные эффекты, находящиеся глубже, чем лаг  $r$ , – равны нулю. Таким образом, при выполнении этого предположения коэффициенты при всех  $r$  распределенных лагах в уравнении (15.3) отражают все ненулевые динамические причинные воздействия. Можно ссылаться на предположение о том, что  $E(u_t | X_t, X_{t-1}, X_{t-2}, \dots) = 0$  как на *прошлую и настоящую экзогенность*, но из-за сходства этого определения и определения экзогенности из главы 12 мы просто используем термин «*экзогенность*».

Вторая концепция экзогенности заключается в том, что ошибка имеет нулевое среднее относительно всех прошлых, настоящего и будущих значений  $X$ , то есть что  $E(u_t | \dots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \dots) = 0$ . Это называется *строгой экзогенностью*; для ясности, мы также называем ее *прошлой, настоящей и будущей экзогенностью*. Причина для введения понятия строгой экзогенности заключается в том, что, когда  $X$  является строго экзогенной, существуют более эффективные оценки динамического причинного влияния, чем МНК-оценки

коэффициентов модели регрессии с распределенными лагами в уравнении (15.3).

## ОСНОВНЫЕ ПОНЯТИЯ

### 15.1

#### Модель с распределенными лагами и экзогенность

В модели с распределенными лагами

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \dots + \beta_{r+1} X_{t-r} + u_t \quad (15.4)$$

существуют два типа экзогенности, то есть два различных условия экзогенности:

Экзогенность в прошлом и настоящем (экзогенность):

$$E(u_t | X_t, X_{t-1}, X_{t-2}, \dots) = 0; \quad (15.5)$$

Экзогенность в прошлом, настоящем и будущем (строгая экзогенность):

$$E(u_t | \dots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \dots) = 0. \quad (15.6)$$

Если случайная величина  $X$  строго экзогенна, то она экзогена, но из экзогенности не следует строгая экзогенность.

Различие между экзогенностью (прошлой и настоящей) и строгой экзогенностью (прошлой, настоящей и будущей) заключается в том, что строгая экзогенность включает будущие значения  $X$  в условное математическое ожидание. Таким образом, из строгой экзогенности следует экзогенность, но не наоборот. Один из способов понять разницу между этими двумя понятиями заключается в рассмотрении последствий этих определений для корреляции между  $X$  и  $u$ . Если  $X$  является экзогенной (в прошлом и настоящем), то  $u$ , не коррелирована с текущим и прошлыми значениями  $X$ . Если  $X$  строго экзогенная, то в дополнение к предыдущему  $u$ , не коррелирована с будущими значениями  $X$ . Например, если  $Y$ , изменяет будущие значения  $X$ , то  $X$ , не является строго экзогенной, хотя она может быть экзогенна (в прошлом и настоящем).

В качестве иллюстрации рассмотрим гипотетический многолетний эксперимент с помидорами и удобрениями, описываемый согласно уравнению (15.3). Так как удобрение добавляется случайным образом, факт добавления удобрения является экзогенным в таком гипотетическом эксперименте. Вследствие того что урожай помидоров сегодня не зависит от количества удобрений, которые будут добавляться в будущем, временной ряд, характеризующий добавление удобрений, также строго экзогенен.

В качестве второй иллюстрации рассмотрим пример с ценами на апельсиновый сок, в котором  $Y_t$  – ежемесячное процентное изменение цены апельсинового сока, а  $X_t$  – число морозных дней в этом месяце. С точки зрения рынков апельсинового сока, мы можем думать о погоде (о количестве морозных дней) как будто бы она случайна в том смысле, что погода находится за пределами человеческого контроля. Если влияние  $FDD$  является линейным и если оно не оказывает никакого влияния на цены через несколько месяцев  $r$  после заморозков, то из этого следует, что погода экзогенна. Но является ли погода строго экзогенной? Если условное среднее  $u_t$  относительно будущих значений

$FDD$  не равно нулю, то  $FDD$  не является строго экзогенной. Ответ на этот вопрос требует серьезных размышлений о том, что точно содержится в  $u_t$ . В частности, если участники рынка апельсинового сока используют прогнозы  $FDD$  для принятия решения о том, сколько апельсинового сока они будут покупать или продавать по определенной цене, то цены на апельсиновый сок, а, следовательно, и ошибка  $u_t$ , могут включать в себя информацию о будущих значениях  $FDD$ , что могло бы сделать  $u_t$  полезной с точки зрения предсказания  $FDD$ . Это означает, что ошибка  $u_t$  будет коррелирована с будущими значениями  $FDD_t$ . Согласно этой логике, так как  $u_t$  включает в себя прогнозы будущей погоды во Флориде,  $FDD$  является экзогенной (в прошлом и настоящем), но не строго экзогенной. Различие между этим примером и примером про помидоры и удобрения заключается в том, что в то время как будущие удобрения не влияют на кусты помидоров, участники рынка апельсинового сока находятся под влиянием прогнозов о будущей погоде во Флориде. Мы вернемся к вопросу о том, является ли переменная  $FDD$  строго экзогенной при анализе цен на апельсиновый сок более подробно в разделе 15.6.

Два определения экзогенности приведены во вставке «Основные понятия 15.1».

### **15.3. Оценка динамического причинного влияния при помощи экзогенных регрессоров**

Если  $X$  является экзогенной переменной, то ее динамическое причинное влияние на  $Y$  может быть оценено по МНК в модели регрессии с распределенными лагами (15.4). В этом разделе представлены условия, в которых эти МНК-оценки приводят к обоснованным статистическим выводам, и вводятся понятия динамических мультиплекторов и совокупных динамических мультиплекторов.

#### ***Предположения модели с распределенными лагами***

Четыре предположения модели регрессии с распределенными лагами похожи на четыре предположения для модели множественной регрессии для случая межъобъектных данных (вставка «Основные понятия 6.4»), модифицированные для временных рядов.

Первое предположение заключается в том, что переменная  $X$  является экзогенной и расширяет предположение о нулевом условном среднем для межъобъектных данных включением всех запаздывающих значений  $X$ . Как отмечалось в разделе 15.2, из этого предположения следует, что  $r$  коэффициентов при распределенных лагах в уравнении (15.3) характеризуют все ненулевое динамическое причинное влияние. В этом смысле теоретическая функция регрессии суммирует весь динамический эффект влияния изменения  $X$  на  $Y$ .

Второе предположение состоит из двух частей: в части (а) требуется, чтобы переменные имели стационарное распределение, а в части (б) требуется, чтобы они были независимо распределенными, когда число моментов времени, отделяющее их друг от друга, стало большим. Это предположение, как

и соответствующее предположение для модели ADL (второе предположение во вставке «Основные понятия 14.6») и вся аргументация в его отношении из раздела 14.4 может быть применена и здесь.

Третье предположение заключается в том, что большие выбросы маловероятны, и переформулированное математически точно, предполагает, что переменные имеют более восьми ненулевых конечных моментов. Это является более сильным предположением, чем предположение о четырех конечных моментах, которое используется в этой книге. Как будет обсуждаться в разделе 15.4, это сильное предположение используется для математических выкладок НАС-оценки дисперсии.

Четвертым является точно такое же предположение, как и для множественной регрессии межобъектных данных, и заключается в предположении об отсутствии совершенной мультиколлинеарности.

Модель регрессии с распределенными лагами и ее предпосылки описаны во вставке «Основные понятия 15.2».

**ОСНОВНЫЕ  
ПОНЯТИЯ  
15.2**

**Предположения модели с распределенными лагами**

В модели с распределенными лагами, описанной во вставке «Основные понятия 15.1» [уравнение (15.4)], где

1.  $X_t$  является экзогенной, то есть  $E(u_t | X_t, X_{t-1}, X_{t-2}, \dots) = 0$ .
2. (а) Случайные величины  $Y_t$  и  $X_t$  имеют стационарные распределения и  
(б)  $(Y_t, X_t)$  и  $(Y_{t-j}, X_{t-j})$  становятся независимыми при достаточно больших  $j$ .
3. Большие выбросы маловероятны:  $Y_t$  и  $X_t$  имеют более восьми ненулевых конечных моментов.
4. Отсутствует совершенная мультиколлинеарность.

**Расширение на случай дополнительных  $X$ -в.** Модель регрессии с распределенными лагами непосредственно расширяется на случай нескольких  $X$ -в: дополнительные  $X$ -ы и их лаги просто включаются в качестве регрессоров в модель распределенных лагов, и предположения из вставки «Основные понятия 15.2» изменяются на случай включения этих дополнительных регрессоров. Несмотря на то что расширение на случай нескольких  $X$ -в концептуально просто, оно усложняет обозначения, скрывая основные идеи, оценки и выводы в модели с распределенными лагами. По этой причине случай нескольких  $X$ -в не рассматривается в явном виде в этой главе, а остается в качестве простого расширения модели с распределенными лагами для одного регрессора  $X$ .

**Автокоррелированные „стандартные ошибки и выводы“**

В модели регрессии с распределенными лагами ошибка  $u_t$  может быть автокоррелирована, то есть  $u_t$  может быть автокоррелирована со своими лагами.

Такая автокорреляция возникает потому, что во временных рядах пропущенные переменные, включенные в  $u_t$ , могут быть автокоррелированы сами по себе. Например, предположим, что спрос на апельсиновый сок также зависит от дохода, следовательно, доходы являются одним из факторов, который влияет на цену на апельсиновый сок, в частности, совокупные доходы потенциальных потребителей апельсинового сока. Тогда совокупный доход является пропущенной переменной в регрессии с распределенными лагами изменения цены на апельсиновый сок от числа морозных дней. Однако совокупный доход является серийно коррелированным: доход имеет тенденцию к снижению в периоды экономического спада и повышению в периоды роста. Таким образом, доход является серийно коррелированным, и поскольку он является частью ошибки  $u_t$ , будет серийно коррелированной. Этот пример типичен: вследствие пропуска переменных, влияющих на  $Y$  и являющихся серийно коррелированными, в общем случае  $u_t$  будет коррелированной в модели регрессии с распределенными лагами.

Автокорреляции  $u_t$  не влияют на состоятельность МНК, а также не вносят смещения. Однако если ошибки автокоррелированы, то в общем случае стандартные ошибки обычного МНК являются неэффективными, и для их расчета должна быть использована другая формула. Таким образом, корреляция ошибок аналогична гетероскедастичности: стандартные ошибки, рассчитываемые в предположении гомоскедастичности, являются «неверными», если ошибки на самом деле гетероскедастичны, в том смысле, что использование стандартных ошибок, рассчитанных в предположении гомоскедастичности, приведет к некорректным статистическим выводам, если ошибки являются гетероскедастичными. Аналогичным образом, если ошибки являются автокоррелированными, стандартные ошибки, основанные на i.i.d. ошибках, «неправильные», в том смысле что они приводят к некорректным статистическим выводам. Решение этой проблемы заключается в использовании стандартных ошибок, устойчивых к гетероскедастичности и автокорреляции (НАС), являющихся предметом рассмотрения в разделе 15.4.

## **Динамические мультиплекторы и совокупные динамические мультиплекторы**

Еще одним названием для динамического причинного влияния является динамический мультиплектор. Совокупные динамические мультиплекторы характеризуют совокупное причинное влияние до данного лага включительно; таким образом, совокупные динамические мультиплекторы измеряют кумулятивный эффект влияния изменения  $X$  на  $Y$ .

**Динамические мультиплекторы.** Эффект влияния единичного изменения  $X$  на  $Y$  через  $h$  периодов, то есть  $\beta_{h+1}$  в уравнении (15.4) называется *динамическим мультиплектором*  $h$ -го периода. Таким образом, динамические мультиплекторы, характеризующие влияние  $X$  на  $Y$ , – это коэффициенты при  $X_t$  и его запаздываниях в уравнении (15.4). Например,  $\beta_2$  – это однопериодный динамический мультиплектор,  $\beta_3$  – это двухпериодный динамический мультиплектор и так далее.

В этой терминологии динамический мультиликатор нулевого периода (или одновременный), или *эффект воздействия* (или *импульсный эффект*), то есть  $\beta_1$ , характеризует влияние изменений  $X$  на  $Y$  в тот же период.

Поскольку динамические мультиликаторы являются МНК-оценками коэффициентов регрессии, их стандартные ошибки являются НАС-стандартными ошибками коэффициентов регрессии МНК.

**Совокупные динамические мультиликаторы.** Совокупный динамический мультиликатор  $h$ -го периода представляет собой накопленный в течение  $h$  периодов эффект влияния единичного изменения  $X$  на  $Y$ . Таким образом, совокупный динамический мультиликатор – это накопленная сумма динамических мультиликаторов. В терминах коэффициентов модели регрессии с распределенными лагами (15.4) совокупный мультиликатор нулевого периода равен  $\beta_1$ , совокупный мультиликатор первого периода –  $\beta_1 + \beta_2$ , а совокупный динамический мультиликатор  $h$ -го периода равен  $\beta_1 + \beta_2 + \dots + \beta_{h+1}$ . Сумма отдельных динамических мультиликаторов  $\beta_1 + \beta_2 + \dots + \beta_{h+1}$  представляет собой совокупное долгосрочное влияние изменения  $X$  на  $Y$  и называется *долгосрочным совокупным динамическим мультиликатором*.

Например, рассмотрим регрессию (15.2). Мгновенный эффект влияния дополнительного морозного дня равен величине изменения цены концентрата апельсинового сока, то есть 0,47 %. Накопленный эффект влияния на изменение цены в следующем месяце равен сумме эффектов воздействия и динамического влияния месяцем раньше; таким образом, накопленное влияние на цену равно начальному увеличению на 0,47 % плюс последующий меньший рост в размере 0,14 %, что в общей сложности равно увеличению на 0,61 %. Аналогичным образом, двухмесячный совокупный динамический мультиликатор равен:  $0,47\% + 0,14\% + 0,06\% = 0,67\%$ .

Совокупный динамический мультиликатор может быть оценен непосредственно с использованием модификации модели регрессии с распределенными лагами (15.4). Эта модифицированная регрессия имеет вид:

$$Y_t = \delta_0 + \delta_1 \Delta X_t + \delta_2 \Delta X_{t-1} + \delta_3 \Delta X_{t-2} + \dots + \delta_r \Delta X_{t-r+1} + \delta_{r+1} X_{t-r} + u_t. \quad (15.7)$$

Коэффициенты  $\delta_0, \delta_1, \dots, \delta_{r+1}$  в уравнении (15.7) на самом деле являются совокупными динамическими мультиликаторами. Это можно показать, используя алгебраические преобразования (упражнение 15.5), что свидетельствует о том, что теоретические регрессии (15.7) и (15.4) эквивалентны, где  $\delta_0 = \beta_0$ ,  $\delta_1 = \beta_1$ ,  $\delta_2 = \beta_1 + \beta_2$ ,  $\delta_3 = \beta_1 + \beta_2 + \beta_3$  и так далее. Коэффициент при  $X_{t-r}$  –  $\delta_{r+1}$  является долгосрочным совокупным динамическим мультиликатором, то есть  $\delta_{r+1} = \beta_1 + \beta_2 + \beta_3 + \dots + \beta_{r+1}$ . Кроме того, МНК-оценки коэффициентов в уравнении (15.7) являются такими же, что и соответствующая накопленная сумма МНК-оценок в уравнении (15.4). Например,  $\hat{\delta}_2 = \hat{\beta}_1 + \hat{\beta}_2$ . Основное преимущество оценки кумулятивных динамических мультиликаторов при использовании спецификации в уравнении (15.7) заключается в том, что, поскольку МНК-оценки коэффициентов регрессии являются оценками совокупных динамических мультиликаторов, НАС-стандартные ошибки коэффициентов в уравнении (15.7) являются НАС-стандартными ошибками совокупных динамических мультиликаторов.

## 15.4. Стандартные ошибки, являющиеся состоятельными при наличии гетероскедастичности и автокорреляции

Если ошибки  $u_t$  автокоррелированы, то МНК-оценки коэффициентов состоятельны, но в целом обычные стандартные ошибки МНК для межобъектных данных таковыми не являются. Это означает, что обычные статистические выводы — проверка гипотез и доверительные интервалы, проводимые с использованием обычных стандартных ошибок МНК, будут, вообще говоря, вводить в заблуждение. Например, доверительные интервалы, построенные как обычная МНК-оценка стандартных ошибок  $\pm 1,96$ , не должна содержать истинного значения коэффициента в 95 % всех повторных выборок, даже если размер выборок является большим. Данный раздел начинается с вывода правильной формулы для дисперсии МНК-оценки с автокоррелированными ошибками, а затем распространяется на случай устойчивых к гетероскедастичности и автокорреляции (НАС) стандартных ошибок.

В этом разделе рассматриваются НАС-стандартные ошибки для регрессии временных рядов. В главе 10 были введены кластеризованные стандартные ошибки, похожие на НАС-стандартные ошибки и подходящие для случая панельных данных. Несмотря на то что кластеризованные стандартные ошибки для панельных данных и НАС-стандартные ошибки для временных рядов имеют одинаковую цель, разные структуры данных приводят к различным формулам. Этот раздел является самодостаточным, а ознакомление с главой 10 не является обязательным условием для ее изучения.

### **Распределение МНК-оценки в случае автокоррелированных ошибок**

Рассмотрим для простоты МНК-оценку  $\hat{\beta}_1$  в модели регрессии с распределенными лагами без запаздываний, то есть модели парной линейной регрессии с одним регрессором  $X_t$ :

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \quad (15.8)$$

где выполняются предположения из вставки «Основные понятия 15.2». В данном разделе показано, что дисперсия  $\hat{\beta}_1$  может быть представлена в виде произведения двух множителей: выражение для  $\text{var}(\hat{\beta}_1)$ , имеющее место, если  $u_t$  не является сериально коррелированной, умноженное на поправочный коэффициент, который возникает из-за автокорреляции в  $u_t$  или, точнее, автокорреляции в  $(X_t - \mu_X)u_t$ .

Как показано в приложении 4.3, формулу для МНК-оценки  $\hat{\beta}_1$  из вставки «Основные понятия 4.2» можно переписать в таком виде:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{T} \sum_{t=1}^T (X_t - \bar{X})u_t}{\frac{1}{T} \sum_{t=1}^T (X_t - \bar{X})^2}, \quad (15.9)$$

где уравнение (15.9) представляет собой уравнение (4.30) с измененными обозначениями, так что  $i$  и  $n$  заменяются на  $t$  и  $T$ . Так как  $\bar{X} \xrightarrow{p} \mu_X$  и  $\frac{1}{T} \sum_{t=1}^T (X_t - \bar{X})^2 \xrightarrow{p} \sigma_X^2$ , в больших выборках  $\hat{\beta}_1 - \beta_1$  приблизительно равно

$$\hat{\beta}_1 - \beta_1 \cong \frac{\frac{1}{T} \sum_{t=1}^T (X_t - \mu_X) u_t}{\sigma_X^2} = \frac{\frac{1}{T} \sum_{t=1}^T v_t}{\sigma_X^2} = \frac{\bar{v}}{\sigma_X^2}, \quad (15.10)$$

где  $v_t = (X_t - \mu_X) u_t$  и  $\bar{v} = \frac{1}{T} \sum_{t=1}^T v_t$ . Таким образом,

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\frac{\bar{v}}{\sigma_X^2}\right) = \frac{\text{var}(\bar{v})}{\left(\sigma_X^2\right)^2}. \quad (15.11)$$

Если  $v_t$  является i.i.d., как предполагается для межобъектных данных во вставке «Основные понятия 4.3»,  $\text{var}(\bar{v}) = \text{var}(v_t)/T$  и формула для дисперсии  $\hat{\beta}_1$  из вставки «Основные понятия 4.4» применима. Однако если  $u_t$  и  $X_t$  не являются независимо распределенными во времени, то в общем случае  $v_t$  будет серийно коррелированной, поэтому  $\text{var}(\bar{v}) \neq \text{var}(v_t)/T$  и формула из вставки «Основные понятия 4.4» неприменима. Вместо этого, если  $v_t$  автокоррелирована, дисперсия  $\bar{v}$  имеет вид:

$$\begin{aligned} \text{var}(\bar{v}) &= \text{var}[(v_1 + v_2 + \dots + v_T)/T] = \\ &= [\text{var}(v_1) + \text{cov}(v_1, v_2) + \dots + \text{cov}(v_1, v_T) + \\ &\quad + \text{cov}(v_1, v_2) + \text{var}(v_2) + \dots + \text{var}(v_T)]/T^2 = \\ &= [T \text{var}(v_t) + 2(T-1) \text{cov}(v_t, v_{t-1}) + 2(T-2) \text{cov}(v_t, v_{t-2}) + \\ &\quad + \dots + 2 \text{cov}(v_t, v_{t-T+1})]/T^2 = \frac{\sigma_v^2}{T} f_T, \end{aligned} \quad (15.12)$$

где

$$f_T = 1 + 2 \sum_{j=1}^{T-1} \left( \frac{T-j}{T} \right) \rho_j \quad (15.13)$$

и где  $\rho_j = \text{corr}(v_t, v_{t-j})$ . В больших выборках  $f_T$  стремится к пределу  $f_T \rightarrow f_\infty = 1 + 2 \sum_{j=1}^{\infty} \rho_j$ .

Комбинируя выражения в уравнении (15.10) для  $\hat{\beta}_1$  и в уравнении (15.12) для  $\text{var}(\bar{v})$ , получаем формулу для дисперсии  $\hat{\beta}_1$ , когда  $v_t$  автокоррелирована:

$$\text{var}(\hat{\beta}_1) = \left[ \frac{1}{T} \frac{\sigma_v^2}{\left(\sigma_X^2\right)^2} \right] f_T, \quad (15.14)$$

где  $f_T$  определено в выражении (15.13).

Уравнение (15.14) выражает дисперсию как произведение двух множителей. Во-первых, это выражение в квадратных скобках, которое является формулой для дисперсии  $\hat{\beta}_1$ , приведенной во вставке «Основные понятия 4.4» и используемой в отсутствие серийной корреляции. Во-вторых, это множитель  $f_T$ , который корректирует эту формулу для случая наличия серийной корреляции. Из-за это-

го дополнительного множителя  $f_t$  в уравнении (15.14) обычные стандартные ошибки МНК, вычисляемые с использованием уравнения (5.4), являются некорректными, если ошибки автокоррелированы: если  $v_t = (X_t - \mu_X)u_t$  серийно коррелирована, то в формуле оценки дисперсии пропущен множитель  $f_t$ .

### **НAC-стандартные ошибки**

Если множитель  $f_t$ , определенный в уравнении (15.13), был бы известен, то дисперсия  $\hat{\beta}_1$  могла бы быть оценена путем умножения обычной межъобъектной оценки дисперсии на  $f_t$ . Однако этот множитель зависит от неизвестных автокорреляций  $v_t$ , поэтому он должен быть оценен. Оценка дисперсии  $\hat{\beta}_1$ , которая включает эту поправку, является состоятельной независимо от того, присутствует или нет гетероскедастичность и является ли  $v_t$  автокоррелированной или нет. Соответственно, эта оценка называется оценкой дисперсии  $\hat{\beta}_1$ , состоятельной при наличии гетероскедастичности и автокорреляции (НAC), и квадратный корень из НAC-оценки дисперсии является НAC-стандартной ошибкой  $\hat{\beta}_1$ .

**НAC-формула для дисперсии.** Оценка дисперсии  $\hat{\beta}_1$ , являющаяся состоятельной при наличии гетероскедастичности и автокоррелированности, имеет вид

$$\tilde{\sigma}_{\hat{\beta}_1}^2 = \hat{\sigma}_{\hat{\beta}_1}^2 \hat{f}_T, \quad (15.15)$$

где  $\hat{\sigma}_{\hat{\beta}_1}^2$  – это оценка дисперсии  $\hat{\beta}_1$  в отсутствие серийной корреляции, приведенная в уравнении (5.4), и где  $\hat{f}_T$  – оценка множителя  $f_t$  в уравнении (15.13).

Задача построения состоятельной оценки  $\hat{f}_t$  довольно сложна. Чтобы понять почему, рассмотрим два крайних случая. В первом из них, имея формулу (15.13), может показаться естественным заменить теоретические автокорреляции  $\rho_j$  выборочными –  $\hat{\rho}_j$  [как определено в уравнении (14.6)], что даст оценку  $1 + 2 \sum_{j=1}^{T-1} \left(\frac{T-j}{T}\right) \hat{\rho}_j$ . Но эта оценка содержит так много оцененных автокорреляций, что является несостоятельной. Интуитивно, из-за того что каждая из оцененных автокорреляций содержит ошибку оценки, то оценивая так много автокорреляций, содержащих ошибки оценки, оценка  $f_t$  остается большой даже в больших выборках. В другом крайнем случае можно использовать всего несколько выборочных автокорреляций, например только выборочную автокорреляцию первого порядка, и игнорировать автокорреляции более высоких порядков. Несмотря на то что такая оценка устраниет проблему оценки слишком большого числа автокорреляций, у нее есть другая проблема: она несостоятельна из-за того, что она игнорирует дополнительные автокорреляции, которые присутствуют в уравнении (15.13). Если говорить коротко, то использование слишком большого числа выборочных автокорреляций приводит к тому, что оценка имеет большую дисперсию, но использование слишком малого числа автокорреляций влечет за собой игнорирование автокорреляций более высокого порядка, так что в любом из этих крайних случаев оценка является несостоятельной.

Оценки  $f_t$ , используемые на практике, представляют собой баланс между этими двумя крайними случаями выбора числа включаемых автокорреляций, в том смысле что зависят от размера выборки  $T$ . Если размер выборки небольшой,

то используется немного автокорреляций, но если выборка большая, то включается много автокорреляций (но все же намного меньше, чем  $T$ ). Более точно, пусть  $\hat{f}_t$  задана формулой:

$$\hat{f}_t = 1 + 2 \sum_{j=1}^{m-1} \left( \frac{m-j}{m} \right) \tilde{\rho}_j, \quad (15.16)$$

где  $\tilde{\rho}_j = \sum_{t=j+1}^T \hat{v}_t \hat{v}_{t-j} / \sum_{t=1}^T \hat{v}_t^2$  и где  $\hat{v}_t = (X_t - \bar{X}) u_t$  (как в определении  $\hat{\sigma}_{\beta_1}^2$ ). Параметр  $m$  в уравнении (15.16) называется *параметром усечения* (или *шириной окна*) НАС-оценки, так как сумма автокорреляции укорочена, или усечена, потому что включается только  $m-1$  автокорреляция вместо  $T-1$  автокорреляций, которые присутствуют в теоретической формуле из уравнения (15.13).

Чтобы  $\hat{f}_t$  была состоятельной,  $m$  должен быть выбран так, чтобы он был большим в больших выборках, но все же по-прежнему гораздо меньше, чем  $T$ . Одним из возможных способов выбора  $m$  на практике является использование формулы:

$$m = 0,75T^{1/3}, \quad (15.17)$$

округленного до целого числа. Эта формула, основанная на предположении о наличии умеренного числа автокорреляций в  $v_t$ , представляет собой критерий для определения  $m$  в зависимости от числа наблюдений в регрессии<sup>1</sup>.

Значение параметра усечения  $m$ , следующее из уравнения (15.17), можно изменить, используя ваши знания о временном ряде, с которым вы работаете. С одной стороны, если  $v_t$  сильно серийно коррелирована, то можно увеличить  $m$  по сравнению со значением из уравнения (15.17). С другой стороны, если  $v_t$  имеет серийные корреляции невысоких порядков, то это позволяет уменьшить  $m$ . Из-за неопределенности, связанной с выбором  $m$ , стандартная практика заключается в выборе одного или двух альтернативных значений  $m$  для, как минимум, одной спецификации, для того чтобы убедиться, что ваши результаты нечувствительны к  $m$ .

НАС-оценка из уравнения (15.15) с  $\hat{f}_t$ , заданным в уравнении (15.16), называется *оценкой дисперсии Ньюи–Веста* (Newey–West), названной так по именам эконометристов Уитни Ньюи (Whitney Newey) и Кеннета Веста (Kenneth West), которые предложили ее. Они показали, что когда используется правило, похожее на правило из уравнения (15.17), при общих предположениях такая оценка является состоятельной оценкой дисперсии  $\hat{\sigma}_{\beta_1}^2$  (Newey, West, 1987). В их доказательстве (и в доказательствах в работе Эндрюса [Andrews, 1991]) предполагается, что  $v_t$  имеет более четырех моментов, из чего, в свою очередь, следует, что  $X_t$  и  $u_t$  имеют более восьми моментов, и это является причиной того, что третья предпосылка во вставке «Основные понятия 15.2» предполагает наличие более восьми моментов у  $X_t$  и  $u_t$ .

<sup>1</sup> Уравнение (15.17) дает возможность выбрать  $m$  «наилучшим» образом, если  $u_t$  и  $X_t$  являются процессами авторегрессии первого порядка с коэффициентами при первых запаздываниях, равными 0,5, где слово «наилучший» означает оценку, минимизирующую  $E(\tilde{\sigma}_{\beta_1}^2 - \sigma_{\beta_1}^2)^2$ . Уравнение (15.17) основано на более общей формуле полученной Эндрюсом [Andrews, 1991, уравнение (5.3)].

**Другие НАС-оценки.** Оценка дисперсии Ньюи–Веста – не единственная НАС-оценка. Например, веса  $(m-j)/m$  в уравнении (15.16) могут быть заменены другими весами. Если используются другие веса, то правило выбора параметра усечения из уравнения (15.17) больше нельзя применять, а вместо него следует использовать другие правила, разработанные для используемых весов. Обсуждение НАС-оценки с использованием других весов выходит за рамки данной книги. Для получения дополнительной информации по этой теме см. Hayashi (2000, раздел 6.6).

**Расширение на случай множественной регрессии.** Все вопросы, обсуждаемые в этом разделе, обобщаются на модель регрессии с распределенными лагами из вставки «Основные понятия 15.1» с несколькими лагами и, в более общем случае, на модель множественной регрессии с серийно коррелированными ошибками. В частности, если ошибка серийно коррелирована, то стандартные ошибки обычного МНК не являются надежными с точки зрения получаемых выводов, и вместо них следует использовать НАС-стандартные ошибки. Если для НАС-оценки дисперсии используется оценка Ньюи–Веста (НАС-оценка дисперсии на основе весов  $(m-j)/m$ ), то параметр усечения  $m$  может быть выбран в соответствии с правилом из уравнения (15.17), независимо от того, один или несколько регрессоров включены в регрессию. Формула для НАС-стандартных ошибок во множественной регрессии включена в современные программные эконометрические пакеты, предназначенные для работы с временными рядами. Из-за того что эта формула основана на матричной алгебре, мы ее опускаем и отсылаем читателя к работе Хаяши (Hayashi, 2000, раздел 6.6) для ознакомления с математическими деталями.

НАС-стандартные ошибки приведены во вставке «Основные понятия 15.3».

## 15.5. Оценка динамического причинного влияния при помощи строго экзогенных регрессоров

Если  $X$  является строго экзогенной величиной, то нужно использовать две альтернативные оценки динамического причинного влияния. Первая такая оценка включает оценку авторегрессионной модели с распределенными лагами (ADL-модель) вместо модели с распределенными лагами и расчет динамических мультипликаторов из оцененных коэффициентов в ADL-модели. Действуя таким образом, мы можем оценивать меньшее число коэффициентов, чем при использовании МНК для оценки модели регрессии с распределенными лагами, потенциально снижая ошибку. Второй способ заключается в оценке коэффициентов модели регрессии с распределенными лагами с использованием *обобщенного метода наименьших квадратов* (ОМНК) вместо МНК. Несмотря на то что при одинаковом числе коэффициентов в модели регрессии с распределенными лагами оценки коэффициентов по ОМНК и МНК совпадают, оценки ОМНК имеют меньшую дисперсию. Чтобы сохранить простоту изложения, мы сначала рассматриваем оба этих метода оценки в контексте модели регрессии с распределенными лагами и ошибками, имеющими структуру модели AR(1). Однако эти две оценки дают большие преимущества, когда в модель регрессии

с распределенными лагами включается много запаздываний, поэтому эти методы оценки затем расширяются на общий случай модели регрессии с распределенными лагами с ошибками, являющимися авторегрессией более высокого порядка.

## ОСНОВНЫЕ ПОНЯТИЯ

### 15.3

#### **НAC-стандартные ошибки**

*Проблема:* ошибка  $u_t$  модели регрессии с распределенными лагами из вставки «Основные понятия 15.1» может быть серийно коррелирована. Если это так, то МНК-оценки коэффициентов состоятельны, но в общем случае обычные (МНК) стандартные ошибки таковыми не являются, что приводит к некорректным результатам проверки гипотез и доверительным интервалам.

*Решение:* стандартные ошибки следует вычислять с использованием устойчивой к гетероскедастичности и автокорреляции оценки (НAC-оценки). НAC-оценка предполагает оценку  $m - 1$  автоковариации, а также дисперсии; соответствующие выражения для случая парной регрессии приведены в формулах (15.15) и (15.16).

На практике, для того чтобы использовать НAC-оценку, необходимо выбрать параметр усечения (ширину окна)  $m$ . Для этого можно использовать формулу (15.17) как базовую, а затем увеличивать или уменьшать  $m$  в зависимости от того, являются или нет ваши регрессоры и ошибки сильно серийно коррелированными.

#### **Модель регрессии с распределенными лагами с ошибками в виде AR(1)**

Предположим, что причинное влияние от изменения  $X$  на  $Y$  длится только в течение двух периодов, то есть оно оказывает импульсный эффект воздействия  $\beta_1$  и эффект воздействия в следующий период  $\beta_2$ , но не влияет после этого. Тогда соответствующая модель регрессии с распределенными лагами представляет собой модель регрессии с распределенными лагами, включающую только текущее и предыдущее значения  $X_t$ :

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t. \quad (15.18)$$

Как отмечалось в разделе 15.2, в общем случае ошибка  $u_t$  в уравнении (15.18) является серийно коррелированной. Одно из следствий этой серийной корреляции заключается в том, что если модель регрессии с распределенными лагами оценивается с помощью МНК, то выводы, основанные на стандартных ошибках, получаемых при использовании обычного МНК, могут быть некорректными. По этой причине в разделах 15.3 и 15.4 подчеркива-

лась необходимость использования НАС-стандартных ошибок, если  $\beta_1$  и  $\beta_2$  из (15.18) оцениваются с помощью МНК.

В этом разделе мы рассмотрим другой подход к серийной корреляции  $u_t$ . Такой подход, применение которого возможно, если  $X_t$  является строго экзогенным регрессором, предполагает введение модели авторегрессии для описания серийной корреляции  $u_t$ , а затем вывод с помощью этой AR-модели оценок, которые могут быть более эффективными, чем МНК-оценки в модели регрессии с распределенными лагами.

Более точно, предположим, что  $u_t$  описываются AR(1)-моделью так:

$$u_t = \phi_1 u_{t-1} + \tilde{u}_t, \quad (15.19)$$

где  $\phi_1$  – параметр авторегрессии,  $\tilde{u}_t$  серийно не коррелированы и константа отсутствует, поскольку  $E(u_t) = 0$ . Из уравнений (15.18) и (15.19) следует модель регрессии с распределенными лагами с серийно коррелированной ошибкой, которая может быть переписана в виде авторегрессионной модели с распределенными лагами с серийно некоррелированной ошибкой. Для этого рассмотрим уравнение (15.18) для случая, когда зависимая переменная является первым запаздыванием, умножим полученное выражение на  $\phi_1$  и вычтем его из уравнения (15.18):

$$\begin{aligned} Y_t - \phi_1 Y_{t-1} &= (\beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t) - \phi_1 (\beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \\ &+ u_{t-1}) = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} - \phi_1 \beta_0 - \phi_1 \beta_1 X_{t-1} - \phi_1 \beta_2 X_{t-2} + \tilde{u}_t, \end{aligned} \quad (15.20)$$

где во втором равенстве используется  $\tilde{u}_t = u_t - \phi_1 u_{t-1}$ . Приводя подобные слагаемые в выражении (15.20), получаем:

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \tilde{u}_t, \quad (15.21)$$

$$\text{где } \alpha_0 = \beta_0(1 - \phi_1), \delta_0 = \beta_1, \delta_1 = \beta_2 - \phi_1 \beta_1, \text{ и } \delta_2 = -\phi_1 \beta_2 \quad (15.22)$$

и где  $\beta_0$ ,  $\beta_1$  и  $\beta_2$  – коэффициенты в уравнении (15.18) и  $\phi_1$  – коэффициент автокорреляции в уравнении (15.19).

Уравнение (15.21) представляет собой ADL-модель, которая включает в себя одновременное значение  $X$  и два его запаздывания. Мы будем называть уравнение (15.21) ADL-представлением модели с распределенными лагами с авторегрессионными ошибками, которая описывается уравнениями (15.18) и (15.19).

Слагаемые в уравнении (15.20) могут быть сгруппированы иначе, чтобы получить выражение, которое эквивалентно уравнениям (15.21) и (15.22). Пусть  $\tilde{Y}_t = Y_t - \phi_1 Y_{t-1}$  обозначает *квазиразность*  $Y_t$  («квази», потому что это не первая разность, т.е. разность между  $Y_t$  и  $Y_{t-1}$ ; точнее, это разность между  $Y_t$  и  $\phi_1 Y_{t-1}$ ). Аналогично, пусть  $\tilde{X}_t = X_t - \phi_1 X_{t-1}$  – квазиразность  $X_t$ , тогда уравнение (15.20) можно записать в таком виде:

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + \tilde{u}_t. \quad (15.23)$$

Мы будем называть уравнение (15.23) представлением модели с распределенными лагами с авторегрессионными ошибками, определяемой уравнениями (15.18)

и (15.19), представлением в виде квазиразностей или квазидифференцированной моделью.

ADL-модель (15.21) [с ограничениями на параметры (15.22)] и квазидифференцированная модель в уравнении (15.23) эквивалентны. В обеих моделях ошибки  $\tilde{u}_t$  серийно коррелированы. Однако эти два представления предполагают разные стратегии их оценки. Но прежде чем обсуждать эти стратегии, мы сформулируем предположения, при которых они дают состоятельные оценки динамических мультипликаторов  $\beta_1$  и  $\beta_2$ .

**Предположение о равенстве нулю условного среднего в ADL(1, 2) и квазидифференцированные модели.** Поскольку уравнения (15.21) [с ограничениями (15.22)] и (15.23) эквивалентны, условия для их оценки одинаковы, поэтому для удобства мы рассмотрим уравнение (15.23).

Квазидифференцированная модель (15.23) представляет собой модель регрессии с распределенными лагами от квазидифференцированных переменных с серийно некоррелированной ошибкой. Соответственно, условия для МНК-оценок коэффициентов в уравнении (15.23) являются предположениями метода наименьших квадратов для модели регрессии с распределенными лагами из вставки «Основные понятия 15.2», выраженные через  $\tilde{u}_t$  и  $\tilde{X}_t$ . Критическим предположением здесь является первое предположение, которое применительно к уравнению (15.23) говорит о том, что переменная  $\tilde{X}_t$  экзогенна, то есть:

$$E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \dots) = 0, \quad (15.24)$$

где мы предполагаем, что условное математическое ожидание зависит от глубоких запаздываний  $\tilde{X}_t$ , что гарантирует отсутствие необходимости включения в теоретическую модель дополнительных лагов  $\tilde{X}_t$  помимо тех, что уже включены в уравнение (15.23).

Из-за того что  $\tilde{X}_t = X_t - \varphi_1 X_{t-1}$ , получаем  $X_t = \tilde{X}_t + \varphi_1 X_{t-1}$ , и тогда рассмотрение условного математического ожидания относительно  $\tilde{X}_t$  и всех его лагов эквивалентно и рассмотрению условного математического ожидания относительно  $X_t$  и всех его лагов. Таким образом, условие (15.24) эквивалентно условию  $E(\tilde{u}_t | X_t, X_{t-1}, \dots) = 0$ . Кроме того, поскольку  $\tilde{u}_t = u_t - \varphi_1 u_{t-1}$ , это условие, в свою очередь, означает:

$$\begin{aligned} 0 &= E(\tilde{u}_t | X_t, X_{t-1}, \dots) = E(u_t - \varphi_1 u_{t-1} | X_t, X_{t-1}, \dots) = \\ &= E(u_t | X_t, X_{t-1}, \dots) - \varphi_1 E(u_{t-1} | X_t, X_{t-1}, \dots). \end{aligned} \quad (15.25)$$

Для того чтобы равенства в уравнении (15.25) имели место в общем случае для различных значений  $\varphi_1$ , должны выполняться оба равенства  $E(u_t | X_t, X_{t-1}, \dots) = 0$  и  $E(u_{t-1} | X_t, X_{t-1}, \dots) = 0$ . Меняя индексы, получаем, что условие  $E(u_{t-1} | X_t, X_{t-1}, \dots) = 0$  может быть переписано в таком виде:

$$E(u_t | X_{t+1}, X_t, X_{t-1}, \dots) = 0, \quad (15.26)$$

из которого (по закону повторного математического ожидания) следует, что  $E(u_t | X_t, X_{t-1}, \dots) = 0$ . Таким образом, выполнение предположения о нулевом

условном среднем (15.24) для произвольных значений  $\varphi_1$  эквивалентно выполнению условия (15.26).

Условие (15.26) следует из строгой экзогенности  $X_t$ , но это не следует из (текущей и прошлой) экзогенности  $X_t$ . Таким образом, предположения метода наименьших квадратов для оценки модели с распределенными лагами (15.23) выполняются, если регрессор  $X_t$  строго экзогенен, но для их выполнения недостаточно (текущей и прошлой) экзогенности  $X_t$ .

Поскольку ADL-представление [уравнения (15.21) и (15.22)] эквивалентно квазидифференцированному представлению [уравнение (15.23)], предположение об условном среднем, необходимое для оценки коэффициентов в квазидифференцированном представлении [т.е.  $E(u_t | X_{t+1}, X_t, X_{t-1}, \dots) = 0$ ], также является предположением об условном среднем для состоятельной оценки коэффициентов в ADL-представлении.

Обратимся теперь к двум стратегиям оценки, предложенным для этих представлений: оценка ADL-коэффициентов и оценка коэффициентов в квазидифференцированной модели.

### **МНК-оценка ADL-модели**

Первая стратегия заключается в использовании МНК для оценки коэффициентов в ADL-модели (15.21). Так как вывод, приводящий к уравнению (15.21), показывает, что включение запаздываний  $Y$  и дополнительных лагов  $X$  в качестве регрессоров приводит к серийной некоррелированности ошибки (в предположении, что ошибки являются процессом авторегрессии первого порядка). Таким образом, обычные МНК-стандартные ошибки могут быть использованы, то есть НАС-стандартные ошибки не нужны, когда коэффициенты ADL-модели (15.21) оцениваются с помощью МНК.

Оценки коэффициентов ADL-модели сами по себе не являются оценками динамических мультипликаторов, но динамические мультипликаторы можно вычислить из коэффициентов ADL-модели. Общий способ вычисления динамических мультипликаторов заключается в выражении оцененной функции регрессии через текущее и прошлые значения  $X_t$ , то есть в удалении  $Y_t$  из оцененной функции регрессии. Для этого необходимо несколько раз подставить выражения для запаздывающих значений  $Y_t$  в оцененную функцию регрессии. Рассмотрим оценку функции регрессии:

$$\hat{Y}_t = \hat{\varphi}_1 Y_{t-1} + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \hat{\delta}_2 X_{t-2}, \quad (15.27)$$

где оценка константы была опущена, поскольку она не включается в выражение для динамических множителей. Рассматривая (15.27) для первого запаздывания объясняемой переменной, получаем:  $\hat{Y}_{t-1} = \hat{\varphi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}$ , поэтому, заменяя  $Y_{t-1}$  в уравнении (15.27) на это выражение для  $\hat{Y}_{t-1}$  и собирая слагаемые, получаем:

$$\begin{aligned} \hat{Y}_t &= \hat{\varphi}_1 (\hat{\varphi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}) + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \\ &+ \hat{\delta}_2 X_{t-2} = \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\varphi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\varphi}_1 \hat{\delta}_1) X_{t-2} + \\ &+ \hat{\varphi}_1 \hat{\delta}_2 X_{t-3} + \hat{\varphi}_1^2 Y_{t-2}. \end{aligned} \quad (15.28)$$

Повторяя этот процесс замены для  $Y_{t-2}$ ,  $Y_{t-3}$  и так далее, получаем:

$$\begin{aligned}\hat{Y}_t = & \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\phi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-2} + \\ & + \hat{\phi}_1 (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-3} + \hat{\phi}_1^2 (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-4} + \dots\end{aligned}\quad (15.29)$$

Коэффициенты в уравнении (15.29) являются оценками динамических мультипликаторов, рассчитанными на основе МНК-оценок коэффициентов в ADL-модели (15.21). Если бы ограничения на коэффициенты в уравнении (15.22) были бы в точности равны оценкам коэффициентов, то динамические мультипликаторы, начиная со второго (т.е. коэффициенты при  $Y_{t-2}$ ,  $Y_{t-3}$  и так далее), все были бы равны нулю<sup>1</sup>. Тем не менее при такой стратегии оценки эти ограничения не будут выполняться в точности, так что оцененные мультипликаторы, начиная со второго, в уравнении (15.29) будут, как правило, отличны от нуля.

## ОМНК-оценка

Вторым методом оценки динамических мультипликаторов при строгой экзогенности  $X_t$  является обобщенный метод наименьших квадратов (ОМНК), что влечет за собой необходимость оценки уравнения (15.23). Для описания ОМНК-оценки мы изначально предполагаем, что  $\varphi_1$  известен. Вследствие того что на практике он неизвестен, получить такую оценку не представляется возможным, поэтому ее называют недоступной (нереализуемой) оценкой ОМНК. Однако недоступная ОМНК-оценка может быть изменена с помощью оценки  $\varphi_1$ , что приводит к так называемой доступной (реализуемой) версии ОМНК-оценки.

**Недоступный (нереализуемый) ОМНК.** Предположим, что  $\varphi_1$  известен, тогда квазидифференцированные переменные  $\tilde{X}_t$  и  $\tilde{Y}_t$  могут быть вычислены непосредственно по определению. Как обсуждалось в контексте уравнений (15.24) и (15.26), если  $X_t$  является строго экзогенным регрессором, то  $E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \dots) = 0$ . Таким образом, если  $X_t$  строго экзогенен и если  $\varphi_1$  известен, то коэффициенты  $\alpha_0$ ,  $\beta_1$  и  $\beta_2$  в уравнении (15.23) могут быть оценены с помощью применения МНК к регрессии  $\tilde{Y}_t$  от  $\tilde{X}_t$  и  $\tilde{X}_{t-1}$  (включая константу). Полученные оценки  $\beta_1$  и  $\beta_2$ , то есть МНК-оценки угловых коэффициентов в уравнении (15.23) при известном  $\varphi_1$ , представляют собой недоступную (нереализуемую) ОМНК-оценку. Получить эту оценку невозможно, поскольку  $\varphi_1$  неизвестен, из-за чего  $\tilde{X}_t$  и  $\tilde{Y}_t$  не могут быть вычислены, и поэтому МНК-оценки фактически не могут быть вычислены.

**Доступный (реализуемый) ОМНК.** Доступная (реализуемая) ОМНК-оценка изменяет нереализуемую ОМНК-оценку с помощью предварительной оценки  $\varphi_1$ ,  $\hat{\phi}_1$ , чтобы вычислить предполагаемые квазиразности. Более точно, доступные ОМНК-оценки  $\beta_1$  и  $\beta_2$  являются МНК-оценками  $\beta_1$  и  $\beta_2$  в уравнении (15.23), вычисленные оценкой регрессии  $\tilde{Y}_t$  на  $\tilde{X}_t$  и  $\tilde{X}_{t-1}$  (с константой), где  $\tilde{X}_t = X_t - \hat{\phi}_1 X_{t-1}$  и  $\tilde{Y}_t = Y_t - \hat{\phi}_1 Y_{t-1}$ .

Предварительная оценка,  $\hat{\phi}_1$ , может быть получена путем оценки модели регрессии с распределенными лагами (15.18) с помощью МНК, затем используя

<sup>1</sup> Подставьте равенства в уравнение (15.22), чтобы показать, что если равенства верны, то  $\delta_2 + \varphi_1 \delta_1 + \varphi_1^2 \delta_0 = 0$ .

МНК, чтобы получить оценку  $\varphi_1$  в регрессии (15.19) МНК-остатков  $\hat{u}_t$ , заменяющих ошибки регрессии  $u_t$ . Эта версия ОМНК-оценки называется оценкой Кохрейна–Оркэтта (Cochrane, Orcutt, 1949).

Расширение метода Кохрейна–Оркэтта заключается в итеративном продолжении этого процесса: используйте ОМНК-оценку коэффициентов  $\beta_1$  и  $\beta_2$ , чтобы переоценить оценку  $u_t$ ; используйте эти новые остатки повторно, чтобы переоценить  $\varphi_1$ ; используйте эту переоценку  $\varphi_1$  для вычисления переоцененных квазиразностей; используйте эти уточненные квазиразности для повторной оценки  $\beta_1$  и  $\beta_2$  и продолжайте этот процесс до тех пор, пока оценки  $\beta_1$  и  $\beta_2$  не сойдутся. Такие оценки называются итеративными оценками Кохрейна–Оркэтта.

**Интерпретация ОМНК-оценки как нелинейной оценки наименьших квадратов.** Эквивалентной интерпретацией ОМНК-оценки является рассмотрение ее как оценки ADL-модели (15.21), на которую накладываются ограничения на параметры (15.22). Эти ограничения являются нелинейными функциями от исходных параметров  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  и  $\varphi_1$ , следовательно, эта оценка не может быть получена с использованием МНК. Вместо этого параметры могут быть оценены при помощи нелинейного метода наименьших квадратов (НМНК). Как отмечалось в приложении 8.1, НМНК минимизирует сумму квадратов остатков, оценивая функцию регрессии, являющуюся нелинейной по оцениваемым параметрам. В общем случае оценка НМНК может требовать сложных вычислительных алгоритмов для минимизации нелинейной функции от неизвестных параметров. Однако в часто встречающихся частных случаях такие сложные алгоритмы оказываются не нужны, точнее, НМНК-оценку можно получить с помощью алгоритма, описанного ранее для итеративного метода Кохрейна–Оркэтта. Таким образом, итеративные ОМНК-оценки Кохрейна–Оркэтта на самом деле являются НМНК-оценкой коэффициентов в ADL-модели с учетом нелинейных ограничений, задаваемых в (15.22).

**Эффективность ОМНК.** Достоинство ОМНК-оценки заключается в том, что когда  $X$  является строго экзогенной переменной и трансформированные ошибки  $\tilde{u}_t$  гомоскедастичны, она является эффективной в классе линейных оценок, по крайней мере в больших выборках. Чтобы убедиться в этом, рассмотрим сначала недоступную ОМНК-оценку. Если  $\tilde{u}_t$  гомоскедастична и  $\varphi_1$  известен (так, что  $\tilde{X}_t$  и  $\tilde{Y}_t$  могут быть рассмотрены как наблюдаемые) и если переменная  $X_t$  строго экзогенна, то из теоремы Гаусса–Маркова следует, что МНК-оценка коэффициентов  $\alpha_0$ ,  $\beta_1$  и  $\beta_2$  в уравнении (15.23) является эффективной в классе всех линейных условно несмещенных оценок, то есть МНК-оценки коэффициентов в уравнении (15.23) являются наилучшими линейными несмещенными оценками, или BLUE (раздел 5.5). Поскольку МНК-оценка регрессии (15.23) является недоступной ОМНК-оценкой, это означает, что недоступная ОМНК-оценка является BLUE. Доступная ОМНК-оценка аналогична недоступной ОМНК-оценке, за исключением того, что  $\varphi_1$  оценивается. Поскольку оценка  $\varphi_1$  состоятельна и ее дисперсия обратно пропорциональна  $T$ , доступная и недоступная ОМНК-оценки имеют одинаковые дисперсии в больших выборках. В этом смысле, если  $X$  строго экзогенен, то доступная ОМНК-оценка является

BLUE в больших выборках. В частности, если  $X$  строго экзогенен, то ОМНК-оценка является более эффективной, чем МНК-оценка коэффициентов при распределенных лагах, обсуждаемых в разделе 15.3.

Оценки Кохрейна–Оркэтта и итеративные оценки Кохрейна–Оркэтта, представленные здесь, являются частными случаями ОМНК-оценки. Вообще, ОМНК-оценка включает в себя преобразование регрессионной модели так, чтобы ошибки были гомоскедастичными и серийно некоррелированными, и оценку коэффициентов такой преобразованной модели регрессии по МНК. В целом ОМНК-оценка является состоятельной и BLUE в больших выборках, если регрессор  $X$  является строго экзогенным, но не является состоятельной, если  $X$  всего лишь экзогенен (в прошлом и настоящем). С точки зрения математики, ОМНК основан на матричной алгебре, поэтому его рассмотрение откладывается до раздела 18.6.

### **Модель с распределенными лагами с дополнительными запаздываниями и AR( $p$ )-ошибками**

Предшествующее обсуждение модели регрессии с распределенными лагами, описываемой уравнениями (15.18) и (15.19), которая включает одно запаздывание  $X$ , и ошибку в виде AR(1), распространяется на общую модель регрессии с распределенными лагами с несколькими запаздываниями и ошибкой в виде AR( $p$ ).

**Общая модель с распределенными лагами для случая авторегрессионных ошибок.** Общая модель регрессии с распределенными лагами с  $r$  лагами и ошибкой в виде AR( $p$ ) имеет вид:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \dots + \beta_{r+1} X_{t-r} + u_r, \quad (15.30)$$

$$u_r = \varphi_1 u_{t-1} + \varphi_2 u_{t-2} + \dots + \varphi_p u_{t-p} + \tilde{u}_t, \quad (15.31)$$

где  $\beta_1, \dots, \beta_{r+1}$  являются динамическими мультипликаторами и  $\varphi_1, \dots, \varphi_p$  являются авторегрессионными коэффициентами в регрессии ошибок. Если ошибки описываются AR( $p$ )-моделью, то  $\tilde{u}_t$  серийно не коррелированы.

Алгебраические преобразования, аналогичные тем, что привели к модели ADL в уравнении (15.21), показывают, что на основе уравнений (15.30) и (15.31)  $Y_t$  можно записать в форме ADL:

$$Y_t = \alpha_0 + \varphi_1 Y_{t-1} + \dots + \varphi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \dots + \delta_q X_{t-q} + \tilde{u}_t, \quad (15.32)$$

где  $q=r+p$  и  $\delta_0, \delta_1, \dots, \delta_q$  являются функциями от коэффициентов  $\beta$  и  $\varphi$  в уравнениях (15.30) и (15.31). Это равносильно тому, что модель, описываемая уравнениями (15.30) и (15.31), может быть записана в квазидифференцированной форме:

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + \dots + \beta_{r+1} \tilde{X}_{t-r} + \tilde{u}_t, \quad (15.33)$$

где  $\tilde{Y}_t = Y_t - \varphi_1 Y_{t-1} - \dots - \varphi_p Y_{t-p}$  и  $\tilde{X}_t = X_t - \varphi_1 X_{t-1} - \dots - \varphi_p X_{t-p}$ .

**Условия состоятельности оценок коэффициентов ADL-модели.** Предшествующее обсуждение условий, необходимых для состоятельности оценки коэф-

фициентов в модели ADL для случая ошибок, имеющих структуру AR(1), распространяется на общую модель с ошибками в форме AR( $p$ ). Предположение о нулевом условном среднем для уравнения (15.33) говорит о том, что:

$$E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \dots) = 0. \quad (15.34)$$

Так как  $\tilde{u}_t = u_t - \varphi_1 u_{t-1} - \varphi_2 u_{t-2} - \dots - \varphi_p u_{t-p}$  и  $\tilde{X}_t = X_t - \varphi_1 X_{t-1} - \dots - \varphi_p X_{t-p}$ , это условие эквивалентно условию:

$$\begin{aligned} E(u_t | X_t, X_{t-1}, \dots) - \varphi_1 E(u_{t-1} | X_t, X_{t-1}, \dots) - \\ - \dots - \varphi_p E(u_{t-p} | X_t, X_{t-1}, \dots) = 0. \end{aligned} \quad (15.35)$$

Для того чтобы уравнение (15.35), выполнялось при различных  $\varphi_1, \dots, \varphi_p$ , необходимо, чтобы каждое условное среднее в уравнении (15.35) было равно нулю или, эквивалентно, должно выполняться равенство:

$$E(u_t | X_{t+p}, X_{t+p-1}, X_{t+p-2}, \dots) = 0. \quad (15.36)$$

Это условие не следует из экзогенности (в прошлом и настоящем)  $X_t$ , но оно вытекает из строгой экзогенности  $X_t$ . На самом деле в пределе при  $p$ , стремящемся к бесконечности (так что ошибка в модели регрессии с распределенными лагами следует модели авторегрессии бесконечного порядка), условие из уравнения (15.36) становится условием строгой экзогенности из вставки «Основные понятия 15.1».

**Оценка ADL-модели при помощи МНК.** Как и в модели регрессии с распределенными лагами с одним запаздыванием и ошибками в виде AR(1), динамические мультиплекторы могут быть оценены из МНК-оценок коэффициентов ADL-модели (15.32). Общие формулы аналогичны формулам (15.29), но являются более сложными, и лучше всего выражаются с использованием обозначений оператора запаздывания; эти формулы приведены в приложении 15.2. На практике современные программные пакеты, предназначенные для регрессионного анализа временных рядов, делают все эти вычисления за вас.

**Оценка при помощи ОМНК.** Кроме того, динамические мультиплекторы могут быть оценены при помощи (реализуемого) ОМНК. Для этого нужно оценить при помощи МНК коэффициенты в квазидифференцированной спецификации (15.33), используя оцененные квазиразности. Оцененные квазиразности можно вычислить с помощью предварительных оценок коэффициентов авторегрессии  $\varphi_1, \dots, \varphi_p$ , как и в случае AR(1). ОМНК-оценки асимптотически являются BLUE, в том же смысле что и для обсуждаемого ранее случая AR(1).

Схема оценки динамических мультиплекторов при наличии строгой экзогенности описана во вставке «Основные понятия 15.4».

**Что использовать: МНК или ОМНК?** Две возможные оценки – МНК-оценка коэффициентов ADL-модели и ОМНК-оценка коэффициентов модели с распределенными лагами – имеют как преимущества, так и недостатки.

Преимущество ADL-подхода заключается в том, что такой подход позволяет уменьшить количество параметров, необходимых для оценки динамических мультиплекторов, по сравнению с МНК-оценками модели с распределенными лагами. Например, оценка ADL-модели, заданной уравнением (15.27), повлечет за собой

необходимость оценки модели с распределенными лагами бесконечным числом запаздываний, что видно из уравнения (15.29). Рассматривая только модель с  $r$  распределенными лагами, мы в действительности рассматриваем лишь приближение модели с большим числом распределенных лагов, и ADL-модель является простым способом оценки такого большого числа запаздываний, используя только несколько неизвестных параметров. Таким образом, на практике можно было бы оценить модель ADL в виде (15.39) со значениями  $p$  и  $q$  много меньшими, чем значение  $r$ , необходимое для получения МНК-оценки коэффициентов модели с распределенными лагами (15.37). Другими словами, ADL-спецификация может дать компактную или экономную альтернативу длинной и сложной модели с распределенными лагами (см. приложение 15.2 для дополнительного обсуждения).

Преимуществом ОМНК-оценки является то, что для данной глубины запаздывания  $r$  в модели с распределенными лагами ОМНК-оценки коэффициентов при распределенных запаздываниях являются более эффективными, чем МНК-оценки, по крайней мере в больших выборках. Таким образом, на практике преимущество использования ADL-подхода возникает потому, что ADL-спецификация позволяет оценивать меньше параметров, чем при оценке с помощью ОМНК.

## ОСНОВНЫЕ ПОНЯТИЯ

### 15.4

#### Оценка динамических мультипликаторов при предположении строгой экзогенности

Модель с распределенными лагами с  $r$ -запаздываниями и ошибкой в форме AR( $p$ ) имеет вид:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \dots + \beta_{r+1} X_{t-r} + u_t \quad (15.37)$$

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \dots + \phi_p u_{t-p} + \tilde{u}_t. \quad (15.38)$$

Если случайная величина  $X_t$  строго экзогенна, то динамические мультипликаторы  $\beta_1, \dots, \beta_{r+1}$  могут быть оценены следующим образом. Сначала при помощи МНК оцениваются коэффициенты ADL-модели:

$$\begin{aligned} Y_t &= \alpha_0 + \varphi_1 Y_{t-1} + \dots + \varphi_p Y_{t-p} + \\ &+ \delta_0 X_t + \delta_1 X_{t-1} + \dots + \delta_q X_{t-q} + \tilde{u}_t, \end{aligned} \quad (15.39)$$

где  $q = r + p$ . А затем вычисляются динамические мультипликаторы, используя эконометрические программные пакеты. Либо, в качестве альтернативы, динамические мультипликаторы могут быть оценены как коэффициенты при распределенных лагах в модели (15.37) при помощи ОМНК.

## 15.6. Цены на апельсиновый сок и заморозки

В данном разделе методы регрессионного анализа временных рядов используются, чтобы получить дополнительную информацию из имеющихся у нас

данных о погоде во Флориде и ценах на апельсиновый сок. Во-первых, насколько длительным является эффект влияния заморозков на цены? Во-вторых, было ли это динамическое влияние стабильным или оно изменилось за 51 год, для которых у нас есть данные, и если да, то как?

Мы начнем наш анализ с оценки динамического влияния методом из раздела 15.3, то есть с МНК-оценки коэффициентов модели регрессии с распределенными лагами процентного изменения цены ( $\%ChgP_t$ ) от числа морозных дней в месяце ( $FDD_t$ ) и его запаздывающих значений. Для того чтобы оценки модели с распределенными лагами были состоятельными, должно выполняться условие экзогенности (в прошлом и настоящем)  $FDD$ . Как отмечалось в разделе 15.2, в данном случае это предположение является разумным. Люди не могут влиять на погоду, поэтому отношение к погоде, как будто ее изменения были случайным экспериментом, является целесообразным. Из-за того что  $FDD$  является экзогенной величиной, мы можем оценить динамическое влияние с помощью МНК-оценки коэффициентов модели с распределенными лагами (15.4) из вставки «Основные понятия 15.1».

Как обсуждалось в разделах 15.3 и 15.4, ошибка может быть серийно автокоррелирована в регрессии с распределенными лагами, поэтому важно использовать НАС-стандартные ошибки, которые устойчивы к наличию автокорреляции. Для начала, параметр усечения для стандартных ошибок Ньюи–Веста ( $t$  в обозначениях раздела 15.4) был выбран с использованием правила из уравнения (15.17): вследствие того что у нас имеется 612 ежемесячных наблюдений, в соответствии с этим правилом  $t = 0,75T^{1/3} = 0,75 \times 612^{1/3} = 6,37$ , но из-за того что  $t$  должно быть целым числом, мы должны округлить его до  $t=7$ ; чувствительность стандартных ошибок к такому выбору параметра усечения исследована ниже.

Результаты МНК-оценки регрессии с распределенными лагами  $\%ChgP_t$  на  $FDD_t, FDD_{t-1}, \dots, FDD_{t-18}$  приведены в столбце (1) таблицы 15.1. Коэффициенты этой регрессии (из которых только некоторые представлены в таблице) являются оценками динамического причинного влияния на изменение цены на апельсиновый сок (в процентах) в течение первых 18 месяцев после увеличения числа морозных дней в месяце на единицу. Например, если в месяце всего один морозный день, то, по оценкам, цена увеличится на 0,50 % в течение месяца, в котором этот морозный день был. Дальнейшее влияние на цены в последующие месяцы, начиная с этого морозного дня, становится меньше: через один месяц оцененное влияние заключается в увеличении цены еще на 0,17 %, а через два месяца оцененное влияние приведет к дополнительному росту цен на 0,07 %.  $R^2$  регрессии равен 0,12, что указывает на то, что большая часть ежемесячного изменения цен на апельсиновый сок не объясняется текущим и прошлыми значениями  $FDD$ .

Графики динамических мультипликаторов могли бы отразить эту информацию более эффективно, чем такие таблицы, как таблица 15.1. Динамические мультипликаторы из столбца (1) таблицы 15.1 представлены на рис 15.2а вместе с их 95 %-м доверительным интервалом, вычисленными как оцененный коэффициент  $\pm 1,96$  НАС-стандартных ошибок. После первоначального резкого

роста цен последующий рост цен меньше, хотя цены, по оценкам, будут немного расти в течение первых 6 месяцев после заморозков. Как можно видеть из рисунка 15.2а, в течение нескольких месяцев, кроме первого, динамические мультиплекторы не отличаются от нуля статистически значимо на уровне значимости 5 %, хотя они оцениваются как положительные вплоть до седьмого месяца.

В столбце (2) таблицы 15.1 приведены совокупные динамические мультиплекторы для данной спецификации, то есть накопленная сумма динамических мультиплекторов из столбца (1). Эти динамические мультиплекторы приведены на рисунке 15.2б вместе с 95 %-м доверительным интервалом. После первого месяца накопленное влияние одного морозного дня выражается в повышении цен на 0,67%; через два месяца цена, по оценкам, повышается на 0,74%, а через 6 месяцев цена, по оценкам, возрастает на 0,90%. Как видно из рисунка 15.2б, совокупные мультиплекторы растут вплоть до седьмого месяца, так как отдельные динамические мультиплекторы положительны в течение первых семи месяцев. Восьмом месяце динамический мультиплектор отрицателен, поэтому цена на апельсиновый сок начинает постепенно падать. После 18 месяцев совокупный рост цен составляет всего 0,37%, то есть долгосрочный совокупный динамический мультиплектор равен только 0,37%. Полученный долгосрочный совокупный динамический мультиплектор не является статистически значимым даже на уровне значимости 10 %.

Таблица 15.1

**Динамическое влияние числа морозных дней (*FDD*) на цену апельсинового сока:  
выборочные оценки динамических мультиплекторов  
и совокупных динамических мультиплекторов**

Глубина запаздывания	(1) Динамические мультиплекторы	(2) Совокупные мультиплекторы	(3) Совокупные мультиплекторы	(4) Совокупные мультиплекторы
0	0,50 (0,14)	0,50 (0,14)	0,50 (0,14)	0,51 (0,15)
1	0,17 (0,09)	0,67 (0,14)	0,67 (0,13)	0,70 (0,15)
2	0,07 (0,06)	0,74 (0,17)	0,74 (0,16)	0,76 (0,18)
3	0,07 (0,04)	0,81 (0,18)	0,81 (0,18)	0,84 (0,19)
4	0,02 (0,03)	0,84 (0,19)	0,84 (0,19)	0,87 (0,20)
5	0,03 (0,03)	0,87 (0,19)	0,87 (0,19)	0,89 (0,20)
6	0,03 (0,05)	0,90 (0,20)	0,90 (0,21)	0,91 (0,21)
.				
.				
.				

## Глава 15. Оценка динамического причинного влияния

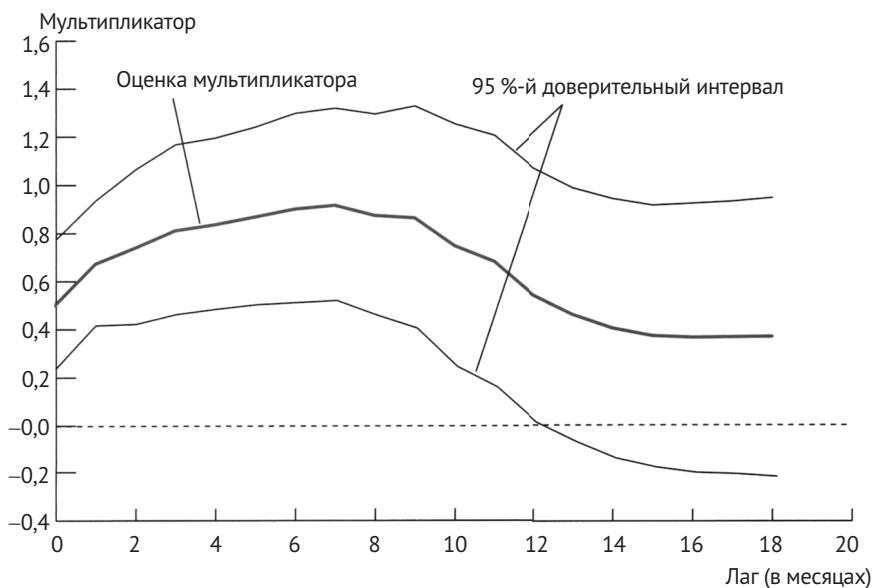
Окончание таблицы 15.1

Глубина запаздывания	(1) Динамические мультипликаторы	(2) Совокупные мультипликаторы	(3) Совокупные мультипликаторы	(4) Совокупные мультипликаторы
12 . . .	-0,14 (0,08)	0,54 (0,27)	0,54 (0,28)	0,54 (0,28)
18	0,00 (0,02)	0,37 (0,30)	0,37 (0,31)	0,37 (0,30)
Месячные индикаторы	Нет	Нет	Нет	Да $F=1,01$ $(p=0,43)$
Параметр усечения НАС-стандартных ошибок ( $m$ )	7	7	14	7

*Примечание.* Все регрессии оценивались с помощью МНК, используя ежемесячные данные (описанные в приложении 15.1) с января 1950 до декабря 2000 года, с общим числом ежемесячных наблюдений  $T=612$ . Зависимой переменной является ежемесячное процентное изменение цены апельсинового сока ( $\%ChgP$ ). Регрессия (1) представляет собой регрессию с распределенными лагами от ежемесячного числа морозных дней и их запаздываний, то есть  $FDD_t, FDD_{t-1}, \dots, FDD_{t-18}$ , а приведенные коэффициенты являются МНК-оценкой динамических мультипликаторов. Совокупные мультипликаторы – это накопленная сумма оцененных динамических мультипликаторов. Все регрессии включают константу, которая пока не приводится. НАС-стандартные ошибки Ньюи–Веста, которые вычисляются с использованием параметра усечения, указанного в последней строке, приведены в скобках.



(a) Оцененные динамические мультипликаторы и 95 %-й доверительный интервал



(б) Оцененные совокупные динамические мультиликаторы и 95 %-й доверительный интервал

**Рисунок 15.2. Динамическое влияние числа морозных дней (FDD) на цену апельсинового сока**

Оцененные динамические мультиликаторы показывают, что заморозки приводят к немедленному росту цен. Будущий рост цен намного меньше, чем первоначальный эффект воздействия. Совокупный мультиликатор показывает, что заморозки оказывают устойчивое влияние на уровень цен на апельсиновый сок с пиком цен через семь месяцев после заморозков.

**Анализ чувствительности.** Как и в любом эмпирическом приложении, здесь важно проверить, являются ли полученные результаты чувствительными к изменениям в деталях эмпирического анализа. Мы рассмотрим три аспекта данного анализа: чувствительность к вычислению НАС-стандартных ошибок; альтернативные спецификации, которые позволяют исследовать потенциальное смещение из-за пропущенных переменных и анализ стабильности во времени оцененных множителей.

Во-первых, мы проверим, все ли стандартные ошибки, приведенные во втором столбце таблицы 15.1, чувствительны к различному выбору НАС-параметра усечения  $t$ . В столбце (3) представлены результаты для  $t = 14$ , то есть для значения параметра усечения в 2 раза больше используемого в столбце (2). Спецификация регрессии такая же, как и в столбце (2), так что оценки коэффициентов и динамических мультиликаторов одинаковы; отличаются только стандартные ошибки, но, как видно, не намного. Мы заключаем, что результаты нечувствительны к изменениям в НАС-параметра усечения.

Во-вторых, мы исследуем возможные источники смещения из-за пропущенных переменных. Заморозки во Флориде не случайным образом распределены в течение года, а, как правило, происходят в зимнее время (конечно). Если спрос на апельсиновый сок является сезонным (спрос на апельсиновый сок больше зимой, чем летом), то сезонный спрос на апельсиновый сок

может быть коррелирован с  $FDD$ , что приводит к смещению из-за пропущенной переменной. Количество апельсинов, проданных для изготовления сока, является эндогенным: цены и количество одновременно определяются спросом и предложением. Таким образом, как обсуждалось в разделе 9.2, включение количества привело бы к смещению одновременных уравнений. Тем не менее сезонная составляющая спроса может быть учтена включением сезонных переменных в качестве регрессоров. По этой причине спецификация в столбце (4) таблицы 15.1 включает 11 бинарных переменных, одна из которых является индикатором января, другая указывает на февраль и так далее (как обычно одна бинарная переменная должна быть опущена, чтобы предотвратить совершенную мультиколлинеарность с константой). Эти ежемесячные бинарные переменные не являются совместно статистически значимыми на уровне значимости 10 % ( $p=0,43$ ), а оцененные совокупные динамические мультипликаторы по сути те же, что и для спецификации, в которой ежемесячные индикаторы исключены. Таким образом, сезонные колебания спроса не являются важным источником смещения из-за пропущенной переменной.

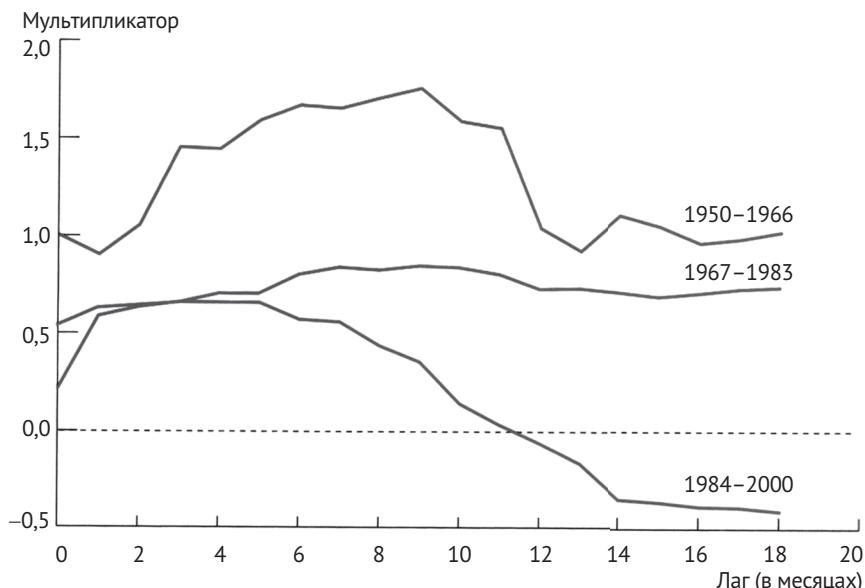
**Стабильны ли динамические мультипликаторы во времени?**<sup>1</sup> Для оценки устойчивости динамических мультипликаторов мы должны проверить, являются ли коэффициенты регрессии с распределенными лагами стабильными во времени. Из-за того что мы не знаем конкретную дату структурного сдвига, для проверки нестабильности коэффициентов регрессии мы используем статистику – отношение правдоподобия Куандта (QLR) (вставка «Основные понятия 14.9»). QLR-статистика (с 15 %-м отсечением и НАС-оценкой дисперсии), вычисленная для регрессии из столбца (1) в предположении наличия структурного сдвига во всех коэффициентах, имеет значение, равное 21,19, с  $q=20$  степенями свободы (коэффициенты при переменной  $FDD$ , ее 18 запаздываний и константа). 1 %-е критическое значение в таблице 14.6 равно 2,43, поэтому QLR-статистика отвергает гипотезу на 1 %-м уровне значимости. Оцененные QLR-регressии содержат по 40 регрессоров, что является большим числом; пересчитав регрессии только для шести лагов (так, чтобы было 16 регрессоров и  $q=8$ ), также отвергаем гипотезу на уровне значимости 1 %. Таким образом, гипотеза о том, что динамические мультипликаторы стабильны, отвергается на 1 %-м уровне значимости.

Одним из способов увидеть, как динамические мультипликаторы менялись во времени, является их расчет на различных подвыборках. На рисунке 15.3 приведены оценки совокупного динамического мультипликатора для первой трети (1950–1966), средней трети (1967–1983) и последней трети (1984–2000) выборки, вычисленные оценкой отдельных регрессий на каждой подвыборке. Результаты оценок интересны и достойны внимания. В 1950-х и в начале 1960-х годов число морозных дней оказывало сильное и устойчивое влияние на цены. Влияние числа морозных дней на цены уменьшилось в 1970-х годах, хотя

<sup>1</sup> Обсуждение стабильности в этом подразделе опирается на материал из раздела 14.7 и может быть пропущено, если этот материал не был изучен.

и осталось весьма устойчивым. В конце 1980-х и в 1990-х годах краткосрочное влияние числа морозных дней было таким же, как и в 1970-х, но оно стало гораздо менее устойчивым и было существенно меньшим через год. Эти оценки показывают, что динамическое причинное влияние заморозков во Флориде на цены апельсинового сока стало меньше и менее устойчивым во второй половине XX века. Во вставке «Миграция апельсиновых деревьев» обсуждается одно из возможных объяснений неустойчивости динамического причинного влияния.

**ADL-модели и ОМНК-оценки.** Как отмечалось в разделе 15.5, если ошибка в модели регрессии с распределенными лагами автокоррелирована и реgres sor  $FDD$  строго экзогенен, можно получить более эффективные оценки динамических мультипликаторов, чем МНК-оценки модели регрессии с распределенными лагами. Однако перед использованием любой ОМНК-оценки или оценки на основе модели ADL необходимо рассмотреть вопрос о том, является ли  $FDD$  на самом деле строго экзогенным. Действительно, люди не могут влиять на погоду, но значит ли это, что погода *строго* экзогенна? Имеет ли ошибка  $u_t$  в модели регрессии с распределенными лагами нулевое условное среднее относительно прошлых, настоящего и будущих значений  $FDD$ ?



**Рисунок 15.3. Оцененные кумулятивные динамические мультипликаторы для различных подвыборок**

Динамическое влияние заморозков на цену апельсинового сока существенно изменилось во второй половине XX века. Заморозки оказывали большее влияние на цены в течение 1950–1966 годов по сравнению с более поздним периодом, и влияние заморозков в 1984–2000 годах было менее устойчивым, чем раньше.

Случайная ошибка в теоретическом аналоге модели регрессии с распределенными лагами из столбца (1) таблицы 15.1 представляет собой расхождение

между ценой и ее предсказанным значением в генеральной совокупности, основанным на информации о погоде в течение последних 18 месяцев. Это несоответствие может возникнуть по многим причинам, одной из которых является то, что трейдеры используют прогнозы погоды в Орландо. Например, если прогнозируется очень холодная зима, то трейдеры будут включать это в цену, так что цена будет выше ее предсказанного на основе теоретической регрессии значения, то есть ошибка будет положительна. Если этот прогноз является точным, то будущая погода действительно будет холоднее, чем обычно. Таким образом, количество будущих морозных дней будет положительным ( $X_{t+1} > 0$ ), когда текущая цена необычайно высока ( $u_t > 0$ ), поэтому  $\text{corr}(X_{t+1}, u_t)$  положительна. Проще говоря, несмотря на то что трейдеры, торгующие фьючерсами на апельсиновый сок, не могут влиять на погоду, они могут (и делают это) предсказать ее (см. вставку). Следовательно, ошибка в регрессии цены от погоды коррелирует с будущей погодой. Другими словами, переменная *FDD* является экзогенной, но если наше рассуждение верно, то она не является строго экзогенной, и ОМНК- и ADL-оценки не будут состоятельными оценками динамических мультипликаторов. Поэтому эти оценки не используются в данном примере.



### **Миграция апельсиновых деревьев**

Почему динамические мультипликаторы, изображенные на рисунке 15.3, изменяются во времени? Одним из возможных объяснений являются рыночные изменения, а другим – то, что деревья начинают расти все южнее.

Как следует из данных Управления по цитрусовым Флориды, несколько заморозков в 1980-х годах, которые видны на рисунке 15.1 в, вынудили производителей цитрусовых искать более теплый климат. Как показано на рисунке 15.4, площадь земель (в акрах) под апельсиновыми деревьями в более подверженных заморозкам северных и западных районах снизилась с 232 тыс. акров в 1981 году до 53 тыс. акров в 1985 году, и, соответственно, в южных и центральных районах площадь земель, на которых выращиваются апельсиновые деревья, возросла с 413 тыс. акров в 1985 году до 588 тыс. акров в 1993 году. Поскольку апельсиновые рощи перемещаются на юг, а северные заморозки уничтожают и без того малую долю урожая, то, как следует из графиков динамических мультиплексоров на рисунке 15.3, цены становятся менее чувствительными к температуре в городах, расположенных севернее Орланда.

Апельсиновые деревья действительно сами по себе не могут перемещаться, но миграция апельсиновых рощ на юг придает новый смысл термину «нестационарность»<sup>1</sup>.

<sup>1</sup> Мы благодарны профессору Джеймсу Коббу из Университета штата Флорида (James Cobbe, Florida State University) за рассказ о перемещении апельсиновых рощ на юг.

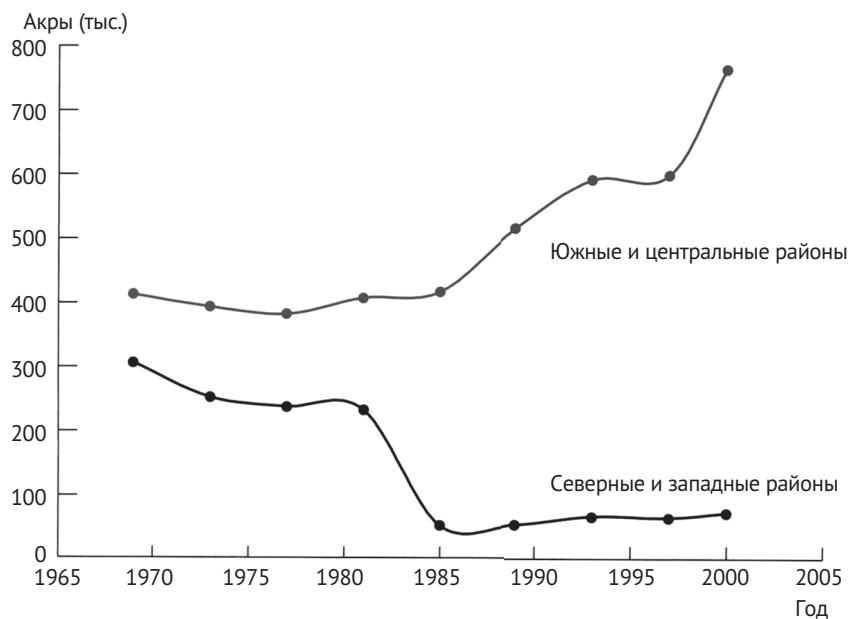


Рисунок 15.4. Площадь апельсиновых рощ в регионах Флориды

◇ ◇ ◇

## 15.7. Является ли экзогенность правдоподобным условием? Некоторые примеры

Как и в регрессии для межобъектных данных, интерпретация коэффициентов регрессии с распределенными лагами как причинного динамического влияния зависит от предположения о том, что регрессор  $X$  является экзогенным. Если  $X_t$  или его запаздывающие значения коррелируют с  $u_t$ , то условное среднее  $u_t$  будет зависеть от  $X_t$  или его лагов, и в этом случае  $X$  не будет экзогенным (в прошлом и настоящем). Регрессоры могут быть коррелированы с ошибкой по нескольким причинам, но для случая экономических временных рядов особенно важной проблемой является наличие одновременной причинности, которая (как обсуждается в разделах 9.2 и 12.1) приводит к эндогенности регрессоров. В разделе 15.6 мы обсуждали в деталях предположения об экзогенности и строгой экзогенности числа морозных дней. В этом разделе мы рассмотрим предположение об экзогенности в четырех других экономических приложениях.

### Доходы в США и австралийский экспорт

Соединенные Штаты Америки являются важным источником спроса на австралийский экспорт. Точнее, вопрос о том, насколько чувствительным является австра-

лийский экспорт к колебаниям совокупных доходов в США, мог бы быть исследован при помощи оценки регрессии австралийского экспорта в Соединенные Штаты на показатель доходов США. Строго говоря, из-за того что мировая экономика сильно интегрирована, существует одновременная причинность в этом соотношении: снижение австралийского экспорта снижает австралийские доходы, что снижает спрос на импорт из Соединенных Штатов и доходы в США. Однако с практической точки зрения этот эффект очень мал, потому что австралийская экономика намного меньше, чем экономика США. Так что доходы в США можно рассматривать как экзогенные в этой регрессии, что выглядит достаточно правдоподобно.

В противоположность этому в регрессии экспорта из стран Европейского союза в Соединенные Штаты на доходы в США, аналогичный аргумент в пользу того, чтобы считать доходы в США экзогенными, менее убедителен, потому что спрос со стороны граждан Европейского союза на американский экспорт составляет значительную долю общего спроса на американский экспорт. Таким образом, снижение американского спроса на европейский экспорт приведет к снижению доходов в ЕС, что в свою очередь уменьшит спрос на американский экспорт и, таким образом, уменьшит доходы в США. Из-за этих связей в международной торговле экспорт ЕС в Соединенные Штаты и доходы в США определяются одновременно, так что в рассматриваемой регрессии доходы в США, возможно, не являются экзогенным показателем. Этот пример иллюстрирует более общую точку зрения, заключающуюся в том, что решение вопроса о том, является ли переменная экзогенной, зависит от контекста: то, что доходы в США являются экзогенным показателем, выглядит правдоподобным в регрессии, объясняющей австралийский экспорт, но не в регрессии, объясняющей экспорт стран Европейского союза.



### ***СРОЧНАЯ НОВОСТЬ: посетители Disney World дрожат из-за продавцов***

Несмотря на то что погода в Disney World в Орландо, штат Флорида, как правило, приятная, время от времени там случаются похолодания. Если вы приехали в Disney World в зимний вечер, должны ли вы захватить с собой теплое пальто? Некоторые люди могут узнать прогноз погоды по телевизору, но те, кто в курсе, могут сделать лучше: проверить цену закрытия на фьючерсы на апельсиновый сок на Нью-Йоркской бирже в этот день!

Финансовый экономист Ричард Ролл (Richard Roll) провел детальное изучение взаимосвязи между ценами на апельсиновый сок и погодой. Ролл (1984) исследовал влияние холодной погоды в Орландо на цены, а также он изучил «эффект» влияния погоды на изменения в ценах фьючерсных контрактов на апельсиновые соки (контрактов на покупку замороженного концентратра апельсинового сока в определенный день в будущем). Ролл использовал ежедневные данные с 1975 по 1981 год о ценах на фьючерсные контракты на апельсиновые соки, которые торгуются на Нью-Йоркской хлопковой бирже, и о дневных иочных температурах

в Орландо. Он обнаружил, что рост цен фьючерсных контрактов в течение торгового дня в Нью-Йорке предсказывал холодную погоду, в частности заморозки в ближайшую ночь в Орландо. В самом деле рынок оказался настолько эффективным в предсказании холодной погоды во Флориде, что рост цен в течение торгового дня фактически предсказывал ошибки официального прогноза погоды в эту ночь.

Исследование Ролла интересно еще и благодаря тому, что он не обнаружил: несмотря на то что его подробные данные о погоде объясняли некоторые изменения ежедневных цен фьючерсных контрактов на апельсиновые соки, большая часть ежедневных изменений в ценах апельсиновых соков оставалась необъясненной. Поэтому он предположил, что рынок фьючерсов контрактов на апельсиновые соки показывает «избыточную волатильность», то есть волатильность, превышающую ту, которая может быть объяснена движениями основных объясняющих факторов. Понимание того, почему (и если) существует избыточная волатильность на финансовых рынках, является в настоящее время важной частью исследований в области финансовой экономики.

Исследование Ролла также иллюстрирует различие между прогнозированием и оценкой динамического причинного влияния. Изменение цен на рынке фьючерсов на апельсиновые соки является полезным фактором для предсказания холодной погоды, но это не означает, что трейдеры обладают достаточной силой, чтобы заставить температуру снизиться. Посетители Disney World могут дрожать после того, как цены на фьючерсные контракты на апельсиновые соки вырастут, но они не дрожат из-за роста цен, если только, конечно, они не связаны с рынком фьючерсных контрактов на апельсиновые соки.



### ***Цены на нефть и инфляция***

С тех пор как произошел рост цен на нефть в 1970-х годах, макроэкономистам интересно оценить динамическое влияние, которое оказывает рост мировых цен на сырую нефть, на уровень инфляции в США. Вследствие того что цены на нефть устанавливаются на мировых рынках в значительной степени иностранными нефтедобывающими странами, на первый взгляд может показаться, что цены на нефть являются экзогенными. Но цены на нефть не похожи на погоду: члены ОПЕК выбирают уровень добычи нефти стратегически, принимая во внимание многие факторы, в том числе состояние мировой экономики. Цены на нефть являются эндогенными в той степени, в которой они (или объемы добычи) устанавливаются на основе оценки текущих и будущих экономических условий в мире, в том числе инфляции в США.

### ***Монетарная политика и инфляция***

Центральным банкам, отвечающим за денежно-кредитную политику, необходимо знать, как денежно-кредитная политика влияет на инфляцию. Из-за того что главным инструментом денежно-кредитной политики является краткосроч-

ная процентная ставка («краткосрочная ставка»), им нужно знать, какое динамическое причинное влияние оказывает изменение краткосрочной ставки на инфляцию. Несмотря на то что краткосрочная ставка определяется центральным банком, она не устанавливается центральными банкирами случайно (как это было бы в идеальном случайному эксперименте), а устанавливается эндогенно: центральный банк определяет краткосрочную ставку на основе оценки текущего и будущего состояния экономики, особенно обращая внимание на текущие и будущие темпы инфляции. Уровень инфляции, в свою очередь, зависит от процентной ставки (более высокие процентные ставки снижают совокупный спрос), но процентная ставка зависит от темпов инфляции, ее прошлых значений и ее (ожидаемого) будущего значения. Таким образом, краткосрочная ставка является эндогенной, и причинное динамическое влияние изменения краткосрочной ставки на будущую инфляцию не может быть состоятельно оценено МНК-регрессией уровня инфляции от текущего и прошлых значений процентной ставки.

### **Кривая Филлипса**

Кривая Филлипса, исследованная в главе 14, представляет собой регрессию изменения темпов инфляции от запаздывающих значений изменения инфляции и лагов уровня безработицы. Поскольку лаги уровня безработицы имели место в прошлом, то поначалу можно было бы думать, что не может существовать обратного влияния текущих темпов инфляции на прошлые значения уровня безработицы, поэтому прошлые значения уровня безработицы могут рассматриваться как экзогенные. Но прошлые значения уровня безработицы не являются случайным результатом эксперимента; напротив, они определяются одновременно с прошлыми значениями инфляции. Поскольку инфляция и безработица определяются одновременно, прочие факторы, определяющие инфляцию и содержащиеся в  $u_t$ , коррелируют с прошлыми значениями уровня безработицы, то есть уровень безработицы не является экзогенным. Отсюда следует, что уровень безработицы не является строго экзогенным, так что динамические мультиплекторы, вычисленные с использованием эмпирической кривой Филлипса [например, ADL-модели (14.17)], не являются состоятельными оценками динамического причинного влияния изменения уровня безработицы на инфляцию.

## **15.8. Заключение**

Временные ряды дают возможность оценить временную траекторию эффекта влияния изменения  $X$  на  $Y$ , то есть динамическое причинное влияние изменения  $X$  на  $Y$ . Однако чтобы оценить динамическое причинное влияние при помощи модели регрессии с распределенными лагами, переменная  $X$  должна быть экзогенной, как это было бы, если бы она была получена посредством идеального случайногого эксперимента. Если  $X$  не только экзогенна, но строго экзогенна, то динамическое причинное влияние можно оценить с помощью авторегрессионной модели с распределенными лагами или с помощью ОМНК.

В некоторых приложениях, таких как оценка динамического причинного влияния заморозков во Флориде на цену апельсинового сока, довольно убедительно показано, что можно утверждать, что регрессор (число морозных дней) является экзогенным, и, таким образом, динамическое причинное влияние может быть оценено с помощью МНК-оценки коэффициентов при распределенных лагах. Однако даже в этом приложении экономическая теория предполагает, что погода не является строго экзогенной, следовательно, ADL- или ОМНК-методы не подходят. Более того, во многих ситуациях, представляющих интерес для эконометристов, присутствует одновременная причинность, так что в этих спецификациях регрессоры не являются экзогенными, ни строго или ни как-то иначе. Выяснение того, является ли регрессор экзогенным (строго экзогенным), в конечном счете требует объединения экономической теории, институциональных знаний и тщательного осмысления.

## **Выходы**

1. Динамическое причинное влияние во временных рядах определяется в контексте случайного эксперимента, где один и тот же объект подвергается различным случайным воздействиям в различное время. Коэффициенты в модели регрессии с распределенными лагами  $Y$  от  $X$  и его лагов можно интерпретировать как динамическое причинное влияние, когда временная траектория  $X$  определяется случайно и независимо от других факторов, которые влияют на  $Y$ .

2. Случайная величина  $X$  является экзогенной (в прошлом и настоящем), если условное математическое ожидание ошибки  $\mu_t$  в модели регрессии с распределенными лагами  $Y$  от текущих и прошлых значений  $X$  не зависит от текущих и прошлых значений  $X$ . Если, кроме того, условное среднее  $\mu_t$  не зависит от будущих значений  $X$ , то  $X$  является строго экзогенной.

3. Если  $X$  является экзогенной, то МНК-оценки коэффициентов в модели регрессии с распределенными лагами  $Y$  от текущих и прошлых значений  $X$  являются состоятельными оценками динамического причинного влияния. В общем случае ошибки  $\mu_t$  в этой регрессии автокоррелированы, так что обычные стандартные ошибки приводят к некорректным выводам и вместо них следует использовать НАС-стандартные ошибки.

4. Если  $X$  строго экзогенна, то динамические мультиплекторы могут быть оценены с помощью МНК-оценки модели ADL или с помощью ОМНК.

5. Экзогенность является сильным предположением, которое часто не выполняется в экономических временных рядах из-за наличия одновременной причинности; а предположение о строгой экзогенности является еще более сильным.

## **Основные понятия**

Динамическое причинное влияние (с. 615).

Модель регрессии с распределенными лагами (с. 622).

Экзогенность (с. 623).

- Строгая экзогенность (с. 623).  
 Динамический мультипликатор (с. 627).  
 Эффект воздействия (импульсный эффект) (с. 628).  
 Совокупный динамический мультипликатор (с. 628).  
 Долгосрочный совокупный динамический мультипликатор (с. 628).  
 Устойчивая к гетероскедастичности и автокорреляции стандартная ошибка (НАС) (с. 631).  
 Параметр усечения (с. 632).  
 Ширина окна (с. 632).  
 Оценка дисперсии Ньюи–Веста (с. 632).  
 Обобщенный метод наименьших квадратов (ОМНК) (с. 633).  
 Квазиразность (с. 635).  
 Недоступная (нереализуемая) ОМНК-оценка (с. 638).  
 Доступная (реализуемая) ОМНК-оценка (с. 638).

### **Вопросы для повторения и закрепления основных понятий**

- 15.1. В 1970-х годах обычной практикой была оценка модели с распределенными лагами, описывающей изменения номинального валового внутреннего продукта ( $Y$ ) в зависимости от текущего и прошлых изменений денежного предложения ( $X$ ). При каких предположениях такая регрессия будет оценивать причинное влияние денежного предложения на номинальный ВВП? Насколько вероятно, что эти предположения будут верны для современной экономики, такой как экономика Соединенных Штатов?
- 15.2. Предположим, что  $X$  строго экзогенная случайная величина. Исследователь оценивает модель ADL(1,1), вычисляет остаток регрессии и обнаруживает сильную серийную автокорреляцию остатков. Следует ли исследователю оценить новую модель ADL с дополнительными лагами или он просто может использовать НАС-стандартные ошибки для коэффициентов оцененной им ADL(1,1)-модели?
- 15.3. Предположим, что оценена модель с распределенными лагами, где вместо  $Y_t$  зависимой переменной является  $\Delta Y_t$ . Объясните, как вы вычислили бы динамические мультипликаторы, характеризующие влияние  $X_t$  на  $Y_t$ .
- 15.4. Предположим, что вы добавили  $FDD_{t+1}$  в качестве дополнительного regressора в уравнение (15.2). Если  $FDD$  строго экзогенен, ожидаете ли вы, что коэффициент при  $FDD_{t+1}$  будет равен нулю или нет? Изменится ли ваш ответ, если  $FDD$  является экзогенным, но не строго экзогенным?

### **Упражнения**

- 15.1. Рост цен на нефть считается причиной нескольких рецессий в развитых странах. Для того чтобы дать количественную оценку влияния цен на нефть на реальную экономическую активность, исследователи оценивали регрессии, похожие на обсуждаемые в данной главе. Пусть  $GDP_t$  обозначает значение квартального валового внутреннего продукта США

и пусть  $Y_t = 100 \ln(GTP_t/GTP_{t-1})$  – это процентное изменение ВВП за текущий квартал. Джеймс Гамильтон, эконометрист и макроэкономист, предположил, что цены на нефть негативно влияют на экономику только тогда, когда они прыгают сильно выше их значений в недавнем прошлом. Пусть  $O_t$  равна максимуму между нулем и разностью в процентных пунктах между ценами на нефть в момент времени  $t$  и их максимальным значением в течение предыдущего года. Оценки модели регрессии с распределенными лагами, связывающей  $Y_t$  и  $O_t$ , на интервале с I квартала 1955 года по IV квартал года имеют вид:

$$\widehat{Y}_t = 1,0 - 0,055O_t - 0,026O_{t-1} - 0,031O_{t-2} - 0,109O_{t-3} - 0,128O_{t-4} + \\ + 0,008O_{t-5} + 0,025O_{t-6} - 0,019O_{t-7} + 0,067O_{t-8}.$$

- a) Предположим, что скачок цены на нефть равен 25 % относительно их предыдущего пикового значения и это есть новый пик (так что  $O_t = 25$  и  $O_{t+1} = O_{t+2} = \dots = 0$ ). Чему равен предсказанный эффект влияния этого скачка на прирост ВВП для каждого квартала в течение ближайших двух лет?
- б) Постройте 95 %-е доверительные интервалы для ваших ответов в пункте (a).
- в) Чему равно предсказанное накопленное изменение прироста ВВП за восемь кварталов?
- г) НАС F-статистика для проверки гипотезы о том, что коэффициенты при  $O_t$  и ее запаздываниях равны нулю, составляет 3,49. Отличаются ли все эти коэффициенты значимо от нуля?

15.2. Макроэкономисты также заметили, что изменения процентных ставок сопровождают скачки цен на нефть. Пусть  $R_t$  обозначает процентную ставку по трехмесячным казначейским векселям (в процентных пунктах в годовом исчислении). Оценки модели регрессии с распределенными лагами, связывающей изменения  $R_t(\Delta R_t)$  и  $O_t$ , на интервале с I квартала 1955 года по IV квартал года, имеют вид:

$$\widehat{\Delta R}_t = 0,07 + 0,062O_t + 0,048O_{t-1} - 0,014O_{t-2} - 0,086O_{t-3} - \\ - 0,000O_{t-4} + 0,023O_{t-5} - 0,010O_{t-6} - 0,100O_{t-7} - 0,014O_{t-8}$$

- а) Предположим, что произошел скачок цен на нефть, равный 25 %, относительно предыдущего пикового значения, и это есть новый пик (так что  $O_t = 25$  и  $O_{t+1} = O_{t+2} = \dots = 0$ ). Чему равно предсказанное изменение процентной ставки для каждого квартала в течение ближайших двух лет?
- б) Постройте 95 %-е доверительные интервалы для ваших ответов в пункте (a).

- в) Каково влияние этого изменения цен на нефть на уровень процентных ставок в период  $t+8$ ? Как связан ваш ответ с кумулятивным мультиликатором?
- г) НАС F-статистика для проверки гипотезы о том, что коэффициенты при  $O_t$  и ее запаздываниях равны нулю, составляет 4,25. Отличаются ли все эти коэффициенты значимо от нуля?
- 15.3. Рассмотрим два случайных эксперимента. В эксперименте А цены на нефть устанавливаются случайным образом, и центральный банк реагирует на это в соответствии с правилами своей обычной политики в ответ на экономические условия, включая изменения цен на нефть. В эксперименте В цены на нефть устанавливаются случайным образом, а центральный банк сохраняет процентные ставки неизменными и, в частности, не реагирует на изменение цены нефти. В обоих экспериментах наблюдается рост ВВП. Теперь предположим, что цены на нефть являются экзогенными в регрессии в упражнении 15.1. К какому из экспериментов, А или В, относится динамическое причинное влияние, оцененное в упражнении 15.1?
- 15.4. Предположим, что цены на нефть строго экзогенны. Обсудите, как вы могли бы улучшить оценки динамических мультиликаторов в упражнении 15.1?
- 15.5. Выполните уравнение (15.7) из уравнения (15.4) и покажите, что  $\delta_0 = \beta_0$ ,  $\delta_1 = \beta_1$ ,  $\delta_2 = \beta_1 + \beta_2$ ,  $\delta_3 = \beta_1 + \beta_2 + \beta_3$  и т.д. (Подсказка: обратите внимание на то, что  $X_t = \Delta X_t + \Delta X_{t-1} + \dots + \Delta X_{t-p+1} + X_{t-p}$ )
- 15.6. Рассмотрим модель регрессии  $Y_t = \beta_0 + \beta_1 X_t + u_t$ , где  $u_t$  описывается стационарной AR(1)-моделью  $u_t = \varphi_1 u_{t-1} + \tilde{u}_t$  с ошибкой  $\tilde{u}_t$ , являющейся i.i.d. с нулевым средним и дисперсией  $\sigma_u^2$  и  $|\varphi_1| < 1$ ; регрессор  $X_t$  также описывается стационарной AR(1)-моделью  $X_t = \gamma_1 X_{t-1} + e_t$  с ошибкой  $e_t$ , являющейся i.i.d. с нулевым средним и дисперсией i.i.d. со средним 0 и дисперсией  $\sigma_e^2$  и  $|\gamma_1| < 1$ , и  $e_t$  не зависит от  $\tilde{u}_t$  для всех  $t$  и  $i$ .
- Покажите, что  $\text{var}(u_t) = \frac{\sigma_u^2}{1 - \varphi_1^2}$  и  $\text{var}(X_t) = \frac{\sigma_e^2}{1 - \gamma_1^2}$ .
  - Покажите, что  $\text{cov}(u_t, u_{t-j}) = \varphi_1^j \text{var}(u_t)$  и  $\text{cov}(X_t, X_{t-j}) = \gamma_1^j \text{var}(X_t)$ .
  - Покажите, что  $\text{corr}(u_t, u_{t-j}) = \varphi_1^j$  и  $\text{corr}(X_t, X_{t-j}) = \gamma_1^j$ .
  - Рассмотрим множители  $\sigma_v^2$  и  $f_t$  в уравнении (15.14).
    - Покажите, что  $\sigma_v^2 = \sigma_X^2 \sigma_u^2$ , где  $\sigma_X^2$  дисперсия  $X$  и  $\sigma_u^2$  дисперсия  $u$ .
    - Выполните выражение для  $f_\infty$ .
- 15.7. Рассмотрим модель регрессии  $Y_t = \beta_0 + \beta_1 X_t + u_t$ , где  $u_t$  описывается стационарной AR(1)-моделью  $u_t = \varphi_1 u_{t-1} + \tilde{u}_t$  с ошибкой  $\tilde{u}_t$ , являющейся i.i.d. с нулевым средним и дисперсией  $\sigma_u^2$  и  $|\varphi_1| < 1$ .
- Предположим, что  $X_t$  не зависит от  $\tilde{u}_t$  при всех  $t$  и  $j$ . Является ли случайная величина  $X_t$  экзогенной (в прошлом и настоящем)? Является ли  $X_t$  строго экзогенной (в прошлом, настоящем и будущем)?

- б) Предположим, что  $X_t = \tilde{u}_{t+1}$ . Является ли  $X_t$  экзогенной? Является ли  $X_t$  строго экзогенной?
- 15.8. Рассмотрим модель регрессии в упражнении 15.7 с  $X_t = \tilde{u}_{t+1}$ .
- Является ли МНК-оценка  $\beta_1$  состоятельной? Объясните.
  - Объясните, почему ОМНК-оценка  $\beta_1$  не является состоятельной?
  - Покажите, что нереализуемая ОМНК-оценка  $\hat{\beta}_1^{GLS} \xrightarrow{p} \beta_1 - \frac{\varphi_1}{1 + \varphi_1^2}$ .
- [Подсказка: используйте формулу смещения из-за пропущенных переменных (6.1) и примените ее к квазидифференцированному уравнению регрессии (15.23)].
- 15.9. Рассмотрим модель регрессии, содержащую только константу  $Y_t = \beta_0 + u_t$ , где  $u_t$  описывается стационарной AR(1)-моделью  $u_t = \varphi_1 u_{t-1} + \tilde{u}_t$  с ошибкой  $\tilde{u}_t$ , являющейся i.i.d. с нулевым средним и дисперсией  $\sigma_{\tilde{u}}^2$  и  $|\varphi_1| < 1$ .
- Покажите, что МНК-оценка  $\hat{\beta}_0 = T^{-1} \sum_{t=1}^T Y_t$ .
  - Покажите, что (недоступная) ОМНК-оценка равна  $\hat{\beta}_0^{GLS} = (1 - \varphi_1)^{-1} (T - 1)^{-1} \sum_{t=2}^{T-1} (Y_t - \varphi_1 Y_{t-1})$ . [Подсказка: ОМНК-оценка  $\beta_0$  равна  $(1 - \varphi_1)^{-1}$ , умноженная на МНК-оценку  $\alpha_0$  из уравнении (15.23). Почему?]
  - Покажите, что  $\hat{\beta}_0^{GLS}$  можно записать в таком виде:  $\hat{\beta}_0^{GLS} = (T - 1)^{-1} \sum_{t=2}^{T-1} Y_t + (1 - \varphi_1)^{-1} (T - 1)^{-1} (Y_T - \varphi_1 Y_1)$ . [Подсказка: перегруппируйте слагаемые в формуле в пункте (б).]
  - Выполните разность  $\hat{\beta}_0 - \hat{\beta}_0^{GLS}$  и обсудите, почему она, вероятно, будет небольшой при больших  $T$ .
- 15.10. Рассмотрим модель ADL:  $Y_t = 3,1 + 0,4Y_{t-1} + 2,0X_t - 0,8X_{t-1} + u_t$ , где регрессор  $X_t$  строго экзогенен.
- Выполните импульсный эффект  $X$  на  $Y$ .
  - Выполните первые пять динамических мультиплексоров.
  - Выполните первые пять совокупных мультиплексоров.
  - Выполните долгосрочный совокупный динамический мультиплексор.

### **Компьютерные упражнения**

- E15.1. В данном упражнении вы будете оценивать влияние цен на нефть на макроэкономическую активность с использованием ежемесячных данных по индексу промышленного производства ( $IP$ ) и ежемесячных данных для  $O_t$ , описанных в упражнении 15.1. Данные можно найти на сайте учебника [www.pearsonhighered.com/stock\\_watson](http://www.pearsonhighered.com/stock_watson) в файле USMacro\_Monthly.
- Вычислите ежемесячные темпы роста  $IP$ , выраженные в процентных пунктах,  $ip\_growth_t = 100 \times \ln(IP_t / IP_{t-1})$ . Чему равны среднее значение и стандартное отклонение  $ip\_growth$  на интервале с января 1952 по декабрь 2009 года?
  - Начертите график  $O_t$ . Почему многие значения  $O_t$  равны нулю? Почему отсутствуют отрицательные значения  $O_t$ ?

- в) Оцените модель регрессии с распределенными лагами  $ip\_growth$  на текущее и 18 запаздывающих значений  $O_t$ . Какое значение параметра усечения  $t$  для расчета НАС-стандартных ошибок вы выбрали? Почему?
- г) Являются ли коэффициенты при  $O_t$  и его лагах совместно статистически значимо отличающимися от нуля?
- д) Постройте графики, аналогичные графикам на рисунке 15.2, показывающие оцененные динамические мультиплекторы, совокупные мультиплекторы и 95 %-е доверительные интервалы. Прокомментируйте полученные значения мультиплекторов.
- е) Предположим, что высокий спрос в США (о чем свидетельствуют большие значения  $ip\_growth$ ) приводит к росту цен на нефть. Является ли  $O_t$  экзогенным? Выглядят ли оцененные мультиплекторы, изображенные на графиках в пункте (д), правдоподобными? Объясните.

E15.2. В файле с данными USMacro\_Monthly вы найдете данные о двух агрегатных временных рядах цен для США: индексе потребительских цен (ИПЦ, CPI) и дефляторе частных потребительских расходов (PCED). Эти временные ряды являются альтернативными мерами потребительских цен в Соединенных Штатах. ИПЦ характеризует цену на товары потребительской корзины, состав которой обновляется каждые 5–10 лет. PCED характеризует взвешенную цену потребительской корзины, состав которой меняется ежемесячно. Экономисты утверждают, что ИПЦ завышает инфляцию, поскольку она не учитывает замещение товаров, происходящее, если относительные цены изменяются. Если такое смещение из-за замещения важно, то средняя инфляция, рассчитанная по ИПЦ, должна быть систематически выше, чем инфляция, рассчитанная по PCED. Пусть  $\pi_t^{CPI} = 1200 \times \ln [CPI(t)/CPI(t-1)]$ ,  $\pi_t^{PCED} = 1200 \times \ln [PCED(t)/PCED(t-1)]$  и  $Y_t = \pi_t^{CPI} - \pi_t^{PCED}$ , таким образом,  $\pi_t^{CPI}$  – ежемесячный темп роста инфляции (измеряется в процентных пунктах в годовом исчислении), рассчитанный на основе индекса потребительских цен,  $\pi_t^{PCED}$  – ежемесячный темп роста инфляции, рассчитанный на основе PCED, и  $Y_t$  их разность. Используя данные на интервале с января 1970 по декабрь 2009 года выполните следующие упражнения:

- а) Вычислите выборочные средние  $\bar{\pi}_t^{CPI}$  и  $\bar{\pi}_t^{PCED}$ . Являются ли эти точечные оценки состоятельными оценками наличия экономически значимого смещения ИПЦ, возникающего из-за эффекта замещения?
- б) Вычислите выборочное среднее  $\bar{Y}_t$ . Объясните, почему оно численно равно разности средних из пункта (а)?
- в) Покажите, что математическое ожидание  $\bar{Y}_t$  в генеральной совокупности равно разности математических ожиданий двух инфляций в генеральной совокупности.
- г) Рассмотрим модель регрессии, содержащую только константу  $Y_t = \beta_0 + u_t$ . Покажите, что  $\beta_0 = E(Y)$ . Является ли ошибка  $u_t$  серийно коррелированной? Объясните.

- д) Постройте 95 %-й доверительный интервал для  $\beta_0$ . Какой параметр усечения  $t$  для расчета НАС-стандартной ошибки вы выбрали? Почему?
- е) Существуют ли статистически значимые доказательства того, что средний уровень инфляции, рассчитанной по ИПЦ, больше, чем уровень инфляции, рассчитанной по PCED?
- ж) Существуют ли признаки нестабильности  $\beta_0$ ? Проведите QLR-тест.

## Приложения

### **Приложение 15.1. База данных по апельсиновому соку**

Данные о ценах на апельсиновый сок используются для расчета сводного индекса цен производителей (PPI) и включаются как компонента замороженного апельсинового сока в группу продуктов питания. Данные собираются американским Бюро статистики труда (BLS<sup>1</sup>, временной ряд wriu02420301). Временной ряд цены на апельсиновый сок делился на общий индекс цен производителей произведенной продукции для корректировки общей инфляции цен. Временной ряд температуры морозных дней был построен на основе данных о ежедневных минимальных температурах, зарегистрированных в районе аэропорта г. Орландо, полученных из Национального управления по океану и атмосфере (NOAA<sup>2</sup>) Министерства торговли США. Временной ряд FDD был построен так, чтобы он и цены на апельсиновый сок были почти синхронизированы во времени. В частности, данные о ценах на замороженный апельсиновый сок собираются с помощью выборочного опроса производителей в середине каждого месяца, хотя точная дата меняется от месяца к месяцу. Соответственно, временной ряд FDD был построен так, чтобы количество морозных дней считалось с 11-го числа одного месяца по 10-е число следующего месяца; то есть FDD является максимумом между нулем и 32 °F<sup>3</sup> минус минимальная дневная температура, суммируемая по всем дням с 11 по 10-е число. Таким образом, %ChgP, за февраль есть процентное изменение реальной цены апельсинового сока с середины января до середины февраля и FDD, за февраль – количество морозных дней с 11 января по 10 февраля.

### **Приложение 15.2. ADL-модель и обобщенный метод наименьших квадратов в обозначениях лагового оператора**

В настоящем приложении рассматривается модель регрессии с распределенными лагами в обозначениях оператора запаздывания, выводятся ADL и ква-

---

<sup>1</sup> The Bureau of Labor Statistics (BLS). – Примеч. науч. ред. перевода.

<sup>2</sup> The National Oceanic and Atmospheric Administration (NOAA). – Примеч. науч. ред. перевода.

<sup>3</sup> 32 °F равно 0 °C. – Примеч. науч. ред. перевода.

зидифференцированное представление модели регрессии с распределенными лагами и обсуждаются условия, при которых модель ADL может иметь меньше параметров, чем исходная модель регрессии с распределенными лагами.

### **Распределенные лаги, ADL и квазидифференцированные модели в обозначениях лагового оператора**

Как определено в приложении 14.3, оператор запаздывания  $L$  обладает тем свойством, что  $L^j X_t = X_{t-j}$  и распределенный лаг  $\beta_1 X_t + \beta_2 X_{t-1} + \dots + \beta_r X_{t-r}$  может быть выражен через  $\beta(L) X_t$ , где  $\beta(L) = \sum_{j=1}^r \beta_j L^j$  и  $L^0 = 1$ . Таким образом, модель регрессии с распределенными лагами из вставки «Основные понятия 15.1» [уравнение (15.4)] можно записать в обозначениях лагового оператора:

$$Y_t = \beta_0 + \beta(L) X_t + u_t. \quad (15.40)$$

Если дополнительно к этому ошибка  $u_t$  является процессом  $AR(p)$ , то можно записать:

$$\varphi(L) u_t = \tilde{u}_t, \quad (15.41)$$

где  $\varphi(L) = \sum_{j=1}^p \varphi_j L^j = 1$ ,  $\varphi_0$  и  $\tilde{u}_t$  серийно некоррелирована [заметим, что  $\varphi_1, \dots, \varphi_p$ , как они определены здесь, равны значениям  $\varphi_1, \dots, \varphi_p$  в обозначениях уравнения (15.31), но взятым со знаком минус].

Чтобы вывести модель ADL, нужно умножить каждую часть выражения (15.40) на  $\varphi(L)$ , чтобы

$$\varphi(L) Y_t = \varphi(L) [\beta_0 + \beta(L) X_t + u_t] = \alpha_0 + \delta(L) X_t + \tilde{u}_t, \quad (15.42)$$

где

$$\alpha_0 = \varphi(1) \beta_0 \text{ и } \delta(L) = \varphi(L) \beta(L), \text{ где } \varphi(1) = \sum_{j=0}^p \varphi_j. \quad (15.43)$$

Для вывода квазидифференцированной модели заметим, что  $\varphi(L) \beta(L) X_t = \beta(L) \varphi(L) X_t = \beta(L) \tilde{X}_t$ , где  $\tilde{X}_t = \varphi(L) X_t$ . Таким образом, преобразовав уравнение (15.42), получаем:

$$\tilde{Y}_t = \alpha_0 + \beta(L) \tilde{X}_t + \tilde{u}_t, \quad (15.44)$$

где  $\tilde{Y}_t$  – квазиразность  $Y_t$ , то есть  $\tilde{Y}_t = \varphi(L) Y_t$ .

### **Оценки ADL и OMHK**

МНК-оценка коэффициентов модели ADL может быть получена МНК-оценкой уравнения (15.42). Коэффициенты исходной модели регрессии с распределенными лагами – это  $\beta(L)$ , которые равны  $\beta(L) = \delta(L)/\varphi(L)$  в терминах оцененных коэффициентов, то есть коэффициенты  $\beta(L)$  удовлетворяют ограничениям, подразумевающим  $\varphi(L)\beta(L) = \delta(L)$ . Таким образом, оценка динамических мультипликаторов на основе МНК-оценки коэффициентов модели ADL,  $\hat{\delta}(L)$  и  $\hat{\varphi}(L)$ , представляет:

$$\hat{\beta}^{ADL}(L) = \frac{\hat{\delta}(L)}{\hat{\varphi}(L)}. \quad (15.45)$$

Выражения для коэффициентов в уравнении (15.29) в тексте выводятся как частный случай уравнения (15.45) при  $r=1$  и  $p=1$ .

Доступная ОМНК-оценка вычисляется при помощи предварительной оценки  $\varphi(L)$  и вычисления оценок квазиразностей, оценивая  $\beta(L)$  в уравнении (15.44), используя эти оцененные квазиразности, а также (по желанию) совершая необходимое число итераций, пока оценки не сойдутся. Итеративная ОМНК-оценка является оценкой НМНК, вычисленной как НМНК-оценка ADL-модели из уравнения (15.42) в соответствии с нелинейными ограничениями на параметры, содержащимися в уравнении (15.43).

Как подчеркивается в обсуждении уравнения (15.36) в тексте, для использования любого из этих методов оценки недостаточно экзогенности (в прошлом и настоящем)  $X_t$ , поскольку экзогенность сама по себе не гарантирует, что выполняется условие (15.36). Однако если регрессор  $X_t$  является строго экзогенным, то условие (15.36) выполняется, и при выполнении предположений со 2 по 4 из вставки «Основные понятия 14.6», эти оценки являются состоятельными и асимптотически нормальными. Более того, обычные (устойчивые к гетероскедастичности в межъобъектных выборках) стандартные ошибки МНК дают надежную основу для статистических выводов.

**Уменьшение числа параметров при помощи ADL.** Предположим, что полином распределенных лагов  $\beta(L)$  может быть записан в виде отношения лаговых полиномов  $\theta_1(L)/\theta_2(L)$ , где  $\theta_1(L)$  и  $\theta_2(L)$  являются лаговыми полиномами более низкой степени. Тогда  $\varphi(L)\beta(L)$  в уравнении (15.43) есть  $\varphi(L)\beta(L) = \varphi(L)\theta_1(L)/\theta_2(L) = [\varphi(L)/\theta_2(L)]\theta_1(L)$ . Если  $\varphi(L) = \theta_2(L)$ , то  $\delta(L) = \varphi(L)\beta(L) = \theta_1(L)$ . Если степень  $\theta_1(L)$  мала, то число запаздываний  $X_t$  в модели ADL, равное  $q$ , может быть значительно меньше, чем  $r$ . Таким образом, в этих предположениях оценка модели ADL потенциально влечет за собой оценку гораздо меньшего числа параметров, чем исходная модель регрессии с распределенными лагами. Это означает, что модель ADL может быть более экономной (т.е., использовать меньшее количество неизвестных параметров), чем модель с распределенными лагами.

Но предположение о том, что  $\varphi(L)$  и  $\theta_2(L)$  одинаковы, кажется совпадением, которое вряд ли может произойти в реальных приложениях. Однако модель ADL способна описать большое число динамических множителей, используя всего несколько коэффициентов.

**ADL или ОМНК: смещение или дисперсия.** Хорошим способом выбора между оценкой динамических мультипликаторов сначала путем оценки модели ADL, а затем вычисления динамических мультипликаторов при помощи ее коэффициентов или, другим способом, путем оценки модели с распределенными лагами непосредственно с помощью ОМНК, является взгляд на принятие такого решения с точки зрения выбора между смещением и дисперсией. Оценка динамических мультипликаторов с использованием приближенной модели ADL

приводит к смещению, однако из-за небольшого числа коэффициентов дисперсия оценки динамических мультипликаторов может быть небольшой. Напротив, оценка содержащей много запаздываний модели с распределенными лагами, проведенная путем использования ОМНК, приводит к меньшему смещению; однако из-за того что оценивается очень много коэффициентов, их дисперсия может быть большей. Если приближение динамических мультиплликаторов при помощи ADL является хорошим, то смещение оцененных динамических мультиплликаторов будет мало, поэтому и подход ADL будет иметь меньшую дисперсию, чем ОМНК-подход с небольшим ростом смещения. По этой причине оценка модели ADL без ограничений с малым числом лагов  $Y$  и  $X$  является привлекательным способом приближения большого числа распределенных лагов при строго экзогенном  $X$ .

# **Глава 16. Модель регрессии временных рядов: дополнительные разделы**

В этой главе рассматривается ряд тем, касающихся модели регрессии временных рядов, начиная с прогнозирования. В главе 14 мы обсуждали вопросы прогнозирования одной переменной. Однако на практике может возникнуть необходимость прогнозирования двух или более переменных, таких как уровень инфляции и темп роста ВВП. В разделе 16.1 мы вводим модель, которую можно использовать для прогнозирования нескольких переменных — модель векторной авторегрессии (VAR-ы), в которой запаздывающие значения двух или более переменных используются для прогнозирования будущих значений этих переменных. В главе 14 также был рассмотрен принцип построения прогноза на один период (например на один квартал) вперед, но также важно уметь строить прогнозы на два, три или более периодов вперед. Соответственно, способы построения многошаговых прогнозов обсуждаются в разделе 16.2.

Разделы 16.3 и 16.4 вернут нас к стохастическим трендам, понятие о которых было введено в разделе 14.6. В разделе 16.3 вводятся дополнительные модели стохастического тренда и альтернативные тесты на единичные авторегрессионные корни. В разделе 16.4 вводится понятие коинтеграции, которое возникает в ситуации, когда две переменные имеют общий стохастический тренд, то есть когда каждая переменная содержит стохастический тренд, но взвешенная разность двух переменных — нет.

В некоторых временных рядах, особенно в финансовых, дисперсия изменяется во времени: иногда временной ряд демонстрирует высокую волатильность, а в другое время волатильность низкая, поэтому данные демонстрируют кластеризованную волатильность. В разделе 16.5 обсуждается кластеризованная волатильность и вводятся модели, в которых дисперсия ошибки изменяется во времени, то есть модели, в которых ошибка является условно гетероскедастичной. У модели условной гетероскедастичности есть несколько приложений. Одним из возможных приложений является вычисление интервальных прогнозов, в которых ширина интервала изменяется во времени с учетом периодов высокой или низкой неопределенности. Другое приложение — прогнозирование неопределенности доходности активов, таких как акции, которые в свою очередь могут быть полезны при оценке риска от владения этим активом.

## **16.1. Векторные авторегрессии**

В главе 14 рассматривался вопрос о прогнозировании темпов инфляции, но в действительности экономическим прогнозистам необходимо строить прогнозы

и других ключевых макроэкономических показателей, например, таких как уровень безработицы, темп роста ВВП и процентные ставки. Один из возможных подходов заключается в разработке отдельных моделей прогнозирования для каждой переменной, используя методы, рассмотренные в разделе 14.4. Другой подход заключается в разработке единой модели, по которой можно получать прогнозы всех переменных и которая может помочь сделать прогнозы взаимно согласованными. Одним из способов спрогнозировать несколько переменных в одной модели является использование модели векторной авторегрессии (VAR). VAR расширяет одномерную авторегрессию на случай нескольких временных рядов, то есть она расширяет одномерную авторегрессию на «вектор», состоящий из нескольких временных рядов.

### Векторные авторегресии

Модель векторной авторегрессии (VAR) представляет собой набор из  $k$  регрессий временных рядов, в которых регрессорами являются запаздывания всех  $k$  временных рядов. VAR расширяет одномерную авторегрессию до случая нескольких временных рядов или «вектора» временных рядов. Если число запаздываний в каждом из уравнений одинаково и равно  $p$ , система уравнений называется VAR( $p$ ).

В случае двух временных рядов  $Y_t$  и  $X_t$  VAR( $p$ ) состоит из двух уравнений:

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \dots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \dots + \gamma_{1p}X_{t-p} + u_{1t}, \quad (16.1)$$

$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \dots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \dots + \gamma_{2p}X_{t-p} + u_{2t}, \quad (16.2)$$

где  $\beta$ -ы и  $\gamma$ -ы – неизвестные коэффициенты, а  $u_{1t}$  и  $u_{2t}$  – случайные ошибки.

Предположения для VAR аналогичны предположениям для модели регрессии временных рядов из вставки «Основные понятия 14.6» применительно к каждому уравнению. Коэффициенты VAR оцениваются с применением МНК к каждому уравнению.

**ОСНОВНЫЕ ПОНЯТИЯ**  
**16.1**

### Модели VAR

Для случая двух временных рядов  $Y_t$  и  $X_t$ , *векторная авторегрессия (VAR)* состоит из двух уравнений: в одном зависимой переменной является  $Y_t$ , а в другом –  $X_t$ . Регрессорами в обоих уравнениях являются запаздывающие значения обеих переменных. В целом для случая  $k$  временных рядов VAR состоит из  $k$  уравнений по одному для каждой из переменных, а регрессорами во всех уравнениях являются лаговые значения всех переменных. Коэффициенты VAR оцениваются путем оценки каждого из уравнений с помощью МНК.

Концепция модели VAR рассматривается во вставке «Основные понятия 16.1».

**Тестирование гипотез в VAR.** При выполнении предположений VAR-модели МНК-оценки состоятельны и имеют совместное нормальное распределение

в больших выборках. Таким образом, статистические выводы можно делать стандартным образом, например, 95 %-е доверительные интервалы для коэффициентов модели могут быть построены как оценка коэффициента  $\pm 1,96$  от стандартной ошибки.

В VAR возникает один новый аспект проверки гипотез, потому что VAR с  $k$  переменными является системой, состоящей из  $k$  уравнений. Таким образом, можно проверить совместные гипотезы, которые включают ограничения на несколько уравнений.

Например, в случае двух переменных VAR( $p$ ) для уравнений (16.1) и (16.2) вы могли бы спросить, правильно ли выбрана длина лага  $p$  или нужно включать  $p-1$  лаг, то есть вы можете спросить, равны ли нулю коэффициенты при  $Y_{t-p}$  и  $X_{t-p}$  в этих двух уравнениях. Нулевая гипотеза о том, что эти коэффициенты равны нулю, имеет вид:

$$H_0 : \beta_{1p} = 0, \beta_{2p} = 0, \gamma_{1p} = 0, \text{ и } \gamma_{2p} = 0. \quad (16.3)$$

Альтернативная гипотеза заключается в том, что по крайней мере один из этих четырех коэффициентов отличен от нуля. Таким образом тестируется гипотеза о коэффициентах из обоих уравнений по два из каждого уравнения.

Так как оцененные коэффициенты имеют совместное нормальное распределение в больших выборках, можно проверить ограничения на эти коэффициенты при помощи  $F$ -статистики. Вывод точной формулы для этой статистики осложняется тем, что мы должны работать с обозначениями для нескольких уравнений, поэтому мы его опускаем. На практике в самые современные программные пакеты встроены процедуры для проверки гипотез о коэффициентах в системе нескольких уравнений.

**Сколько переменных следует включать в VAR?** Число коэффициентов в каждом уравнении VAR пропорционально количеству переменных в VAR. Например, VAR с пятью переменными и четырьмя лагами будет включать 21 коэффициент (четыре лага каждой из пяти переменных и константа) в каждом из пяти уравнений, что в общей сложности составляет 105 коэффициентов! Оценка всех этих коэффициентов увеличивает ошибку оценки, которая, в свою очередь, входит в прогноз, что может привести к ухудшению точности прогноза.

Практическим следствием этого является необходимость включения в VAR небольшого числа переменных и, что особенно важно, наличие уверенности в том, что переменные действительно связаны друг с другом так, что они будут полезны для прогнозирования друг друга. Например, как мы знаем из сочетания эмпирических данных (как, например, в главе 14) и экономической теории, уровень инфляции, уровень безработицы, а также краткосрочные процентные ставки связаны друг с другом, что предполагает, что эти переменные могут помочь при прогнозировании друг друга в VAR. Включение в VAR несвязанных переменных, однако, вносит погрешности в оценку без улучшения предсказательной силы, тем самым снижая точность прогноза.

**Определение числа запаздываний в VAR<sup>1</sup>.** Глубина запаздывания в VAR может быть определена при помощи  $F$ -тестов или информационных критериев.

---

<sup>1</sup> В этом разделе используются матричные обозначения и он может быть пропущен, чтобы не усложнять математические выкладки.

Информационный критерий для системы уравнений расширяет понятие информационного критерия для одного уравнения из раздела 14.5. Чтобы определить такой информационный критерий, мы должны ввести матричные обозначения. Пусть  $\Sigma_u$  является ковариационной матрицей ошибок VAR размера  $k \times k$  и пусть  $\hat{\Sigma}_u$  – оценка этой ковариационной матрицы, где элемент  $(i,j)$  матрицы  $\hat{\Sigma}_u$  равен  $\frac{1}{T} \sum_{t=1}^T \hat{u}_{it} \hat{u}_{jt}$ , где  $\hat{u}_{it}$  – МНК-остаток  $i$ -го уравнения и  $\hat{u}_{jt}$  – МНК-остаток  $j$ -го уравнения. Тогда ВIC для модели VAR будет иметь вид:

$$\text{BIC}(p) = |\ln[\det(\hat{\Sigma}_u)] + k(kp+1) \frac{\ln(T)}{T}|, \quad (16.4)$$

где  $\det(\hat{\Sigma}_u)$  – определитель матрицы  $\hat{\Sigma}_u$ . AIC вычисляется при помощи выражения (16.4), модифицированного заменой множителя « $\ln(T)$ » на «2».

Выражение (16.4) для информационного критерия BIC для  $k$  уравнений в VAR расширяет аналогичное выражение для случая одного уравнения, приведенного в разделе 14.5. В случае одного уравнения первое слагаемое упрощается до  $\ln[SSR(p)/T]$ . Второй член в уравнении (16.4) представляет собой штраф за включение дополнительных регрессоров;  $k(kp+1)$  – это общее число коэффициентов регрессии в VAR ( $k$  уравнений, каждое из которых содержит константу и  $p$  запаздываний каждого из  $k$  временных рядов).

Выбор оценки глубины запаздывания в VAR при помощи BIC аналогичен случаю одного уравнения: среди множества возможных значений  $p$  оценка глубины запаздывания  $\hat{p}$  – это значение  $p$ , которое минимизирует  $\text{BIC}(p)$ .

**Использование VAR для анализа причинных зависимостей.** До сих пор мы обсуждали VAR с точки зрения прогнозирования. Еще одно приложение VAR-моделей заключается в использовании их для анализа причинных зависимостей между экономическими временными рядами; действительно, модели VAR впервые были введены в экономику эконометристом и макроэкономистом Кристофером Симсом<sup>1</sup> (Sims, 1980) с этой целью. Использование VAR'ов для анализа причинных зависимостей известно как структурное VAR-моделирование: «структурное» потому, что в этом приложении VAR'ы используются для моделирования основной структуры экономики. Анализ на основе структурных VAR'ов использует методы, представленные в этом разделе в контексте прогнозирования, а также некоторые дополнительные инструменты. Однако самая большая концептуальная разница между использованием VAR'ов для прогнозирования и использованием их для структурного моделирования заключается в том, что структурное моделирование требует наличия конкретных предположений об экзогенности переменных, основанных на экономической теории и институциональных знаниях. Обсуждение структурных VAR'ов лучше всего проводить в контексте рассмотрения методов оценки систем одновременных уравнений, что выходит за рамки данной книги. Для ознакомления с основами прогнозирования по VAR-моделям

<sup>1</sup> В 2011 году Кристофер Симс (совместно с Томасом Сарджентом) был удостоен Нобелевской премии в области экономики за «статистическое моделирование монетарной политики в экономике». Более подробно см. <http://www.nobelprize.org>. – Примеч. науч. ред. перевода.

и использования их для анализа экономической политики используйте Stock, Watson (2001). Для ознакомления с дополнительными математическими деталями, касающимися структурных VAR, см.: Hamilton (1994) или Watson (1994).

### **VAR-модель уровней инфляции и безработицы**

В качестве иллюстрации рассмотрим модель VAR от двух переменных: уровня инфляции,  $Inf_t$ , и уровня безработицы,  $Unemp_t$ . Как и в главе 14, мы рассматриваем уровень инфляции как временной ряд, содержащий стохастический тренд, поэтому целесообразно преобразовать его, вычислив первую разность,  $\Delta Inf_t$ .

VAR-модель для  $\Delta Inf_t$ , и  $Unemp_t$ , состоит из двух уравнений: в одном зависимой переменной является  $\Delta Inf_t$ , в другом –  $Unemp_t$ . В качестве регрессоров в обоих уравнениях используются запаздывающие значения  $\Delta Inf_t$  и  $Unemp_t$ . Из-за очевидного структурного сдвига в кривой Филлипса в начале 1980-х годов, обнаруженного в разделе 14.7 помошью QLR-теста, VAR оценивается с использованием данных с I квартала 1982 по IV квартал 2004 года.

Первым уравнением VAR является уравнение инфляции:

$$\begin{aligned} \widehat{\Delta Inf}_t = & 1,47 - 0,64 \Delta Inf_{t-1} - 0,64 \Delta Inf_{t-2} - 0,13 \Delta Inf_{t-3} - \\ & - 0,13 \Delta Inf_{t-4} - 3,49 Unemp_{t-1} + 2,80 Unemp_{t-2} + \\ & + 2,44 Unemp_{t-3} - 2,03 Unemp_{t-4}. \end{aligned} \quad (16.5)$$

Скорректированный  $R^2$  равен  $\bar{R}^2 = 0,44$ .

Второе уравнение VAR является уравнением безработицы, в котором регрессоры такие же, как и в уравнении инфляции, но зависимой переменной является уровень безработицы:

$$\begin{aligned} \widehat{Unemp}_t = & 0,22 + 0,005 \Delta Inf_{t-1} + 0,004 \Delta Inf_{t-2} - 0,007 \Delta Inf_{t-3} - \\ & - 0,003 \Delta Inf_{t-4} + 1,52 Unemp_{t-1} - 0,29 Unemp_{t-2} - \\ & - 0,43 Unemp_{t-3} + 0,16 Unemp_{t-4}. \end{aligned} \quad (16.6)$$

Скорректированный  $R^2$  равен  $\bar{R}^2 = 0,982$ .

Уравнения (16.5) и (16.6), взятые вместе, представляют VAR(4) – модель изменения уровня инфляции,  $\Delta Inf_t$ , и уровня безработицы,  $Unemp_t$ .

Эти VAR-уравнения могут быть использованы для проверки наличия причинности по Грейндджеру.  $F$ -статистика для тестирования нулевой гипотезы о том, что коэффициенты при  $Unemp_{t-1}$ ,  $Unemp_{t-2}$ ,  $Unemp_{t-3}$  и  $Unemp_{t-4}$  в уравнении инфляции [уравнение (16.5)] равны нулю, равна 11,04 и имеет  $p$ -значение меньше, чем 0,001. Таким образом, нулевая гипотеза отвергается, так что можно сделать вывод о том, что уровень безработицы является полезным фактором для прогнозирования изменений в инфляции с учетом лагов инфляции (т.е. уровень безработицы является причиной изменений инфляции по Грейндджеру).

$F$ -статистика для проверки гипотезы о том, что коэффициенты при четырех лагах  $\Delta Inf_t$ , в уравнении безработицы [уравнение (16.6)] равны нулю, составляет 0,16 и имеет  $p$ -значение 0,96. Таким образом, изменение инфляции не является причиной по Грейндджеру безработицы на 10 %-м уровне значимости.

Прогнозы уровней инфляции и безработицы на один период вперед получаются точно такие же, как описано в разделе 14.4. Основываясь на уравнении (16.5), получаем, что прогноз изменения инфляции в I квартале 2005 года, сделанный в IV квартале 2004 года, равен:  $\widehat{\Delta Inf}_{2005:1|2004:IV} = -0,1\%$ . Аналогичные расчеты с использованием уравнения (16.6) дают прогноз уровня безработицы в I квартале 2005 года на основе данных до IV квартала 2004 года включительно, который равен:  $\widehat{Unemp}_{2005:1|2004:IV} = 5,4\%$ , что очень близко к его реальному значению  $Unemp_{2005:1} = 5,3\%$ .

## 16.2. Многошаговые прогнозы

До сих пор мы обсуждали лишь одношаговые прогнозы. Однако прогнозистам часто необходимы длительные прогнозы. В этом разделе описываются два метода получения многошаговых прогнозов. Обычный метод заключается в построении итеративных прогнозов, в которых для расчета прогнозов на более чем один период вперед используются прогнозы, полученные на предыдущих шагах. Второй метод заключается в построении «прямых» прогнозов с помощью регрессии, в которой зависимой переменной является переменная, которую мы хотим прогнозировать, а ее прогноз строится только на основе имеющихся данных. По причинам, изложенным в конце этого раздела, в большинстве приложений рекомендуется использовать итеративный метод<sup>1</sup>.

### Итеративные многошаговые прогнозы

Основная идея, лежащая в основе итеративного прогнозирования, заключается в том, что оцененная модель используется для прогнозирования, чтобы сделать прогноз на один период вперед, то есть для момента  $T+1$ , используя данные до момента  $T$  включительно. После этого модель используется, чтобы сделать прогноз на момент  $T+2$  с учетом данных до момента  $T$  включительно, где прогнозное значение на момент  $T+1$  рассматривается как данные, которые используются для того, чтобы сделать прогноз на момент  $T+2$ . Таким образом, прогноз на один шаг вперед (который также упоминается как одношаговый прогноз) используется в качестве промежуточного шага, чтобы сделать двухшаговый прогноз. Этот процесс повторяется до тех пор, пока не построен прогноз для желаемого горизонта прогнозирования  $h$ .

**Итеративное прогнозирование по AR-модели: AR(1).** Итеративный AR(1) прогноз использует модель AR(1) для построения одношагового прогноза. Рассмотрим, например, модель авторегрессии первого порядка для  $\Delta Inf_t$  [уравнение (14.7)]:

$$\widehat{\Delta Inf}_t = 0,02 - 0,24 \Delta Inf_{t-1}. \quad (16.7)$$

<sup>1</sup> Итеративные прогнозы часто называют динамическими. – Примеч. науч. ред. перевода.

Первым шагом в вычислении прогноза  $\Delta Inf_{2005:\text{II}}$  на два квартала вперед на основе уравнения (16.7) с использованием данных до IV квартала 2004 года является вычисление прогноза  $\Delta Inf_{2005:\text{I}}$  на один квартал на основе данных до IV квартала 2004 года:  $\widehat{\Delta Inf}_{\text{:I}|2004:\text{IV}} = 0,02 - 0,24\Delta Inf_{2004:\text{IV}} = 0,02 - 0,24 \times 1,9 = -0,4$ . На втором шаге мы подставляем этот прогноз в уравнение (16.7) так, чтобы  $\widehat{\Delta Inf}_{2005:\text{II}|2004:\text{IV}} = 0,02 - 0,24\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{IV}} = 0,02 - 0,24 \times (-0,4) = 0,1$ . Таким образом, полученный на основе информации до четвертого квартала 2004 года прогноз говорит, что во втором квартале 2005 года уровень инфляции увеличится на 0,1 процентного пункта по сравнению с первым кварталом 2005 года.

**Итеративное прогнозирование по AR-модели: AR( $p$ )**. Метод построения итеративного прогноза на основе модели AR(1) распространяется на случай модели AR( $p$ ), заменяя  $Y_{t+1}$  на его прогноз  $\widehat{Y}_{t+1|T}$  и используя этот одношаговый прогноз для построения прогноза  $\widehat{Y}_{t+2}$  по модели AR( $p$ ). Рассмотрим, например, итеративный двухшаговый прогноз инфляции на основе модели AR( $p$ ) из раздела 14.3 [уравнение (14.13)]:

$$\widehat{\Delta Inf}_t = 0,02 - 0,26\Delta Inf_{t-1} - 0,32\Delta Inf_{t-2} + 0,16\Delta Inf_{t-3} - 0,03\Delta Inf_{t-4}. \quad (16.8)$$

Прогноз  $\Delta Inf_{2005:\text{I}}$  на основе данных до IV квартала 2004 года с использованием этой AR(4)-модели, вычисленный в разделе 14.3, равен:  $\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{IV}} = 0,4$ . Таким образом, итеративный прогноз на два квартала вперед на основе модели AR(4) равен:  $\widehat{\Delta Inf}_{2005:\text{II}|2004:\text{IV}} = 0,02 - 0,26\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{IV}} - 0,32\Delta Inf_{2004:\text{IV}} + 0,16\Delta Inf_{2004:\text{III}} - 0,03\Delta Inf_{2004:\text{II}} = 0,02 - 0,26 \times 0,4 - 0,32 \times 1,9 + 0,16 \times (-2,8) - 0,08 \times 0,6 = -1,1$ . В соответствии с этим итеративный прогноз, построенный по модели AR(4) и основанный на данных до четвертого квартала 2004 года, говорит о том, что уровень инфляции во втором квартале 2005 года сократится на 1,1 процентных пункта по сравнению с первым кварталом 2005 года.

**Итеративное многомерное прогнозирование по итеративной VAR-модели.** Итеративные многомерные прогнозы можно вычислить с помощью VAR во многом таким же образом, как итеративные одномерные прогнозы вычисляются с использованием авторегрессии. Основной новой особенностью итеративного многомерного прогнозирования является то, что двухшаговый прогноз (для момента  $T+2$ ) для одной переменной зависит от прогнозов на момент  $T+1$  всех переменных в VAR. Например, чтобы вычислить прогноз изменения инфляции в момент времени  $T+2$  по сравнению с  $T+1$  с использованием VAR-модели для двух переменных  $\Delta Inf_t$  и  $Unemp_t$ , сначала, как промежуточный шаг в прогнозировании  $\Delta Inf_{t+2}$ , надо построить прогноз и  $\Delta Inf_{t+1}$  и  $Unemp_{t+1}$  с использованием данных до момента  $T$ . В целом для вычисления многошаговых итеративных VAR-прогнозов на  $h$  периодов вперед, необходимо вычислить прогнозы всех переменных для всех моментов времени между  $T$  и  $T+h$ .

В качестве примера вычислим итеративный VAR-прогноз  $\Delta Inf_{2005:\text{II}}$  на основе данных до IV квартала 2004 года с использованием VAR(4) для  $\Delta Inf_t$  и  $Unemp_t$  из раздела 16.1 [уравнения (16.5) и (16.6)]. На первом шаге вычислим прогнозы  $\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{V}}$  по этой VAR. Прогноз  $\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{V}}$  на основе уравнения (16.5) был вычислен в разделе 14.3 и составляет  $-0,1\%$  [уравнение (14.18)]. Аналогичные расчеты с использованием уравнения (16.6) показывают, что  $\widehat{Unemp}_{2005:\text{I}|2004:\text{V}} = 5,4\%$ .

На втором этапе эти прогнозы подставляются в уравнения (16.5) и (16.6) для получения прогнозов на два квартала вперед  $\widehat{\Delta Inf}_{2005:\text{II}|2004:\text{V}}$ :

$$\begin{aligned}
 \widehat{\Delta Inf}_{2005:\text{II}|2004:\text{IV}} &= 1,47 - 0,64\widehat{\Delta Inf}_{2005:\text{I}|2004:\text{IV}} - 0,64\Delta Inf_{2004:\text{IV}} - \\
 &- 0,13\Delta Inf_{2004:\text{III}} - 0,13\Delta Inf_{2004:\text{II}} - 3,49Unemp_{2005:\text{I}|2004:\text{IV}} + \\
 &+ 2,80Unemp_{2004:\text{IV}} + 2,44Unemp_{2004:\text{III}} - 2,03Unemp_{2004:\text{II}} = \\
 &= 1,47 - 0,64 \times (-0,1) - 0,64 \times 1,9 - 0,13 \times (-2,8) - 0,13 \times \\
 &\times 0,6 - 3,49 \times 5,4 + 2,80 \times 5,4 + 2,44 \times 5,4 - \\
 &- 2,03 \times 5,6 = -1,1. \tag{16.9}
 \end{aligned}$$

Таким образом, итеративный VAR(4)-прогноз, основанный на данных до четвертого квартала 2004 года включительно, говорит о том, что во втором квартале 2005 года инфляция снизится на 1,1 процентных пункта по сравнению с первым кварталом 2005 года.

Концепция итеративных многошаговых прогнозов представлена во вставке «Основные понятия 16.2».

### Итеративные многошаговые прогнозы

*Итеративный многошаговый AR-прогноз* рассчитывается на основе следующей схемы: сначала вычисляется прогноз на один шаг вперед, а затем он используется для вычисления прогноза на два шага вперед и так далее. Двух- и трехшаговые итеративные прогнозы, основанные на AR( $p$ ), вычисляются следующим образом:

$$\hat{Y}_{T+2|T} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{T+1|T} + \hat{\beta}_2 Y_T + \hat{\beta}_3 Y_{T-1} + \dots + \hat{\beta}_p Y_{T-p+2} \tag{16.10}$$

$$\hat{Y}_{T+3|T} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{T+2|T} + \hat{\beta}_2 \hat{Y}_{T+1|T} + \hat{\beta}_3 Y_T + \dots + \hat{\beta}_p Y_{T-p+3}, \tag{16.11}$$

где  $\hat{\beta}$ ’ы являются МНК-оценками коэффициентов AR( $p$ ). Продолжая этот процесс (совершая «итерации»), получаем прогнозы для больших горизонтов прогнозирования.

*Итеративный многошаговый VAR-прогноз* вычисляется по аналогичной схеме: сначала вычисляется одношаговый прогноз всех переменных, включенных в VAR, а затем эти прогнозы используются для вычисления двухшаговых прогнозов. Этот процесс продолжается итеративно до желаемого горизонта прогнозирования. Двухшаговый итеративный прогноз  $\hat{Y}_{T+2}$  на основе VAR( $p$ ) из вставки «Основные понятия 16.1», включающей две переменные, имеет вид:

$$\begin{aligned}
 \hat{Y}_{T+2|T} &= \hat{\beta}_{10} + \hat{\beta}_{11} \hat{Y}_{T+1|T} + \hat{\beta}_{12} Y_T + \hat{\beta}_{13} Y_{T-1} + \dots + \hat{\beta}_{1p} Y_{T-p+2} + \\
 &+ \hat{\gamma}_{11} \hat{X}_T + \hat{\gamma}_{12} X_T + \hat{\gamma}_{13} X_{T-1} + \dots + \hat{\gamma}_{1p} X_{T-p+2}, \tag{16.12}
 \end{aligned}$$

где коэффициенты в уравнении (16.12) являются МНК-оценками коэффициентов VAR. Повторяя эту процедуру, получаем прогнозы для более длинных горизонтов прогнозирования.

## ОСНОВНЫЕ ПОНЯТИЯ

### 16.2

## Прямые многошаговые прогнозы

Прямые многошаговые прогнозы вычисляются без использования итераций с помощью одной регрессии, в которой зависимая переменная является переменной, для которой нужно построить многошаговый прогноз, а регрессоры являются независимыми переменными, используемыми для прогнозирования интересующей нас переменной. Прогнозы, вычисляемые таким образом, называются прямыми прогнозами, потому что коэффициенты регрессии могут быть непосредственно использованы для того, чтобы сделать многошаговый прогноз.

**Прямой метод построения многошаговых прогнозов.** Предположим, что вы хотите сделать прогноз  $\widehat{Y}_{T+2}$ , используя данные до момента  $T$ . Для этого нужно использовать ADL-модель в качестве отправной точки, но лаги объясняющих переменных должны быть взяты соответствующим образом. Например, если в модели использовано два лага объясняющих переменных, то зависимой переменной будет  $Y_t$ , а объясняющими —  $Y_{t-2}$ ,  $Y_{t-3}$ ,  $X_{t-2}$  и  $X_{t-3}$ . Коэффициенты из этой регрессии могут быть использованы непосредственно для вычисления прогноза  $\widehat{Y}_{T+2}$  с помощью данных о  $Y_T$ ,  $Y_{T-1}$ ,  $X_T$  и  $X_{T-1}$  без каких-либо итераций. Другими словами, при прямом  $h$ -шаговом прогнозировании все объясняющие переменные регрессии лагируются на  $h$  периодов для получения прогноза на  $h$  шагов вперед.

Например, прогноз  $\Delta\text{Inf}_t$  на два квартала вперед, построенный с использованием четырех лагов каждой из переменных  $\Delta\text{Inf}_{t-2}$  и  $\text{Unemp}_{t-2}$ , рассчитывается путем оценки регрессии так:

$$\begin{aligned} \widehat{\Delta\text{Inf}}_{t|t-2} = & -0,15 - 0,25\Delta\text{Inf}_{t-2} + 0,16\Delta\text{Inf}_{t-3} - 0,15\Delta\text{Inf}_{t-4} - \\ & - 0,10\Delta\text{Inf}_{t-5} - 0,17\text{Unemp}_{t-2} - 1,82\text{Unemp}_{t-3} - \\ & - 3,53\text{Unemp}_{t-4} + 1,89\text{Unemp}_{t-5}, \end{aligned} \quad (16.13)$$

тогда двухшаговый прогноз изменения инфляции во II квартале 2005 года по сравнению с I кварталом 2005 года вычисляется подстановкой значений  $\Delta\text{Inf}_{2004:\text{IV}}, \dots, \Delta\text{Inf}_{2004:\text{I}}$ ,  $\text{Unemp}_{2004:\text{IV}}, \dots, \text{Unemp}_{2004:\text{I}}$  в уравнение (16.13):

$$\begin{aligned} \widehat{\Delta\text{Inf}}_{2005:\text{II}|2004:\text{IV}} = & 0,15 - 0,25\Delta\text{Inf}_{2004:\text{IV}} + 0,16\Delta\text{Inf}_{2004:\text{III}} - \\ & - 0,15\Delta\text{Inf}_{2004:\text{II}} - 0,10\Delta\text{Inf}_{2004:\text{I}} - 0,17\text{Unemp}_{2004:\text{IV}} + \\ & + 1,82\text{Unemp}_{2004:\text{III}} - 3,53\text{Unemp}_{2004:\text{II}} + \\ & + 1,89\text{Unemp}_{2004:\text{I}} = -1,38. \end{aligned} \quad (16.14)$$

Прямой прогноз на три квартала вперед,  $\Delta\text{Inf}_{T+3}$ , может быть получен лагируя все регрессоры в уравнении (16.13) еще на один квартал, оценивая эту регрессию, а затем вычисляя прогноз. Прямой прогноз на  $h$  кварталов вперед,  $\Delta\text{Inf}_{T+h}$ ,

вычисляют по регрессии зависимой переменной  $\Delta Inf_t$  и регрессорами  $\Delta Inf_{t-h}$  и  $Unemp_{t-h}$ , а также включая дополнительные лаги  $\Delta Inf_{t-h}$  и  $Unemp_{t-h}$ , если нужно.

**Стандартные ошибки в прямой многошаговой регрессии.** Поскольку зависимая переменная в многошаговой регрессии оценивается для двух или более периодов в будущем, ее ошибка является серийно коррелированной. Чтобы убедиться в этом, рассмотрим прогноз инфляции на два периода вперед и предположим, что неожиданный скачок цен на нефть произойдет в следующем квартале. Построенный сегодня прогноз инфляции на два шага вперед будет слишком низким, так как он не включает это неожиданное событие. Поскольку рост цен на нефть был также неизвестен в предыдущем квартале, то двухшаговый прогноз, сделанный в прошлом квартале, также будет слишком мал. Таким образом, из-за неожиданного роста цен в следующем квартале оба двухшаговых прогноза, сделанных в этом квартале и в прошлом квартале, будут заниженными. Из-за таких неожиданных событий ошибка в многошаговой регрессии является серийно коррелированной.

Как отмечалось в разделе 15.4, если ошибка серийно коррелирована, не-корректно использовать обычные стандартные МНК-ошибки или, точнее, они не являются надежной основой для статистических выводов. Следовательно, в прямой многошаговой регрессии нужно использовать стандартные ошибки, являющиеся состоятельными при наличии гетероскедастичности и автокорреляции (НАС). Таким образом, стандартные ошибки в уравнении (16.13) для прямой многошаговой регрессии являются НАС-стандартными ошибками Ньюи-Веста, где параметр усечения  $t$  выбирается в соответствии с формулой (15.17); для имеющихся данных (для которых  $T = 92$ ) из формулы (15.17) получаем  $t = 3$ . Для более длинного горизонта прогнозирования размер совпадения — и, таким образом, степень серийной корреляции ошибок — увеличивается: в общем случае первые  $h - 1$  коэффициенты автокорреляции ошибок в регрессии, используемой для прямого  $h$ -шагового прогнозирования, отличны от нуля. Таким образом, значения  $t$ , большие чем рассчитываемые по формуле (15.17), являются более подходящими для многошаговой регрессии с длинными горизонтами прогнозирования.

Прямые многошаговые прогнозы представлены во вставке «Основные понятия 16.3».

### Прямые многошаговые прогнозы

Прямые многошаговые прогнозы на  $h$  периодов вперед, построенные на основе  $p$ -запаздываний переменной  $Y_t$  и дополнительного регрессора  $X_p$ , вычисляются при помощи оценки регрессии так:

$$Y_t = \delta_0 + \delta_1 Y_{t-h} + \dots + \delta_p Y_{t-p-h+1} + \\ + \delta_{p+1} X_{t-h} + \dots + \delta_{2p} X_{t-p-h+1} + u_t. \quad (16.15)$$

Это на первом шаге, а затем — при помощи оценок коэффициентов непосредственно, чтобы рассчитать прогноз  $Y_{T+h}$  с использованием данных до периода  $T$  включительно.

## ОСНОВНЫЕ ПОНЯТИЯ

### 16.3

## **Какой метод использовать?**

Существует две причины того, почему в большинстве случаев рекомендуется итеративный метод многошагового прогнозирования. Во-первых, с теоретической точки зрения, если модель, на основе которой строится одношаговый прогноз (AR или VAR, используемая для итеративного прогнозирования), специфицирована правильно, то коэффициенты оцениваются более эффективно, если они оцениваются на основе обычной регрессии (а затем – через необходимое количество итераций), чем если бы они были получены на основе многошаговой регрессии. Во-вторых, с практической точки зрения прогнозисты, как правило, заинтересованы в получении прогнозов не на каком-то конкретном горизонте прогнозирования, а на разных. Из-за того что эти прогнозы получаются с использованием одной той же модели, итеративные прогнозы сходятся к некоторой временной траектории и поэтому являются менее неустойчивыми для различных горизонтов прогнозирования, чем прямые прогнозы. Поскольку для прямых прогнозов используются различные модели на каждом горизонте прогнозирования, ошибка выборки в оценках коэффициентов может добавлять случайные колебания во временную траекторию последовательности прямых многошаговых прогнозов.

Однако в некоторых случаях прямые прогнозы предпочтительнее итеративных. Одним из таких обстоятельств является ситуация, когда у вас есть основания полагать, что обычная модель (AR или VAR) специфицирована неверно. Например, вы можете считать, что уравнение для переменной, которую вы пытаетесь спрогнозировать, в VAR'e специфицировано правильно, но одно или более из других уравнений в VAR'e указаны неправильно, например из-за пропуска нелинейных регрессоров. Если обычная модель указана неправильно, то в общем случае итеративный многошаговый прогноз будет смешенным и MSFE итеративного прогноза может превышать MSFE прямого прогноза, даже несмотря на то что прямой прогноз имеет большую дисперсию. Вторая ситуация, при которой прямой прогноз может быть желательным, возникает в многомерной модели прогнозирования с многими предикторами, в этом случае модель VAR, определенная для всех переменных, может быть ненадежной, поскольку она будет содержать очень много оцененных коэффициентов.

### **16.3. Порядок интегрированности и DF-GLS-тест на единичные корни**

В данном разделе мы расширяем концепцию стохастических трендов из раздела 14.6, рассматривая еще два вопроса. Во-первых, тренды, содержащиеся в некоторых временных рядах, плохо описываются моделью случайного блуждания, поэтому мы введем расширение этой модели и обсудим ее приложения к регрессионному анализу таких рядов. Во-вторых, мы продолжим обсуждение методов тестирования наличия единичного корня во временных рядах и, среди прочего, рассмотрим второй тест на единичный корень – DF-GLS.

## Другие модели трендов и порядок интегрированности

Напомним, что модель случайного блуждания, введенная в разделе 14.6, говорит о том, что тренд в момент  $t$  равен тренду в момент  $t-1$  плюс случайная ошибка. Если  $Y_t$  является процессом случайного блуждания со сносом (с дрейфом)  $\beta_0$ , то получаем:

$$Y_t = \beta_0 + Y_{t-1} + u_t, \quad (16.16)$$

где  $u_t$  серийно не коррелирована. Также напомним из раздела 14.6, что если временной ряд описывается моделью случайного блуждания, то он содержит авторегрессионный корень, равный единице.

Несмотря на то что модели случайных блужданий описывают долгосрочные тенденции многих экономических временных рядов, некоторые экономические временные ряды содержат тренды, которые являются более гладкими, то есть меньше меняются от одного периода к другому, чем это следует из уравнения (16.16). Для описания трендов, содержащихся в таких рядах, необходимы другие модели.

### Порядок интегрированности, взятие разностей и стационарность

- Если  $Y_t$  является интегрированным первого порядка, то есть если  $Y_t$  является  $I(1)$ , то  $Y_t$  содержит единичный авторегрессионный корень, а его первая разность  $\Delta Y_t$  является стационарной.
- Если  $Y_t$  является интегрированным второго порядка, то есть если  $Y_t$  является  $I(2)$ , то  $\Delta Y_t$  содержит единичный авторегрессионный корень, а его вторая разность  $\Delta^2 Y_t$  является стационарной.
- Если  $Y_t$  является *интегрированным порядка d*, то есть если  $Y_t$  является  $I(d)$ , то для устранения его стохастического тренда необходимо взять  $d$  разностей  $Y_t$ , то есть  $\Delta^d Y_t$  является стационарным.

**ОСНОВНЫЕ  
ПОНЯТИЯ**  
**16.4**

Одной из моделей гладкого тренда является модель случайного блуждания для первой разности:

$$\Delta Y_t = \beta_0 + \Delta Y_{t-1} + u_t, \quad (16.17)$$

где  $u_t$  серийно не коррелирована. Таким образом, если  $Y_t$  описывается уравнением (16.17), то  $\Delta Y_t$  является случайным блужданием, и поэтому разность  $\Delta Y_t - \Delta Y_{t-1}$  стационарна. Разность первых разностей  $\Delta Y_t - \Delta Y_{t-1}$  называется *второй разностью*  $Y_t$  и обозначается  $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$ . В этой терминологии, если  $Y_t$  описывается уравнением (16.17), то его вторая разность является стационарной. Если временной ряд содержит тренд в виде, описываемом уравнением (16.17), то его первая разность содержит авторегрессионный корень, равный единице.

**Терминология «порядка интегрированности».** Для того чтобы различать между собой две модели тренда, полезно ввести некоторые дополнительные термины. Временной ряд, содержащий тренд, описываемый моделью случайного блуждания, называют *интегрированным первого порядка*, или  $I(1)$ . Временной ряд, содержащий тренд в форме, описываемой уравнением (16.17), называют *интегрированным второго порядка*, или  $I(2)$ . Временной ряд, не содержащий стохастический тренд, называют *интегрированным нулевого порядка*, или  $I(0)$ .

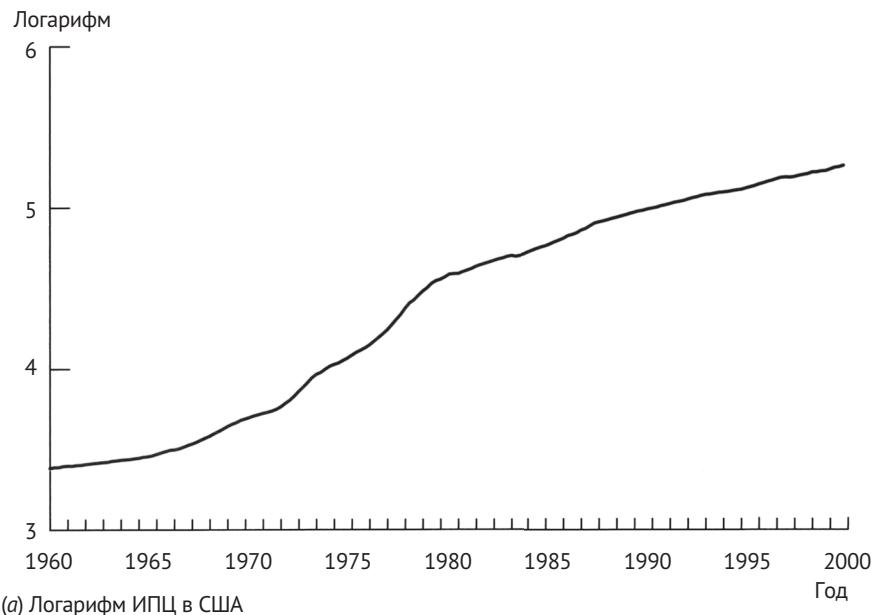
**Порядок интегрированности** в терминах  $I(1)$  и  $I(2)$  – это число разностей, которые нужно взять, чтобы временной ряд стал стационарным: если  $Y_t$  является  $I(1)$ , то его первая разность  $\Delta Y_t$  является стационарной, и если  $Y_t$  является  $I(2)$ , то его вторая разность  $\Delta^2 Y_t$  является стационарной. Если  $Y_t$  является  $I(0)$ , то  $Y_t$  стационарен.

Определение порядка интегрированности приведено во вставке «Основные понятия 16.4».

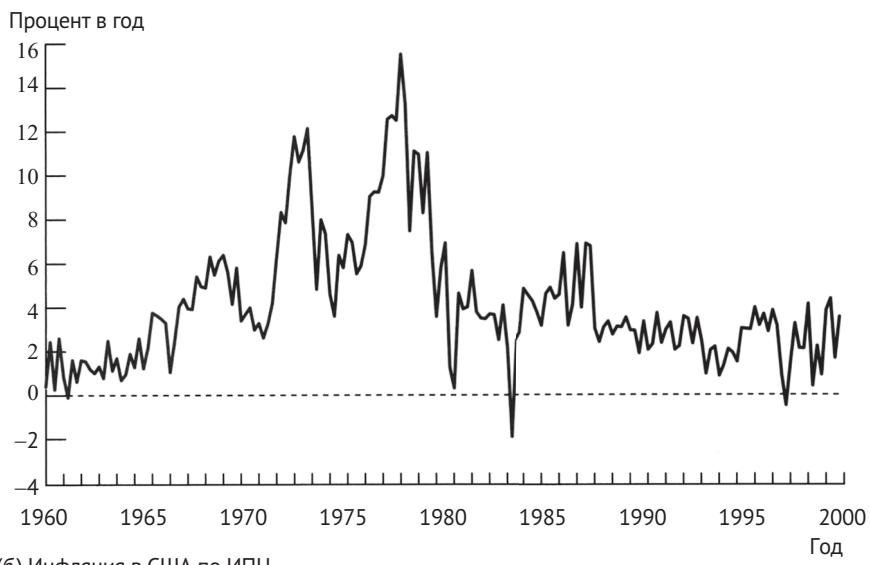
**Как выяснить, является ли временной ряд  $I(2)$  или  $I(1)$ .** Если  $Y_t$  является  $I(2)$ , то  $\Delta Y_t$  является  $I(1)$ , поэтому  $\Delta Y_t$  имеет авторегрессионный корень, равный единице. Однако если  $Y_t$  является  $I(1)$ , то  $\Delta Y_t$  является стационарным. Таким образом, гипотеза о том, что  $Y_t$  является  $I(2)$ , может быть протестирована против альтернативной гипотезы о том, что  $Y_t$  является  $I(1)$ , с помощью проверки наличия единичного авторегрессионного корня у  $\Delta Y_t$ . Если гипотеза о том, что  $\Delta Y_t$  имеет единичный авторегрессионный корень, отвергается, то гипотеза о том, что  $Y_t$  является  $I(2)$ , отвергается в пользу альтернативы о том, что  $Y_t$  является  $I(1)$ .

**Примеры временных рядов типа  $I(2)$  или  $I(1)$ :** уровень цен и инфляция. В главе 14 мы пришли к выводу, что уровень инфляции в Соединенных Штатах содержит стохастический тренд и описывается моделью случайного блуждания, то есть уровень инфляции является  $I(1)$ . Если инфляция является  $I(1)$ , то ее стохастический тренд удаляется взятием первых разностей, так что  $\Delta Inf_t$  является стационарной. Напомним из раздела 14.2 [уравнение (14.2)], что квартальная инфляция в годовом исчислении является первой разностью логарифма уровня цен, умноженной на 400, то есть  $Inf_t = 400\Delta p_t$ , где  $p_t = \ln(CPI_t)$  и  $CPI_t$  обозначает значение индекса потребительских цен в квартале  $t$ . Таким образом, если мы считаем, что уровень инфляции является  $I(1)$ , то эквивалентно этому –  $\Delta p_t$  является  $I(1)$ , но это, в свою очередь, эквивалентно тому, что  $p_t$  является  $I(2)$ . Следовательно, все это время мы рассматривали логарифм уровня цен как  $I(2)$ , хотя и не использовали эту терминологию.

Графики логарифма уровня цен  $p_t$  и уровня инфляции приведены на рисунке 16.1. Долгосрочный тренд в логарифме уровня цен (рис. 16.1а) изменяется более плавно, чем долгосрочный тренд в уровне инфляции (рис. 16.1б). Плавно меняющийся тренд в логарифме уровня цен характерен для временных рядов типа  $I(2)$ .



(a) Логарифм ИПЦ в США



(б) Инфляция в США по ИПЦ

**Рисунок 16.1. Логарифм уровня цен и инфляция в Соединенных Штатах, 1960–2004 годы**

Тренд в логарифме цен (рис. 16.1а) гораздо более гладкий, чем тренд в инфляции (рис. 16.1б).

### DF-GLS-тест на единичный корень

В этом разделе продолжается обсуждение методов тестирования наличия единичного авторегрессионного корня. Сначала мы рассмотрим еще один тест

на единичный авторегрессионный корень, так называемый DF-GLS-тест. Далее, в дополнительном математическом разделе мы обсудим, почему получаемые тестовые статистики не имеют нормальных распределений даже в больших выборках.

**DF-GLS-тест.** ADF-тест был первым тестом, разработанным для тестирования нулевой гипотезы о наличии единичного корня, и является тестом, наиболее широко используемым на практике. Однако впоследствии были предложены и другие тесты, многие из которых имеют более высокую мощность (вставка «Основные понятия 3.5»), чем ADF-тест. Тест с более высокой мощностью, чем ADF-тест, с большей вероятностью отвергнет нулевую гипотезу о наличии единичного корня против стационарной альтернативы, если альтернативная гипотеза верна, поэтому более мощный тест лучше различает единичный AR-корень и корень, являющийся большим, но меньшим единицами.

В данном разделе рассматривается один из таких тестов – *DF-GLS-тест*, – разработанный Эллиотом, Розенбергом и Стоком (Elliott, Rothenberg, Stock, 1996). Тест рассматривается для случая, когда в условиях нулевой гипотезы  $Y_t$  является случайным блужданием, возможно, с дрейфом, а в условиях альтернативной гипотезы  $Y_t$  является стационарным вокруг линейного тренда.

DF-GLS-тест состоит из двух шагов. На первом шаге константа и тренд оцениваются при помощи обобщенного метода наименьших квадратов (ОМНК, см. раздел 15.5). ОМНК-оценки получаются при помощи трех новых переменных  $V_1$ ,  $X_{1t}$  и  $X_{2t}$ , где  $V_1 = Y_t$  и  $V_t = Y_t - \alpha^* Y_{t-1}$ ,  $t=2, \dots, T$ ,  $X_{1t} = 1$  и  $X_{2t} = t - \alpha^*(t-1)$ ,  $t=2, \dots, T$ , и  $X_{21} = 1$  и  $X_{2T} = T - \alpha^*(T-1)$ , где  $\alpha^*$  вычисляется по формуле:  $\alpha^* = 1 - 13,5/T$ . Затем оценивается регрессия  $V_t$  от  $X_{1t}$  и  $X_{2t}$ , то есть МНК используется для оценки коэффициентов теоретической модели регрессии вида:

$$V_t = \delta_0 X_{1t} + \delta_1 X_{2t} + e_t \quad (16.18)$$

с помощью наблюдений  $t=1, \dots, T$ , где  $e_t$  является рядом ошибок регрессии. Обратите внимание, что константа в регрессии (16.18) отсутствует. МНК-оценки  $\hat{\delta}_0$  и  $\hat{\delta}_1$  используются затем для вычисления «детрендированного» временного ряда  $Y_t^d$ ,  $Y_t^d = Y_t - (\hat{\delta}_0 + \hat{\delta}_1 t)$ .

На втором шаге используется тест Дики–Фуллера для проверки наличия единичного авторегрессионного корня в  $Y_t^d$ , причем регрессия в teste Дики–Фуллера не включает константу и временной тренд. То есть оценивается регрессия  $\Delta Y_t^d$  от  $Y_{t-1}^d$ ,  $\Delta Y_{t-2}^d, \dots, \Delta Y_{t-p}^d$ , где число запаздываний  $p$  определяется как обычно, либо на основе экспертного мнения, либо с использованием имеющихся данных на основе информационных критериев, таких как AIC или BIC, как было описано в разделе 14.5.

Если альтернативная гипотеза заключается в том, что  $Y_t$  является стационарным со средним значением, которое может быть ненулевым, но без временного тренда, то предыдущие шаги должны были изменены. Более точно,  $\alpha^*$  вычисляется по формуле:  $\alpha^* = 1 - 75/T$ ,  $X_{2t}$  исключается из регрессии (16.18), а временной ряд  $Y_t^d$  вычисляется как  $Y_t^d = Y_t - \hat{\delta}_0$ .

ОМНК-регрессия на первом шаге DF-GLS-теста делает этот тест сложнее, чем обычный ADF-тест, но также улучшает его способность отвергать нулевую гипотезу.

тезу о наличии единичного авторегрессионного корня в пользу альтернативной, если  $Y_t$  является стационарным. Это улучшение может быть значительным. Например, предположим, что  $Y_t$  на самом деле является стационарной авторегрессией AR(1) с авторегрессионным коэффициентом  $\beta_1 = 0,95$ , имеется  $T=200$  наблюдений и тест на наличие единичного корня проводится без временного тренда [т.е.  $t$  исключается из регрессии в teste Дики–Фуллера и  $X_{2t}$  опущен в уравнении (16.18)]. Тогда вероятность того, что ADF-тест отвергнет нулевую гипотезу на уровне значимости 5 %, составляет приблизительно 31 % по сравнению с 75 % для DF-GLS теста.

**Критические значения DF-GLS-теста.** Так как в ADF- и DF-GLS-тестах коэффициенты при детерминированных переменных оцениваются по-разному, эти тесты имеют различные критические значения. Критические значения для DF-GLS-теста приведены в таблице 16.1. Если тестовая статистика DF-GLS-теста ( $t$ -статистика при  $Y_{t-1}^d$  в регрессии, оцениваемой на втором шаге) меньше, чем критическое значение (т.е. более отрицательная, чем критическое значение), то нулевая гипотеза о том, что  $Y_t$  содержит единичный корень, отвергается. Как и для критических значений теста Дики–Фуллера, соответствующие критические значения зависят от того, какая из спецификаций теста используется, то есть от наличия или отсутствия временного тренда [включен или нет  $X_{2t}$  в уравнение (16.18)].

Таблица 16.1

## Критические значения DF-GLS-теста

Детерминированные регрессоры [Регрессоры в уравнении (16.18)]	10%	5%	1%
Только константа (только $X_{1t}$ )	-1,62	-1,95	-2,58
Константа и временной тренд ( $X_{1t}$ и $X_{2t}$ )	-2,57	-2,89	-3,48

Источник: Fuller (1976) и Elliot, Rothenberg, Stock (1996, табл. 1)

**Приложение к инфляции.** DF-GLS-статистика, вычисленная для временного ряда темпов инфляции ИПЦ,  $Inf_t$ , за период с I квартала 1962 по IV квартал 2004 года, с включением константы, но без временного тренда, равна -2,06 при включении трех запаздываний  $\Delta Y_t^d$  в регрессию Дики–Фуллера на втором этапе. Это значение меньше 5 %-го критического значения из таблицы 16.1, которое равно -1,95, поэтому использование DF-GLS-теста с тремя лагами приводит к отверждению нулевой гипотезы о наличии единичного корня на 5 %-м уровне значимости. Три лага были выбраны на основе информационного критерия AIC (максимальное число лагов предполагалось равным шести).

Вследствие того что DF-GLS-тест способен лучше различать нулевую гипотезу о наличии единичного корня и ее стационарную альтернативу, одна из интерпретаций этого результата состоит в том, что инфляция на самом деле стационарна, но тест Дики–Фуллера, реализованный в разделе 14.6, не смог обнаружить этого (на уровне значимости 5 %). Однако этот вывод следует смягчить,

отметив, что в этом приложении результаты DF-GLS-теста отвергают нулевую гипотезу и очень чувствительны к выбору глубины включаемых запаздывающих разностей. Если в тестовую регрессию включить два лага, что соответствует количеству лагов, выбранных критерием ВIC, то он отвергает нулевую гипотезу на 10%-м уровне значимости, но не на 5%-м. Результаты также чувствительны к размеру выборки; если вычислить соответствующую статистику на интервале с I квартала 1963 по IV квартал 2004 года (т.е. исключив всего лишь первый год), тест отвергает нулевую гипотезу на 10%-м уровне значимости, но не на уровне значимости 5 %, если использовать число лагов, выбранное по критерию AIC. Поэтому общая картина довольно неоднозначна [так же, как и на основе теста ADF, что было описано в дискуссии после уравнения (14.34)] и требует от прогнозиста принятия обоснованного решения о том, какая модель инфляции лучше:  $I(1)$  или стационарная.

### **Почему распределения в тестах на единичный корень не являются нормальными?**

В разделе 14.6 было подчеркнуто, что асимптотическое нормальное распределение, на которое опирается регрессионный анализ, не может быть использовано, если регрессоры являются нестационарными. В условиях нулевой гипотезы о том, что временной ряд содержит единичный корень, регрессор  $Y_{t-1}$  в регрессии теста Дики–Фуллера (и регрессор  $Y_{t-1}^d$  в модифицированной регрессии Дики–Фуллера, используемой на втором этапе DF-GLS-теста) нестационарен. Ненормальное распределение статистик теста на наличие единичного корня является следствием этой нестационарности.

Для того чтобы увидеть некоторое математическое обоснование этой ненормальности, рассмотрим простейшую регрессию теста Дики–Фуллера, в которой  $\Delta Y_t$  оценивается только на один регрессор  $Y_{t-1}$  и константа не включена. В обозначениях из вставки «Основные понятия 14.8» МНК-оценка коэффициента в этой регрессии имеет вид:  $\hat{\delta} = \sum_{t=1}^{T-1} Y_{t-1} \Delta Y_t / \sum_{t=1}^T Y_{t-1}^2$ , поэтому

$$T\hat{\delta} = \frac{\frac{1}{T} \sum_{t=1}^T Y_{t-1} \Delta Y_t}{\frac{1}{T^2} \sum_{t=1}^T Y_{t-1}^2}. \quad (16.19)$$

Рассмотрим числитель в формуле (16.19). При дополнительном предположении о том, что  $Y_0=0$ , используя алгебраические преобразования (упражнение 16.5), можно показать, что

$$\frac{1}{T} \sum_{t=1}^T Y_{t-1} \Delta Y_t = \frac{1}{2} \left[ \left( \frac{Y_T}{\sqrt{T}} \right)^2 - \frac{1}{T} \sum_{t=1}^T (\Delta Y_t)^2 \right]. \quad (16.20)$$

В условиях нулевой гипотезы  $\Delta Y_t = u_t$  является серийно некоррелированной и имеет конечную дисперсию, поэтому второй член в уравнении (16.20) имеет предел по вероятности, равный  $\frac{1}{T} \sum_{t=1}^T (\Delta Y_t)^2 \xrightarrow{P} \sigma_u^2$ . В предположении о том, что

$Y_0 = 0$ , первый член выражения (16.20) может быть записан как  $Y_T / \sqrt{T} = \sqrt{\frac{1}{T} \sum_{t=1}^T \Delta Y_t} = \sqrt{\frac{1}{T} \sum_{t=1}^T u_t}$ , что подчиняется центральной предельной теореме, то есть  $Y_T / \sqrt{T} \xrightarrow{d} N(0, \sigma_u^2)$ . Таким образом,  $(Y_T / \sqrt{T})^2 - \frac{1}{T} \sum_{t=1}^T (\Delta Y_t)^2 \xrightarrow{d} \sigma_u^2 (Z^2 - 1)$ , где  $Z$  является стандартной нормальной случайной величиной. Вспомним, однако, что квадрат стандартного нормального распределения имеет хи-квадрат распределение с одной степенью свободы. Таким образом, из уравнения (16.20) следует, что в условиях нулевой гипотезы числитель выражения (16.19) имеет предельное распределение:

$$\frac{1}{T} \sum_{t=1}^T Y_{t-1} \Delta Y_t \xrightarrow{d} \frac{\sigma_u^2}{2} (\chi^2 - 1). \quad (16.21)$$

Асимптотическое распределение из выражения (16.21) отличается от обычного асимптотического нормального распределения для случая стационарного регрессора. Действительно, числитель МНК-оценки коэффициента при  $Y_{t-1}$  в регрессии в teste Дики–Фуллера имеет распределение, пропорциональное распределению хи-квадрат (с 1 степенью свободы) минус один.

Эта дискуссия рассматривает только числитель  $T \hat{\delta}$ . Знаменатель также ведет себя необычно в условиях нулевой гипотезы: из-за того что  $Y_t$  следует случайному блужданию в условиях нулевой гипотезы,  $\frac{1}{T} \sum_{t=1}^T Y_{t-1}^2$  не сходится по вероятности к константе. Вместо этого знаменатель в выражении (16.19) является случайной величиной даже в больших выборках: в условиях нулевой гипотезы для  $\frac{1}{T^2} \sum_{t=1}^T Y_{t-1}^2$ , так же, как и для числителя, имеет место сходимость по распределению. Необычные распределения числителя и знаменателя в выражении (16.19) являются источником нестандартного распределения тестовой статистики Дики–Фуллера и причиной того, что ADF-статистика имеет свои собственные таблицы критических значений.

## 16.4. Коинтеграция

Иногда два или более временных ряда имеют общий стохастический тренд. В этом особом случае, называемом коинтеграцией, регрессионный анализ может выявить наличие долгосрочных соотношений между временными рядами, но для этого необходимо знать некоторые новые методы.

### Коинтеграция и коррекция ошибками

Два или более временных ряда, содержащих стохастические тренды, могут двигаться вместе, находясь близко друг к другу на протяжении длительного времени так, что кажется, будто они имеют одинаковый тренд или *общий тренд*. Например,

на рисунке 16.2 изображены графики двух процентных ставок по государственным облигациям в США. Одной из ставок является процентная ставка по 90-дневным казначейским векселям США в годовом исчислении ( $R90$ ), а другой – процентная ставка по годовым американским казначейским облигациям ( $R1yr$ ); эти процентные ставки описаны в приложении 16.1. Процентные ставки демонстрируют одинаковые долгосрочные тенденции или тренды: обе были низкими в 1960-х годах, обе росли в 1970-е годы и достигли своего пика в начале 1980-х, а затем обе упали в 1990-х. Кроме того, разность между этими двумя временными рядами ( $R1yr_t - R90_t$ ), которая называется «спредом» между двумя процентными ставками и также изображена на рисунке 16.2, кажется, не содержит тренда. То есть вычитая 90-дневную процентную ставку из годовой процентной ставки, удается устранить тренды, присущие в каждой из двух ставок. Иначе говоря, несмотря на то что эти две процентные ставки различны, судя по всему, они имеют общий стохастический тренд: так как тренды в каждом временном ряде устраниются, если вычесть один временной ряд из другого, оба временных ряда должны иметь одинаковый тренд, то есть они должны иметь общий стохастический тренд.

Говорят, что два или более временных ряда, имеющих общий стохастический тренд, являются коинтегрированными. Формальное определение *коинтеграции* (введенное эконометристом Клайвом Грейнджером (Clive Granger) в 1983 году; см. вставку «Клайв Грейнджер и Роберт Энгл») приведено во вставке «Основные понятия 16.5». В этом разделе мы рассмотрим тест на отсутствие коинтеграции, обсудим оценку коэффициентов при коинтегрированных переменных в регрессии и расскажем, как использовать коинтеграционные соотношения для прогнозирования. Мы будем рассматривать случай только двух переменных,  $X_t$  и  $Y_t$ .



**Рисунок 16.2. Годовая процентная ставка, трехмесячная процентная ставка и «спред» между процентными ставками**

Годовые и трехмесячные процентные ставки имеют общий стохастический тренд. «Спред», или разница, между двумя ставками не образует тренда. Эти две процентные ставки кажутся коинтегрированными.

### Коинтеграция

Предположим, что временные ряды  $X_t$  и  $Y_t$  являются интегрированными первого порядка. Если существует коэффициент  $\theta$  такой, что разность  $Y_t - \theta X_t$  является интегрированной нулевого порядка, то говорят, что  $X_t$  и  $Y_t$  коинтегрированы. Коэффициент  $\theta$  называется *коинтегрирующим коэффициентом*<sup>1</sup>.

Если  $X_t$  и  $Y_t$  коинтегрированы, то они имеют один и тот же – или общий – стохастический тренд. Вычисление разности  $Y_t - \theta X_t$  устраивает этот общий стохастический тренд.

### ОСНОВНЫЕ ПОНЯТИЯ

16.5

**Векторная модель коррекции ошибками.** До сих пор мы устранили стохастический тренд во временном ряде  $Y_t$  типа  $I(1)$ , вычисляя его первые разности  $\Delta Y_t$ ; и тогда мы избегали проблем, возникающих из-за наличия стохастических трендов, используя  $\Delta Y_t$  вместо  $Y_t$  в регрессиях. Однако если  $X_t$  и  $Y_t$  коинтегрированы, то устранить тренды можно еще одним способом: нужно вычислить  $Y_t - \theta X_t$ , где  $\theta$  выбирается так, чтобы устранить общий тренд из разности. Поскольку разность  $Y_t - \theta X_t$  является стационарной, она может быть использована в регрессионном анализе.

В самом деле, если  $X_t$  и  $Y_t$  коинтегрированы, то первые разности  $X_t$  и  $Y_t$  можно моделировать при помощи VAR-модели, дополненной разностью  $Y_{t-1} - \theta X_{t-1}$  в качестве дополнительного регрессора:

$$\begin{aligned} \Delta Y_t = & \beta_{10} + \beta_{11}\Delta Y_{t-1} + \dots + \beta_{1p}\Delta Y_{t-p} + \gamma_{11}\Delta X_{t-1} + \dots + \gamma_{1p}\Delta X_{t-p} + \\ & + \alpha_1(Y_{t-1} - \theta X_{t-1}) + u_{1t} \end{aligned} \quad (16.22)$$

$$\begin{aligned} \Delta X_t = & \beta_{20} + \beta_{21}\Delta Y_{t-1} + \dots + \beta_{2p}\Delta Y_{t-p} + \gamma_{21}\Delta X_{t-1} + \dots + \gamma_{2p}\Delta X_{t-p} + \\ & + \alpha_2(Y_{t-1} - \theta X_{t-1}) + u_{2t}. \end{aligned} \quad (16.23)$$

Член  $Y_{t-1} - \theta X_{t-1}$  называется *корректирующим членом*. Комбинированная модель, заданная выражениями (16.22) и (16.23), называется *векторной моделью коррекции ошибками* (VECM)<sup>2</sup>. В VECM прошлые значения  $Y_t - \theta X_t$  помогают предсказать будущие значения  $\Delta Y_t$  и / или  $\Delta X_t$ .

### Как выяснить, коинтегрированы ли две случайные величины?

Существует три способа, которые могут помочь принять решение о том, можно ли моделировать две переменные как коинтегрированные: можно использовать экспертные оценки и экономическую теорию, начертить графики временных рядов и убедиться, что они содержат общий стохастический тренд, а также провести статистические тесты на коинтеграцию. Все три метода могут быть использованы на практике.

<sup>1</sup> Или коинтеграционным коэффициентом. – Примеч. науч. ред. перевода.

<sup>2</sup> Отметим, что в русскоязычной литературе также общепринятым считается и другое название – векторная модель коррекции ошибок. – Примеч. науч. ред. перевода.

Во-первых, вы должны использовать свои экспертные знания об этих переменных, чтобы решить, выглядит ли действительно правдоподобно наличие коинтеграции. Например, две процентные ставки, изображенные на рисунке 16.2, связаны между собой так называемой теорией ожиданий временной структуры процентных ставок. Согласно этой теории, процентная ставка по годовым казначейским облигациям на 1 января равна средней процентной ставке по 90-дневным казначейским векселям в первом квартале текущего года и ожидаемым значениям процентной ставки по будущим 90-дневным казначейским векселям во втором, третьем и четвертом кварталах года; а если это не так, то инвесторы могут иметь возможность заработать деньги, удерживая либо годовые облигации Казначейства либо последовательность из четырех 90-дневных казначейских векселей и взвинчивая цены, пока ожидаемые доходности не станут равны. Если 90-дневная процентная ставка является процессом случайного блуждания, то теория предполагает, что годовая процентная ставка также будет содержать этот стохастический тренд и что разность между этими двумя показателями, (т.е. спред) является стационарной. Таким образом, теория ожиданий временной структуры процентных ставок означает, что если процентные ставки являются  $I(1)$ , то они будут коинтегрированы с коэффициентом  $\theta = 1$  (упражнение 16.2).



### ***Нобелевские лауреаты Роберт Энгл и Клайв Грейндженер***

В 2003 году два эконометриста, Роберт Энгл и Клайв Грейндженер, получили Нобелевскую премию по экономике за фундаментальные теоретические исследования в эконометрике временных рядов, которые они сделали в конце 1970-х и начале 1980-х годов.

Работа Грейндженера касалась проблемы устранения стохастических трендов в экономических временных рядах. Из более ранних работ, как собственных, так и других эконометристов, он знал, что два несвязанных временных ряда, содержащих стохастические тренды, могут ошибочно казаться связанными, если использовать обычные статистические критерии, основанные на  $t$ -статистиках и  $R^2$  регрессии (или «ложной»), что называется проблемой «кажущейся» регрессии. В 1970-х годах стандартная практика заключалась в использовании разностей временных рядов для того, чтобы избежать риска возникновения ложной регрессии. По этой причине Грейндженер был скептически настроен относительно недавней работы некоторых британских эконометристов (Davidson, Hendry, Srba and Yeo, 1978), которые утверждали, что лагированная разность между логарифмом потребления и логарифмом дохода ( $\ln C_{t-1} - \ln Y_{t-1}$ ) является ценным предиктором темпа роста потребления ( $\Delta \ln C_t$ ). Из-за того что каждый из  $\ln C_t$  и  $\ln Y_t$  имеет еди-

ничный корень, обычная мудрость заключается в том, что они должны быть включены в регрессию в первых разностях, так как включение их в уровнях приведет к оценке ложной регрессии.

Грейндженер хотел доказать математически, что британская команда совершила ошибку, но вместо этого он доказал, что их спецификация была правильной, потому что существует представление, четко определенное математически, – векторная модель коррекции ошибками – для временных рядов, каждый из которых является  $I(1)$ , но линейная комбинация которых является  $I(0)$ . Он назвал это явление коинтеграцией. В последующей работе со своим коллегой из Калифорнийского университета в Сан-Диего Робертом Энглом Грейндженер предложил несколько тестов на коинтеграцию, в первую очередь ADF-тест Энгла–Грейндженера. Методы анализа коинтеграции теперь являются основными в современной макроэкономике.

Примерно в то же время Роберт Энгл размышлял о впечатляющем росте волатильности инфляции в США в конце 1970-х годов (см. рис. 16.1б). Если волатильность инфляции возросла, рассуждал он, то интервалы прогнозирования инфляции должны быть шире, чем можно получить по дневным данным, потому что в таких моделях (по которым получаются более широкие прогнозные интервалы) дисперсия инфляции постоянна. Но как же именно можно прогнозировать изменяющуюся во времени дисперсию (которую вы не наблюдаете) остаточного члена (который вы также не наблюдаете)?

Ответом Энгла была разработка авторегрессионных моделей условной гетероскедастичности (ARCH), описанных в разделе 16.5. Модель ARCH и ее расширения, разработанные главным образом Энглом и его учениками, оказались особенно полезными для моделирования волатильности доходностей финансовых инструментов, и получаемые прогнозы волатильности могут быть использованы для моделирования цен финансовых деривативов и оценки изменений во времени рисков владения финансовыми активами. Сегодня измерение и прогнозы волатильности являются одним из основных компонентов финансовой эконометрики, а ARCH-модель и ее расширения являются рабочей лошадкой среди инструментов, используемых для моделирования волатильности.



Во-вторых, визуальный анализ графика временного ряда помогает определить случаи, в которых наличие коинтеграции выглядит правдоподобным. Например, графики двух процентных ставок на рисунке 16.2 показывают, что каждый из рядов выглядит как  $I(1)$ , но спред выглядит как  $I(0)$ , поэтому эти два ряда кажутся коинтегрированными.

В-третьих, процедуры тестирования наличия единичного корня, введенные выше, могут быть расширены для тестирования коинтеграции. Основная идея,

на которой основаны эти тесты, заключается в том, что если  $X_t$  и  $Y_t$  коинтегрированы с коэффициентом  $\theta$ , то их линейная комбинация  $Y_t - \theta X_t$  стационарна; в противном случае  $Y_t - \theta X_t$  является нестационарной [является  $I(1)$ ]. Следовательно, гипотеза о том, что  $X_t$  и  $Y_t$  не коинтегрированы [т.е. что  $Y_t - \theta X_t$  является  $I(1)$ ], может быть проверена при помощи тестирования нулевой гипотезы о том, что временной ряд  $Y_t - \theta X_t$  имеет единичный корень; если эта гипотеза отвергается, то  $X_t$  и  $Y_t$  могут быть смоделированы как коинтегрированные. Детали такого теста зависят от того, известен или нет коинтегрирующий коэффициент  $\theta$ .

**Тестирование коинтеграции, если  $\theta$  известен.** В некоторых случаях экспертные оценки, или экономическая теория, предполагает знание значений  $\theta$ . Если  $\theta$  известен, то тест Дики–Фуллера и DF-GLS-тест на наличие единичного корня могут быть использованы для тестирования на коинтеграцию, создавая сначала временной ряд  $z_t = Y_t - \theta X_t$ , а затем проверяя нулевую гипотезу о том, что  $z_t$  содержит авторегрессионный единичный корень.

**Тестирование коинтеграции, если  $\theta$  неизвестен.** Если коинтеграционный коэффициент  $\theta$  неизвестен, то он должен быть оценен до начала тестирования на наличие единичного корня корректирующего члена. Этот предварительный этап вынуждает нас использовать иные критические значения в последующих тестах на наличие единичного корня.

В частности, на первом этапе коинтегрирующий коэффициент оценивается при помощи МНК-регрессии:

$$Y_t = \alpha + \theta X_t + z_t. \quad (16.24)$$

На втором этапе для тестирования на наличие единичного корня в остатках этой регрессии  $z_t$  может быть использован  $t$ -тест Дики–Фуллера (с константой, но без временного тренда). Эта двухступенчатая процедура называется тестом Энгла–Грейнджа на коинтеграцию, основанным на расширенном teste Дики–Фуллера, или *EG-ADF-тестом* (Engle, Granger, 1987).

Критические значения EG-ADF-статистики приведены в таблице 16.2<sup>1</sup>. Критические значения в первой строке используются, если в регрессии (16.24) существует один регрессор, и поэтому существуют две коинтегрированные переменные ( $X_t$  и  $Y_t$ ). Следующие строки таблицы используются в случае нескольких коинтегрированных переменных, который обсуждается в конце этого раздела.

Таблица 16.2

## Критические значения ADF-статистики теста Энгла–Грейнджа

Число $X'$ в уравнении (16.24)	10 %	5 %	1 %
1	-3,12	-3,41	-3,96
2	-3,52	-3,80	-4,36
3	-3,84	-4,16	-4,73
4	-4,20	-4,49	-5,07

<sup>1</sup> Критические значения, представленные в таблице 16.2, взяты из работ Фуллера (Fuller, 1976) и Филиппса и Улиариса (Phillips, Ouliaris, 1990). Следуя предположению Хансена (Hansen, 1992), критические значения в таблице 16.2 выбраны так, что они могут быть использованы независимо от того, содержат они или нет дрейф  $X_t$  и  $Y_t$ .

## Оценка коэффициентов коинтеграционного соотношения

Если  $X_t$  и  $Y_t$  коинтегрированы, то МНК-оценка коэффициента в коинтеграционной регрессии (16.24) является состоятельной. Однако в общем случае МНК-оценка имеет ненормальное распределение, и выводы, основанные на ее  $t$ -статистиках, могут быть некорректными независимо от того, вычисляются ли эти  $t$ -статистики с использованием НАС-стандартных ошибок или нет. Из-за этих недостатков МНК-оценки  $\theta$  эконометристы разработали ряд других оценок коинтегрирующего коэффициента.

Одной из таких оценок  $\theta$ , которая проста в практическом использовании, является *оценка динамического МНК (DOLS)* (Stock, Watson, 1993). DOLS-оценка основана на модифицированной версии регрессии (16.24), которая включает прошлые, настоящее и будущие значения изменений  $X_t$ :

$$Y_t = \beta_0 + \theta X_t + \sum_{j=-p}^p \delta_j \Delta X_{t-j} + u_t. \quad (16.25)$$

Таким образом, в уравнении (16.25) регрессорами являются  $X_t$ ,  $\Delta X_{t+p}$ , ...,  $\Delta X_{t-p}$ . DOLS-оценкой коэффициента  $\theta$  является МНК-оценка  $\theta$  в регрессии (16.25).

Если  $X_t$  и  $Y_t$  коинтегрированы, то оценка DOLS эффективна в больших выборках. Более того, статистические выводы о  $\theta$  и  $\delta'$  в регрессии (16.25), основанные на НАС-стандартных ошибках, являются обоснованными. Например,  $t$ -статистика, построенная с помощью DOLS-оценки с НАС-стандартными ошибками, имеет в больших выборках стандартное нормальное распределение.

Для того чтобы интерпретировать регрессию (16.25), можно вспомнить из раздела 15.3 о том, что совокупный динамический мультиплитатор может быть вычислен путем изменения модели регрессии с распределенными лагами  $Y_t$  от  $X_t$  и его запаздываний. Более точно, в уравнении (15.7) совокупные динамические мультиплитаторы были вычислены при помощи оценки регрессии  $Y_t$  на  $\Delta X_p$ , лаги  $\Delta X_t$  и  $X_{t-r}$ ; коэффициент при  $X_{t-r}$  в этой спецификации является долгосрочным совокупным динамическим мультиплитатором. Аналогично, если  $X_t$  строго экзогенен, то в уравнении (16.25) коэффициент при  $X_t - \theta$  будет долгосрочным динамическим мультиплитатором, то есть будет характеризовать долгосрочное влияние на  $Y$  от изменения  $X$ . Если  $X_t$  не является строго экзогенным, то коэффициенты не имеют такой интерпретации. Тем не менее, из-за того что  $Y_t$  и  $X_t$  имеют общий стохастический тренд, если они коинтегрированы, то оценки DOLS являются состоятельными, даже если регрессор  $X_t$  эндогенен.

Оценка DOLS не является единственной эффективной оценкой коинтегрирующего коэффициента. Первая такая оценка была разработана Сореном Йохансеном (Johansen, 1988). Для знакомства с методом Йохансена и другими способами оценки коинтегрирующего коэффициента см. книгу Гамильтона (Hamilton, 1994, гл. 20).

Даже если экономическая теория не позволяет сделать каких-либо предположений о конкретном значении коинтеграционного коэффициента, важно

проверить, имеет ли предполагаемая коинтеграционная связь смысл на практике. Вследствие того что тесты на коинтеграцию могут приводить к некорректным статистическим выводам (они могут неправильно отвергать нулевую гипотезу об отсутствии коинтеграции чаще, чем следовало бы, и часто они ошибочно ее не отвергают), особенно важно опираться на экономическую теорию, институциональные знания и здравый смысл при оценке и использовании коинтеграционных соотношений.

### ***Расширение на случай нескольких коинтегрированных переменных***

Все, что обсуждалось выше (определения, тесты и оценки), может быть распространено на случай более чем двух переменных. Например, если имеется три переменные  $Y_t, X_{1t}$  и  $X_{2t}$ , каждая из которых является  $I(1)$ , то они будут коинтегрированы с коинтегрирующими коэффициентами  $\theta_1$  и  $\theta_2$ , если временной ряд  $Y_t - \theta_1 X_{1t} - \theta_2 X_{2t}$  стационарен. При наличии трех или более переменных может существовать несколько коинтеграционных соотношений. Рассмотрим, например, моделирование отношений между тремя процентными ставками: трехмесячной ставкой, годовой ставкой и пятилетней ставкой ( $R3yr$ ). Если они  $I(1)$ , то теория ожиданий временной структуры процентных ставок предполагает, что все они будут коинтегрированы. Одно коинтеграционное соотношение, предлагаемое теорией, имеет вид:  $R1yr_t - R90_t$ , а второе соотношение —  $R5yr_t - R90_t$ . (Соотношение  $R5yr_t - R1yr_t$  также является коинтеграционным, но оно не содержит никакой дополнительной информации, кроме той, что присутствует в других соотношениях, потому что оно совершенно мультиколлинеарно с двумя другими коинтеграционными соотношениями.)

EG-ADF-процедура для тестирования наличия (точнее, отсутствия) одного коинтеграционного соотношения между несколькими переменными ничем не отличается от случая двух переменных, за исключением того, что регрессия (16.24) модифицируется так, чтобы оба  $X_{1t}$  и  $X_{2t}$  были регрессорами; критические значения EG-ADF-теста приведены в таблице 16.2, где соответствующие строки зависят от количества регрессоров, включенных в коинтеграционное соотношение, оцениваемое на первом шаге при помощи МНК-регрессии. Оценка DOLS для одного коинтеграционного соотношения между несколькими  $X$ ’ми включает оценку регрессии от уровня каждого  $X$  и их опережающих и запаздывающих разностей. Тесты для нескольких коинтеграционных соотношений могут быть проведены с использованием системы методов, таких как метод Йохансена (Johansen, 1988), а DOLS-оценка может быть распространена на случай нескольких коинтеграционных соотношений при помощи оценки нескольких уравнений по одному для каждого коинтеграционного соотношения. Дополнительное обсуждение коинтеграционных методов для нескольких переменных см.: Hamilton (1994).

**Предостережение.** Если две или более переменных коинтегрированы, то корректирующий член может помочь прогнозировать эти переменные и, возможно, другие связанные с ними переменные. Тем не менее коинтеграция требует, чтобы переменные содержали одинаковые стохастические тренды. Тренды в экономических переменных, как правило, возникают из-за сложного взаимодействия

разнородных сил, и тесно связанные временные ряды могут иметь различные тренды по очень специфическим причинам. Если переменные, не являющиеся коинтегрированными, неправильно моделируются при помощи VECM, то коинтегрирующий член будет являться  $I(1)$ ; это приводит к введению тренда в прогноз, из чего может следовать низкое качество вневыборочного прогноза. Таким образом, использование VECM для целей прогнозирования должно быть основано на сочетании убедительных теоретических аргументов в пользу коинтеграции при тщательном эмпирическом анализе.

### **Приложение к ставкам процента**

Как обсуждалось ранее, теория ожиданий временной структуры процентных ставок означает, что если две процентные ставки с различными сроками погашения являются  $I(1)$ , то они будут коинтегрированы с коинтегрирующим коэффициентом  $\theta = 1$ , то есть спред между двумя ставками будет стационарным. Анализ рисунка 16.2 качественно подтверждает гипотезу о том, что годовая и трехмесячная процентные ставки коинтегрированы. Сначала мы используем тесты на наличие единичного корня и коинтеграцию, чтобы формально обосновать доказательство этой гипотезы, а затем оценим векторную модель коррекции ошибками для этих двух процентных ставок.

**Тесты на единичные корни и коинтеграцию.** Статистики различных тестов на наличие единичных корней и коинтеграцию для этих двух временных рядов приведены в таблице 16.3. Тестовые статистики тестов на единичный корень, расположенные в первых двух строках, рассматривают нулевую гипотезу о том, что две процентные ставки, трехмесячная ставка ( $R90$ ) и годовая ставка ( $R1_{yr}$ ), содержат по одному единичному корню. Две из четырех статистик в первых двух строках не в состоянии отвергнуть эту гипотезу на уровне 10 %, а три из четырех не в состоянии отвергнуть на уровне 5 %. Исключением является ADF-статистика для 90-дневных казначейских векселей ( $-2,96$ ), которая отвергает гипотезу о наличии единичного корня на уровне значимости 5 %. ADF- и DF-GLS-статистики приводят к разным выводам для этой переменной.

Таблица 16.3

#### **Тестовые статистики тестов на единичные корни и коинтеграцию для двух ставок процента**

Временной ряд	ADF-статистика	DF-GLS-статистика
$R90$	$-2,96^*$	$-1,88$
$R1_{yr}$	$-2,22$	$-1,37$
$R1_{yr} - R90$	$-6,31^{**}$	$-5,59^{**}$
$R1_{yr} - 1,046 R90$	$-6,97^{**}$	—

*Примечание.*  $R90$  — ставка процента по 90-дневным казначейским векселям в годовом исчислении и  $R1_{yr}$  — ставка процента по годовым казначейским облигациям в США. Регрессии были оценены на основе квартальных данных с I квартала 1962 по IV квартал 1999 года. Число запаздываний в тестах на единичные корни выбиралось при помощи критерия Акаике (с максимальным допустимым числом лагов, равным 6). Тестовые статистики тестов на единичные корни значимы на уровнях значимости \*5 % или \*\*1 %.

(ADF-тест отвергает гипотезу о наличии единичного корня на уровне значимости 5 %, в то время как DF-GLS-тест – нет), что означает, что мы должны каким-то образом обосновать, почему мы считаем правдоподобным моделировать этих переменные как  $I(1)$ . Взятые вместе, эти результаты показывают, что рассмотрение процентных ставок как процессов типа  $I(1)$  выглядит правдоподобным.

Тестовая статистика тестов на единичные корни для спреда  $R1yr_t - R90_t$ , проверяет гипотезу о том, что эти переменные не коинтегрированы, против альтернативы о том, что коинтеграция есть. Нуевая гипотеза о том, что спред содержит единичный корень, отвергается на 1%-м уровне значимости с использованием обоих тестов на единичный корень. Таким образом, мы отвергаем гипотезу о том, что временные ряды не являются коинтегрированными против альтернативной о том, что они коинтегрированы с коинтегрирующим коэффициентом  $\theta = 1$ . Рассмотренные вместе, результаты первых трех строк таблицы 16.3 предполагают, что эти переменные могут быть смоделированы как коинтегрированные с  $\theta = 1$ .

Вследствие того что в этом примере экономическая теория предполагает известное значение  $\theta$  (теория ожиданий временной структуры предполагает, что  $\theta = 1$ ) и поскольку корректирующий член является  $I(0)$ , когда это ограничение накладывается (спред является стационарным), в принципе не стоит использовать EG-ADF-тест, в котором  $\theta$  нужно оценивать. Тем не менее мы проведем расчеты в качестве иллюстрации. Первым шагом в EG-ADF-тесте является оценка  $\theta$  при помощи МНК-регрессии одной переменной на другую, в результате получаем:

$$\widehat{R1yr}_t = 0,361 + 1,046 R90_t, \quad \overline{R^2} = 0,973. \quad (16.26)$$

Второй шаг состоит в вычислении ADF-статистики для остаточного члена  $\hat{z}_t$  из этой регрессии. Результат, приведенный в последней строке таблицы 16.3, меньше, чем 1 %-е критическое значение, равное  $-3,96$ , из таблицы 16.2, так что нулевая гипотеза о том, что  $\hat{z}_t$  содержит авторегрессионный единичный корень, отвергается. Эта статистика также указывает на то, что эти две процентные ставки коинтегрированы. Следует отметить, что стандартные ошибки не представлены в уравнении (16.26), потому что, как обсуждалось ранее, МНК-оценка коинтегрирующего коэффициента имеет ненормальное распределение и ее  $t$ -статистика распределена ненормально, поэтому представление стандартных ошибок (НАС или других) может привести к некорректным выводам.

**Векторная модель коррекции ошибками для двух ставок процента.** Если  $Y_t$  и  $X_t$  являются коинтегрированными, то прогнозы  $\Delta Y_t$  и  $\Delta X_t$  могут быть улучшены расширением VAR для  $\Delta Y_t$  и  $\Delta X_t$  при помощи включения значения корректирующего члена, то есть вычисляя прогнозы с помощью VECM, задаваемой уравнениями (16.22) и (16.23). Если  $\theta$  известен, то неизвестные коэффициенты VECM могут быть оценены с помощью МНК, включая  $z_{t-1} = Y_{t-1} - \theta X_{t-1}$  в качестве дополнительного регрессора. Если  $\theta$  неизвестен, тогда VECM можно оценить, используя  $\hat{z}_t$  как регрессор, где  $\hat{z}_t = Y_t - \hat{\theta} X_t$  и  $\hat{\theta}$  является оценкой  $\theta$ .

В приложении к двум процентным ставкам теория предполагает, что  $\theta = 1$ , и тесты на наличие единичного корня подтверждают, что две процентные ставки можно моделировать как коинтегрированные с коинтегрирующим коэффициентом, равным единице. Поэтому мы рассмотрим VECM, используя предлагаемое теорией значение  $\theta = 1$ , то есть добавляя запаздывающее значение спреда  $R1yr_{t-1} - R90_{t-1}$  в VAR для  $\Delta R1yr_t$  и  $\Delta R90_t$ . Рассматривая VECM с двумя запаздывающими разностями, получаем:

$$\begin{aligned}\widehat{\Delta R90}_t &= 0,14 - 0,24 \Delta R90_{t-1} - 0,44 \Delta R90_{t-2} - 0,01 \Delta R1yr_{t-1} + \\ &+ 0,15 \Delta R1yr_{t-2} - 0,18 (R1yr_{t-1} - R90_{t-1})\end{aligned}\quad (16.27)$$

$$\begin{aligned}\widehat{\Delta R1yr}_t &= 0,36 - 0,14 \Delta R90_{t-1} - 0,33 \Delta R90_{t-2} - 0,11 \Delta R1yr_{t-1} + \\ &+ 0,10 \Delta R1yr_{t-2} - 0,52 (R1yr_{t-1} - R90_{t-1}).\end{aligned}\quad (16.28)$$

В первом уравнении ни один из коэффициентов не является значимым на уровне значимости 5 %, а коэффициенты при лагах первых разностей процентных ставок не являются совместно значимыми на уровне значимости 5 %. Во втором уравнении коэффициенты при первых запаздывающих разностях не являются совместно значимыми, но коэффициент при запаздывающем спреде (корректирующем члене), который, по оценкам равен  $-0,52$ , имеет  $t$ -статистику, равную  $-2,17$ , так что является статистически значимым на уровне значимости 5 %. Несмотря на то что запаздывающие значения первых разностей процентных ставок не являются полезными для прогнозирования будущих процентных ставок, запаздывание спреда действительно помогает предсказать изменения годовой процентной ставки по казначейским облигациям. Когда годовая ставка процента превышает 90-дневную ставку процента, прогнозируется, что годовая ставка процента будет падать в будущем.

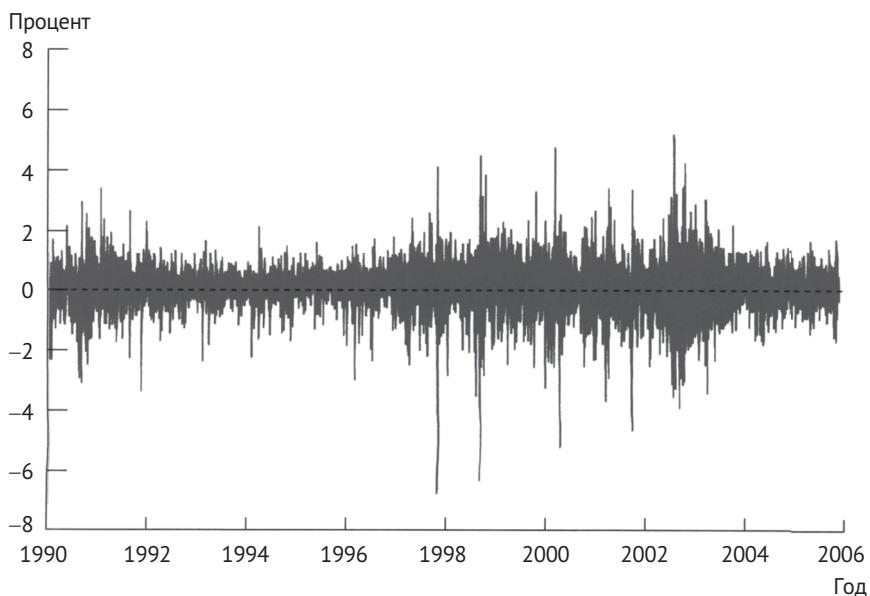
## 16.5. Кластеризованная волатильность и авторегрессионные модели условной гетероскедастичности

Феномен, выражющийся в том, что в одни периоды времени временные ряды изменяются не сильно, а в другие, напротив, сильно, то есть кластеризованная волатильность наблюдается во многих экономических временных рядах. В этом разделе мы рассматриваем пару моделей, позволяющих измерить кластеризованную волатильность (или условную гетероскедастичность) количественно.

### Кластеризованная волатильность

Волатильность многих финансовых и макроэкономических переменных изменяется во времени. Например, ежедневные процентные изменения сводного индекса Нью-Йоркской фондовой биржи (NYSE), изображенные на рисунке 16.3,

демонстрируют периоды высокой волатильности, такие как в 1990 и 2003 годах, а также другие периоды – периоды низкой волатильности, например в 1993 году. Временные ряды с некоторыми периодами низкой волатильности и некоторыми периодами высокой волатильности, как принято полагать, характеризуются *кластеризованной волатильностью*. Поскольку волатильность появляется в кластерах, дисперсию ежедневного процентного изменения индекса NYSE можно прогнозировать, даже несмотря на то что изменение цены прогнозировать очень трудно.



**Рисунок 16.3. Ежедневные процентные изменения сводного индекса Нью-Йоркской фондовой биржи, 1990–2005 года**

Ежедневные процентные изменения индекса Нью-Йоркской фондовой биржи (NYSE) демонстрируют кластеризованную волатильность, в которой есть некоторые периоды высокой волатильности, например в конце 1990-х годов, и другие периоды относительного спокойствия, например в середине 1990-х годов.

Прогнозирование дисперсии временных рядов представляет интерес по нескольким причинам. Во-первых, дисперсия цен на активы является мерой риска владения этим активом: чем больше дисперсия ежедневного изменения цены акции, тем больше участников фондового рынка заработают (или потеряют) в обычный день. Инвестор, обеспокоенный наличием риска, скорее всего, воздержался бы от игры на фондовой бирже в период высокой (но не низкой) волатильности.

Во-вторых, стоимость некоторых производных финансовых инструментов, таких как опционы, зависит от дисперсии базового актива. Трейдер, торгующий опционами, хочет знать наилучшие доступные прогнозы волатильности в будущем, чтобы они помогли ему или ей узнать цену, по которой следует покупать или продавать опционы.

В-третьих, прогнозирование дисперсии делает возможным построение аккуратных интервальных прогнозов. Предположим, что вы прогнозируете

уровень инфляции. Если дисперсия ошибки прогноза постоянна, то приблизительный прогнозный доверительный интервал может быть построен так, как обсуждалось в разделе 14.4, то есть как значение прогноза плюс или минус кратное  $SE$ . Однако если дисперсия ошибки прогноза изменяется во времени, то ширина интервального прогноза должна меняться во времени: в периоды, когда инфляция подвергается особо сильным шокам, этот интервал должен быть широким, в периоды относительного спокойствия интервал должен быть более узким.

Кластеризованную волатильность можно рассматривать как кластеризацию дисперсии остаточного члена во времени: если ошибка регрессии имеет небольшую дисперсию в один период, его дисперсия, как правило, будет небольшой и в следующий период. Другими словами, кластеризованная волатильность означает, что ошибка характеризуется изменяющейся во времени гетероскедастичностью.

### **Авторегрессионные модели условной гетероскедастичности**

Двумя моделями кластеризованной волатильности являются *авторегрессионная модель условной гетероскедастичности (ARCH)* и ее расширение – *обобщенная ARCH (GARCH)-модель*.

**ARCH.** Рассмотрим модель регрессии ADL (1, 1):

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_1 X_{t-1} + u_t. \quad (16.29)$$

В ARCH-модели, разработанной эконометристом Робертом Энглом (Engle, 1982; см. вставку «Клайв Грейндже и Роберт Энгл»), ошибка  $u_t$  моделируется как нормально распределенная с нулевым средним и дисперсией  $\sigma_t^2$ , где  $\sigma_t^2$  зависит от квадрата прошлого значения  $u_{t-1}$ . В частности, модель ARCH-порядка  $p$ , которая обозначается как ARCH( $p$ ), имеет вид:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \dots + \alpha_p u_{t-p}^2, \quad (16.30)$$

где  $\alpha_0, \alpha_1, \dots, \alpha_p$  – неизвестные коэффициенты. Если эти коэффициенты являются положительными и квадраты недавних ошибок принимают большие значения, то прогноз по ARCH-модели говорит о том, что квадрат текущей ошибки будет большим по величине, в том смысле что ее дисперсия  $\sigma_t^2$  является большой.

Несмотря на то что здесь мы используем для описания ADL(1, 1)-модель, задаваемую уравнением (16.29), модель ARCH может быть применена к дисперсии ошибки любой модели регрессии временных рядов с ошибкой, которая имеет нулевое условное среднее, в том числе к модели ADL более высокого порядка, авторегрессиям и моделям временных рядов с несколькими регрессорами.

**GARCH.** Обобщенная ARCH (GARCH)-модель, разработанная эконометристом Тимом Боллерслев (Bollerslev, 1986), расширяет ARCH-модель, позволяя  $\sigma_t^2$  зависеть от ее собственных лагов, а также лагов квадрата ошибки. GARCH ( $p, q$ ) -модель имеет вид:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \dots + \alpha_p u_{t-p}^2 + \phi_1 \sigma_{t-1}^2 + \dots + \phi_q \sigma_{t-q}^2, \quad (16.31)$$

где  $\alpha_0, \alpha_1, \dots, \alpha_p, \phi_1, \dots, \phi_q$  – некоторые неизвестные коэффициенты.

Модель ARCH аналогична модели регрессии с распределенными лагами, а GARCH-модель аналогична модели ADL. Как уже говорилось в приложении 15.2, ADL-модель (при необходимости) дает более простую модель динамических мультипликаторов, чем модель регрессии с распределенными лагами. Аналогичным образом, включая лаги  $\sigma_t^2$ , GARCH-модель может учесть медленно меняющиеся отклонения при помощи меньшего числа параметров, чем ARCH-модели.

Важным приложением ARCH- и GARCH-моделей является измерение и прогнозирование изменяющейся во времени волатильности доходностей финансовых активов, в частности активов, наблюдаемых с высокой частотностью, таких как ежедневные доходности акций на рисунке 16.3. В таких случаях сама доходность часто моделируется как непредсказуемая, поэтому регрессия (16.29) включает в себя только константу.

**Оценка и качество.** ARCH- и GARCH-модели оцениваются методом максимального правдоподобия (приложение 11.2). Оценки ARCH- и GARCH-коэффициентов нормально распределены в больших выборках, поэтому в больших выборках  $t$ -статистики имеют стандартное нормальное распределение, и доверительные интервалы могут быть построены как оценка максимального правдоподобия  $\pm 1,96$  от стандартной ошибки.

### Приложение к волатильности (колебаниям) цен акций

Модель GARCH(1,1) сводного индекса NYSE, характеризующего ежедневное процентное изменение курсов акций  $R_t$ , оценивается с использованием данных по всем торговым дням с 2 января 1990 по 11 ноября 2005 года и имеет вид:

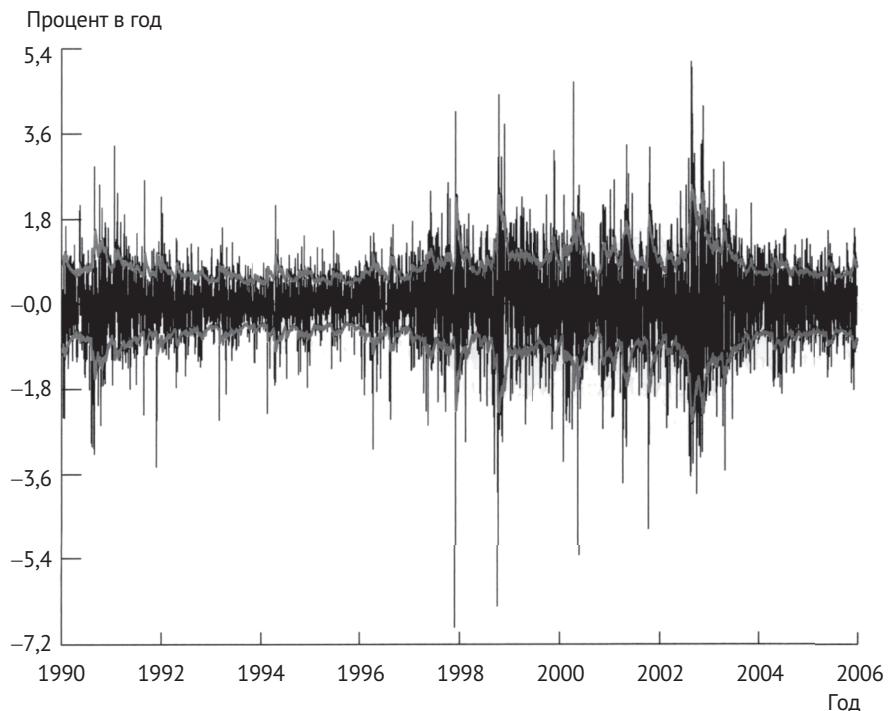
$$\hat{R}_t = 0,049; \quad (0,012) \quad (16.32)$$

$$\hat{\sigma}_t^2 = 0,0079 + 0,072 u_{t-1}^2 + 0,919 \sigma_{t-1}^2. \quad (0,0014) \quad (0,005) \quad (0,006) \quad (16.33)$$

В уравнении (16.32) нет лагов регрессоров, потому что ежедневные изменения индекса NYSE являются по существу непредсказуемыми. Два коэффициента в модели GARCH (коэффициенты при  $u_{t-1}$  и  $\sigma_{t-1}^2$ ) статистически значимы на уровне значимости 5 %. Одной из мер устойчивости дисперсии является сумма коэффициентов при  $u_{t-1}$  и  $\sigma_{t-1}^2$  в модели GARCH (упражнение 16.9). Эта сумма (0,991) велика, что указывает на то, что изменения в условной дисперсии являются постоянными. Говоря иначе, оценки модели GARCH означают, что периоды высокой волатильности индекса NYSE будут длительными. Этот вывод согласуется с длительными периодами кластеризованной волатильности, которые мы видели на рисунке 16.3.

Оцененная условная дисперсия в момент  $t - \hat{\sigma}_t^2$  может быть вычислена с использованием остатков уравнения (16.32) и коэффициентов из уравнения (16.33). На рисунке 16.4 изображены границы, равные плюс или минус одно условное стандартное отклонение (т.е.  $\pm \hat{\sigma}_t$ ), вычисленное на основе GARCH(1,1)-модели, вместе с отклонениями процентного изменения сводного индекса от своего

среднего. Границы условного стандартного отклонения количественно измеряют изменяющуюся во времени волатильность ежедневных изменений цен. В середине 1990-х годов границы условного стандартного отклонения близки друг к другу, что указывает на низкий риск для инвесторов, владеющих акциями компаний, включаемых в индекс NYSE. В противоположность этому на рубеже веков границы этого условного стандартного отклонения широкие, что говорит о периоде больших ежедневных изменений цен акций.



**Рисунок 16.4. Границы ежедневных процентных изменений сводного индекса Нью-Йоркской фондовой биржи и GARCH (1, 1)**

GARCH (1, 1)-границы... – полосы, рассчитанные по модели GARCH (1,1), которые равны  $\pm \hat{\sigma}_t$ , где  $\hat{\sigma}_t$  вычисляется с помощью уравнения (16.33), являются узкими, когда условная дисперсия мала, и широкими, когда она является большой. Условная волатильность изменений цен акций существенно меняется в период 1990–2005 годов.

## 16.6. Заключение

В данной части книги мы рассмотрели некоторые из наиболее часто используемых инструментов и понятий регрессии временных рядов. Помимо этих инструментов разработано множество других инструментов, используемых для анализа экономических временных рядов в конкретных приложениях. Если вы хотите ознакомиться с дополнительной информацией об экономическом прогнозировании, см. учебники вводного уровня Эндерса (Enders, 1995) и Диболда (Diebold, 2007). Для более продвинутого изучения эконометрики временных рядов см. Hamilton (1994).

## Выводы

1. В модели векторной авторегрессии «вектора», состоящего из  $k$  временных рядов, каждый временной ряд зависит от своих собственных запаздываний и лагов  $k-1$  других временных рядов. Прогнозы каждого из временных рядов, рассчитанные по VAR, являются согласованными в том смысле, что они основаны на одинаковой информации.
2. Прогнозы на два и более периодов вперед можно вычислить, либо итеративно переоценивая модели AR или VAR и строя одношаговые прогнозы, либо как многошаговый прогноз, построенный на основе одной регрессии.
3. Два временных ряда, содержащих общий стохастический тренд, являются коинтегрированными, то есть  $X_t$  и  $Y_t$  коинтегрированы, если  $X_t$  и  $Y_t$  являются  $I(1)$ , но  $Y_t - \theta X_t$  является  $I(0)$ . Если  $X_t$  и  $Y_t$  коинтегрированы, то корректирующий член  $Y_t - \theta X_t$  может помочь предсказать  $\Delta Y_t$  и/или  $\Delta X_t$ . Векторная модель коррекции ошибками представляет собой VAR-модель  $\Delta Y_t$  и  $\Delta X_t$ , расширенную включением первого запаздывания корректирующего члена.
4. Кластеризованная волатильность, то есть ситуация, когда дисперсия временного ряда велика в некоторые периоды времени и низка в другие, распространена в экономических временных рядах, особенно в финансовых временных рядах.
5. ARCH-модель кластеризованной волатильности выражает условную дисперсию ошибки регрессии как зависимость от запаздываний квадратов ошибок регрессии. Модель GARCH является расширением ARCH-модели и дополнительно включает запаздывания условной дисперсии. Оценив ARCH- и GARCH-модели, можно получить интервальные прогнозы с границами, которые зависят от волатильности наиболее близких остатков регрессии.

## Основные понятия

- Векторная авторегрессия (VAR) (с. 665).  
Итеративный многошаговый AR-прогноз (с. 669).  
Итеративный многошаговый VAR-прогноз (с. 671).  
Прямой многошаговый прогноз (с. 672).  
Интегрированный порядка  $d$ ,  $I(d)$  (с. 675).  
Вторая разность (с. 675).  
Интегрированный нулевого порядка [ $I(0)$ ], первого порядка [ $I(1)$ ] или второго порядка [ $I(2)$ ] (с. 676).  
Порядок интегрированности (с. 676).  
DF-GLS-тест (с. 678).  
Общий тренд (с. 681).  
Коинтеграция (с. 681).  
Коинтегрирующий коэффициент (с. 683).  
Корректирующий член (с. 683).  
Векторная модель коррекции ошибками (с. 683).  
EG-ADF-тест (с. 686).

Оценка динамического МНК (DOLS) (с. 687).

Кластеризованная волатильность (с. 691).

Авторегрессионная модель условной гетероскедастичности (ARCH) (с. 693).

Обобщенная ARCH (GARCH) (с. 693).

### **Вопросы для повторения и закрепления основных понятий**

- 16.1. Макроэкономист хочет построить прогнозы для следующих макроэкономических показателей: ВВП, потребления, инвестиций, государственных закупок, экспорта, импорта, краткосрочных процентных ставок, долгосрочных процентных ставок и уровня ценовой инфляции. У него есть ежеквартальные временные ряды для каждой из этих переменных с 1970 по 2010 год. Должен ли он оценить модель VAR для этих переменных и использовать ее для прогнозирования? Почему да или почему нет? Можете ли вы предложить альтернативный подход?
- 16.2. Пусть  $Y_t$  описывается стационарной AR(1)-моделью с  $\beta_0 = 0$  и  $\beta_1 = 0,7$ . Если  $Y_t = 5$ , каков ваш прогноз для  $Y_{t+2}$  (т.е. чему равен  $Y_{t+2|t}$ )? Чему равен  $Y_{t+h|t}$  для  $h = 30$ ? Считаете ли вы этот прогноз для  $h = 30$  разумным?
- 16.3. Теория постоянного дохода потребителя предполагает, что логарифм реального ВВП ( $Y$ ) и логарифм реального потребления ( $C$ ) коинтегрированы с коинтегрирующим коэффициентом, равным единице. Объясните, как вы проверяли бы это утверждение (a) при помощи графического анализа данных и (б) с помощью статистических тестов.
- 16.4. Рассмотрим ARCH-модель  $\sigma_t^2 = 1,0 + 0,8u_{t-1}^2$ . Объясните, почему эта модель описывает кластеризованную волатильность? (Подсказка: что происходит, когда  $u_{t-1}^2$  является необычно большим?)
- 16.5. DF-GLS-тест на единичный корень имеет более высокую мощность, чем тест Дики–Фуллера. Почему следует использовать более мощный тест?

### **Упражнения**

- 16.1. Пусть  $Y_t$  задается стационарной AR(1)-моделью  $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$ .
  - a) Покажите, что прогноз  $Y_t$  на  $h$  шагов вперед рассчитывается по формуле:  $Y_{t+h|t} = \mu_Y + \beta_1^h (Y_t - \mu_Y)$ , где  $\mu_Y = \beta_0(1 - \beta_1)$ .
  - b) Предположим, что  $X_t$  выражается через  $Y_t$  как  $X_t = \sum_{i=0}^{\infty} \delta^i Y_{t+i|t}$ , где  $|\delta| < 1$ . Покажите, что  $X_t = \mu_Y / (1 - \delta) + (Y_t - \mu_Y) / (1 - \beta_1 \delta)$ .
- 16.2. Теория ожиданий временной структуры процентных ставок утверждает, что долгосрочная ставка процента равна среднему значению ожидаемых краткосрочных процентных ставок в будущем плюс некоторая премия, являющаяся  $I(0)$ . Более точно, пусть  $Rk_t$  обозначает процентную ставку на  $k$  периодов, пусть  $R1_t$  обозначает процентную ставку на один период, и пусть  $e_t$  обозначает премию, которая является  $I(0)$ . Обозначим  $Rk_t = \frac{1}{k} \sum_{i=0}^{k-1} R1_{t+i|t} + e_t$ , где  $R1_{t+i|t}$  – прогноз  $R1$ , сделанный в момент времени  $t$  на момент  $t+1$ . Предположим, что  $R1_t$  описывается моделью случайного блуждания, так что:  $R1_t = R1_{t-1} + u_t$ .

- a) Покажите, что  $Rk_t = Rl_t + e_t$ .
- б) Покажите, что  $Rk_t$  и  $Rl_t$  коинтегрированы. Чему равен коинтегрирующий коэффициент?
- в) Теперь предположим, что  $\Delta Rl_t = 0,5\Delta Rl_{t-1} + u_t$ . Как изменится ваш ответ на вопрос (б)?
- г) Теперь предположим, что  $Rl_t = 0,5Rl_{t-1} + u_t$ . Как изменится ваш ответ на вопрос (б) в этом случае?
- 16.3. Предположим, что  $u_t$  является ARCH-процессом вида  $\sigma_t^2 = 1,0 + 0,5u_{t-1}^2$ .
- а) Пусть  $E(u_t^2) = \text{var}(u_t)$  обозначает безусловную дисперсию  $u_t$ . Покажите, что  $\text{var}(u_t) = 2$ . [Подсказка: используйте закон повторного математического ожидания  $E(u_t^2) = E[E(u_t^2 | u_{t-1})]$ ].
- б) Предположим, что условное распределение  $u_t$  относительно своих запаздывающих значений является  $N(0, \sigma_t^2)$ . Чему равна вероятность  $\Pr(-3 \leq u_t \leq 3)$ , если  $u_{t-1} = 0,2$ ? Чему равна вероятность  $\Pr(-3 \leq u_t \leq 3)$ , если  $u_{t-1} = 2,0$ ?
- 16.4. Предположим, что  $Y_t$  задается стационарной AR( $p$ )-моделью  $Y_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} + u_t$ , где  $E(u_t | Y_{t-1}, Y_{t-2}, \dots) = 0$ . Пусть  $Y_{t+h|t} = E(Y_{t+h} | Y_t, Y_{t-1}, \dots)$ . Покажите, что  $Y_{t+h|t} = \beta_0 + \beta_1 Y_{t-1+h|t} + \dots + \beta_p Y_{t-p+h|t}$  для  $h > p$ .
- 16.5. Докажите выражение (16.20). [Подсказка: используйте  $\sum_{t=1}^T Y_t^2 = \sum_{t=1}^T (Y_{t-1} + \Delta Y_t)^2$ , чтобы показать, что  $\sum_{t=1}^T Y_t^2 = \sum_{t=1}^T Y_{t-1}^2 + 2 \sum_{t=1}^T Y_{t-1} \Delta Y_t + \sum_{t=1}^T \Delta Y_t^2$ , и решите его для  $\sum_{t=1}^T Y_{t-1} \Delta Y_t$ .]
- 16.6. Регрессия  $Y_t$  на текущее, прошлое и будущее значения  $X_t$  имеет вид:
- $$Y_t = 3,0 + 1,7X_{t+1} + 0,8X_t - 0,2X_{t-1} + u_t.$$
- а) Перепишите регрессию так, чтобы она имела вид из уравнения (16.25). Чему равны значения  $\theta$ ,  $\delta_{-1}$ ,  $\delta_0$  и  $\delta_1$ ?
- б) (i) Предположим, что  $X_t$  является  $I(1)$  и  $u_t$  является  $I(1)$ . Являются ли  $X$  и  $Y$  коинтегрированными?
- (ii) Предположим, что  $X_t$  является  $I(0)$ , а  $u_t$  является  $I(1)$ . Являются ли  $X$  и  $Y$  коинтегрированными?
- (iii) Предположим, что  $X_t$  является  $I(1)$ , а  $u_t$  является  $I(0)$ . Являются ли  $X$  и  $Y$  коинтегрированными?
- 16.7. Предположим, что  $\Delta Y_t = u_t$ , где  $u_t$  является i.i.d.  $N(0, 1)$ , и рассмотрим регрессию  $Y_t = \beta X_t + \text{error}$ , где  $X_t = \Delta Y_{t+1}$  и  $\text{error}$  является ошибкой регрессии. Покажите, что  $\hat{\beta} \xrightarrow{d} \frac{1}{2}(\chi_1^2 - 1)$ . [Подсказка: проанализируйте числитель  $\hat{\beta}$ , используя метод, аналогичный тому, который использовался в уравнении (16.21). Проанализируйте знаменатель, используя закон больших чисел.]
- 16.8. Рассмотрим следующую VAR-модель с двумя переменными, одним лагом и без константы:
- $$Y_t = \beta_{11}Y_{t-1} + \gamma_{11}X_{t-1} + u_{1t};$$
- $$X_t = \beta_{21}Y_{t-1} + \gamma_{21}X_{t-1} + u_{2t}.$$

- a) Покажите, что итеративный двухшаговый прогноз  $\hat{Y}$  можно записать в таком виде:

$$Y_{t|t-2} = \delta_1 Y_{t-2} + \delta_2 X_{t-2}$$

и выведите значения для  $\delta_1$  и  $\delta_2$  через коэффициенты VAR.

- б) В свете вашего ответа на вопрос (a) можно ли сказать, что итеративные многошаговые прогнозы отличаются от прямых многошаговых прогнозов? Объясните.

- 16.9. a) Предположим, что  $E(u_t | u_{t-1}, u_{t-2}, \dots) = 0$ , что  $\text{var}(u_t | u_{t-1}, u_{t-2}, \dots)$  описывается моделью ARCH(1)  $\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$  и что процесс  $u_t$  является стационарным. Покажите, что  $\text{var}(u_t) = \alpha_0 / (1 - \alpha_1)$ . (Подсказка: используйте закон повторного математического ожидания  $E(u_t^2) = E[E(u_t^2 | u_{t-1})]$ .)  
 б) Расширьте результат из пункта (a) на случай модели ARCH( $p$ ).  
 в) Покажите, что для стационарной ARCH( $p$ ) -модели  $\sum_{i=1}^p \alpha_i < 1$ .  
 г) Распространите результат из пункта (a) на случай GARCH(1, 1)-модели.  
 д) Покажите, что  $\alpha_1 + \phi_1 < 1$  для стационарной GARCH(1, 1)-модели.

- 16.10. Рассмотрим коинтеграционную модель  $Y_t = \theta X_t + v_{1t}$  и  $X_t = X_{t-1} + v_{2t}$ , где  $v_{1t}$  и  $v_{2t}$  являются серийно некоррелированными случайными величинами с нулевыми средними и с  $E(v_{1t} v_{2t}) = 0$  для всех  $t$  и  $i$ . Выведите векторную модель коррекции ошибками [уравнения (16.22) и (16.23)] для  $X$  и  $Y$ .

## Компьютерные упражнения

Рассмотренные ниже упражнения основаны на временных рядах из баз данных USMacro\_Quarterly и USMacro\_Monthly, описанных в разделе «Компьютерные упражнения» в главах 14 и 15. Пусть  $Y_t = \ln(GDP_t)$ ,  $R_t$  обозначает трехмесячную ставку процента по казначейским векселям,  $\pi_t^{CPI}$  и  $\pi_t^{CPE}$  обозначают инфляцию по ИПЦ и дефлятор расходов на личное потребление (PCE), соответственно.

- Е16.1. Используя квартальные данные с I квартала 1955 по IV квартал 2009 года, оцените модель VAR(4) (VAR с четырьмя лагами) для  $\Delta Y_t$  и  $\Delta R_t$ .

- а) Является ли  $\Delta R$  причиной по Грейндженеру для  $\Delta Y$ ? Является ли  $\Delta Y$  причиной по Грейндженеру для  $\Delta R$ ?

- б) Следует ли включить в VAR более четырех лагов?

- Е16.2. В этом упражнении нужно вычислить псевдовневыборочные прогнозы на два квартала вперед для  $\Delta Y$ , начиная с IV квартала 1989 года и до конца выборки. (То есть нужно вычислить  $\Delta Y_{1990:\text{III}|1989:\text{IV}}$ ,  $\Delta Y_{1990:\text{III}|1990:\text{I}}$  и так далее.)

- а) Постройте итеративный псевдовневыборочный прогноз на два квартала вперед с использованием модели AR(1).

- б) Постройте итеративный псевдовневыборочный прогноз на два квартала вперед с использованием модели VAR(4) для  $\Delta Y$  и  $\Delta R$ .

- в) Постройте итеративный псевдовневыборочный прогноз на два квартала вперед, используя наивный прогноз  $\Delta Y_{t+2|t} = (\Delta Y_t + \Delta Y_{t-1} + \Delta Y_{t-2} + \Delta Y_{t-3}) / 4$ .

- г) Какая модель имеет наименьшую среднеквадратическую ошибку прогноза?

E16.3. Используйте DF-GLS-тест для проверки наличия единичного авторегрессионного корня у  $Y_t$ . В качестве альтернативы предположим, что  $Y_t$  является стационарным вокруг детерминированного тренда. Сравните полученные результаты с результатами, полученными в упражнении E14.3.

E16.4. В упражнении E15.2 вы изучали поведение  $\pi_t^{CPI} - \pi_t^{PCE}$  в январе 1970 – декабре 2009 годов. Этот анализ был основан на предположении, что  $\pi_t^{CPI} - \pi_t^{PCE}$  является процессом  $I(0)$ .

а) Проверьте наличие единичного авторегрессионного корня у  $\pi_t^{CPI} - \pi_t^{PCE}$ .

Проведите ADF-тест, включив в тестовую регрессию 12 запаздываний первой разности  $\pi_t^{CPI} - \pi_t^{PCE}$  и константу. Также проведите тест, используя DF-GLS-тест.

б) Проверьте наличие единичного корня у рядов  $\pi_t^{CPI} - \pi_t^{PCE}$ . Как и в пункте (а), используйте оба теста ADF и DF-GLS, включая константу и 12 запаздывающих разностей.

в) Что говорят результаты из пунктов (а) и (б) о коинтеграции между этими двумя показателями инфляции? Какой вывод о значении коинтегрирующего коэффициента ( $\theta$ ) можно сделать на основе ответов в пунктах (а) и (б)?

г) Предположим, мы знаем, что коинтегрирующий коэффициент равен  $\theta = 1$ . Как вы проверили бы наличие коинтеграции? Проведите тест. Как вы оценили бы  $\theta$ ? Оцените значение  $\theta$ , используя DOLS и оценивая регрессию  $\pi_t^{CPI}$  на  $\pi_t^{PCE}$  и на шесть опережающих и запаздывающих значений  $\Delta\pi_t^{PCE}$ . Близко ли оцененное значение  $\theta$  к единице?

E16.5. Используя данные о  $\Delta Y$  (темперы роста ВВП) с I квартала 1955 по IV квартал 2009 года, оцените модель AR(1) с остатками в форме GARCH(1, 1).

а) Постройте график остатков модели AR(1) вместе границей, равной  $\pm\hat{\sigma}_e$ , как на рисунке 16.4.

б) Некоторые макроэкономисты утверждали, что произошло резкое падение дисперсии  $\Delta Y$  в районе 1983 года, которое они называют великой умеренностью. Можно ли увидеть эту великую умеренность на графике, который вы начертigli в пункте (б)?

## Приложения

### Приложение 16.1. Финансовые данные США, используемые в главе 16

Как сообщает Совет управляющих Федеральной резервной системы США, процентные ставки по трехмесячным казначейским векселям США и по годовым казначейским облигациям США являются среднемесячными значениями их дневной ставки, преобразованные к годовому виду. Квартальные данные, используемые в этой главе, являются среднемесячной процентной ставкой для последнего месяца квартала.

Часть V

ТЕОРЕТИЧЕСКИЕ  
ОСНОВЫ  
РЕГРЕССИОННОГО  
АНАЛИЗА



# Глава 17. Теория парной линейной регрессии

Почему прикладные эконометристы должны тратить время на изучение эконометрической теории? Существует несколько объяснений этой необходимости. Изучая теорию, вы превращаете используемые вами статистические программные пакеты из «черного ящика» в гибкий набор инструментов, в котором вы можете найти подходящий инструмент для текущей работы. Знание эконометрической теории помогает вам понять, почему эти инструменты работают и какие предположения необходимы для того, чтобы каждый из них работал правильно. Возможно, самая важная причина заключается в том, что знание эконометрической теории помогает разобраться в ситуациях, когда этот инструмент работает неправильно и когда вам необходимо искать другой эконометрический подход.

В данной главе мы рассматриваем введение в эконометрическую теорию парной линейной регрессии. Это введение должно дополнить, а не заменить, материалы из глав 4 и 5, которые следует прочесть в первую очередь.

Данная глава дополняет материалы глав 4 и 5 следующим образом.

Во-первых, в ней приводятся математические обоснования выборочных распределений МНК-оценок и  $t$ -статистики как в больших выборках в рамках трех предположений МНК из вставки «Основные понятия 4.3», так и в конечных выборках в рамках двух дополнительных предположений о гомоскедастичности и нормальности ошибок. Эти пять дополнительных предположений МНК, приведенные в разделах 17.1–17.3 и в приложении 17.2, являются основой для доказательства нормальности распределения МНК-оценок и  $t$ -статистики в больших выборках при первых трех предположениях (предположения МНК из вставки «Основные понятия 4.3»). В разделе 17.4 содержится вывод точных распределений МНК-оценок и  $t$ -статистики при двух дополнительных предположениях о гомоскедастичности и нормальности ошибок.

Во-вторых, мы рассматриваем дополнительный метод оценки регрессий, который используется при наличии гетероскедастичности ошибок. Подход, рассмотренный в главах 4 и 5, заключался в использовании стандартных ошибок, устойчивых к гетероскедастичности, для того чтобы гарантировать корректность статистических выводов даже в случае наличия гетероскедастичности ошибок. Но этот метод имеет свои издержки: если ошибки гетероскедастичны, то в теории существует метод более эффективный, чем МНК. Этот метод оценки, называемый взвешенным методом наименьших квадратов, рассмотрен в разделе 17.5. Взвешенный метод наименьших квадратов (ВМНК) требует наличия точных априорных знаний о природе гетероскедастичности, то есть об условной

дисперсии и относительно  $X$ . Если такие знания доступны, взвешенный метод наименьших квадратов позволяет получить более точные оценки по сравнению с МНК. Однако во многих эмпирических работах такая информация недоступна; и в таких случаях более предпочтительным является использование МНК со стандартными ошибками, устойчивыми к гетероскедастичности.

## 17.1. Расширенные предположения метода наименьших квадратов и оценка МНК

В данном разделе мы вводим ряд предположений, дополняющих и усиливающих три предположения метода наименьших квадратов из главы 4. Эти более сильные предположения использованы в последующих разделах для вывода более сильных теоретических результатов об МНК-оценке, чем это было возможно при более слабых (хотя и более реалистичных) предположениях из главы 4.

### *Расширенные предположения метода наименьших квадратов*

**Расширенные предположения метода наименьших квадратов № 1, № 2 и № 3.** Первые три расширенных предположения МНК приведены во вставке «Основные понятия 4.3»: условное среднее  $u_i$  при заданных  $X_i$  равно нулю;  $(X_i, Y_i), i=1, \dots, n$ , являются независимыми и одинаково распределенными (i.i.d.) элементами выборки из их совместного распределения;  $X_i$  и  $u_i$  имеют моменты до четвертого порядка включительно.

При этих трех предположениях МНК-оценки являются несмещенными, состоятельными и асимптотически нормальными. Если эти три предположения выполняются, то методы, которые приведены в главе 4 и с помощью которых мы делаем выводы (проверка гипотез с использованием  $t$ -статистики и построение 95 %-го доверительного интервала как  $\pm 1,96$  стандартной ошибки) доказаны в больших выборках. Чтобы разработать теорию эффективных оценок с использованием МНК или чтобы охарактеризовать точные выборочные распределения МНК-оценок, нам понадобятся дополнительные предположения.

**Расширенное предположение метода наименьших квадратов № 4.** Четвертое расширенное предположение заключается в том, что  $u_i$  является гомоскедастичной, то есть  $\text{var}(u_i | X_i) = \sigma_u^2$  где  $\sigma_u^2$  является постоянной величиной. Как было показано в разделе 5.5, если это дополнительное предположение выполнено, то МНК-оценка является эффективной в классе всех линейных оценок, являющихся условно несмещенными относительно  $X_1, \dots, X_n$ .

**Расширенное предположение метода наименьших квадратов № 5.** Пятое расширенное предположение заключается в том, что условное распределение  $u_i$  при заданном  $X_i$  является нормальным.

При выполнении предположений МНК № 1 и № 2 и расширенных предположений № 4 и № 5  $u_i$  является i.i.d.  $N(0, \sigma_u^2)$ , а  $u_i$  и  $X_i$  являются независимо распределенными. Чтобы увидеть это, достаточно заметить, что пятое расши-

ренное предположение МНК утверждает, что условное распределение  $u_i | X_i$  равно  $N(0, \text{var}(u_i | X_i))$ , причем равенство нулю среднего данного распределения следует из расширенного предположения МНК № 1. Однако по четвертому предположению МНК  $\text{var}(u_i | X_i) = \sigma_u^2$ , так что условное распределение  $u_i | X_i$  равно  $N(0, \sigma_u^2)$ . Поскольку это условное распределение не зависит от  $X_i, u_i$  и  $X_i$  независимо распределены. По второму предположению МНК  $u_i$  распределено независимо от  $u_j$  для всех  $j \neq i$ . Следовательно, при расширенных предположениях МНК № 1, № 2, № 4 и № 5  $u_i$  и  $X_i$  распределены независимо, а  $u_i$  является i.i.d.  $N(0, \sigma_u^2)$ .

В разделе 17.4 показано, что если все пять расширенных предположений МНК выполнены, то МНК-оценка имеет точное нормальное распределение, и  $t$ -статистика, рассчитанная в предположении гомоскедастичности, имеет точное распределение Стьюдента.

Четвертое и пятое расширенные предположения МНК гораздо более сильные, чем первые три. Мы зачастую можем полагать, что первые три предположения выполняются в выборке, последние два предположения не столь реалистичны. Тем не менее даже если эти два предположения не выполняются на практике, они представляют теоретический интерес, поскольку если хотя бы одно из них выполняется, то МНК-оценка имеет дополнительные свойства помимо свойств, рассмотренных в главах 4 и 5. Следовательно, мы можем лучше понять свойства МНК-оценок и в целом теории оценки модели линейной регрессии, рассмотрев оценки при этих строгих предположениях.

Пять расширенных предположений МНК для случая модели парной регрессии приведены во вставке «Основные понятия 17.1».

### Расширенные предположения МНК для парной регрессии

Модель парной линейной регрессии имеет следующий вид:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i=1, \dots, n. \quad (17.1)$$

Расширенные предположения МНК:

- $E(u_i | X_i) = 0$  (условное среднее равно нулю);
- $(X_i, Y_i), i=1, \dots, n$  являются независимыми и одинаково распределенными (i.i.d.) случайными величинами;
- $X_i$  и  $u_i$  имеют ненулевые конечные четвертые моменты;
- $\text{var}(u_i | X_i) = \sigma_u^2$  (гомоскедастичность);
- условное распределение  $u_i$  при заданном  $X_i$  является нормальным (нормальные ошибки).

**ОСНОВНЫЕ ПОНЯТИЯ**

17.1

### МНК-оценки

Для удобства вновь приведем формулы МНК-оценок коэффициентов  $\beta_0$  и  $\beta_1$ :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}; \quad (17.2)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (17.3)$$

Уравнения (17.2) и (17.3) выведены в приложении 4.2.

## 17.2. Основные понятия асимптотической теории

Асимптотическая теория – это теория о распределении статистик – оценок, тестовых статистик и доверительных интервалов – в выборках больших размеров. Формально эта теория включает в себя описание поведения выборочного распределения статистик для последовательности выборок растущего размера. Эта теория асимптотическая в том смысле, что она описывает поведение статистики в пределе, когда  $n \rightarrow \infty$ .

Несмотря на то что размер выборки никогда не бывает бесконечным, асимптотическая теория играет центральную роль в эконометрике и статистике по следующим причинам. Во-первых, если число наблюдений в выборке, используемое в эмпирической работе, велико, то асимптотическое распределение является очень хорошим приближением для распределения в конечной выборке. Во-вторых, асимптотические распределения обычно гораздо проще устроены и, следовательно, их проще использовать на практике, чем точные распределения в конечных выборках. Вместе эти две причины означают, что надежные и простые методы, используемые для статистических выводов, – тесты с использованием  $t$ -статистики и 95 %-е доверительные интервалы, вычисленные как  $\pm 1,96$  от стандартной ошибки, – могут быть основаны на приблизительных распределениях, полученных в рамках асимптотической теории.

Две основы асимптотической теории: закон больших чисел и центральная предельная теорема – сформулированы в разделе 2.6. Мы начинаем этот раздел с продолжения обсуждения закона больших чисел и центральной предельной теоремы, включая доказательство закона больших чисел. Затем мы введем два новых инструмента, теорему Слуцкого и теорему о непрерывном отображении, расширяющих возможности использования закона больших чисел и центральной предельной теоремы. Позднее в качестве иллюстрации мы используем для доказательства результат о том, что распределение  $t$ -статистики, основанное на  $\bar{Y}$  и используемое для тестирования гипотезы  $E(Y) = \mu_0$ , в условиях нулевой гипотезы распределено по стандартному нормальному распределению.

### **Сходимость по вероятности и закон больших чисел**

Понятие о сходимости по вероятности и закон больших чисел были введены в разделе 2.6. Здесь мы приводим точное математическое определение

ление сходимости по вероятности, формулировку и доказательство закона больших чисел.

**Состоительность и сходимость по вероятности.** Пусть  $S_1, S_2, \dots, S_n, \dots$  является последовательностью случайных величин. К примеру,  $S_n$  может быть выборочным средним  $\bar{Y}$  для выборки из  $n$  наблюдений случайной величины  $Y$ . Эта последовательность случайных величин  $S_n$  называется *сходящейся по вероятности* к пределу  $\mu$  (т.е.  $S_n \xrightarrow{P} \mu$ ), если для любой положительной константы  $\delta$  вероятность того, что  $S_n$  находится в пределах  $\pm\delta$  от  $\mu$  стремится к 1 при  $n \rightarrow \infty$ . То есть:

$$S_n \xrightarrow{P} \mu \text{ тогда и только тогда, если } \Pr(|S_n - \mu| \geq \delta) \rightarrow 0 \quad (17.4)$$

при  $n \rightarrow \infty$  для любого  $\delta > 0$ . Если  $S_n \xrightarrow{P} \mu$ , то считается, что  $S_n$  является *состоительной оценкой*  $\mu$ .

**Закон больших чисел.** Закон больших чисел утверждает, что при определенных условиях на  $Y_1, Y_2, \dots, Y_n$  выборочное среднее  $\bar{Y}$  сходится по вероятности к среднему в генеральной совокупности. Математики разработали множество версий закона больших чисел, соответствующих различным условиям на  $Y_1, Y_2, \dots, Y_n$ . Версия закона больших чисел, используемая в данной книге, соответствует i.i.d. реализациям  $Y_1, Y_2, \dots, Y_n$ , взятым из распределения с конечной дисперсией. Закон больших чисел (также приведенный во вставке «Основные понятия 2.6») имеет вид

Если  $Y_1, \dots, Y_n$  являются i.i.d.,  $E(Y_i) = \mu_Y$ , и  $\text{var}(Y_i) < \infty$ ,

$$\text{то } \bar{Y} \xrightarrow{P} \mu_Y. \quad (17.5)$$

Идея закона больших чисел изображена на рисунке 2.8: с ростом выборки выборочное распределение  $\bar{Y}$  сгущается около генерального среднего  $\mu$ . Одной из особенностей выборочного распределения является то, что дисперсия  $\bar{Y}$  уменьшается по мере увеличения размера выборки. Другой особенностью является то, что вероятность того, что  $\bar{Y}$  окажется за пределами интервала  $\pm\delta$  от  $\mu$ , становится пренебрежимо малой с ростом  $n$ . Эти две черты выборочного распределения связаны между собой, и при доказательстве закона больших чисел используется эта связь.

**Доказательство закона больших чисел.** Связь между дисперсией  $\bar{Y}$  и вероятностью того, что  $\bar{Y}$  попадет в интервал  $\pm\delta$  от  $\mu$ , вытекает из неравенства Чебышева, которое приведено и доказано в приложении 17.2 [см. уравнение (17.42)]. Записанное в терминах  $\bar{Y}$ , неравенство Чебышева имеет вид:

$$\Pr(|\bar{Y} - \mu_Y| \geq \delta) \leq \frac{\text{var}(\bar{Y})}{\delta^2} \quad (17.6)$$

для любой положительной константы  $\delta$ . Поскольку  $Y_1, Y_2, \dots, Y_n$  являются i.i.d. с дисперсией  $\sigma_Y^2$ , то  $\text{var}(\bar{Y}) = \sigma_Y^2/n$ ; тогда для любого  $\delta > 0$  выполняется  $\text{var}(\bar{Y})/\delta^2 = \sigma_Y^2/(\delta^2 n)$ . Из уравнения (17.6) следует, что  $\Pr(|\bar{Y} - \mu_Y| \geq \delta) \rightarrow 0$  (для любого  $\delta > 0$ ), что и доказывает утверждение закона больших чисел.

*Примеры.* Состоятельность является основным понятием асимптотической теории, поэтому мы рассмотрим некоторые примеры состоятельных и несостоятельных оценок генерального среднего  $\mu_Y$ . Предположим, что  $Y_i, i = 1, \dots, n$  являются i.i.d. с положительной и конечной дисперсией  $\sigma_Y^2$ . Рассмотрим следующие три оценки  $\mu_Y$ : (1)  $m_a = Y_1$ ; (2)  $\mu_b = \left(\frac{1-a^n}{1-a}\right)^{-1} \sum_{i=1}^n a^{i-1} Y_i$ , где  $0 < a < 1$ , и (3)  $\mu_c = \bar{Y} + 1/n$ . Являются ли эти оценки состоятельными?

Первая оценка,  $\mu_a = Y_1$ , — это просто первое наблюдение, так что  $E(m_a) = E(Y_1) = \mu_Y$ , то есть  $\mu_a = Y_1$  является несмешенной оценкой. Однако  $\mu_a = Y_1$  не является состоятельной оценкой:  $\Pr(|m_a - \mu_Y| \geq \delta) = \Pr(|Y_1 - \mu_Y| \geq \delta)$ , что является положительной величиной для достаточно малых  $\delta$  (поскольку  $\sigma_Y^2 > 0$ ) и, следовательно,  $\Pr(|m_a - \mu_Y| \geq \delta)$  не стремится к нулю при  $n \rightarrow \infty$ , следовательно,  $\mu_a = Y_1$  несостоительна. Эта несостоятельность не должна вас удивлять: поскольку  $\mu_a = Y_1$  использует информацию только о первом наблюдении, ее распределение не может концентрироваться вокруг  $\mu_Y$  при росте числа наблюдений выборки.

Вторая оценка  $m_b$  является несмешенной, но несостоятельной. Она является несмешенной, поскольку

$$E(m_b) = E\left[\left(\frac{1-a^n}{1-a}\right)^{-1} \sum_{i=1}^n a^{i-1} Y_i\right] = \left(\frac{1-a^n}{1-a}\right)^{-1} \sum_{i=1}^n a^{i-1} \mu_Y = \mu_Y,$$

$$\text{так как } \sum_{i=1}^n a^{i-1} = 1 - a^n \sum_{i=1}^{\infty} a^i = \frac{1-a^n}{1-a}.$$

Дисперсия  $m_b$  равна:

$$\text{var}(m_b) = \left(\frac{1-a^n}{1-a}\right)^{-2} \sum_{i=1}^n a^{2(i-1)} \sigma_Y^2 = \sigma_Y^2 \frac{(1-a^{2n})(1-a)^2}{(1-a^2)(1-a^n)^2} = \sigma_Y^2 \frac{(1+a^n)(1-a)}{(1-a^n)(1+a)},$$

что в пределе дает  $\text{var}(m_b) \rightarrow \sigma_Y^2 (1-a)/(1+a)$  при  $n \rightarrow \infty$ . Следовательно, дисперсия этой оценки не стремится к нулю, и распределение не концентрируется вокруг  $\mu_Y$ , и оценка, несмотря на несмешенность, не является состоятельной. Возможно, это выглядит странным, поскольку все наблюдения входят в оценку. Однако большинство наблюдений имеют очень маленький вес (вес  $i$ -го наблюдения пропорционален  $a^{i-1}$  — очень малой величине при больших  $i$ ), и по этой причине не происходит достаточного взаимного сокращения ошибок для состоятельности оценки.

Третья оценка,  $m_c$ , является смешенной, но состоятельной. Ее смещение равно  $1/n$ :  $E(m_c) = E(\bar{Y} + 1/n) = \mu_Y + 1/n$ . Однако это смещение стремится к нулю при увеличении размера выборки, и поэтому  $m_c$  является состоятельной:  $\Pr(|m_c - \mu_Y| \geq \delta) = \Pr(|\bar{Y} + 1/n - \mu_Y| \geq \delta)$ . Из уравнения (17.43) приложения 17.2, являющегося обобщением неравенства Чебышева, следует, что для любой случайной величины  $W$  верно  $\Pr(|W| \geq \delta) \leq E(W^2)/\delta^2$  для любой положительной константы  $\delta$ . Следовательно,  $\Pr(|\bar{Y} + 1/n - \mu_Y| \geq \delta) \leq E[(\bar{Y} + 1/n - \mu_Y)^2]/\delta^2$ . Но  $E[(\bar{Y} + 1/n - \mu_Y)^2] = \text{var}(\bar{Y}) + 1/n^2 = \sigma^2/n + 1/n^2 \rightarrow 0$  при больших  $n$ . Следовательно,  $\Pr(|\bar{Y} + 1/n - \mu_Y| \geq \delta) \rightarrow 0$ , и  $m_c$  состоятельна. Этот пример иллюстрирует

общий результат о том, что оценка может быть смещенной в конечных выборках, но если смещение исчезает с ростом выборки, то оценка может быть состоятельной (упражнение 17.10).

### **Центральная предельная теорема и сходимость по распределению**

Если распределения последовательности случайных величин сходятся к пределу при  $n \rightarrow \infty$ , то такая последовательность случайных величин называется сходящейся по распределению. Центральная предельная теорема утверждает, что при более общих условиях стандартизированное выборочное среднее сходится по распределению к нормальному распределенной случайной величине.

**Сходимость по распределению.** Пусть  $F_1, F_2, \dots, F_n, \dots$  является последовательностью функций распределения, соответствующих последовательности случайных величин  $S_1, S_2, \dots, S_n, \dots$ . Например,  $S_n$  может быть стандартизованным выборочным средним, то есть  $(\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$ . Тогда последовательность случайных величин  $S_n$  называется *сходящейся по распределению* к  $S$  (обозначается  $S_n \xrightarrow{d} S$ ), если последовательность функций распределения  $\{F_n\}$  сходится к распределению  $F$ , являющемуся распределением случайной величины  $S$ , то есть:

$$S_n \xrightarrow{d} S \text{ тогда и только тогда, когда } \lim_{n \rightarrow \infty} F_n(t) = F(t), \quad (17.7)$$

где сходимость имеет место для всех точек  $t$ , в которых функция распределения  $F$  непрерывна. Распределение  $F$  называется *асимптотическим распределением*  $S_n$ .

Полезно сравнить понятия сходимости по вероятности ( $\xrightarrow{P}$ ) и сходимости по распределению ( $\xrightarrow{d}$ ). Если  $S_n \xrightarrow{P} \mu$ , то  $S_n$  становится близкой к  $\mu$  с большой вероятностью при росте  $n$ . Напротив, если  $S_n \xrightarrow{d} S$ , то распределение  $S_n$  приближается к распределению  $S$  при росте  $n$ .

**Центральная предельная теорема.** Мы приведем здесь центральную предельную теорему, используя понятие сходимости по распределению. Центральная предельная теорема из вставки «Основные понятия 2.7» утверждает, что если  $Y_1, Y_2, \dots, Y_n$  – независимые одинаково распределенные случайные величины и  $0 < \sigma_Y^2 < \infty$ , то асимптотическое распределение  $(\bar{Y} - \mu_Y) / \sigma_{\bar{Y}}$  является стандартным нормальным распределением  $N(0, 1)$ . Поскольку  $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$ , то  $(\bar{Y} - \mu_Y) / \sigma_{\bar{Y}} = \sqrt{n}(\bar{Y} - \mu_Y) / \sigma_Y$ . Следовательно, центральная предельная теорема может быть сформулирована иначе:  $\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{d} \sigma_Y Z$ , где  $Z$  является стандартной нормальной случайной величиной. Это означает, что распределение  $\sqrt{n}(\bar{Y} - \mu_Y)$  сходится к  $N(0, \sigma_Y^2)$  при  $n \rightarrow \infty$ . Это принято обозначать так:

$$\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2). \quad (17.8)$$

То есть если  $Y_1, Y_2, \dots, Y_n$  являются i.i.d. с  $0 < \sigma_Y^2 < \infty$ , то распределение  $\sqrt{n}(\bar{Y} - \mu_Y)$  сходится к нормальному распределению с нулевым средним и дисперсией  $\sigma_Y^2$ .

**Обобщение на временные ряды.** Закон больших чисел и центральная предельная теорема, сформулированные в разделе 2.6, применимы к случаю i.i.d.

наблюдений. Как было показано в главе 14, предположение об i.i.d. не подходит для временных рядов, поэтому упомянутые теоремы необходимо обобщить, прежде чем применять их к случаю временных рядов. Эти обобщения носят технический характер в том смысле, что все выводы остаются прежними (варианты закона больших чисел и центральной предельной теоремы применяются к временным рядам) однако условия, при которых эти теоремы работают, отличаются от предыдущих вариантов теорем. Все это было коротко рассмотрено в разделе 16.4, но математическое обоснование асимптотической теории для временных рядов выходит за рамки данной книги, и заинтересованный читатель может обратиться к работе Хаяши (Hayashi, 2000. Ch. 2).

### **Теорема Слуцкого и теорема о непрерывном отображении**

Теорема Слуцкого объединяет понятия состоятельности и сходимости по распределению. Предположим, что  $a_n \xrightarrow{d} a$ , где  $a$  является некоторой константой, и  $S_n \xrightarrow{d} S$ . Тогда

$$\begin{aligned} a_n + S_n &\xrightarrow{d} a + S, \quad a_n S_n \xrightarrow{d} aS, \text{ и если } a \neq 0, \\ \text{то } S_n/a_n &\xrightarrow{d} S/a. \end{aligned} \tag{17.9}$$

Эти три результата все вместе называются теоремой Слуцкого.

В теореме о непрерывном отображении рассматриваются асимптотические свойства непрерывной функции  $g$  от последовательности случайных величин  $S_n$ . Эта теорема состоит из двух частей. Первая часть заключается в следующем утверждении: если  $S_n$  сходится по вероятности к константе  $a$ , то  $g(S_n)$  сходится по вероятности к  $g(a)$ ; вторая часть говорит о том, что если  $S_n$  сходится по распределению к  $S$ , то  $g(S_n)$  сходится по распределению к  $g(S)$ . То есть если  $g$  является непрерывной функцией, то

- (i) Если  $S_n \xrightarrow{p} a$ , то  $g(S_n) \xrightarrow{p} g(a)$
- и (ii) Если  $S_n \xrightarrow{d} S$ , то  $g(S_n) \xrightarrow{d} g(S)$ .

В качестве примера применения пункта (i) рассмотрим следующий: если  $s_Y^2 \xrightarrow{p} \sigma_Y^2$ , то  $\sqrt{s_Y^2} = s_Y \xrightarrow{p} \sigma_Y$ . Для второго случая рассмотрим такой пример: предположим, что  $S_n \xrightarrow{d} Z$ , где  $Z$  – стандартное нормальное распределение, и пусть  $g(S_n) = S_n^2$ . Поскольку  $g$  является непрерывной функцией, теорема о непрерывном отображении применима, и  $g(S_n) \xrightarrow{d} g(Z)$ , то есть  $S_n^2 \xrightarrow{d} Z^2$ . Другими словами, распределение  $S_n^2$  сходится к распределению квадрата нормальной случайной величины, что является распределением  $\chi_1^2$ , то есть  $S_n^2 \xrightarrow{d} \chi_1^2$ .

### **Приложения к t-статистике, основанной на выборочном среднем**

Сейчас мы применим центральную предельную теорему, закон больших чисел и теорему Слуцкого, чтобы показать, что в условиях нулевой гипотезы t-статистика, основанная на  $\bar{Y}$ , имеет стандартное нормальное распределение, если  $Y_1, Y_2, \dots, Y_n$  являются независимыми одинаково распределенными случайными величинами с  $0 < E(Y_i^4) < \infty$ .

$t$ -статистика для тестирования нулевой гипотезы о том, что  $E(Y_i) = \mu_0$ , построенная на основе выборочного среднего  $\bar{Y}$ , задана уравнениями (3.8) и (3.11) и может быть записана в таком виде:

$$t = \frac{\bar{Y} - \mu_0}{s_Y / \sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma_Y} \div \frac{s_Y}{\sigma_Y}, \quad (17.11)$$

где во втором равенстве используется стандартный прием деления числителя и знаменателя на  $\sigma_Y$ .

Поскольку  $Y_1, Y_2, \dots, Y_n$  имеют два момента (что следует из наличия конечного четвертого момента; см., упражнение 17.5) и поскольку  $Y_1, Y_2, \dots, Y_n$  являются i.i.d., первый член в последнем равенстве уравнения (17.11) подчиняется центральной предельной теореме: в условиях нулевой гипотезы  $\sqrt{n}(\bar{Y} - \mu_0) / \sigma_y \xrightarrow{d} N(0, 1)$ . Кроме того,  $s_Y^2 \xrightarrow{p} \sigma_Y^2$  (как было показано в приложении 3.3), поэтому  $s_Y^2 / \sigma_Y^2 \xrightarrow{p} 1$  и соотношение в делителе в уравнении (17.11) стремится к единице (упражнение 17.4). Следовательно, выражение в последнем равенстве уравнения (17.11) имеет форму последнего выражения в уравнении (17.9), где [в обозначениях уравнения (17.9)]  $S_n = \sqrt{n}(\bar{Y} - \mu_0) / \sigma_y \xrightarrow{d} N(0, 1)$  и  $a_n = s_Y / \sigma_Y \xrightarrow{p} 1$ . Тогда, по теореме Слуцкого, получаем, что  $t \xrightarrow{d} N(0, 1)$ .

### 17.3. Асимптотические распределения МНК-оценки и $t$ -статистики

Как было показано в главе 4, при выполнении предположений из вставки «Основные понятия 4.3» (первых трех предположений из вставки «Основные понятия 17.1») МНК-оценка  $\hat{\beta}_1$  состоятельна и  $\sqrt{n}(\hat{\beta}_1 - \beta_1)$  асимптотически нормально распределена. Более того,  $t$ -статистика, используемая для тестирования нулевой гипотезы о том, что  $\beta_1 = \beta_{1,0}$ , имеет стандартное нормальное распределение в условиях нулевой гипотезы. В данном разделе приведены все эти результаты, а также дополнительные детали их доказательств.

#### Состоятельность и асимптотическая нормальность МНК-оценок

Асимптотическое распределение оценки  $\hat{\beta}_1$ , изначально рассмотренное во вставке «Основные понятия 4.4», имеет вид:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\text{var}(v_i)}{[\text{var}(X_i)]^2}\right), \quad (17.12)$$

где  $v_i = (X_i - \mu_X)u_i$ . Доказательство этого результата было схематически описано в приложении 4.3, однако в этом доказательстве пропущены некоторые детали и оно содержит приближение, не доказанное формально. Пропущенные детали приведены в качестве упражнения 17.3.

Как следствие уравнения (17.12), оценка  $\hat{\beta}_1$  состоятельна (упражнение 17.4).

## Состоятельность стандартных ошибок, устойчивых к гетероскедастичности

При условии выполнения первых трех предположений МНК-оценки стандартные ошибки  $\hat{\beta}_1$ , устойчивые к гетероскедастичности, дают возможность делать правильные статистические выводы. В особенности,

$$\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} \xrightarrow{P} 1, \quad (17.13)$$

где  $\sigma_{\hat{\beta}_1}^2 = \text{var}(v_i) / \left\{ n \left[ \text{var}(X_i) \right]^2 \right\}$  и  $\hat{\sigma}_{\hat{\beta}_1}^2$  является квадратом устойчивой к гетероскедастичности стандартной ошибки, определенной в уравнении (5.4), то есть

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n-2} \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2}. \quad (17.14)$$

Чтобы доказать результат из уравнения (17.13), для начала используем определения  $\sigma_{\hat{\beta}_1}^2$  и  $\hat{\sigma}_{\hat{\beta}_1}^2$  и перепишем отношение в уравнении (17.13) следующим образом:

$$\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} = \left[ \frac{n}{n-2} \right] \left[ \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\text{var}(v_i)} \right] \div \left[ \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}{\text{var}(X_i)} \right]. \quad (17.15)$$

Нам необходимо показать, что каждый из трех членов в скобках в правой части уравнения (17.15) сходится по вероятности к единице. Очевидно, что первый член сходится к 1, и по свойству состоятельности выборочной дисперсии (приложение 3.3) последний член также сходится по вероятности к 1. Следовательно, все, что остается показать, – это то, что второй член сходится по вероятности к 1, то есть  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2 \xrightarrow{P} \text{var}(v_i)$ .

Доказательство того, что  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2 \xrightarrow{P} \text{var}(v_i)$ , состоит из двух шагов. На первом шаге нужно показать, что  $\frac{1}{n} \sum_{i=1}^n v_i^2 \xrightarrow{P} \text{var}(v_i)$ , на втором –  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2 - \frac{1}{n} \sum_{i=1}^n v_i^2 \xrightarrow{P} 0$ .

Предположим, что  $X_i$  и  $u_i$  имеют восемь моментов [т.е.  $E(X_i^8) < \infty$  и  $E(u_i^8) < \infty$ ], что является более сильным предположением, чем предположение о существовании четырех моментов, требуемое третьим предположением МНК. Чтобы доказать первый шаг, мы должны показать, что  $\frac{1}{n} \sum_{i=1}^n v_i^2$  удовлетворяет закону больших чисел из уравнения (17.5). Чтобы это выполнялось,  $v_i^2$  должны быть i.i.d. случайными величинами (что выполняется в силу второго предположения МНК), а  $\text{var}(v_i^2)$  – являться конечной величиной. Чтобы показать, что  $\text{var}(v_i^2) < \infty$ ,

применим неравенство Коши–Шварца (приложение 17.2):  $\text{var}(v_i^2) \leq E(v_i^4) = E[(X_i - \mu_X)^4 u_i^4] \leq \left\{E[(X_i - \mu_X)^8 u_i^8]\right\}^{1/2}$ . Следовательно, если  $X_i$  и  $u_i$  имеют восемь моментов, то  $v_i^2$  имеет конечную дисперсию и, следовательно, удовлетворяет закону больших чисел из уравнения (17.5).

Второй шаг заключается в том, чтобы доказать  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2 - \frac{1}{n} \sum_{i=1}^n v_i^2 \xrightarrow{P} 0$ . Поскольку  $v_i = (X_i - \mu_X)u_i$ , этот второй шаг эквивалентен доказательству такого утверждения:

$$\frac{1}{n} \sum_{i=1}^n \left[ (X_i - \bar{X})^2 \hat{u}_i^2 - (X_i - \mu_X)^2 u_i^2 \right] \xrightarrow{P} 0. \quad (17.16)$$

Для доказательства этого результата нужно подставить  $\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)X_i$ , раскрыть скобки в уравнении (17.16), несколько раз применить неравенство Коши–Шварца и использовать свойство состоятельности оценок  $\hat{\beta}_0$  и  $\hat{\beta}_1$ . Алгебраические выкладки мы оставляем в качестве упражнения 17.9.

Предыдущее рассуждение предполагает, что у  $X_i$  и  $u_i$  существует восемь моментов. Однако это не является необходимым, и результат  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2 \xrightarrow{P} \text{var}(v_i)$  может быть получен и при более слабом предположении о том, что у  $X_i$  и  $u_i$  есть четыре момента, следующих из третьего предложения МНК. Доказательство этого, однако, выходит за рамки данного учебника (см. Hayashi, 2000. Раздел 2.5)).

### Асимптотическая нормальность $t$ -статистики, устойчивой к гетероскедастичности

Теперь мы покажем, что при нулевой гипотезе устойчивая к гетероскедастичности  $t$ -статистика МНК-оценки, используемая для проверки гипотезы  $\beta_1 = \beta_{1,0}$ , имеет асимптотическое стандартное нормальное распределение, если выполнены предположения МНК № 1, № 2 и № 3.

$t$ -статистика, построенная с использованием устойчивых к гетероскедастичности стандартных ошибок  $SE(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$  [определенных в уравнении (17.14)], имеет вид:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\sqrt{n}(\hat{\beta}_1 - \beta_{1,0})}{\sqrt{n\hat{\sigma}_{\hat{\beta}_1}^2}} \div \sqrt{\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\hat{\sigma}_{\hat{\beta}_1}^2}}. \quad (17.17)$$

Из уравнения (17.12) следует, что первый член после второго знака равенства в уравнении (17.17) сходится по распределению к стандартной нормальной случайной величине. В дополнение к этому, поскольку устойчивая к гетероскедастичности стандартная ошибка является состоятельной [уравнение (17.13)],  $\sqrt{\hat{\sigma}_{\hat{\beta}_1}^2 / \sigma_{\hat{\beta}_1}^2} \xrightarrow{P} 1$  (упражнение 17.4). Из теоремы Слуцкого следует, что  $t \xrightarrow{d} N(0, 1)$ .

## 17.4. Точные выборочные распределения при нормально распределенных ошибках

В небольших выборках распределения МНК-оценки и  $t$ -статистики зависят от распределения ошибок и, как правило, имеют сложный вид. Однако, как уже говорилось в разделе 5.6, при нормально распределенных и гомоскедастичных ошибках вид этих распределений достаточно простой. В частности, если все пять расширенных предположений МНК из вставки «Основные понятия 17.1» выполнены, то МНК-оценка имеет условное нормальное распределение в малых выборках относительно  $X_1, \dots, X_n$ . Кроме того,  $t$ -статистика имеет распределение Стьюдента. Ниже мы выведем эти результаты для  $\hat{\beta}_1$ .

### Распределение $\hat{\beta}_1$ в случае нормальных ошибок

Если ошибки независимы и одинаково нормально распределены и не зависят от regressоров, то условным распределением  $\hat{\beta}_1$  относительно  $X_1, \dots, X_n$  является  $N(\beta_1, \sigma_{\hat{\beta}_{1|x}}^2)$ , где

$$\sigma_{\hat{\beta}_{1|x}}^2 = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (17.18)$$

Вывод условного нормального распределения  $N(\beta_1, \sigma_{\hat{\beta}_{1|x}}^2)$  относительно  $X_1, \dots, X_n$  требует (i) показать, что распределение является нормальным, (ii) доказать, что  $E(\hat{\beta}_1 | X_1, \dots, X_n) = \beta_1$ , и (iii) проверить уравнение (17.18).

Чтобы показать (i), отметим, что  $\hat{\beta}_1 - \beta_1$  относительно  $X_1, \dots, X_n$  представляет собой взвешенное среднее от  $u_1, \dots, u_n$ :

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) u_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (17.19)$$

Эта формула была выведена в приложении 4.3 [уравнение (4.30)] и приведена здесь для удобства. В расширенных предположениях МНК № 1, № 2, № 4 и № 5  $u_i$  является i.i.d.  $N(0, \sigma_u^2)$ , а  $u_i$  и  $X_i$  независимо распределены. Из того что взвешенные средние нормально распределенных случайных величин являются нормально распределенными, следует, что  $\hat{\beta}_1$  имеет условное нормальное распределение относительно  $X_1, \dots, X_n$ .

Чтобы доказать (ii), возьмем условные математические ожидания от обеих частей уравнения (17.19):  $E(\hat{\beta}_1 - \beta_1 | X_1, \dots, X_n) = E\left[\sum_{i=1}^n (X_i - \bar{X}) u_i / \sum_{i=1}^n (X_i - \bar{X})^2 | X_1, \dots, X_n\right] = \sum_{i=1}^n (X_i - \bar{X}) E(u_i | X_1, \dots, X_n) / \sum_{i=1}^n (X_i - \bar{X})^2 = 0$ , где последнее равенство следует из того, что  $E(u_i | X_1, \dots, X_n) = E(u_i | X_i) = 0$ . Таким образом,  $\hat{\beta}_1$  является условно несмещенной, то есть

$$E(\hat{\beta}_1 | X_1, \dots, X_n) = \beta_1. \quad (17.20)$$

Чтобы доказать (iii), используем тот факт, что ошибки независимо условно распределены относительно  $X_1, \dots, X_n$ , и вычислим условную дисперсию  $\hat{\beta}_1$ , используя уравнение (17.19):

$$\begin{aligned} \text{var}(\hat{\beta}_1 | X_1, \dots, X_n) &= \text{var}\left[\frac{\sum_{i=1}^n (X_i - \bar{X}) u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} | X_1, \dots, X_n\right] = \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \text{var}(u_i | X_1, \dots, X_n)}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} = \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_u^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2}. \end{aligned} \quad (17.21)$$

Сокращение члена в числителе в последнем выражении в уравнении (17.21) дает формулу для условной дисперсии из уравнения (17.18).

### Распределение t-статистики в случае гомоскедастичности ошибок

t-статастика для тестирования гипотезы  $\beta_1 = \beta_{1,0}$  в случае гомоскедастичности ошибок имеет вид:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}, \quad (17.22)$$

где  $SE(\hat{\beta}_1)$  вычисляется с использованием стандартных ошибок  $\hat{\beta}_1$ , вычисленных при условии гомоскедастичности. Подставляя формулу для  $SE(\hat{\beta}_1)$  [уравнение (5.29) из приложения 5.1] в уравнение (17.22), получаем:

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s_u^2 / \sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma_u^2 / \sum_{i=1}^n (X_i - \bar{X})^2}} \div \\ &\div \sqrt{\frac{s_u^2}{\sigma_u^2}} = \frac{(\hat{\beta}_1 - \beta_{1,0}) / \sigma_{\hat{\beta}_1|X}^2}{\sqrt{W/(n-2)}}, \end{aligned} \quad (17.23)$$

где  $s_u^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$  и  $W = \sum_{i=1}^n \hat{u}_i^2 / \sigma_u^2$ . В соответствии с нулевой гипотезой  $\hat{\beta}_1$  распределена согласно  $N(\beta_{1,0}, \sigma_{\hat{\beta}_1|X}^2)$  относительно  $X_1, \dots, X_n$ , поэтому распределение

числителя в последнем выражении в уравнении (17.23) является  $N(0, 1)$ . Как показано в разделе 18.4,  $W$  имеет распределение хи-квадрат с  $n-2$  степенями свободы и, кроме того,  $W$  распределена независимо от стандартизованных МНК-оценок в числителе формулы (17.23). Из формулировки распределения Стьюдента (приложение 17.1) следует, что при выполнении пяти расширенных предположений МНК  $t$ -статистика в случае гомоскедастичности ошибок регрессии имеет распределение Стьюдента с  $n-2$  степенями свободы.

**Зачем нужно делать корректировку числа степеней свободы?** Корректировка степеней свободы в  $s_u^2$  гарантирует, что  $s_u^2$  будет несмещенной оценкой  $\sigma_u^2$  и что  $t$ -статистика будет иметь распределение Стьюдента в случае, когда ошибки распределены нормально.

Так как  $W = \sum_{i=1}^n \hat{u}_i^2 / \sigma_u^2$  является случайной величиной, распределенной по закону хи-квадрат с  $n-2$  степенями свободы, ее среднее равно  $E(W)=n-2$ . Таким образом,  $E[W/(n-2)]=(n-2)/(n-2)=1$ . Сделав необходимые подстановки в определение  $W$ , мы получаем, что  $E\left(\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2\right) = \sigma_u^2$ . Следовательно, коррекция степеней свободы делает  $s_u^2$  несмещенной оценкой  $\sigma_u^2$ . Также, благодаря делению на  $n-2$  вместо  $n$ , знаменатель последней части уравнения (17.23) удовлетворяет определению случайной величины, распределенной по закону Стьюдента, приведенному в приложении 17.1. То есть благодаря корректировке степеней свободы при расчете стандартной ошибки  $t$ -статистика имеет распределение Стьюдента при условии, что ошибки распределены нормально.

## 17.5. Взвешенный метод наименьших квадратов

В рамках первых четырех расширенных предположений МНК-оценки являются эффективными в классе линейных (по  $Y_1, \dots, Y_n$ ) условных (относительно  $X_1, \dots, X_n$ ) несмешенных оценок, то есть МНК-оценка является BLUE (т.е. наилучшей линейной несмешенной оценкой). Это результат из теоремы Гаусса–Маркова, который обсуждался в разделе 5.5 и доказан в приложении 5.2. Теорема Гаусса–Маркова дает теоретическое обоснование для использования МНК-оценки. Главное ограничение теоремы Гаусса–Маркова состоит в том, что она требует наличия гомоскедастичности ошибок. Если же, как часто бывает на практике, ошибки гетероскедастичны, то теорема Гаусса–Маркова не выполняется и МНК-оценки не являются BLUE.

В данном разделе представлена модификация МНК-оценки, называемая *взвешенным методом наименьших квадратов* (ВМНК), являющейся более эффективной по сравнению с МНК-оценкой в случае, когда ошибки гетероскедастичны.

ВМНК требует наличия информации о виде функции условной дисперсии,  $\text{var}(u_i | X_i)$ . Рассмотрим два случая. В первом случае  $\text{var}(u_i | X_i)$  известна с точностью до множителя, а ВМНК-оценка является BLUE. Во втором случае функциональная форма  $\text{var}(u_i | X_i)$  известна, но эта функциональная форма включает некоторые неизвестные параметры, подлежащие оценке. При дополнительных условиях асимптотическое распределение ВМНК во втором случае такое же,

как если бы параметры функции условной дисперсии на самом деле были известны, и в этом смысле оценка ВМНК является BLUE асимптотически. Раздел заканчивается обсуждением практических преимуществ и недостатков учета гетероскедастичности с использованием ВМНК и устойчивых к гетероскедастичности стандартных ошибок.

### **ВМНК при известной форме гетероскедастичности**

Предположим, что условная дисперсия  $\text{var}(u_i | X_i)$  известна с точностью до множителя; то есть

$$\text{var}(u_i / X_i) = \lambda h(X_i), \quad (17.24)$$

где  $\lambda$  является постоянной, а  $h$  – известной функцией. В этом случае оценка ВМНК получается путем деления зависимой переменной и регрессора на квадратный корень  $h$  и, затем, оценки регрессии этой преобразованной зависимой переменной на преобразованный регрессор с помощью МНК. Другими словами, разделим обе части регрессионной модели с одной переменной на  $\sqrt{h(X_i)}$  и получим:

$$\tilde{Y}_i = \beta_0 \tilde{X}_{0i} + \beta_1 \tilde{X}_{1i} + \tilde{u}_i, \quad (17.25)$$

где  $\tilde{Y}_i = Y_i / \sqrt{h(X_i)}$ ,  $\tilde{X}_{0i} = 1 / \sqrt{h(X_i)}$ ,  $\tilde{X}_{1i} = X_i / \sqrt{h(X_i)}$  и  $\tilde{u}_i = u_i / \sqrt{h(X_i)}$ .

ВМНК-оценка является МНК-оценкой для  $\hat{\beta}_1$  в уравнении (17.25), то есть она является оценкой, полученной из МНК-регрессии  $\tilde{Y}_i$  на  $\tilde{X}_{0i}$  и  $\tilde{X}_{1i}$ , где коэффициент при  $\tilde{X}_{0i}$  соответствует свободному члену в невзвешенной регрессии.

В рамках первых трех предположений МНК из вставки «Основные понятия 17.1» и при предположении о том, что вид формы гетероскедастичности в уравнении (17.24) известен, ВМНК является наилучшей линейной несмещенной оценкой (BLUE). Причина, по которой оценка ВМНК является BLUE, заключается в том, что взвешивание переменных сделало ошибку во взвешенной регрессии гомоскедастичной, то есть

$$\text{var}(\tilde{u}_i / X_i) = \text{var}\left[\frac{u_i}{\sqrt{h(X_i)}} | X_i\right] = \frac{\text{var}(u_i | X_i)}{h(X_i)} = \frac{\lambda h(X_i)}{h(X_i)} = \lambda, \quad (17.26)$$

так что условная дисперсия  $\tilde{u}_i$ ,  $\text{var}(u_i | X_i)$ , является постоянной. Таким образом, первые четыре предположения МНК относятся к уравнению (17.25). Строго говоря, теорема Гаусса–Маркова была доказана в приложении 5.2 для уравнения (17.1), которое включает в себя свободный член  $\beta_0$ , поэтому она неприменима напрямую к уравнению (17.25), в котором вместо свободного члена стоит  $\beta_0 \tilde{X}_{0i}$ . Тем не менее обобщение теоремы Гаусса–Маркова для множественной регрессии (см. раздел 18.5) распространяется на оценку  $\beta_1$  во взвешенной теоретической регрессии, то есть в уравнении (17.25). Соответственно, МНК-оценка  $\beta_1$  в уравнении (17.25), то есть ВМНК-оценка  $\beta_1$  является BLUE.

На практике функция  $h$  обычно неизвестна, так что ни взвешенные переменные в уравнении (17.25), ни ВМНК-оценки не могут быть вычислены. По этой

причине ВМНК-оценку, описанную выше, иногда называют недоступной ВМНК-оценкой. Для реализации ВМНК на практике функция  $h$  должна быть оценена; к обсуждению этой оценки мы сейчас и переходим.

### **ВМНК при неизвестной форме гетероскедастичности**

Если гетероскедастичность имеет известную функциональную форму, то функцию  $h$  можно оценить, и ВМНК-оценка может быть рассчитана с помощью данной оцененной функции.

**Пример № 1: дисперсия и является квадратичной функцией от  $X$ .** Предположим, что условная дисперсия является квадратичной функцией:

$$\text{var}(u_i | X_i) = \theta_0 + \theta_1 X_i^2, \quad (17.27)$$

где  $\theta_0$  и  $\theta_1$  – неизвестные параметры, а  $\theta_0 > 0$  и  $\theta_1 \geq 0$ . Поскольку  $\theta_0$  и  $\theta_1$  неизвестны, невозможно построить взвешенные переменные  $\tilde{Y}_i$ ,  $\tilde{X}_{0i}$  и  $\tilde{X}_{1i}$ . Однако можно оценить  $\theta_0$  и  $\theta_1$  и использовать эти оценки для вычисления оценки

$\text{var}(u_i | X_i)$ . Пусть  $\hat{\theta}_0$  и  $\hat{\theta}_1$  являются оценками  $\theta_0$  и  $\theta_1$ , и пусть  $\widehat{\text{var}}(u_i | X_i) = \hat{\theta}_0 + \hat{\theta}_1 X_i^2$ .

Определим взвешенные регрессоры  $\tilde{Y}_i = Y_i / \sqrt{\widehat{\text{var}}(u_i | X_i)}$ ,  $\tilde{X}_{0i} = 1 / \sqrt{\widehat{\text{var}}(u_i | X_i)}$

и  $\tilde{X}_{1i} = X_i / \sqrt{\widehat{\text{var}}(u_i | X_i)}$ . ВМНК-оценка является МНК-оценкой коэффициентов

в регрессии  $\tilde{Y}_i$  на  $\tilde{X}_{0i}$  и  $\tilde{X}_{1i}$  (где  $\beta_0 \tilde{X}_{0i}$  играет роль константы  $\beta_0$ ).

Для получения этой оценки нужно оценить функцию условной дисперсии, то есть оценить  $\theta_0$  и  $\theta_1$  в уравнении (17.27). Один из способов состоятельно оценить  $\theta_0$  и  $\theta_1$  – вычислить регрессию  $\hat{u}_i^2$  на  $X_i^2$ , используя МНК, где  $\hat{u}_i^2$  – это квадрат  $i$ -го остатка МНК-регрессии.

Предположим, что условная дисперсия имеет вид, описываемый уравнением (17.27), и что  $\hat{\theta}_0$  и  $\hat{\theta}_1$  являются состоятельными оценками  $\theta_0$  и  $\theta_1$ . При предположениях № 1–3 из вставки «Основные понятия 17.1», а также при выполнении дополнительных условий на моменты, которые возникают из-за того, что  $\theta_0$  и  $\theta_1$  заменяются их оценками, асимптотическое распределение ВМНК-оценки такое же, как если бы  $\theta_0$  и  $\theta_1$  были известны. Таким образом, ВМНК-оценка, полученная с использованием оценок  $\hat{\theta}_0$  и  $\hat{\theta}_1$ , имеет приблизительно такое же асимптотическое распределение, как недоступная ВМНК-оценка, и в этом смысле асимптотически является BLUE.

Поскольку ВМНК может быть реализован путем оценки неизвестных параметров условной функции дисперсии, этот метод иногда называют *доступным ВМНК* или *оцененным ВМНК*.

**Пример № 2: дисперсия зависит от третьей переменной.** ВМНК также может быть использован, если условная дисперсия зависит от третьей переменной,  $W_i$ , которая не включена в функцию регрессии. В частности, предположим, что собираются данные о трех переменных,  $Y_i$ ,  $X_i$  и  $W_i$ ,  $i=1, \dots, n$ ; теоретическая функция регрессии зависит от  $X_i$ , но не от  $W_i$ , а условная дисперсия зависит от  $W_i$ , но не от  $X_i$ . Другими словами, теоретическая регрессия имеет вид:  $E(Y_i | X_i, W_i) = \beta_0 + \beta_1 X_i$ , а условная дисперсия задана формулой:  $\text{var}(u_i | X_i, W_i) = \lambda h(W_i)$ ,

где  $\lambda$  является постоянной величиной, а  $h$  является функцией, которую необходимо оценить.

Например, предположим, что исследователь заинтересован в моделировании связи между уровнем безработицы в государстве и государственной экономической политикой, описываемой переменной ( $X_i$ ). Однако измеренный уровень безработицы ( $Y_i$ ) является оценкой истинного уровня безработицы, полученной в результате обследования населения ( $Y_i^*$ ). Таким образом,  $Y_i$  отражает  $Y_i^*$  с ошибкой, причем источником ошибки является случайная погрешность обследования, то есть  $Y_i = Y_i^* + v_i$ , где  $v_i$  – это случайная погрешность измерения, возникшая во время опроса. В этом примере вполне вероятно, что размер выборки данного обследования,  $W_i$ , сам по себе не определяет истинного уровня безработицы. Таким образом, теоретическая функция регрессии не зависит от  $W_i$ , то есть  $E(Y_i^*|X_i, W_i) = \beta_0 + \beta_1 X_i$ . Следовательно, мы имеем два уравнения:

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i^* \quad (17.28)$$

$$\text{и } Y_i = Y_i^* + v_i, \quad (17.29)$$

где уравнение (17.28) моделирует связь между переменной, характеризующей государственную экономическую политику, и истинным уровнем безработицы в государстве, а уравнение (17.29) представляет собой связь между измеренным уровнем безработицы  $Y_i$  и истинным уровнем безработицы  $Y_i^*$ .

Модель, описываемую уравнениями (17.28) и (17.29), можно привести к теоретической регрессии, в которой условная дисперсия ошибки зависит от  $W_i$ , но не от  $X_i$ . Ошибка  $u_i^*$  в равенстве (17.28) представляет различные пропущенные факторы в регрессии, в то время как остаточный член  $v_i$  в уравнении (17.29) представляет собой погрешности измерений, возникающие при измерении безработицы. Если  $u_i^*$  является гомоскедастичной, то дисперсия  $\text{var}(u_i^*|X_i, W_i) = \sigma_{u^*}^2$  постоянна. Однако дисперсия ошибки измерения находится в обратной зависимости от размера выборки  $W_i$  проведенного опроса, то есть  $\text{var}(v_i|X_i, W_i) = a/W_i$ , где  $a$  является постоянной величиной. Поскольку  $v_i$  является случайной ошибкой опроса, можно с уверенностью утверждать, что она не коррелирует с  $u_i^*$ , так что  $\text{var}(u_i^* + v_i|X_i, W_i) = \sigma_{u^*}^2 + a/W_i$ . Таким образом, подставляя уравнение (17.28) в уравнение (17.29), мы получим регрессионную модель с гетероскедастичными ошибками:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad (17.30)$$

$$\text{var}(u_i|X_i, W_i) = \theta_0 + \theta_1 \left( \frac{1}{W_i} \right), \quad (17.31)$$

где  $u_i = u_i^* + v_i$ ,  $\theta_0 = \sigma_{u^*}^2$ ,  $\theta_1 = a$  и  $E(u_i|X_i, W_i) = 0$ .

Если бы  $\theta_0$  и  $\theta_1$  были известны, то функция условной дисперсии в уравнении (17.31) могла бы быть использована для оценки  $\beta_0$  и  $\beta_1$  по ВМНК. В этом примере  $\theta_0$  и  $\theta_1$  неизвестны, но они могут быть оценены при помощи регрессии квадратов МНК-остатков [получаемых в результате оценки МНК-регрессии (17.30)]

на  $1/W_i$ . Тогда оцененная функция условной дисперсии может быть использована для построения весов в доступном ВМНК.

Следует подчеркнуть важность выполнения  $E(u_i|X_i, W_i) = 0$ : в противном случае взвешенные ошибки будут иметь ненулевое условное среднее и ВМНК будет несостоятельным. Другими словами, если  $W_i$  в самом деле связана с  $Y_i$ , то уравнение (17.30) должно быть уравнением множественной регрессии, включающим в себя в качестве регрессоров как  $X_i$ , так и  $W_i$ .

**Общий алгоритм доступного ВМНК.** В общем случае алгоритм доступного ВМНК можно разбить на пять этапов:

1. Оцените регрессию  $Y_i$  на  $X_i$  при помощи МНК и вычислите остатки  $\hat{u}_i$ ,  $i=1,\dots,n$ .  
Оцените модель функции условной дисперсии  $\text{var}(u_i|X_i)$ . Например, если функция условной дисперсии имеет вид, как в уравнении (17.27), для этого потребуется оценить регрессию  $\hat{u}_i^2$  на  $X_i^2$ . В общем случае этот шаг требует оценки функции условной дисперсии  $\text{var}(u_i|X_i)$ .
2. Используйте оцененную функцию для расчета предсказанных значений функции условной дисперсии,  $\widehat{\text{var}}(u_i|X_i)$ .
3. Умножьте зависимую переменную и регрессоры (в том числе и свободный член) на величину, обратную квадратному корню из оцененной функции условной дисперсии.
4. Оцените коэффициенты взвешенной регрессии при помощи МНК; полученные оценки являются ВМНК-оценками.

Программные статистические пакеты обычно включают в себя набор дополнительных команд для взвешенного метода наименьших квадратов, в которых автоматизированы четвертый и пятый этапы.

### **Устойчивые к гетероскедастичности стандартные ошибки или ВМНК?**

Существует два способа корректировки гетероскедастичности: можно оценить  $\beta_0$  и  $\beta_1$  с помощью ВМНК, а можно оценить  $\beta_0$  и  $\beta_1$  с помощью МНК и использовать устойчивые к гетероскедастичности стандартные ошибки. Чтобы решить, какой из подходов следует использовать на практике, требуется сравнить преимущества и недостатки каждого из них.

Преимуществом ВМНК является то, что он дает более эффективные, чем МНК, оценки коэффициентов с использованием сырых (некорректированных) регрессоров, по крайней мере асимптотически. Недостатком ВМНК является то, что для его использования необходимо знать функцию условной дисперсии и оценки ее параметров. Если функция условной дисперсии имеет квадратичную форму, как в уравнении (17.27), то сделать это легко. На практике, однако, функциональная форма условной дисперсии редко бывает известна. Более того, если функциональная форма окажется неправильной, то стандартной ошибки регрессии, вычисленные при помощи ВМНК, являются неверными в том смысле, что они приводят к неправильным статистическим выводам (тесты имеют неправильный размер).

Преимущество использования устойчивых к гетероскедастичности стандартных ошибок заключается в том, что они дают асимптотически верные выводы, даже если вы не знаете вида функции условной дисперсии. Дополнительным преимуществом является то, что устойчивые к гетероскедастичности стандартные ошибки легко вычисляются при помощи специальной опции в современных регрессионных пакетах, так что для их использования не нужно прилагать никаких дополнительных усилий. Недостатком устойчивых к гетероскедастичности стандартных ошибок является то, что МНК-оценка будет иметь большую дисперсию по сравнению с ВМНК-оценкой (полученной на основе верной функции условной дисперсии), по крайней мере асимптотически.

На практике функциональная форма  $\text{var}(u_i | X_i)$  редко бывает известной (практически никогда), что создает проблемы для использования ВМНК в реальных эмпирических приложениях. Решение этой проблемы уже достаточно сложно в случае с одним регрессором, но в приложениях с несколькими регрессорами оно становится еще более трудоемким. По этой причине практическое применение ВМНК требует решения ряда дополнительных проблем. Напротив, в современных статистических пакетах очень легко вычислить устойчивые к гетероскедастичности стандартные ошибки, и получающиеся выводы являются надежными при весьма общих предположениях, в частности, устойчивые к гетероскедастичности стандартные ошибки можно использовать, даже не зная функциональной формы условной дисперсии. По этим причинам, по нашему мнению, несмотря на теоретическую привлекательность ВМНК, устойчивые к гетероскедастичности стандартные ошибки являются лучшим способом справиться с возможной проблемой гетероскедастичности в большинстве эмпирических приложений.

## **Выводы**

1. Асимптотическая нормальность МНК-оценки в сочетании с состоятельностью устойчивых к гетероскедастичности стандартных ошибок гарантирует, что устойчивая к гетероскедастичности  $t$ -статистика имеет асимптотическое стандартное нормальное распределение в условиях нулевой гипотезы, если выполнены первые три предположения МНК из вставки «Основные понятия 17.1».
2. Если ошибки в регрессии являются i.i.d. и условно нормально распределены относительно регрессоров, то  $\hat{\beta}_1$  имеет точное условное нормальное распределение относительно регрессоров. Кроме того,  $t$ -статистика, вычисленная в предположении гомоскедастичности ошибок, имеет точное выборочное распределение Стьюдента  $t_{n-2}$  при нулевой гипотезе.
3. Оценка метода взвешенных наименьших квадратов (ВМНК) является МНК-оценкой, примененной к взвешенной регрессии, где все переменные взвешиваются при помощи квадратного корня из величины, обратной по отношению к условной дисперсии  $\text{var}(u_i | X_i)$ , или ее оценки. Несмотря на то что ВМНК-оценка асимптотически более эффективна по сравнению с методом наименьших квадратов, для реализации ВМНК вы должны знать функциональную форму функции условной дисперсии, что случается очень редко.

## **Основные понятия**

Сходимость по вероятности (с. 707).  
Состоятельная оценка (с. 707).  
Сходимость по распределению (с. 709).  
Асимптотическое распределение (с. 709).  
Теорема Слуцкого (с. 710).  
Теорема о непрерывном отображении (с. 710).  
Взвешенный метод наименьших квадратов (ВМНК) (с. 716).  
ВМНК-оценки (с. 717).  
Недоступный ВМНК (с. 718).  
Доступный ВМНК (с. 718).  
Плотность нормального распределения (с. 726).  
Плотность двухмерного нормального распределения (с. 727).

## **Вопросы для повторения и закрепления основных понятий**

- 17.1. Предположим, что выполнено предположение № 4 из вставки «Основные понятия 17.1», но в большой выборке вы строите 95 %-й доверительный интервал для  $\beta_1$  с помощью устойчивой к гетероскедастичности стандартной ошибки. Будет ли этот доверительный интервал асимптотически надежным в том смысле, что будет содержать истинное значение  $\beta_1$  в 95 % всех повторных выборок большого размера  $n$ ? Будем считать, вместо этого, что предположение № 4 из вставки «Основные понятия 17.1» ложно, но вы строите 95 %-й доверительный интервал для  $\beta_1$  с помощью стандартной ошибки в предположении гомоскедастичности ошибок регрессии в большой выборке. Будет ли этот доверительный интервал надежным асимптотически?
- 17.2. Допустим, что  $A_n$  – это случайная величина, которая сходится по вероятности к 3. Предположим, что  $B_n$  является случайной величиной, которая сходится по распределению к стандартной нормальной случайной величине. Каким будет асимптотическое распределение  $A_n B_n$ ? Используйте полученное асимптотическое распределение для вычисления приближенного значения  $\Pr(A_n B_n) < 2$ .
- 17.3. Предположим, что переменные  $Y$  и  $X$  связаны уравнением регрессии  $Y = 1,0 + 2,0X + u$ . У исследователя есть наблюдения  $Y$  и  $X$ , где  $0 \leq X \leq 20$ , причем условная дисперсия имеет вид:  $\text{var}(u_i | X_i = x) = 1$  при  $0 \leq x \leq 10$  и  $\text{var}(u_i | X_i = x) = 16$  при  $10 \leq x \leq 20$ . Начертите гипотетический график наблюдений  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ . Придает ли ВМНК больший вес наблюдениям с  $x \leq 10$  или  $x > 10$ ? Почему?
- 17.4. Вместо ВМНК исследователь в предыдущей задаче решает использовать МНК-оценку, используя только наблюдения с  $x \leq 10$ , а затем только с  $x > 10$ , после чего рассчитывает среднее двух МНК-оценок. Является ли эта оценка более эффективной, чем ВМНК?

## Упражнения

- 17.1. Рассмотрим модель регрессии без свободного члена  $Y_i = \beta_1 X_i + u_i$  (так что истинное значение константы  $\beta_0$  равно нулю).
- Выполните оценки наименьших квадратов коэффициента  $\beta_1$  для регрессии с ограничением  $Y_i = \beta_1 X_i + u_i$ . Такая оценка называется оценкой наименьших квадратов с ограничениями ( $\hat{\beta}_1^{RLS}$ ) для параметра  $\beta_1$ , поскольку она получается при дополнительном ограничении, которым в данном случае является  $\beta_0 = 0$ .
  - Выполните асимптотическое распределение  $\hat{\beta}_1^{RLS}$  при выполнении предположений № 1–3 из вставки «Основные понятия 17.1».
  - Покажите, что  $\hat{\beta}_1^{RLS}$  является линейной [уравнение (5.24)] и, при выполнении предположений № 1 и № 2 из вставки «Основные понятия 17.1», условно несмещенной [уравнение (5.25)].
  - Выполните условную дисперсию для  $\hat{\beta}_1^{RLS}$  в предположениях теоремы Гаусса–Маркова (предположения № 1–4 из вставки «Основные понятия 17.1»).
  - Сравните условную дисперсию  $\hat{\beta}_1^{RLS}$  из пункта (г) с условной дисперсией МНК-оценки и  $\hat{\beta}_1$  (из регрессии со свободным членом) при выполнении предположений Гаусса–Маркова. Какая оценка является более эффективной? Используйте формулы для дисперсии, чтобы объяснить ваш ответ.
  - Выполните точное выборочное распределение  $\hat{\beta}_1^{RLS}$ , предполагая, что выполняются предположения № 1–5 из вставки «Основные понятия 17.1».
  - Теперь рассмотрим оценки  $\tilde{\beta}_1 = \sum_{i=1}^n Y_i / \sum_{i=1}^n X_i$ . Выполните выражение для  $\text{var}(\tilde{\beta}_1 | X_1, \dots, X_n) - \text{var}(\hat{\beta}_1^{RLS} | X_1, \dots, X_n)$  в предположениях Гаусса–Маркова и используйте это выражение, чтобы показать, что  $\text{var}(\tilde{\beta}_1 | X_1, \dots, X_n) \geq \text{var}(\hat{\beta}_1^{RLS} | X_1, \dots, X_n)$ .
- 17.2. Пусть  $X_i, Y_i$  являются i.i.d. с конечным четвертым моментом. Докажите, что выборочная ковариация является состоятельной оценкой теоретической ковариации, то есть  $s_{XY} \xrightarrow{P} \sigma_{XY}$ , где  $s_{XY}$  определена в уравнении (3.24). (Подсказка: используйте идею из приложения 3.3 и неравенство Коши–Шварца).
- 17.3. В данном упражнении присутствуют детали вывода асимптотического распределения  $\hat{\beta}_1$ , приведенного в приложении 4.3.
- Используйте уравнение (17.19), чтобы получить выражение:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n v_i}}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} - \frac{(\bar{X} - \mu_X) \sqrt{\frac{1}{n} \sum_{i=1}^n u_i}}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2},$$

где  $v_i = (X_i - \mu_X)u_i$ .

- б) Используйте центральную предельную теорему, закон больших чисел и теорему Слуцкого, чтобы показать, что последний член в уравнении из пункта (а) сходится по вероятности к нулю.
- в) С помощью неравенства Коши–Буняковского и третьего предположения МНК из вставки «Основные понятия 17.1» докажите, что  $\text{var}(v_i) < \infty$ . Удовлетворяет ли член  $\sqrt{\frac{1}{n} \sum_{i=1}^n v_i^2} / \sigma_v$  условиям центральной предельной теоремы?
- г) Примените центральную предельную теорему и теорему Слуцкого, чтобы получить результат из уравнения (17.12).
- 17.4. Докажите следующие результаты:
- Покажите, что если  $\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, a^2)$ , где  $a^2$  является константой, то  $\hat{\beta}_1$  состоятельна. (Подсказка: используйте теорему Слуцкого.)
  - Покажите, что если  $s_u^2 / \sigma_u^2 \xrightarrow{p} 1$ , то  $s_u / \sigma_u \xrightarrow{p} 1$ .
- 17.5. Предположим, что  $W$  является случайной величиной с  $E(W^4) < \infty$ . Покажите, что  $E(W^2) < \infty$ .
- 17.6. Покажите, что если  $\hat{\beta}_1$  является условно несмещенной, то она несмещенная, то есть покажите, что если  $E(\hat{\beta}_1 | X_1, \dots, X_n) = \beta_1$ , то  $E(\hat{\beta}_1) = \beta_1$ .
- 17.7. Предположим, что  $X$  и  $u$  – непрерывные случайные величины и  $(X_i, u_i)$ ,  $i=1, \dots, n$  являются i.i.d.
- Покажите, что совместная функция плотности вероятности для  $(u_i, u_j, X_i, X_j)$  может быть записана в таком виде:  $f(u_i, X_i)f(u_j, X_j)$  для  $i \neq j$ , где  $f(u_i, X_i)$  является функцией совместной плотности распределения  $u_i$  и  $X_i$ .
  - Покажите, что  $E(u_i u_j | X_i, X_j) = E(u_i | X_i)E(u_j | X_j)$  для  $i \neq j$ .
  - Покажите, что  $E(u_i u_j | X_1, \dots, X_n) = E(u_i | X_i)E(u_j | X_j)$  для  $i \neq j$ .
- 17.8. Рассмотрим модель регрессии из вставки «Основные понятия 17.1» и предположим, что предпосылки № 1–5 выполнены. Предположим, что предпосылка № 4 заменяется на предположение о том, что  $\text{var}(u_i | X_i) = \theta_0 + \theta_1 |X_i|$ , где  $|X_i|$  обозначает абсолютную величину  $X_i$ ,  $\theta_0 > 0$  и  $\theta_1 \geq 0$ .
- Являются ли МНК-оценки  $\hat{\beta}_1$  наилучшими в классе линейных несмешанных оценок?
  - Предположим, что  $\theta_0$  и  $\theta_1$  известны. Что является наилучшей линейной несмешанной оценкой для  $\beta_1$ ?
  - Выведите точное выборочное условное распределение для МНК-оценки  $(\hat{\beta}_1)$  относительно  $X_1, \dots, X_n$ .
  - Выведите точное выборочное условное распределение ВМНК-оценки  $\hat{\beta}_1$  (рассматривая  $\theta_0$  и  $\theta_1$  как известные величины) относительно  $X_1, \dots, X_n$ .
- 17.9. Докажите равенство (17.16), предполагая, что выполняются предпосылки № 1 и № 2 из вставки «Основные понятия 17.1» и предположение о том, что  $X_i$  и  $u_i$  имеют по восемь моментов.
- 17.10. Пусть  $\hat{\theta}$  это оценка параметра  $\theta$ , причем  $\hat{\theta}$  может быть смещенной. Покажите, что если  $E[(\hat{\theta} - \theta)^2] \rightarrow 0$  при  $n \rightarrow \infty$  (т.е. среднеквадратичная

ошибка  $\hat{\theta}$  стремится к нулю), то тогда  $\hat{\theta} \xrightarrow{p} \theta$ . [Подсказка: используйте уравнение (17.43) с  $W = \hat{\theta} - \theta$ .]

17.11. Предположим, что  $X$  и  $Y$  распределены согласно двумерному нормальному закону распределения с функцией плотности, заданной в уравнении (17.38).

а) Покажите, что функция плотности  $Y$  при условии  $X=x$  может быть представлена в таком виде:

$$f_{Y|X=x}(y) = \frac{1}{\sigma_{Y|X}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_{Y|X}}{\sigma_{Y|X}}\right)^2\right],$$

где  $\sigma_{Y|X} = \sqrt{\sigma_Y^2(1-\rho_{XY}^2)}$  и  $\mu_{Y|X} = \mu_Y - (\sigma_{XY}/\sigma_X^2)(x-\mu_X)$  [Подсказка: используйте определение плотности условной вероятности  $f_{Y|X=x}(y) = [g_{X,Y}(x,y)]/[f_X(x)]$ , где  $g_{X,Y}$  является функцией плотности совместного распределения  $X$  и  $Y$ , а  $f_X$  является функцией плотности безусловного распределения вероятностей случайной величины  $X$ .]

б) Используя результат из части (а), покажите, что  $Y|X=x \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$ .

в) Используя результат из части (б), покажите, что  $E(Y|X=x) = a + bx$  для правильно подобранных  $a$  и  $b$ .

17.12. а) Предположим, что  $u \sim N(0, \sigma_u^2)$ . Покажите, что  $E(e^u) \sim e^{\frac{1}{2}\sigma_u^2}$ .

б) Предположим, что условное распределение  $u$  при условии  $X=x$  имеет вид  $N(0, a+bx^2)$ , где  $a$  и  $b$  положительные числа. Покажите, что

$$E(e^u|X=x) = e^{\frac{1}{2}(a+bx^2)}.$$

17.13. Рассмотрим неоднородную регрессионную модель  $Y_i = \beta_{0i} + \beta_{1i}X_i + u_i$ , где  $\beta_{0i}$  и  $\beta_{1i}$  являются случайными переменными, изменяющимися от наблюдения к наблюдению. Предположим, что  $E(u_i|X_i) = 0$  и  $(\beta_{0i}, \beta_{1i})$  распределены независимо от  $X_i$ .

а) Пусть  $\hat{\beta}_1^{OLS}$  обозначает МНК-оценку  $\beta_1$ , заданную уравнением (17.2).

Покажите, что  $\hat{\beta}_1^{OLS} \xrightarrow{p} E(\beta_1)$ , где  $E(\beta_1)$  это среднее значение  $\beta_{1i}$  в генеральной совокупности. [Подсказка: см. уравнение (13.10).]

б) Предположим, что  $\text{var}(u_i|X_i) = \theta_0 + \theta_1 X_i^2$ , где  $\theta_0$  и  $\theta_1$  – известные положительные числа. Пусть  $\hat{\beta}_1^{WLS}$  обозначает ВМНК-оценку. Верно ли, что  $\hat{\beta}_1^{WLS} \xrightarrow{p} E(\beta_1)$ ? Объясните.

## Приложения

### Приложение 17.1. Нормальное и связанные с ним распределения и моменты непрерывных случайных величин

В данном приложении вводятся определения и рассматриваются нормальное и связанные с ним распределения. Определения распределений хи-квадрат,  $F$  и распределения Стьюдента, приведенные в разделе 2.4, приводятся здесь

повторно для удобства. Мы начнем с определений вероятностных распределений и моментов для непрерывных случайных величин.

### **Вероятности непрерывных случайных величин и их моменты**

Как отмечалось в разделе 2.1, если  $Y$  является непрерывной случайной величиной, то вероятность наступления события, связанного с ней, можно описать с помощью функции плотности распределения (ф.п.р.). Вероятность того, что  $Y$  находится между двумя значениями – это площадь под функцией плотности распределения между этими двумя числами. Однако поскольку  $Y$  является непрерывной случайной величиной, математические выражения для вероятностей описываются интегралами, а не суммами, используемыми для описания дискретных случайных величин.

Пусть  $f_Y$  обозначает функцию плотности распределения  $Y$ . Так как вероятность не может быть отрицательной,  $f_Y(y) \geq 0$  для всех  $y$ . Вероятность того, что  $Y$  находится между  $a$  и  $b$  (где  $a < b$ ) задается формулой:

$$\Pr(a \leq Y \leq b) = \int_a^b f_Y(y) dy. \quad (17.32)$$

Поскольку  $Y$  должна принимать значения на числовой прямой, то  $\Pr(-\infty \leq Y \leq \infty) = 1$ , что подразумевает выполнение равенства  $\int_{-\infty}^{\infty} f_Y(y) dy = 1$ .

Математические ожидания и моменты непрерывных случайных величин, как и в случае дискретных случайных величин, являются их значениями, взвешенными по вероятности, за исключением того, что суммы [например суммирование в уравнении (2.3)] нужно заменить на интегралы. Таким образом, математическое ожидание  $Y$  является

$$E(Y) = \mu_Y = \int y f_Y(y) dy, \quad (17.33)$$

где область интегрирования есть множество значений, для которых  $f_Y$  отлична от нуля. Дисперсия – это ожидаемое значение  $(Y - \mu_Y)^2$ , и  $r$ -й момент случайной величины – это ожидаемое значение  $Y^r$ . Таким образом,

$$\text{var}(Y) = E(Y - \mu_Y)^2 = \int (y - \mu_Y)^2 f_Y(y) dy \text{ и} \quad (17.34)$$

$$E(Y^r) = \int y^r f_Y(y) dy. \quad (17.35)$$

### **Нормальное распределение**

**Одномерное нормальное распределение.** Функция плотности распределения нормально распределенной случайной величины (нормальная ф.п.р.) имеет вид:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], \quad (17.36)$$

где  $\exp(x)$  обозначает показательную функцию от  $x$ . Множитель  $1/(\sigma\sqrt{2\pi})$  в уравнении (17.36) гарантирует, что  $\Pr(-\infty \leq Y \leq \infty) = \int_{-\infty}^{\infty} f_Y(y) dy = 1$ .

Среднее значение нормального распределения равно  $\mu$ , а дисперсия —  $\sigma^2$ . Нормальное распределение является симметричным, поэтому все нечетные центральные моменты третьего порядка и выше равны нулю. Четвертый центральный момент равен  $3\sigma^4$ . В общем случае, если  $Y$  распределена согласно  $N(\mu, \sigma^2)$ , то ее четные центральные моменты задаются формулой:

$$E(Y - \mu)^k = \frac{k!}{2^{k/2} (k/2)!} \sigma^k \quad (k \text{ — четное}). \quad (17.37)$$

Нормальное распределение с параметрами  $\mu = 0$  и  $\sigma^2 = 1$  называют стандартным нормальным распределением. Стандартная нормальная ф.п.р. обозначается  $\varphi$ , а стандартная нормальная функция распределения —  $\Phi$ . Таким образом, функция плотности стандартной нормальной случайной величины имеет вид:  $\varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$ , а распределение —  $\Phi(y) = \int_{-\infty}^y \varphi(s) ds$ .

**Двухмерное нормальное распределение.** Двухмерная нормальная ф.п.р. для двух случайных величин  $X$  и  $Y$  имеет вид:

$$\begin{aligned} g_{X,Y}(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \times \\ &\times \exp\left\{ \frac{1}{-2(1-\rho_{XY}^2)} \left[ \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - \right. \right. \\ &\left. \left. - 2\rho_{XY} \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}, \end{aligned} \quad (17.38)$$

где  $\rho_{XY}$  является коэффициентом корреляции между  $X$  и  $Y$ .

Если  $X$  и  $Y$  не коррелированы ( $\rho_{XY} = 0$ ),  $g_{X,Y}(x, y) = f_X(x)f_Y(y)$ , где  $f$  — это функция плотности нормального распределения, заданная уравнением (17.36). Из этого следует, что если  $X$  и  $Y$  совместно нормально распределены и не коррелированы, то они независимо распределены. Это свойство является особым для нормального распределения, которое, как правило, неверно для других распределений.

Многомерное нормальное распределение является обобщением двухмерного нормального распределения для случая более чем двух случайных величин. Это распределение наиболее удобно определять с использованием матриц. Оно представлено в приложении 18.1.

**Условное нормальное распределение.** Предположим, что  $X$  и  $Y$  являются совместно нормально распределенными. Тогда условное распределение  $Y$  относительно  $X$  будет  $N(\mu_{Y|X}, \sigma_{Y|X}^2)$  со средним  $\mu_{Y|X} = \mu_Y + (\sigma_{XY}/\sigma_X)(X - \mu_X)$  и дисперсией  $\sigma_{Y|X}^2 = (1 - \rho_{XY}^2)\sigma_Y^2$ . Среднее значение этого условного распределения относительно  $X=x$  является линейной функцией  $x$ , а дисперсия не зависит от  $x$  (упражнение 17.11).

## Связанные распределения

**Распределение хи-квадрат.** Пусть  $Z_1, \dots, Z_n$  являются нормальными случайными величинами. Случайная величина

$$W = \sum_{i=1}^n Z_i^2 \quad (17.39)$$

имеет распределение хи-квадрат с  $n$  степенями свободы. Это распределение обозначается  $\chi^2_n$ . Так как  $E(Z_i^2) = 1$  и  $E(Z_i^4) = 3$ , то  $E(W) = n$  и  $\text{var}(W) = 2n$ .

**Распределение Стьюдента.** Пусть  $Z$  имеет стандартное нормальное распределение, пусть  $W$  распределена по закону  $\chi^2_m$  и  $Z$  и  $W$  независимо распределены. Тогда случайная величина

$$t = \frac{Z}{\sqrt{W/m}} \quad (17.40)$$

имеет распределение Стьюдента с  $m$  степенями свободы и обозначается  $t_m$ . Распределение  $t_\infty$  является стандартным нормальным распределением.

**F-распределение.** Пусть  $W_1$  и  $W_2$  – две независимые случайные величины с распределениями хи-квадрат со степенями свободы  $n_1$  и  $n_2$ , соответственно. Тогда случайная величина

$$F = \frac{W_1/n_1}{W_2/n_2} \quad (17.41)$$

имеет F-распределение с  $n_1$  и  $n_2$  степенями свободы. Это распределение обозначается  $F_{n_1, n_2}$ .

F-распределение зависит от числа степеней свободы  $n_1$  числителя и числа степеней свободы  $n_2$  знаменателя. При увеличении числа степеней свободы в знаменателе  $F_{n_1, \infty}$  распределение может быть хорошо приближено при помощи распределения  $\chi^2_{n_1}$ , деленного на  $n_1$ , то есть  $\chi^2_{n_1} / n_1$ .

## Приложение 17.2. Два неравенства

В данном приложении содержится утверждение и доказательство неравенств Чебышева и Коши–Шварца.

### Неравенство Чебышева

В неравенстве Чебышева дисперсия случайной величины  $V$  используется, чтобы ограничить вероятность того, что  $V$  отклоняется от среднего больше, чем на  $\pm \delta$ , где  $\delta$  является положительной константой:

$$\Pr(|V - \mu_V| \geq \delta) \leq \frac{\text{var}(V)}{\delta^2} \quad (\text{неравенство Чебышева}). \quad (17.42)$$

Для доказательства уравнения (17.42) обозначим:

$W = V - \mu_V$ , пусть  $f$  – ф.п.р.  $W$ , а  $\delta$  – произвольное положительное число. Тогда

$$\begin{aligned} E(W^2) &= \int_{-\infty}^{\infty} w^2 f(w) dw = \int_{-\infty}^{-\delta} w^2 f(w) dw + \int_{-\delta}^{\delta} w^2 f(w) dw + \\ &+ \int_{\delta}^{\infty} w^2 f(w) dw \geq \int_{-\infty}^{-\delta} w^2 f(w) dw + \int_{\delta}^{\infty} w^2 f(w) dw \geq \\ &\geq \delta^2 \left[ \int_{-\infty}^{-\delta} f(w) dw + \int_{\delta}^{\infty} f(w) dw \right] = \\ &= \delta^2 \Pr(|W| \geq \delta), \end{aligned} \quad (17.43)$$

где первое равенство является определением  $E(W^2)$ ; второе равенство справедливо, так как пределы интегрирования вместе составляют всю числовую прямую; первое неравенство верно, поскольку член, который был пропущен, неотрицателен; второе неравенство верно, потому что  $w^2 \geq \delta^2$  в пределах интегрирования; финальное равенство верно по определению  $\Pr(|W| \geq \delta)$ . Подставляем  $W = V - \mu_V$  в последнее выражение, заметим, что  $E(W^2) = E[(V - \mu_V)^2] = \text{var}(V)$ , и преобразуем его, чтобы получить неравенство из уравнения (17.42). Если  $V$  является дискретной величиной, то доказательство аналогично приведенному с заменой интегралов на суммы.

### Неравенство Коши–Шварца

Неравенство Коши–Шварца обобщает корреляционное неравенство  $|\rho_{xy}| \leq 1$  на случай ненулевого среднего. Неравенство Коши–Шварца имеет вид:

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)} \quad (\text{неравенство Коши–Шварца}). \quad (17.44)$$

Доказательство неравенства (17.44) аналогично доказательству корреляционного неравенства в приложении 2.1. Пусть  $W = Y + bX$ , где  $b$  является константой. Тогда  $E(W^2) = E(Y^2) + 2bE(XY) + b^2E(X^2)$ . Пусть теперь  $b = -E(XY)/E(X^2)$ , так что (после упрощения) выражение становится таким:  $E(W^2) = E(Y^2) - [E(XY)]^2/E(X^2)$ . Поскольку  $E(W^2) \geq 0$  (так как  $W^2 \geq 0$ ), то верно, что  $[E(XY)]^2 \leq E(X^2)E(Y^2)$ , и неравенство Коши–Шварца получается, если извлечь квадратный корень из этого выражения.

# Глава 18. Теория множественного регрессионного анализа

В данной главе мы рассматриваем основы теории множественного регрессионного анализа. Глава преследует четыре цели. Первая цель – показать, как модель множественной регрессии представляется в матричной форме, что даст нам компактную запись для формул МНК-оценок и  $t$ -статистик. Вторая цель – описать характеристики распределения МНК-оценок как в случае большой выборки (используя асимптотическую теорию), так и в случае малой выборки (если ошибки гомоскедастичны и нормально распределены). Третья цель состоит в том, чтобы изучить методы получения эффективных оценок коэффициентов модели множественной регрессии и описать обобщенный метод наименьших квадратов (ОМНК – GLS). Четвертая цель – кратко рассмотреть асимптотическую теорию для линейной регрессии с инструментальными переменными (IV-регрессии), включая, в том числе, введение в обобщенный метод моментов (GMM) в линейной IV-регрессии с гетероскедастичными ошибками.

Глава начинается с описания модели множественной регрессии и оценок МНК в матричной форме в разделе 18.1. В этом разделе также будут представлены расширенные предположения МНК для модели множественной регрессии. Первые четыре предположения аналогичны предпосылкам МНК, рассмотренным во вставке «Основные понятия 6.4», и относятся к асимптотическим распределениям, обосновывающим процедуры, использованные в главах 6 и 7. Оставшиеся две предпосылки являются более сильными и позволяют изучить подробнее теоретические свойства оценок МНК в модели множественной регрессии.

В следующих трех разделах рассматриваются выборочные распределения МНК-оценок и тестовых статистик. В разделе 18.2 представлены асимптотические распределения МНК-оценок и  $t$ -статистик, учитывающие предположения МНК из вставки «Основные понятия 6.4». В разделе 18.3 объединяются и обобщаются методы проверки гипотез, включая методы тестирования множественных гипотез о коэффициентах модели, представленные в разделах 7.2 и 7.3, и приводится асимптотическое распределение результирующей  $F$ -статистики. В разделе 18.4 мы рассматриваем точное выборочное распределение оценок МНК и статистик в конкретном случае при гомоскедастичных ошибках и их нормальном распределении. Хотя предположения о гомоскедастичности и нормальности ошибок неправдоподобны в большинстве эконометрических прикладных работ, точное выборочное распределение предоставляет теоретический интерес, и  $p$ -значения, вычисленные с помощью этих распределений, часто появляются в результатах разных программных продуктов.

В следующих двух подразделах рассматриваются методы получения эффективных оценок коэффициентов модели множественной регрессии. В разделе 18.5 обобщается теорема Гаусса–Маркова на случай множественной регрессии. В разделе 18.6 разбирается обобщенный метод наименьших квадратов (GLS).

В последнем разделе разбирается модель линейной регрессии с инструментальными переменными при наличии допустимых и сильных инструментов. В разделе приводятся асимптотическое распределение 2МНК-оценки, когда ошибки гетероскедастичны, и формула для расчета стандартных ошибок двухшаговой МНК-оценки. Двухшаговая МНК-оценка – одна из возможных оценок обобщенного метода моментов, и в последнем разделе мы рассматриваем основы обобщенного метода моментов для модели линейной регрессии с инструментальными переменными. Показано, что 2МНК-оценка – это эффективная оценка обобщенного метода моментов, если ошибки гомоскедастичны.

**Математические требования.** Для изложения линейной модели в данной главе используются матричные обозначения и основные инструменты линейной алгебры и предполагается, что читатель уже прослушал курс введения в линейную алгебру. В приложении 18.1 приводится обзор теории векторов, матриц и матричных операторов, используемых в данной главе. Кроме того, многомерное дифференциальное исчисление используется в разделе 18.1 для получения МНК-оценок.

## 18.1. Линейная модель множественной регрессии и МНК-оценки в матричной форме

Линейная модель множественной регрессии и оценки МНК могут быть записаны компактно в матричном представлении.

### Модель множественной регрессии в матричном представлении

Теоретическая модель множественной регрессии (вставка «Основные понятия 6.2») записывается так:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i, \quad i = 1, \dots, n. \quad (18.1)$$

Для того чтобы записать модель множественной регрессии в матричной форме, определим следующие матрицы и векторы:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \vdots \\ \mathbf{X}'_n \end{pmatrix}, \quad \text{и } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad (18.2)$$

где  $\mathbf{Y}$  – это вектор размерности  $n \times 1$ ,  $\mathbf{X}$  – матрица размерности  $n \times (k + 1)$ ,  $\mathbf{U}$  – вектор размерности  $n \times 1$  и  $\boldsymbol{\beta}$  – вектор размерности  $(k + 1) \times 1$ . Здесь и далее векторы и матрицы будут выделены жирным шрифтом. В этих обозначениях:

- $\mathbf{Y}$  – вектор, содержащий  $n$  наблюдений наблюданной переменной, размерности  $n \times 1$ ;

- $X$  – матрица размерности  $n \times (k + 1)$ , которая содержит  $n$  наблюдений  $k+1$  регрессоров (включая константу в качестве свободного члена);
- вектор-столбец размерности  $(k + 1) \times 1$  матрицы  $X_i$  – это  $i$ -е наблюдение  $k+1$  регрессоров, то есть  $X'_i = (1X_{1i}, \dots, X_{ki})$ , где  $X'_i$  – транспонированная матрица  $X_i$ ;
- $U$  – вектор  $n$  ошибок размерности  $n \times 1$ ;
- $\beta$  – вектор размерности  $(k + 1) \times 1$ , который содержит  $k+1$  неизвестный коэффициент регрессии.

Модель множественной регрессии, представленная в уравнении (18.1) для  $i$ -ого наблюдения, записанная с помощью  $\beta$  и  $X_i$  выглядит следующим образом:

$$Y_i = X'_i \beta + u_i, i = 1, \dots, n. \quad (18.3)$$

## ОСНОВНЫЕ ПОНЯТИЯ

### 18.1

#### Расширенные предпосылки МНК в модели множественной регрессии

Линейная модель множественной регрессии:

$$Y_i = X'_i \beta + u_i, i = 1, \dots, n. \quad (18.4)$$

Расширенные предпосылки метода наименьших квадратов:

1.  $E(u_i | X_i) = 0$  ( $u_i$  имеет нулевое условное математическое ожидание).
2.  $(X_i, Y_i) i=1, \dots, n$  – независимо и одинаково распределенные (i.i.d.) случайные величины, взятые из их совместного распределения.
3.  $X_i$  и  $u_i$  имеют ненулевые конечные моменты четвертого порядка.
4.  $X$  имеет полный ранг (нет совершенной мультиколлинеарности).
5.  $\text{var}(u_i | X_i) = \sigma_u^2$  (гомоскедастичность).
6. Условное распределение  $u_i$  относительно  $X_i$  является нормальным (нормальные ошибки).

В уравнении (18.3) первый регрессор является константой, которая всегда равна 1, и коэффициент при этом регрессоре является свободным членом. Таким образом, свободный член не появляется отдельно в уравнении (18.3), а является первым элементом вектора коэффициентов  $\beta$ .

Учитывая все  $n$  наблюдений в уравнении (18.3), получаем модель множественной регрессии в матричной форме:

$$Y = X\beta + U. \quad (18.5)$$

#### Расширенные предпосылки МНК

В расширенные предпосылки МНК для модели множественной регрессии входят четыре предпосылки МНК для множественной регрессии из вставки

«Основные понятия 6.4» и две дополнительные предпосылки о гомоскедастичности и нормальности ошибок. Предпосылка о гомоскедастичности ошибок будет использована при выводе эффективности МНК-оценок, а предпосылка о нормальности применяется для получения конкретных результатов для распределений оценок МНК и их статистик.

Расширенные предпосылки МНК приведены во вставке «Основные понятия 18.1».

Исключая разницу в обозначениях, первые три предпосылки, описанные во вставке «Основные понятия 18.1», идентичны первым трем предпосылкам из вставки «Основные понятия 6.4».

Четвертые предпосылки во вставках «Основные понятия 6.4» и «Основные понятия 18.1» могут показаться разными, но на самом деле это одно и то же: в них просто по-разному сформулировано условие о том, что в модели не должно быть совершенной мультиколлинеарности. Напомним, что совершенная мультиколлинеарность возникает, когда один регрессор можно записать в виде линейной комбинации других регрессоров. В матричных обозначениях уравнения (18.2) совершенная мультиколлинеарность означает, что один столбец  $X$  является линейной комбинацией других столбцов  $X$ , и если это так, тогда матрица  $X$  не имеет полного ранга. Таким образом, когда говорится, что  $X$  имеет ранг  $k+1$ , то есть ранг равен количеству столбцов матрицы  $X$ , то имеется в виду, что регрессоры не являются совершенно мультиколлинеарными, но другими словами.

В пятой предпосылке МНК во вставке «Основные понятия 18.1» говорится о том, что ошибки условно гомоскедастичны, а в шестой — о том, что  $u_i$  имеет нормальное условное распределение относительно  $X_i$ . Эти две предпосылки аналогичны двум последним предпосылкам во вставке «Основные понятия 17.1», исключая тот момент, что теперь они относятся к модели множественной регрессии.

**Применение для вектора математических ожиданий и ковариационной матрицы вектора  $U$ .** Предпосылки МНК во вставке «Основные понятия 18.1» позволяют записать простое выражение для вектора математических ожиданий и ковариационной матрицы для условного распределения  $U$  при данной матрице регрессоров  $X$ . (Вектор средних значений и ковариационная матрица определены в приложении 18.2.) Точнее говоря, из первой и второй предпосылок вставки «Основные понятия 18.1» следует, что  $E(u_i | X_i) = E(u_i | X_i) = 0$  и что  $\text{cov}(u_i, u_j | X) = E(u_i u_j | X) = E(u_i u_j | X_i, X_j) = E(u_i | X_i) E(u_j | X_j) = 0$  для  $i \neq j$  (упражнение 17.7). Из первого, второго и пятого предположений следует, что  $E(u_i^2 | X) = E(u_i^2 | X_i) = \sigma_u^2$ . Объединяя эти результаты, получаем:

$$\text{при выполнении предположений \#1 и \#2, } E(U | X) = \mathbf{0}_n \quad (18.6)$$

и при выполнении предположений \#1, \#2 и \#5,

$$E(UU' | X) = \sigma_u^2 I_n, \quad (18.7)$$

где  $\mathbf{0}_n$  — это нулевой вектор размерности  $n$  и  $I_n$  — единичная матрица размерности  $n \times n$ .

Аналогично, первое, второе, пятое и шестое предположения из вставки «Основные понятия 18.1» позволяют записать, что условное распределение  $n$ -мерного случайного вектора  $U$ , условного по  $X$ , – это многомерное нормальное распределение (определяется в приложении 18.2). Другими словами,

при выполнении предположений #1, #2, #5 и #6

условное распределение  $U$  относительно  $X$  есть

$$N(\mathbf{0}_n, \sigma_u^2 I_n). \quad (18.8)$$

### **MНK-оценка**

МНК-оценка минимизирует сумму квадратов остатков,  $\sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2$  [уравнение (6.8)]. Формула МНК-оценки получается путем взятия первой производной суммы квадратов остатков по каждому элементу вектора коэффициентов, приравнивая производные к нулю, и решая эти уравнения для получения оценки  $\hat{\beta}$ .

Производная суммы квадратов остатков по  $j$ -му коэффициенту  $b_j$  выглядит следующим образом:

$$\begin{aligned} \frac{\partial}{\partial b_j} \sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki})^2 = \\ = -2 \sum_{i=1}^n X_{ji} (Y_i - b_0 - b_1 X_{1i} - \dots - b_k X_{ki}) \end{aligned} \quad (18.9)$$

для  $j = 0, \dots, k$ , где для  $j=0$ ,  $X_{0i} = 1$  для всех  $i$ . Производная с правой стороны уравнения (18.9) – это  $j$ -ый элемент вектора размерности  $k+1$ ,  $-2X'(Y - Xb)$ , где  $b$  – это  $k+1$  вектор, состоящий из  $b_0, \dots, b_k$ . Существует  $k+1$  таких производных, каждая из которых соответствует элементу из  $b$ . Приравнивая производные к нулю и объединяя уравнения в систему из  $k+1$  уравнения, получаем условия первого порядка для МНК-оценки  $\hat{\beta}$ . Другими словами,  $\hat{\beta}$  есть решение системы из  $k+1$  уравнений:

$$X'(Y - X\hat{\beta}) = \mathbf{0}_{k+1} \quad (18.10)$$

или, эквивалентно,  $X'Y = X'X\hat{\beta}$ .

Решение системы (18.10) дает МНК-оценку  $\hat{\beta}$  в матричной форме:

$$\hat{\beta} = (X'X)^{-1} X'Y, \quad (18.11)$$

где  $(X'X)^{-1}$  матрица, обратная к  $X'X$ .

**Роль отсутствия «совершенной мультиколлинеарности».** Четвертая предпосылка МНК из вставки «Основные понятия 18.1» гласит, что  $X$  имеет полный ранг. В свою очередь это означает, что матрица  $X'X$  имеет полный ранг, то есть не вырождена. Так как  $X'X$  не вырождена, она обратима. Таким образом, предпосылка об отсутствии совершенной мультиколлинеарности гарантирует, что  $(X'X)^{-1}$  существует, и поэтому система (18.10) имеет единственное решение

МНК-оценка (18.11) может быть вычислена. Говоря иначе, если бы  $X$  не имела полного ранга, то матрица  $(X'X)^{-1}$  была бы вырождена и решение системы (18.10) не было бы единственным. Следовательно,  $(X'X)^{-1}$  не могла бы быть вычислена, также как  $\hat{\beta}$ .

## 18.2. Асимптотическое распределение МНК-оценок и $t$ -статистик

Если размер выборки большой и выполнены первые четыре предпосылки из вставки «Основные понятия 18.1», то оценки МНК имеют асимптотически нормальное совместное распределение, устойчивые к гетероскедастичности оценки ковариационной матрицы состоятельны, устойчивые к гетероскедастичности  $t$ -статистики оценок МНК имеют асимптотическое стандартное нормальное распределение. Это является результатом приложения концепции многомерного нормального распределения и обобщения центральной предельной теоремы на случай многомерного нормального распределения.

### *Центральная предельная теорема для многомерной случайной величины*

Центральная предельная теорема, рассмотренная во вставке «Основные понятия 2.7», применяется к одномерной случайной величине. Для вывода совместного асимптотического распределения элементов вектора  $\hat{\beta}$  нам нужно расширение центральной предельной теоремы на многомерный случай для применения к вектору случайных величин.

#### Центральная предельная теорема для многомерной случайной величины

Предположим, что  $W_1, \dots, W_n$  –  $m$ -мерный вектор независимых одинаково распределенных случайных величин, вектор средних значений которых равен  $E(W_i) = \mu_W$ , а ковариационная матрица равна  $E[(W_i - \mu_W)(W_i - \mu_W)'] = \Sigma_W$ , где  $\Sigma_W$  – положительно определена и конечна. Пусть  $\bar{W} = \frac{1}{n} \sum_{i=1}^n W_i$ . Тогда  $\sqrt{n}(\bar{W} - \mu_W) \xrightarrow{d} N(\mathbf{0}_m, \Sigma_W)$ .

#### ОСНОВНЫЕ ПОНЯТИЯ

#### 18.2

Центральная предельная теорема для многомерной случайной величины обобщает центральную предельную теорему для одномерной случайной величины на случай вектора выборочных средних значений случайных величин  $\bar{W}$ , где  $\bar{W}$  – вектор размерности  $m$ . Разница между центральной предельной теоремой для скалярной случайной величины и для вектора случайных величин состоит в ограничениях на отклонения. В одномерном случае из вставки «Основные понятия 2.7» требовалось, чтобы дисперсия была одновременно ненулевой

и конечной. В случае вектора случайных величин требование заключается в том, чтобы ковариационная матрица была положительно определена и конечна. Если вектор случайных величин  $\mathbf{W}$  имеет конечную положительно определенную ковариационную матрицу, тогда  $0 < \text{var}(\mathbf{c}'\mathbf{W}) < \infty$  для любого ненулевого вектора  $\mathbf{c}$  размерности  $m$  (в упражнении 18.3).

Центральная предельная теорема для многомерной случайной величины, которая будет использована, сформулирована во вставке «Основные понятия 18.2».

### Асимптотическая нормальность $\hat{\beta}$

В больших выборках оценки МНК имеют многомерное нормальное асимптотическое распределение:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(\mathbf{0}_{k+1}, \Sigma_{\sqrt{n}(\hat{\beta} - \beta)}\right), \text{ где } \Sigma_{\sqrt{n}(\hat{\beta} - \beta)} = \mathbf{Q}_X^{-1} \Sigma_V \mathbf{Q}_X^{-1}, \quad (18.12)$$

где  $\mathbf{Q}_X$  – матрица вторых моментов регрессоров размерности  $(k+1) \times (k+1)$ , то есть  $\mathbf{Q}_X = E(\mathbf{X}'_i \mathbf{X}_i)$ , и  $\Sigma_V$  ковариационная матрица случайной величины  $V_i = X_i u_i$ , размерности  $(k+1) \times (k+1)$ , то есть  $\Sigma_V = E(V_i V_i')$ . Заметим, что вторая предпосылка МНК из вставки «Основные понятия 18.1» означает, что случайные величины  $V_i$ ,  $i = 1, \dots, n$  независимо одинаково распределены.

Записав уравнение (18.12) в терминах  $\hat{\beta}$ , а не в терминах  $\sqrt{n}(\hat{\beta} - \beta)$ , получаем его нормальное приближение.

В больших выборках случайный вектор  $\hat{\beta}$  распределен как  $N(\beta, \Sigma_{\hat{\beta}})$ ,

$$\text{где } \Sigma_{\hat{\beta}} = \frac{\Sigma_{\sqrt{n}(\hat{\beta} - \beta)}}{n} = \mathbf{Q}_X^{-1} \Sigma_V \mathbf{Q}_X^{-1} / n. \quad (18.13)$$

Ковариационная матрица  $\Sigma_{\hat{\beta}}$  в уравнении (18.13) – это ковариационная матрица приближенного нормального распределения  $\hat{\beta}$ , в то время как  $\Sigma_{\sqrt{n}(\hat{\beta} - \beta)}$  в уравнении (18.12) – это ковариационная матрица асимптотически нормального распределения  $\sqrt{n}(\hat{\beta} - \beta)$ . Эти две ковариационные матрицы отличаются на множитель  $n$ , в зависимости от того, масштабируются ли оценки МНК на  $\sqrt{n}$ .

**Выход уравнения (18.12).** Чтобы вывести уравнение (18.12), сначала используем выражения (18.4) и (18.11) и запишем:  $\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{U}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' (\mathbf{X} \beta + \mathbf{U})$ , так что

$$\hat{\beta} = \beta + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{U}. \quad (18.14)$$

Таким образом,  $\hat{\beta} - \beta = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{U}$ , из чего следует:

$$\sqrt{n}(\hat{\beta} - \beta) = \left( \frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}' \mathbf{U}}{\sqrt{n}} \right). \quad (18.15)$$

Выход уравнения (18.12) основан на двух предположениях: во-первых, что матрица в «знаменателе» в (18.15),  $\mathbf{X}' \mathbf{X}/n$ , состоятельная, и, во-вторых, что матрица в числителе,  $\mathbf{X}' \mathbf{U}/\sqrt{n}$ , удовлетворяет условиям многомерной центральной предельной теоремы, сформулированной во вставке «Основные понятия 18.2». Более подробный вывод этого соотношения приведен в приложении 18.3.

### **Устойчивые к гетероскедастичности стандартные ошибки**

Устойчивая при гетероскедастичности оценка  $\Sigma_{\sqrt{n}(\hat{\beta}-\beta)}$  получена путем замены теоретических моментов в определении [уравнение (18.12)] на выборочные моменты. Соответственно, устойчивая к гетероскедастичности оценка ковариационной матрицы случайной величины  $\sqrt{n}(\hat{\beta}-\beta)$  равна:

$$\hat{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)} = \left( \frac{X'X}{n} \right)^{-1} \hat{\Sigma}_{\hat{v}} \left( \frac{X'X}{n} \right), \text{ где } \hat{\Sigma}_{\hat{v}} = \frac{1}{n-k-1} \sum_{i=1}^n X_i X_i' \hat{u}_i^2. \quad (18.16)$$

Оценка  $\hat{\Sigma}_{\hat{v}}$  включает в себя те же поправки к степеням свободы, что и стандартная ошибка (SER) в модели множественной регрессии (раздел 6.4), чтобы скорректировать потенциальное смещение в сторону уменьшения из-за оценки  $k+1$  коэффициента регрессии.

Доказательство того, что  $\hat{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)} \xrightarrow{p} \Sigma_{\sqrt{n}(\hat{\beta}-\beta)}$ , идейно похоже на доказательство состоятельности стандартных ошибок в модели парной регрессии, которые являются устойчивыми при наличии гетероскедастичности, и приведено в разделе 17.3.

**Устойчивые к гетероскедастичности стандартные ошибки.** Устойчивая при наличии гетероскедастичности, оценка ковариационной матрицы  $\hat{\beta}, \hat{\Sigma}_{\hat{\beta}}$  равна:

$$\hat{\Sigma}_{\hat{\beta}} = n^{-1} \hat{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)}. \quad (18.17)$$

Стандартная ошибка  $j$ -ого коэффициента регрессии равна квадратному корню  $j$ -го диагонального элемента матрицы  $\hat{\Sigma}_{\hat{\beta}}$ . Другими словами, стандартная ошибка  $j$ -го коэффициента равна:

$$SE(\hat{\beta}_j) = \sqrt{(\hat{\Sigma}_{\hat{\beta}})_{jj}}, \quad (18.18)$$

где  $(\hat{\Sigma}_{\hat{\beta}})_{jj}$  – элемент  $(j,j)$  матрицы  $\hat{\Sigma}_{\hat{\beta}}$ .

### **Доверительные интервалы для предсказанных изменений**

В разделе 8.1 были описаны два метода для вычисления стандартных ошибок предсказанных изменений, включающие в себя изменение двух или более регрессоров. Существует компактная матричная запись для этих стандартных ошибок и, таким образом, для доверительных интервалов для предсказанных изменений.

Рассмотрим изменение в значении регрессора для  $i$ -ого наблюдения от некоторого начального уровня, скажем  $X_{i,0}$ , до нового значения,  $X_{i,0} + d$ , так что изменение  $X_i$  – это  $\Delta X_i = d$ , где  $d$  – это вектор размерности  $k+1$ . Такое изменение в  $X$  может включать изменение нескольких регрессоров (т.е. нескольких элементов матрицы  $X$ ). Например, если регрессорами являются сама независимая переменная и ее квадрат, то  $d$  будет разностью между последующим и начальным значениями этих двух переменных.

Ожидаемый эффект влияния от изменения в  $X_j$  – это  $d'\beta$ , и оценка этого изменения будет  $d'\hat{\beta}$ . Так как линейная комбинация нормально распределенных случайных величин нормально распределена, то  $\sqrt{n}(d'\hat{\beta} - d'\beta) = d'\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, d'\Sigma_{\sqrt{n}(\hat{\beta} - \beta)} d)$ . Таким образом, стандартная ошибка прогнозируемого эффекта равна  $(d'\hat{\Sigma}_{\hat{\beta}} d)^{1/2}$ . 95 %-й доверительный интервал для прогнозируемого эффекта равен:

$$d'\hat{\beta} \pm 1,96 \sqrt{d'\hat{\Sigma}_{\hat{\beta}} d}. \quad (18.19)$$

### **Асимптотическое распределение t-статистик**

*t*-статистика для проверки гипотезы  $\beta_j = \beta_{j,0}$ , построенная с использованием устойчивых к гетероскедастичности стандартных ошибок из уравнения (18.18), приведена во вставке «Основные понятия 7.1». Доказательство того, что *t*-статистика имеет асимптотически нормальное распределение аналогично доказательству, приведенному в разделе 17.3 для случая модели парной регрессии.

### **18.3. Проверка совместных гипотез**

В разделе 7.2 рассматриваются методы проверки совместных гипотез, включающих несколько ограничений, где каждое ограничение включает в себя один коэффициент, и в разделе 7.3 рассматривается одно ограничение, в котором участвует два или более коэффициента. Введение матриц в разделе 18.1 позволяет объединить представление этих двух типов гипотез в виде линейного ограничения на вектор коэффициентов, где каждое ограничение может включать несколько коэффициентов. При выполнении первых четырех предпосылок МНК из вставки «Основные понятия 18.1» устойчивая к гетероскедастичности МНК *F*-статистика для тестирования этих гипотез имеет  $F_{q,\infty}$  асимптотическое распределение.

#### **Совместные гипотезы в матричном обозначении**

Рассмотрим совместную гипотезу, которая линейна по коэффициентам, и введем  $q$  ограничений, где  $q \leq k+1$ . Каждое из этих ограничений может включать один или более коэффициентов регрессии. Совместная нулевая гипотеза может быть записана в матричной форме:

$$R\beta = r, \quad (18.20)$$

где  $R$  – неслучайная матрица размерности  $q \times (k+1)$ , имеющая полный ранг, и  $r$  – неслучайный вектор размерности  $q \times 1$ . Количество строк в  $R$  равно  $q$ , то есть числу ограничений в нулевой гипотезе.

Нулевая гипотеза в уравнении (18.20) включает в себя все виды нулевых гипотез, рассмотренных в разделах 7.2 и 7.3. Например, совместная гипотеза, рассмотренная в разделе 7.2, имеет вид:  $\beta_0 = 0, \beta_1 = 0, \dots, \beta_{q-1} = 0$ . Для того что-

бы записать эту совместную гипотезу в виде уравнения (18.20), положим  $\mathbf{R} = \begin{bmatrix} \mathbf{I}_q & \mathbf{0}_{q \times (k+1-q)} \end{bmatrix}$  и  $\mathbf{r} = \mathbf{0}_q$ .

Используя запись (18.20), можно сформулировать ограничения из раздела 7.3, включающие несколько коэффициентов регрессии. Например, если  $k = 2$ , тогда гипотеза  $\beta_1 + \beta_2 = 1$  может быть записана в форме выражения (18.20), если положить  $\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$ ,  $\mathbf{r} = 1$  и  $q = 1$ .

### Асимптотическое распределение F-статистики

Устойчивая к гетероскедастичности F-статистика для тестирования совместной гипотезы в выражении (18.20) имеет вид:

$$F = (\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}\hat{\Sigma}_{\hat{\beta}}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) / q. \quad (18.21)$$

Если первые четыре предпосылки из вставки «Основные понятия 18.1» выполняются, то в условиях нулевой гипотезы

$$F \xrightarrow{d} F_{q, \infty}. \quad (18.22)$$

Этот результат следует из объединения свойств асимптотической нормальности  $\hat{\beta}$  и состоятельности устойчивой к гетероскедастичности оценки  $\hat{\Sigma}_{\sqrt{n}(\hat{\beta} - \beta)}$  ковариационной матрицы. Точнее говоря, заметим, во-первых, что из выражений (18.12) и (18.74) в приложении 18.2 вытекает, что в условиях нулевой гипотезы  $\sqrt{n}(\mathbf{R}\hat{\beta} - \mathbf{r}) = \sqrt{n}\mathbf{R}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{R}\Sigma_{\sqrt{n}(\hat{\beta} - \beta)}\mathbf{R}')$ . Из выражения (18.77) следует, что в условиях нулевой гипотезы должно выполняться  $(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}\hat{\Sigma}_{\hat{\beta}}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) = [\sqrt{n}\mathbf{R}(\hat{\beta} - \beta)]' [\mathbf{R}\Sigma_{\sqrt{n}(\hat{\beta} - \beta)}\mathbf{R}']^{-1} [\sqrt{n}\mathbf{R}(\hat{\beta} - \beta)] \xrightarrow{d} \chi_q^2$ . Однако, так как  $\hat{\Sigma}_{\sqrt{n}(\hat{\beta} - \beta)} \xrightarrow{p} \Sigma_{\sqrt{n}(\hat{\beta} - \beta)}$ , из теоремы Слуцкого следует, что  $[\sqrt{n}\mathbf{R}(\hat{\beta} - \beta)]' [\mathbf{R}\hat{\Sigma}_{\sqrt{n}(\hat{\beta} - \beta)}\mathbf{R}']^{-1} [\sqrt{n}\mathbf{R}(\hat{\beta} - \beta)] \xrightarrow{d} \chi_q^2$ , или, что эквивалентно,  $F \xrightarrow{d} \chi_q^2 / q$  (так как  $\Sigma_{\hat{\beta}} = \Sigma_{\beta \sqrt{n}(\hat{\beta} - \beta)} / n$ ), которая в свою очередь распределена как  $F_{q, \infty}$ .

### Доверительные интервалы для нескольких переменных

Как говорилось в разделе 7.4, асимптотически состоятельные доверительные интервалы для двух или более элементов вектора  $\beta$  могут быть получены как набор значений, которые не отвергаются F-статистикой, когда не отвергается нулевая гипотеза. В принципе этот набор значений может быть вычислен повторяющимися вычислениями F-статистики для многих значений  $\beta$ , но, в этом случае, также как и в случае с доверительным интервалом для одного коэффициента, проще вывести формулу для доверительного интервала в явном виде.

Далее описана процедура построения доверительного интервала для двух или более элементов вектора  $\beta$ . Пусть  $\delta$  будет  $q$ -мерным вектором, состоящим из коэффициентов, для которых мы хотим получить доверительное множество.

Например, если мы хотим получить доверительное множество для коэффициентов регрессии  $\beta_1$  и  $\beta_2$ , тогда  $q = 2$  и  $\delta = (\beta_1 \ \beta_2)'$ . В общем случае мы можем записать:  $\delta = R\beta$ , где матрица  $R$  состоит из нулей и единиц [как отмечалось в обсуждении после выражения (18.20)].  $F$ -статистика для проверки гипотезы  $\delta = \delta_0$  равна:  $F = (\hat{\delta} - \delta_0)' [R\hat{\Sigma}_{\hat{\beta}} R']^{-1} (\hat{\delta} - \delta_0) / q$ , где  $\hat{\delta} = R\hat{\beta}$ . 95 %-е доверительное множество для  $\delta$  – это множество из таких  $\delta_0$ , которые не отвергаются при помощи  $F$ -статистики. Иными словами, при  $\delta = R\beta$ , 95 %-е доверительное множество для  $\delta$  – это

$$\left\{ \delta : (\hat{\delta} - \delta_0)' [R\hat{\Sigma}_{\hat{\beta}} R']^{-1} (\hat{\delta} - \delta_0) / q \leq c \right\}, \quad (18.23)$$

где  $c$  – это 95-й процентиль (5 %-е критическое значение)  $F_{q, \infty}$ -распределения.

Множество, описанное выражением (18.23), состоит из всех точек, находящихся внутри эллипса, полученного при выполнении неравенства (18.23) как равенства (при  $q > 2$  получается эллипсоид). Таким образом, доверительное множество для  $\delta$  может быть получено вычислением решения для уравнения (18.23) на границе эллипса.

## 18.4. Распределение статистик регрессии с нормальными ошибками

Распределения, представленные в разделах 18.2 и 18.3, полученные при помощи закона больших чисел и центральной предельной теоремы, используются в больших выборках. Если, однако, ошибки гомоскедастичны и нормально распределены относительно  $X$ , тогда МНК-оценки имеют многомерное нормальное распределение в конечной выборке относительно  $X$ . В дополнение к этому распределение квадрата стандартной ошибки в конечной выборке пропорционально распределению хи-квадрат с  $n-k-1$  степенями свободы, МНК  $t$ -статистики, рассчитанная в предположении гомоскедастичности, имеет распределение Стьюдента с  $n-k-1$  степенями свободы и  $F$ -статистика, рассчитанная в предположении гомоскедастичности, имеет  $F_{q, n-k-1}$ -распределение. В доказательствах используются некоторые специальные матричные формулы, которые представлены вначале.

### Матричное представление статистик в МНК

Прогнозируемые значения, остатки и сумма квадратов остатков имеют компактное матричное представление. Эти представления используют две матрицы,  $P_x$  и  $M_x$ .

**Матрицы  $P_x$  и  $M_x$ .** Алгебра МНК в модели множественной регрессии основана на использовании двух симметричных матриц размерности  $n \times n$ ,  $P_x$  и  $M_x$ :

$$P_x = X(X'X)^{-1}X' \text{ и} \quad (18.24)$$

$$M_x = I_n - P_x. \quad (18.25)$$

Матрица  $\mathbf{C}$  называется идемпотентной, если  $\mathbf{C}$  квадратная и  $\mathbf{CC}=\mathbf{C}$  (см. приложение 18.1). Так как  $\mathbf{P}_x = \mathbf{P}_x \mathbf{P}_x$  и  $\mathbf{M}_x = \mathbf{M}_x \mathbf{M}_x$  (упражнение 18.5) и так как  $\mathbf{P}_x$  и  $\mathbf{M}_x$  – симметричные матрицы, получается, что  $\mathbf{P}_x$  и  $\mathbf{M}_x$  – симметричные идемпотентные матрицы.

Матрицы  $\mathbf{P}_x$  и  $\mathbf{M}_x$  обладают несколькими полезными свойствами, которые следуют прямо из определений, приведенных в (18.24) и (18.25):

$$\begin{aligned} \mathbf{P}_x \mathbf{X} &= \mathbf{X} \text{ и } \mathbf{M}_x \mathbf{X} = \mathbf{0}_{n \times (k+1)}; \\ \text{rank}(\mathbf{P}_x) &= k+1 \text{ и } \text{rank}(\mathbf{M}_x) = n-k-1, \end{aligned} \quad (18.26)$$

где  $\text{rank}(\mathbf{P}_x)$  это ранг матрицы  $\mathbf{P}_x$ .

Матрицы  $\mathbf{P}_x$  и  $\mathbf{M}_x$  могут быть использованы, чтобы провести декомпозицию  $n$ -мерного вектора  $\mathbf{Z}$  на две части: на часть, которая натянута на столбцы матрицы  $\mathbf{X}$ , и на часть, которая ортогональна столбцам матрицы  $\mathbf{X}$ . Другими словами,  $\mathbf{P}_x \mathbf{Z}$  проекция вектора  $\mathbf{Z}$  на пространство, образованное столбцами  $\mathbf{X}$ ,  $\mathbf{M}_x \mathbf{Z}$ , это часть  $\mathbf{Z}$ , которая ортогональна столбцам  $\mathbf{X}$  и  $\mathbf{Z} = \mathbf{P}_x \mathbf{Z} + \mathbf{M}_x \mathbf{Z}$ .

**Предсказанные значения и остатки МНК.** Матрицы  $\mathbf{P}_x$  и  $\mathbf{M}_x$  дают возможность простым образом записать выражения для предсказанных значений МНК и остатков. Предсказанные значения МНК  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$  и остатки  $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}}$  могут быть выражены как показано ниже (упражнение 18.5):

$$\hat{\mathbf{Y}} = \mathbf{P}_x \mathbf{Y} \quad (18.27)$$

$$\text{и } \hat{\mathbf{U}} = \mathbf{M}_x \mathbf{Y} = \mathbf{M}_x \mathbf{U}. \quad (18.28)$$

Выражения (18.27) и (18.28) дают простое доказательство того, что остатки МНК и предсказанные значения ортогональны, то есть что выражение (4.37) выполняется:  $\hat{\mathbf{Y}}' \hat{\mathbf{U}} = \mathbf{Y}' \mathbf{P}_x' \mathbf{M}_x \mathbf{Y} = 0$ , где второе равенство следует из  $\mathbf{P}_x' \mathbf{M}_x = \mathbf{0}_{n \times n}$ , которое в свою очередь следует из  $\mathbf{M}_x \mathbf{X} = \mathbf{0}_{n \times (k+1)}$  из (18.26).

**Стандартные ошибки регрессии.** SER, определенные в 4.3, это  $s_{\hat{u}}$ ,

$$\text{где } s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-k-1} \hat{\mathbf{U}}' \hat{\mathbf{U}} = \frac{1}{n-k-1} \mathbf{U}' \mathbf{M}_x \mathbf{U}, \quad (18.29)$$

где последнее равенство следует из  $\hat{\mathbf{U}}' \hat{\mathbf{U}} = (\mathbf{M}_x \mathbf{U})' (\mathbf{M}_x \mathbf{U}) = \mathbf{U}' \mathbf{M}_x \mathbf{M}_x \mathbf{U} = \mathbf{U}' \mathbf{M}_x \mathbf{U}$  (так как  $\mathbf{M}_x$  – симметричная и идемпотентная матрица).

### Распределение $\boldsymbol{\beta}$ при нормальных ошибках

Так как  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{U}$  (18.14) и так как условное распределение  $\mathbf{U}$  относительно  $\mathbf{X}$ , по предположению, является  $N(0_n, \sigma_u^2 I_n)$  (выражение 18.8), условное распределение  $\hat{\boldsymbol{\beta}}$  при данном  $\mathbf{X}$  является многомерным нормальным распределением со средним  $\boldsymbol{\beta}$ . Ковариационная матрица  $\hat{\boldsymbol{\beta}}$ , условная относительно  $\mathbf{X}$ , равна  $\Sigma_{\hat{\boldsymbol{\beta}}|\mathbf{X}} = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X}] = E[(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{U} \mathbf{U}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} | \mathbf{X}] = = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' (\sigma_u^2 I_n) \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} = \sigma_u^2 (\mathbf{X}' \mathbf{X})^{-1}$ . Соответственно, при выполнении всех

предпосылок из вставки «Основные понятия 18.1», в конечной выборке условное распределение  $\hat{\beta}$  при данном  $X$  имеет вид:

$$\hat{\beta} \sim N\left(\beta, \Sigma_{\hat{\beta}|X}\right), \text{ где } \Sigma_{\hat{\beta}|X} = \sigma_u^2 (X'X)^{-1}. \quad (18.30)$$

### **Распределение $s_{\hat{\beta}}^2$**

Если выполняются все предположения из вставки «Основные понятия 18.1», то  $s_{\hat{\beta}}^2$  имеет точное выборочное распределение, пропорциональное распределению хи-квадрат с  $n - k - 1$  степенями свободы:

$$s_{\hat{\beta}}^2 \sim \frac{\sigma_u^2}{n - k - 1} \times \chi_{n-k-1}^2. \quad (18.31)$$

Доказательство выражения (18.31) начинается с (18.29). Так как  $U$  условно нормально распределена относительно  $X$  и так как  $M_X$  симметрична и идемпотентна, квадратичная форма  $U'M_XU/\sigma_u^2$  имеет точное распределение хи-квадрат со степенями свободы, равными рангу  $M_X$  [выражение (18.78) из приложения 18.2]. Из выражения (18.26) следует, что ранг матрицы  $M_X$  равен  $n - k - 1$ . Таким образом,  $U'M_XU/\sigma_u^2$  имеет  $\chi_{n-k-1}^2$  распределение, из которого следует (18.31).

Корректировка на степени свободы гарантирует несмещенность  $s_{\hat{\beta}}^2$ . Математическое ожидание случайной величины, распределенной как  $\chi_{n-k-1}^2$ , равно  $n - k - 1$ ; поэтому  $E(U'M_XU) = (n - k - 1)\sigma_u^2$ , так что  $E(s_{\hat{\beta}}^2) = \sigma_u^2$ .

### **Стандартные ошибки, рассчитанные при наличии гомоскедастичности остатков**

Оценка  $\tilde{\Sigma}_{\hat{\beta}}$  ковариационной матрицы  $\hat{\beta}$ , условной по  $X$ , при гомоскедастичности получена подстановкой выборочной дисперсии вместо теоритической дисперсии  $\sigma_u^2$  в выражение для  $\Sigma_{\hat{\beta}|X}$  в формулу (18.30). Получаем:

$$\tilde{\Sigma}_{\hat{\beta}} = s_u^2 (X'X)^{-1} \text{ (при гомоскедастичности).} \quad (18.32)$$

Оценка дисперсии нормально условно распределенного  $\hat{\beta}_j$  при заданной  $X$  это элемент  $(j,j)$  матрицы  $\tilde{\Sigma}_{\hat{\beta}}$ . Таким образом, стандартная ошибка при наличии гомоскедастичности  $\hat{\beta}_j$  – это квадратный корень  $j$ -го диагонального элемента  $\tilde{\Sigma}_{\hat{\beta}}$ . Получаем, что стандартная ошибка  $\hat{\beta}_j$  при наличии гомоскедастичности

$$\widetilde{SE}\left(\hat{\beta}_j\right) = \sqrt{\left(\tilde{\Sigma}_{\hat{\beta}}\right)_{jj}} \quad (\text{для случая гомоскедастичности}). \quad (18.33)$$

### **Распределение $t$ -статистики**

Пусть  $\hat{t}$  является  $t$ -статистикой для проверки гипотезы  $\beta_j = \beta_{j,0}$ , вычисленной в предположении гомоскедастичности ошибок, то есть пусть:

$$\hat{t} = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\left(\tilde{\Sigma}_{\hat{\beta}}\right)_{jj}}}. \quad (18.34)$$

При выполнении всех шести расширенных предпосылок МНК из вставки «Основные понятия 18.1» точным распределением  $\tilde{t}$  является распределение Стьюдента с  $n - k - 1$  степенями свободы:

$$\tilde{t} \sim t_{n-k-1}. \quad (18.35)$$

Доказательство (18.35) приведено в приложении 18.4.

### **Распределение F-статистики**

Если выполняются все предпосылки из вставки «Основные понятия 18.1», то  $F$ -статистика для проверки гипотезы, выписанной в (18.20), построенная с использованием оценки ковариационной матрицы, рассчитанной в предположении наличия гомоскедастичности, имеет точное  $F_{q, n-k-1}$ -распределение в условиях нулевой гипотезы.

**$F$ -статистика при гомоскедастичности.**  $F$ -статистика, рассчитанная в предположении гомоскедастичности ошибок, похожа на  $F$ -статистику, устойчивую к гетероскедастичности, из (18.21), за исключением того факта, что оценка  $\hat{\Sigma}_{\hat{\beta}}$  при наличии гомоскедастичности используется вместо оценки  $\hat{\Sigma}_{\hat{\beta}}$ , устойчивой к гетероскедастичности. Подставляя выражение  $\tilde{\Sigma}_{\hat{\beta}} = s_u^2 (\mathbf{X}' \mathbf{X})^{-1}$  в выражение для  $F$ -статистики (18.21), получаем выражение для  $F$ -статистики для тестирования нулевой гипотезы (18.20) при наличии гомоскедастичности:

$$\tilde{F} = \frac{(\mathbf{R}\hat{\beta} - r)' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'] (\mathbf{R}\hat{\beta} - r) / q}{s_u^2}. \quad (18.36)$$

Если все предпосылки из вставки «Основные понятия 18.1» выполняются, тогда в условиях выполнения нулевой гипотезы

$$\tilde{F} \sim F_{q, n-k-1}. \quad (18.37)$$

Доказательство выражения (18.37) приведено в приложении 18.4.

$F$ -статистика в (18.36) называется  $F$ -статистикой Вальда (в честь математика Абрахама Вальда). Несмотря на то что формула для  $F$ -статистики при наличии гомоскедастичности ошибок, приведенная в (7.13), на первый взгляд отличается от формулы для статистики Вальда в (18.36),  $F$ -статистика, рассчитанная при наличии гомоскедастичности, и статистика Вальда представляют собой две версии одной и той же статистики. Другими словами, эти два выражения эквивалентны, что показано в упражнении 18.13.

### **18.5. Эффективность МНК-оценки при наличии гомоскедастичности ошибок**

При выполнении условий Гаусса–Маркова для множественной регрессии, МНК-оценка  $\hat{\beta}$  эффективна в классе линейных условно несмешанных оценок, то есть оценка МНК является BLUE.

#### **Условия Гаусса–Маркова для множественной регрессии**

Условия Гаусса–Маркова для множественной регрессии имеют вид:

$$(i) E(\mathbf{U} | \mathbf{X}) = \mathbf{0}_n,$$

$$(ii) E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$$

и (iii)  $\mathbf{X}$  имеет полный ранг системы столбцов. (18.38)

Условия Гаусса–Маркова для множественной регрессии, в свою очередь, вытекают из первых пяти предпосылок, описанных во вставке «Основные понятия 18.1» [см. (18.6) и (18.7)]. Условия, приведенные в (18.38), обобщают условия теоремы Гаусса–Маркова со случая парной модели до множественной регрессии. [При использовании матричных обозначений второе и третье условия теоремы Гаусса–Маркова из (5.31) переходят в единственное условие (ii) в (18.38).]

### **Линейные условно несмешенные оценки**

Мы начинаем с описания класса линейных несмешенных оценок и доказательства того, что оценки принадлежат этому классу.

**Класс линейных условно несмешенных оценок.** Считается, что оценка вектора  $\beta$  линейна, если она является линейной функцией от  $Y_1, \dots, Y_n$ . Соответственно, оценка  $\tilde{\beta}$  линейна по  $Y$ , если она может быть записана в таком виде:

$$\tilde{\beta} = \mathbf{A}' \mathbf{Y}, \quad (18.39)$$

где  $\mathbf{A}$  – матрица весов размерности  $n \times (k+1)$ , которые могут зависеть от  $X$  и неслучайной константы, но не от  $Y$ .

Оценка является условно несмешенной, если математическое ожидание его условного выборочного распределения относительно данного  $X$  равно  $\beta$ . Иначе говоря,  $\tilde{\beta}$  не смешена условно, если  $E(\tilde{\beta}|X) = \beta$ .

**МНК-оценка линейна и условно несмешена.** Сравнение выражений (18.11) и (18.39) показывает, что МНК-оценки линейны по  $Y$ ; более точно,  $\hat{\beta} = \mathbf{A}' \mathbf{Y}$ , где  $\hat{\mathbf{A}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ . Чтобы показать, что  $\hat{\beta}$  условно несмешен, вспомним из (18.14), что  $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{U}$ . Взяв математическое ожидание от обеих сторон выражения, получаем:  $E(\hat{\beta}|X) = \beta + E[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{U}|X] = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' E(\mathbf{U}|X) = \beta$ , где последнее равенство вытекает из первого условия Гаусса–Маркова  $E(\mathbf{U}|X) = 0$ .

### **Теорема Гаусса–Маркова для множественной регрессии**

**Теорема Гаусса–Маркова для множественной регрессии** дает формулировку условий, при которых оценка МНК эффективна в классе линейных условно несмешенных оценок. Однако здесь появляется щекотливый момент, так как  $\hat{\beta}$  является вектором и его «дисперсия» – это ковариационная матрица. Если «дисперсия» оценки является матрицей, что мы имеем в виду, когда говорим, что одна оценка имеет дисперсию меньше, чем другая?

В теореме Гаусса–Маркова справляются с этой проблемой путем сравнения дисперсии возможной оценки линейной комбинации элементов  $\beta$  с дисперсией, соответствующей линейной комбинации  $\tilde{\beta}$ . Если говорить точнее, то пусть  $c$  – это вектор размерности  $k+1$ , и рассмотрим проблему оценки линейной комбинации  $c'\beta$ , используя в качестве оценки один из возможных вариантов  $c'\tilde{\beta}$ .

(где  $\tilde{\beta}$  – линейная несмешенная оценка), с одной стороны, и  $c'\hat{\beta}$  – с другой. Так как обе величины,  $c'\tilde{\beta}$  и  $c'\hat{\beta}$ , являются скалярными величинами и обе являются линейными условно несмешенными оценками  $c'\beta$ , имеет смысл сравнивать их дисперсии.

В теореме Гаусса–Маркова для множественной регрессии утверждается, что МНК-оценка  $c'\beta$  эффективна, то есть МНК-оценка  $c'\hat{\beta}$  имеет наименьшую дисперсию в классе линейных условно несмешенных оценок  $c'\tilde{\beta}$ . Замечательно, что это верно при любой линейной комбинации. В этом смысле МНК-оценка множественной регрессии является BLUE.

Теорема Гаусса–Маркова сформулирована во вставке «Основные понятия 18.3» и доказана в приложении 18.5.

### Теорема Гаусса–Маркова для множественной регрессии

Предположим, что условия теоремы для множественной регрессии (18.38) выполнены. В этом случае МНК-оценка  $\hat{\beta}$  является BLUE. Другими словами, пусть  $\tilde{\beta}$  – линейная условно несмешенная оценка  $\beta$  и пусть  $c$  – неслучайный вектор размерности  $k+1$ . Тогда  $\text{var}(c'\hat{\beta}|X) \leq \text{var}(c'\tilde{\beta}|X)$  для любого ненулевого вектора  $c$ , и неравенство выполняется как равенство при всех  $c$  тогда и только тогда, когда  $\tilde{\beta} = \hat{\beta}$ .

**ОСНОВНЫЕ  
ПОНЯТИЯ**

**18.3**

## 18.6. Обобщенный метод наименьших квадратов<sup>1</sup>

Во многих приложениях используется предположение о независимости и одинаковой распределенности выборки. Например, предположим, что  $Y_i$  и  $X_i$  содержат данные об индивидах, такие как их заработка, образование и личные характеристики, и индивиды были выбраны из генеральной совокупности простым случайным образом. В данном случае, вследствие использования простого случайного выбора  $(X_i, Y_i)$ , как и необходимо, независимо одинаково распределены. Так как  $(X_i, Y_i)$  и  $(X_j, Y_j)$  независимо распределены при  $i \neq j$ , ошибки  $u_i$  и  $u_j$  также независимо распределены при  $i \neq j$ . Это, в свою очередь, означает, что  $u_i$  и  $u_j$  некоррелированы при  $i \neq j$ . В терминах предпосылок теоремы Гаусса–Маркова, предпосылка о том, что  $E(UU'|X)$  диагональна, выполняется, если данные собраны так, что наблюдения независимо распределены.

<sup>1</sup> Обобщенный метод наименьших квадратов (generalized least square, далее – GLS) был представлен в разделе 15.5 в контексте модели регрессии с распределенными лагами. Данный раздел (раздел 18.6) содержит математические выкладки и может изучаться независимо от раздела 15.5, но его предварительное изучение поможет лучше понять данный раздел.

Однако некоторые схемы сбора данных, встречающиеся в эконометрике, не позволяют получать независимые наблюдения и могут приводить к ошибкам  $u_i$ , которые будут коррелированы от одного наблюдения к другому. Самый яркий пример – это временные ряды, то есть данные, которые собираются в течение какого-то времени для одного и того же объекта. Как говорилось в разделе 15.3, в регрессиях, включающих временные ряды, многие невключенные факторы коррелируют от одного периода к другому, и это может отражаться в ошибках регрессии (отражающих эти пропущенные факторы), которые коррелированы от одного периода к другому. Другими словами, вектор ошибок в одном периоде, в общем случае, не будет распределен независимо от вектора ошибок в следующем периоде. Наоборот, ошибки в одном периоде могут быть коррелированы с ошибками в другом периоде.

Присутствие коррелированных ошибок создает две проблемы для использования результатов, основанных на МНК. Во-первых, ни устойчивые при гетероскедастичности ошибки, ни ошибки при наличии гомоскедастичности, полученные при помощи МНК, не обеспечивают надежной основы для выводов. Решение проблемы заключается в использовании стандартных ошибок, которые устойчивы и к гетероскедастичности и к корреляции ошибок по наблюдениям. Этот вопрос – получение состоятельной при наличии гетероскедастичности и коррелированности (НАС) оценки ковариационной матрицы – рассматривался в разделе 15.4, и мы не будем касаться его здесь.

Во-вторых, если ошибки коррелированы по наблюдениям, то  $E(UU' | X)$  не будет диагональной, и второе требование теоремы Гаусса–Маркова из (18.38) не выполняется, и оценка МНК не является BLUE. В этом разделе мы рассматриваем оценку *обобщенного метода наименьших квадратов* (ОМНК), которая является BLUE (по крайней мере асимптотически), когда матрица условных ковариаций ошибок более не пропорциональна единичной матрице. Отдельным случаем ОМНК является взвешенный метод наименьших квадратов, рассмотренный в разделе 17.5, матрица условных ковариаций является диагональной и  $i$ -й диагональный элемент – это функция от  $X_i$ . Как и ВМНК, ОМНК преобразовывает модель регрессии так, чтобы ошибки измененной модели удовлетворяли условиям Гаусса–Маркова. ОМНК-оценка – это МНК-оценка преобразованной модели.

### **Предпосылки обобщенного метода наименьших квадратов**

Существует четыре предпосылки, при выполнении которых ОМНК является эффективным. Первая предпосылка ОМНК подразумевает, что  $u_i$  имеет нулевое условное среднее относительно  $X_1, \dots, X_n$ :

$$E(U|X) = \mathbf{0}_n. \quad (18.40)$$

Эта предпосылка вытекает из двух первых предпосылок МНК из вставки «Основные понятия 18.1», то есть если  $E(u_i | X_i) = 0$  и  $(X_i, Y_i), i=1, \dots, n$ , независимо одинаково распределены, то  $E(U|X) = \mathbf{0}_n$ . Тем не менее в ОМНК мы не требуем выполнения предпосылки о независимости и одинаковости распределений; в конце концов цель ОМНК – справиться с последствиями того, что

ошибки коррелированы по наблюдениям. Мы обсудим важность предпосылки (18.40) после изучения ОМНК-оценки.

Вторая предпосылка ОМНК говорит о том, что матрица условных ковариаций вектора  $\mathbf{U}$  – это некая функция от  $\mathbf{X}$ :

$$E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \Omega(\mathbf{X}), \quad (18.41)$$

где  $\Omega(\mathbf{X})$  – это  $n \times n$  матричная функция от  $\mathbf{X}$ .

### Предпосылки обобщенного метода наименьших квадратов

В линейной регрессионной модели  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} = \mathbf{U}$  предпосылки ОМНК следующие:

1.  $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}_n$ .
2.  $E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \Omega(\mathbf{X})$ , где  $\Omega(\mathbf{X})$  положительно определенная матрица  $n \times n$ , которая может зависеть от  $\mathbf{X}$ .
3.  $X_i$  и  $u_i$  удовлетворяют необходимым моментным условиям.
4.  $\mathbf{X}$  имеет полный ранг системы столбцов (нет явной мультиколлинеарности).

### ОСНОВНЫЕ ПОНЯТИЯ

18.4

Существует два главных приложения ОМНК, которые описываются этими двумя предпосылками. Первое приложение – это независимая случайная выборка с гетероскедастичными ошибками, в этом случае  $\Omega(\mathbf{X})$  – это диагональная матрица с диагональным элементом  $\lambda h(X_i)$ , где  $\lambda$  константа, а  $h$  функция. В этом случае, который мы обсудили в разделе 17.5, ОМНК – это ВМНК.

Вторая предпосылка отражает случай гомоскедастичных ошибок, которые серийно коррелированы. На практике в таком случае разрабатывается модель для серийной корреляции. Например, одной из моделей может быть модель, в которой коррелированы только соседние ошибки, так что  $\text{corr}(u_i, u_{i-1}) = \rho \neq 0$ , но  $\text{corr}(u_i, u_{i-1}) = 0$ , если  $|i - j| \geq 2$ . В этом случае  $\Omega(\mathbf{X})$  имеет  $\sigma_u^2$  на главной диагонали, элементы  $\rho\sigma_u^2$ , располагающиеся над и под главной диагональю, и нули в остальной части матрицы. Таким образом,  $\Omega(\mathbf{X})$  не зависит от  $\mathbf{X}$ ,  $\Omega_{ii} = \sigma_u^2$ ,  $\Omega_{ij} = \rho\sigma_u^2$  при  $|i - j| = 1$ ,  $\Omega_{ij} = 0$  при  $|i - j| > 1$ . Другие модели серийной корреляции, в том числе авторегрессионные модели первого порядка, рассмотрены в контексте ОМНК в разделе 15.5 (см. также упражнение 18.8).

Одна из предпосылок, которая упоминалась во всех предыдущих перечнях предположений МНК для межъобъектных данных, говорит о том, что  $X_i$  и  $u_i$  имеют ненулевые конечные четвертые моменты. В случае с ОМНК специфические требования к моментам требуются для вывода асимптотических результатов, зависящих от вида функции  $\Omega(\mathbf{X})$ , известна ли она или оценивается, и рассматриваемых статистик (ОМНК-оценки,  $t$ -статистики и т.д.). Так как требования зависят от ситуации и модели, мы не приводим специфичные моментные предпосылки здесь, и обсуждение свойств ОМНК в больших выборках предполагает, что такие моментные условия применяются к конкретному рассматриваемому

случаю. Таким образом, третье предположение ОМНК говорит о том, что  $X_i$  и  $u_i$  удовлетворяют подходящим моментным условиям.

Четвертое требование – это требование того, что  $X$  имеет полный ранг по столбцам, то есть в данных отсутствует совершенная мультиколлинеарность.

Предпосылки к ОМНК представлены во вставке «Основные понятия 18.4».

Мы рассматриваем ОМНК-оценку для двух случаев. В первом случае  $\Omega(X)$  известна. Во втором случае  $\Omega(X)$  известна с точностью до некоторых параметров, которые можно оценить. Для упрощения обозначений мы будем ссылаться на  $\Omega(X)$  как на матрицу  $\Omega$ , так что зависимость  $\Omega$  от  $X$  подразумевается.

### **ОМНК при известной $\Omega$**

Когда  $\Omega$  известна, то в ОМНК-оценке она используется, чтобы преобразовать регрессионную модель в модель, которая удовлетворяла бы условиям теоремы Гаусса–Маркова. Более точно, пусть матрица  $F$  равна квадратному корню из  $\Omega^{-1}$ ; то есть пусть  $F$  – матрица, которая удовлетворяет условию  $F'F=\Omega^{-1}$  (см. приложение 18.1). Матрица  $F$  обладает свойством  $F\Omega F'=I_n$ . Теперь умножим слева обе части выражения (18.4) на  $F$ , чтобы получить:

$$\tilde{Y} = \tilde{X}\beta + \tilde{U}, \quad (18.42)$$

где  $\tilde{Y} = FY$ ,  $\tilde{X} = FX$ ,  $\tilde{U} = FU$ .

Ключевая идея ОМНК заключается в том, что при выполнении четырех требований ОМНК условия Гаусса–Маркова выполняются для измененной регрессии (18.42). Иначе говоря, изменения все переменные умножением на квадратный корень из обратной матрицы  $\Omega$ , ошибки в измененной модели имеют нулевое условное среднее и ковариационную матрицу, которая равна единичной матрице. Для того чтобы показать это математически, во-первых, заметим, что  $E(\tilde{U}|\tilde{X}) = E(FU|FX) = FE(U|FX) = 0_n$ , по первой предпосылке ОМНК [см. (18.40)]. В дополнение к этому  $E(\tilde{U}\tilde{U}'|\tilde{X}) = E[(FU)(FU)'|FX] = FE(UU'|FX)F' = F\Omega F' = I_n$ , где второе равенство следует из  $(FU) = UF'$  и последнее равенство следует из определения  $F$ . Следовательно, измененная регрессионная модель в (18.42) удовлетворяет условиям теоремы Гаусса–Маркова из раздела «Основные понятия 18.3».

ОМНК-оценка  $\tilde{\beta}^{GLS}$  – это МНК-оценка  $\beta$  в уравнении (18.42), то есть  $\tilde{\beta}^{GLS} = (\tilde{X}'\tilde{X})^{-1}(\tilde{X}'\tilde{Y})$ . Так как измененная регрессия удовлетворяет условиям теоремы Гаусса–Маркова, ОМНК-оценка является лучшей условно несмещенной оценкой, которая линейна по  $\tilde{Y}$ . С учетом того что  $\tilde{Y} = FY$  и  $F$  (здесь) по нашим предположениям известна и так как  $F$  обратима (так как  $\Omega$  положительно определена), класс оценок, которые линейны по  $\tilde{Y}$ , совпадает с классом оценок, линейных по  $Y$ . Таким образом, МНК-оценка  $\beta$  в (18.42) также является лучшей условно несмещенной оценкой в классе линейных оценок по  $Y$ . Другими словами, при предпосылках ОМНК оценка ОМНК является BLUE.

ОМНК-оценка может быть выражена прямо через  $\Omega$ , так что в принципе нет необходимости вычислять значение  $F$ . Так как  $\tilde{X} = FX$  и  $\tilde{Y} = FY$ ,  $\tilde{\beta}^{GLS} = (X'F'FX)^{-1}(X'F'FY)$ . Но  $F'F = \Omega^{-1}$ , поэтому

$$\tilde{\beta}^{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y). \quad (18.43)$$

На практике  $\Omega$  обычно неизвестна, поэтому ОМНК-оценка в (18.43) не может быть вычислена, и, таким образом, эту оценку часто называют *недоступной ОМНК-оценкой*. Если, тем не менее,  $\Omega$  имеет известную функциональную форму, но параметры функции неизвестны, то  $\Omega$  может быть оценена, и может быть вычислена доступная версия ОМНК-оценки.

### **GLS в случае, когда $\Omega$ содержит неизвестные параметры**

Если функциональная форма  $\Omega$  известна и зависит от некоторых параметров, которые, в свою очередь, могут быть оценены, то эти оцененные параметры могут быть использованы для вычисления оценки ковариационной матрицы  $\Omega$ . Например, рассмотрим случай временных рядов. Как было упомянуто выше после (18.41), в этом случае  $\Omega(X)$  не зависит от  $X$ ,  $\Omega_{ii} = \sigma_u^2$ ,  $\Omega_{ij} = \rho\sigma_u^2$  для  $|i - j| = 1$  и  $\Omega_{ij} = 0$  для  $|i - j| > 1$ . В этом случае  $\Omega$  имеет два неизвестных параметра,  $\sigma_u^2$  и  $\rho$ . Эти параметры могут быть оценены как остатки от предварительной МНК-регрессии; более точно, вместо  $\sigma_u^2$  может быть использована оценка  $s_u^2$ , а  $\rho$  может быть оценена как коэффициент выборочной корреляции между всеми соседними парами остатков МНК. Эти оцененные параметры, в свою очередь, могут быть использованы для вычисления оценки для  $\Omega$ ,  $\hat{\Omega}$ .

В общем случае предполагаем, что у нас есть оценка  $\hat{\Omega}$  для  $\Omega$ . Тогда ОМНК-оценка, основанная на  $\hat{\Omega}$ , равна:

$$\hat{\beta}^{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y). \quad (18.44)$$

ОМНК-оценка в (18.44) иногда называется *доступной ОМНК-оценкой*, потому что она может вычислена, если ковариационная матрица содержит несколько неизвестных параметров, которые могут быть оценены.

### **Предпосылка об условном нулевом среднем и ОМНК**

Чтобы МНК-оценка была состоятельной, должно выполняться первое требование МНК, то есть  $E(u_i | X_i) = 0$  должно быть равно нулю. В отличие от этого, первая предпосылка ОМНК выглядит как  $E(u_i | X_1, \dots, X_n) = 0$ . Выражаясь иначе, первая предпосылка МНК говорит о том, что ошибки для  $i$ -го наблюдения имеют нулевое условное среднее при данном значении регрессора для этого наблюдения, тогда как первая предпосылка ОМНК заключается в том, что  $u_i$  имеет нулевое условное среднее при заданных значениях регрессоров для всех наблюдений.

Как обсуждалось в разделе 18.1, из  $E(u_i | X_i) = 0$  и из того, что выборка является i.i.d., следует, что  $E(u_i | X_1, \dots, X_n) = 0$ . Таким образом, когда выборка является i.i.d., и поэтому ОМНК является ВМНК, первое предположение ОМНК подразумевает первое предположение метода наименьших квадратов из вставки «Основные понятия 18.1».

Однако если выборка не является i.i.d., первое предположение ОМНК не предполагает, что  $E(u_i | X_i) = 0$ , то есть первое предположение ОМНК является

более строгим. Несмотря на то что различие между этими двумя условиями может показаться небольшим, оно может быть очень важным при применении к временным рядам. Это различие обсуждается в разделе 15.5 в контексте вопроса о том, является ли регрессор экзогенным в «настоящем и прошлом» или «строго» экзогенным; предпосылка о том, что  $E(u_i | X_1, \dots, X_n) = 0$ , соответствует строгой экзогенности. Здесь мы обсудим это различие на более общем уровне, используя матричные обозначения. Для того чтобы сделать это, мы сфокусируем внимание на случае, когда  $\mathbf{U}$  гомоскедастична,  $\Omega$  известна и  $\Omega$  имеет ненулевые внедиагональные элементы.

**Роль первой предпосылки ОМНК.** Чтобы увидеть источник различий между этими двумя предпосылками, полезно сравнить состоятельность ОМНК и МНК.

Сначала мы схематически покажем состоятельность ОМНК-оценки, приведенной в (18.43). Подставим выражение (18.4) в выражение (18.43) и получим:  $\tilde{\beta}^{GLS} = \beta + (X'\Omega^{-1}X/n)^{-1}(X'\Omega^{-1}U/n)$ . С учетом первой предпосылки ОМНК  $E(X'\Omega^{-1}U) = E[X'\Omega^{-1}E(U|X)] = \mathbf{0}_n$ . Если, кроме того, дисперсия  $X'\Omega^{-1}U/n$  стремится к нулю и  $X'\Omega^{-1}X/n \xrightarrow{p} \tilde{Q}$ , где  $\tilde{Q}$  – некая обратимая матрица, то  $\tilde{\beta}^{GLS} \xrightarrow{p} \beta$ . Когда  $\Omega$  имеет внедиагональные элементы, член  $X'\Omega^{-1}U = \sum_{i=1}^n \sum_{j=1}^n X_i (\Omega^{-1})_{ij} u_j$  включает произведение  $X_i$  и  $u_i$  для разных  $i, j$ , где  $(\Omega^{-1})_{ij}$  обозначает элемент  $(i, j)$  матрицы  $\Omega^{-1}$ . Таким образом, чтобы матрица  $X'\Omega^{-1}U$  имела нулевое среднее, недостаточно, чтобы  $E(u_i | X_i) = 0$ , скорее,  $E(u_i | X_i)$  должно быть равно нулю для всех пар  $i, j$  – таких, при которых соответствующие  $(\Omega^{-1})_{ij}$  являются ненулевыми. В зависимости от ковариационной структуры ошибок только некоторые элементы или все элементы  $(\Omega^{-1})_{ij}$  могут быть ненулевыми. Например, если  $u_i$  следует процессу авторегрессии первого порядка (как обсуждалось в разделе 15.5), единственные ненулевые элементы  $(\Omega^{-1})_{ij}$  те, для которых  $|i - j| \leq 1$ . В целом, тем не менее, все элементы матрицы  $\Omega^{-1}$  могут быть ненулевыми, так что в общем случае, для того чтобы  $X'\Omega^{-1}U/n \xrightarrow{p} \mathbf{0}_{(k+1)\times 1}$  (и, следовательно,  $\tilde{\beta}^{GLS}$  была состоятельна), нам необходимо, чтобы  $E(U|X) = \mathbf{0}_n$ , то есть должно выполняться первое требование ОМНК.

Напротив, напомним обсуждение состоятельности ОМНК-оценок. Перепишем выражение (18.14) как  $\hat{\beta} = \beta + (X'X/n)^{-1} \frac{1}{n} \sum_{i=1}^n X_i u_i$ . Если  $E(u_i | X_i) = 0$ , то член  $\frac{1}{n} \sum_{i=1}^n X_i u_i$  имеет нулевое среднее, и если этот член имеет дисперсию, сходящуюся к нулю, то он стремится к нулю по вероятности, если в дополнение к этому  $X'X/n \xrightarrow{p} Q_X$ , то  $\hat{\beta} \xrightarrow{p} \beta$ .

**Является ли первое требование ОМНК ограничивающим?** Первая предпосылка ОМНК требует, чтобы ошибки  $i$ -го наблюдения не коррелировали с регрессорами для всех наблюдений. Это ограничение вызывает опасения для применения к некоторым временным рядам. Данный вопрос обсуждался в разделе 15.6 в контексте эмпирического примера взаимосвязи между изменением в цене контракта будущей поставки концентратра замороженного апельсинового сока и погоды

во Флориде. Как объяснялось ранее, правдоподобным выглядит предположение о том, что вектор ошибок в регрессии изменения цены от погоды не коррелирует с текущими и прошлыми значениями погоды, так что первое требование МНК выполняется. Тем не менее этот вектор ошибок явно коррелирует с будущими значениями погоды, так что первая предпосылка ОМНК не выполняется.

Этот пример иллюстрирует распространенный в экономических временных рядах феномен, возникающий, когда сегодняшнее значение переменной частично зависит от будущих ожиданий: эти будущие ожидания обычно подразумевают, что вектор ошибок сегодня зависит от прогноза регрессоров завтра, которые в свою очередь коррелированы с реальными значениями регрессоров на завтра. По этой причине первая предпосылка ОМНК намного строже, чем первая предпосылка МНК. Соответственно, при применении к временным рядам ОМНК-оценка не является состоятельной, даже если МНК-оценка является таковой.

## 18.7. Инструментальные переменные и обобщенный метод моментов

В данном разделе мы рассматриваем введение в теорию инструментальных переменных (далее IV) и асимптотическое распределение IV-оценок. Предполагается, что на протяжении всего раздела предпосылки метода инструментальных переменных из вставок «Основные понятия 12.3 и 12.4» выполняются, и, более того, инструменты сильны. Эти требования применяются к межобъектным данным с одинаковыми независимо распределенными наблюдениями. При определенных условиях результаты, полученные в этом разделе, также применимы и к временным рядам, и приложение к временным рядам кратко освещается в конце раздела. Все асимптотические результаты в этом разделе были получены в условиях предпосылки использования сильных инструментов.

Этот раздел начинается с представления регрессионной модели с инструментальными переменными, оценки двухшагового метода наименьших квадратов (2МНК, TSLS), его асимптотических распределений в общем случае при наличии гетероскедастичности в матричной форме. Далее показано, что в частном случае гомоскедастичности 2МНК-оценка является асимптотически эффективной в классе IV-оценок, в которых инструменты представляют собой линейную комбинацию экзогенных переменных. Более того,  $J$ -статистика имеет асимптотическое хи-квадрат распределение, в котором количество степеней свободы равно количеству сверхидентифицируемых ограничений. Раздел заканчивается обсуждением эффективной IV-оценки и теста на сверхидентифицируемые ограничения, когда ошибки гетероскедастичны – ситуация, когда эффективная IV-оценка известна как эффективная оценка обобщенного метода моментов (GMM).

### **Матричная запись метода инструментальных переменных**

Пусть в этом разделе  $X$  обозначает матрицу регрессоров размерности  $n \times (k + r + 1)$  в интересующем нас уравнении, так что  $X$  содержит включенные эндогенные регрессоры ( $X'$  из вставки «Основные понятия 12.1») и включенные

экзогенные регрессоры ( $W$ 'ы из вставки «Основные понятия 12.1»). Иначе говоря, в терминах вставки «Основные понятия 12.1»,  $i$ -я строка матрицы  $X$  – это  $X_i = (1 \ X_{1i} \ X_{2i} \dots \ X_{ki} \ W_{1i} \ W_{2i} \dots \ W_{ri})$ . Также, пусть  $Z$  обозначает матрицу  $n \times (m+r+1)$ , содержащую все экзогенные регрессоры, и входящие в интересующее нас уравнение ( $W$ 'ы), и те, что исключены из интересующего нас уравнения (инструменты). То есть в обозначениях вставки «Основные понятия 12.1»,  $i$ -я строка  $Z$  – это  $Z_i = (1 \ Z_{1i} \ Z_{2i} \dots \ Z_{ki} \ W_{1i} \ W_{2i} \dots \ W_{ri})$ .

В этих обозначениях регрессия с инструментальными переменными из вставки «Основные понятия 12.1», записанная в матричной формуле:

$$Y = X\beta + U, \quad (18.45)$$

где  $U$  – это вектор ошибок  $n \times 1$  в интересующем уравнении с  $i$ -м элементом  $u_i$ .

Матрица  $Z$  содержит все экзогенные регрессоры, так что при выполнении предпосылок IV регрессии из вставки «Основные понятия 12.4» получаем:

$$E(Z_i u_i) = \mathbf{0} \text{ (инструментальная экзогенность).} \quad (18.46)$$

Так как в регрессию включено  $k$  эндогенных регрессоров, на первом шаге мы имеем  $k$  уравнений.

**2МНК-оценка.** 2МНК-оценка – это оценка метода инструментальных переменных, в которой инструменты – это предсказанные значения  $X$ , основанные на МНК-оценке регрессии на первом шаге. Пусть  $\hat{X}$  обозначает матрицу предсказанных значений, так что  $i$ -я строка  $\hat{X}$  – это  $(1 \ \hat{X}_{1i} \ \hat{X}_{2i} \dots \ \hat{X}_{ki} \ W_{1i} \ W_{2i} \dots \ W_{ri})$ , где  $\hat{X}_{1i}$  является предсказанным значением из регрессии  $X_{1i}$  на  $Z$ , и так далее. Так как элементы  $W$  содержатся в  $Z$ , предсказанные из регрессии  $W_{1i}$  на  $Z$  значения это просто  $W_{1i}$ , и так далее, так что  $\hat{X} = P_z X$ , где  $P_z = Z(Z'Z)^{-1}Z'$  [см. (18.27)]. Соответственно, 2МНК-оценка имеет вид:

$$\hat{\beta}^{TSLS} = (\hat{X}' \hat{X})^{-1} \hat{X}' Y. \quad (18.47)$$

Так как  $\hat{X} = P_z X$ ,  $\hat{X}' \hat{X} = X' P_z X$ , и  $\hat{X}' Y = X' P_z Y$ , 2МНК-оценка может быть переписана так:

$$\hat{\beta}^{TSLS} := (X' P_z X)^{-1} X' P_z Y. \quad (18.48)$$

### Асимптотическое распределение 2МНК-оценки

Подставив выражение (18.45) в (18.48), перегруппировав его и умножив на  $\sqrt{n}$ , получаем выражение для центрированной и нормированной 2МНК-оценки:

$$\begin{aligned} \sqrt{n}(\hat{\beta}^{TSLS} - \beta) &= \left( \frac{X' P_z X}{n} \right)^{-1} \frac{X' P_z U}{\sqrt{n}} = \\ &= \left[ \frac{X' Z}{n} \left( \frac{Z' Z}{n} \right)^{-1} \frac{Z' X}{n} \right]^{-1} \left[ \frac{X' Z}{n} \left( \frac{Z' Z}{n} \right)^{-1} \frac{Z' u}{\sqrt{n}} \right], \end{aligned} \quad (18.49)$$

где во втором равенстве используется определение  $P_z$ . При выполнении предпосылок метода инструментальных переменных имеем:  $X' Z / n \xrightarrow{\rho} Q_{xz}$

и  $\mathbf{Z}'\mathbf{Z}/n \xrightarrow{\rho} \mathbf{Q}_{zz}$ , где  $\mathbf{Q}_{xz} = E(\mathbf{X}_i\mathbf{Z}'_i)$  и  $\mathbf{Q}_{zz} = E(\mathbf{Z}_i\mathbf{Z}'_i)$ . В дополнение к этому при выполнении IV-предпосылок, величина  $\mathbf{Z}_i u_i$  является независимо одинаково распределенной с нулевым средним [см. (18.46)] и ненулевой конечной дисперсией, так что их сумма, деленная на  $\sqrt{n}$ , удовлетворяет условиям центральной предельной теоремы:

$$\mathbf{Z}'\mathbf{U}/\sqrt{n} \xrightarrow{d} \mathbf{\Psi}_{zu}, \text{ где } \mathbf{\Psi}_{zu} \sim N(0, \mathbf{H}) \text{ и } \mathbf{H} = E(\mathbf{Z}_i\mathbf{Z}'_i u_i^2) \quad (18.50)$$

и  $\mathbf{\Psi}_{zu}$  вектор размерности  $(m+r+1) \times 1$ .

Применяя уравнение (18.50) и пределы  $\mathbf{X}'\mathbf{Z}/n \xrightarrow{\rho} \mathbf{Q}_{xz}$  и  $\mathbf{Z}'\mathbf{Z}/n \xrightarrow{\rho} \mathbf{Q}_{zz}$  к выражению (18.49), получаем, что при выполнении IV-предпосылок 2МНК-оценка асимптотически нормально распределена:

$$\sqrt{n}(\hat{\beta}^{TSLS} - \beta) \xrightarrow{d} (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{\Psi}_{zu} \sim N(\mathbf{0}, \Sigma^{TSLS}), \quad (18.51)$$

$$\text{где } \Sigma^{TSLS} = (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{H}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx} (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \quad (18.52)$$

и  $\mathbf{H}$  определена в выражении (18.50).

**Стандартные ошибки для 2МНК.** Формула в выражении (18.52) выглядит пугающе. Тем не менее она предоставляет способ оценить  $\Sigma^{TSLS}$ , подставляя выборочные моменты вместо теоретических моментов. Результирующая оценка дисперсии равна

$$\hat{\Sigma}^{TSLS} = (\hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx})^{-1} \hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{H}}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx} (\hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx})^{-1}, \quad (18.53)$$

$$\text{где } \hat{\mathbf{Q}}_{xz} = \mathbf{X}'\mathbf{Z}/n, \hat{\mathbf{Q}}_{zz} = \mathbf{Z}'\mathbf{Z}/n, \hat{\mathbf{Q}}_{zx} = \mathbf{Z}'\mathbf{X}/n$$

$$\text{и } \hat{\mathbf{H}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}'_i \hat{u}_i^2, \text{ где } \hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{TSLS}, \quad (18.54)$$

так что  $\hat{\mathbf{U}}$  – вектор 2МНК-остатков и  $\hat{u}_i$  – это  $i$ -й элемент этого вектора (2МНК-остаток для  $i$ -го наблюдения).

2МНК-стандартные ошибки равны квадратному корню из диагонального элемента матрицы  $\hat{\Sigma}^{TSLS}$ .

### Свойства 2МНК при наличии гомоскедастичности ошибок

Когда ошибки гомоскедастичны, тогда 2МНК-оценка асимптотически эффективна в классе инструментальных оценок, в которых инструменты являются линейной комбинацией строк  $\mathbf{Z}$ . Этот результат для метода инструментальных переменных является аналогом теоремы Гаусса – Маркова и представляет собой важный аргумент в пользу использования 2МНК.

**2МНК-распределение при наличии гомоскедастичности ошибок.** Если ошибки гомоскедастичны, то есть если  $E(u_i^2 | \mathbf{Z}_i) = \sigma_u^2$ , то  $\mathbf{H} = E(\mathbf{Z}_i\mathbf{Z}'_i u_i^2) = E[E(\mathbf{Z}_i\mathbf{Z}'_i u_i^2 | \mathbf{Z}_i)] = E[\mathbf{Z}_i\mathbf{Z}'_i E(u_i^2 | \mathbf{Z}_i)] = \mathbf{Q}_{zz}\sigma_u^2$ . В этом случае дисперсия асимптотического распределения 2МНК-оценки в выражении (18.52) упрощается до вида:

$$\Sigma^{TSLS} = (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \sigma_u^2 \text{ (при наличии гомоскедастичности).} \quad (18.55)$$

2МНК-оценка ковариационной матрицы при наличии гомоскедастичности равна:

$$\tilde{\Sigma}^{TSLS} = (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \hat{\sigma}_u^2,$$

где  $\hat{\sigma}_u^2 = \frac{\hat{\mathbf{U}}' \hat{\mathbf{U}}}{n - k - r - 1}$  (при наличии гомоскедастичности), (18.56)

и стандартные ошибки 2МНК при наличии гомоскедастичности равны квадратному корню из диагональных элементов  $\tilde{\Sigma}^{TSLS}$ .

**Класс IV-оценок, которые используют линейную комбинацию  $Z$ .** Класс IV-оценок, использующих линейную комбинацию  $Z$  в качестве инструментов, может быть описан двумя эквивалентными способами.

В первом случае проблема оценки рассматривается как задача минимизации квадратичной целевой функции, так же как и МНК-оценка получается путем минимизации суммы квадратов остатков. При предпосылке об экзогенности инструментов ошибки  $\mathbf{U} = \mathbf{Y} - \mathbf{X}\beta$  не коррелированы с экзогенными регрессорами, то есть при истинном значении  $\beta$  из выражения (18.46) следует, что

$$E[(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{Z}] = 0. \quad (18.57)$$

Выражение (18.57) представляет систему из  $m + r + 1$  уравнений, включающих  $k + r + 1$  неизвестных элементов из  $\beta$ . Когда  $m > k$ , эти уравнения становятся избыточными, в том смысле что все удовлетворяют истинному значению  $\beta$ . Когда эти теоретические моменты заменяются на их выборочные аналоги, система уравнений  $(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{Z} = 0$  может быть решена для  $b$  в случае точной идентификации ( $m = k$ ). Это значение  $b$  и есть IV-оценка  $\beta$ . Однако в случае сверхидентификации ( $m > k$ ) система уравнений обычно не может быть решена относительно некоторого значения  $b$ , так как у нас есть больше уравнений, чем неизвестных, и в общем случае эта система не имеет решения.

Одним из подходов к решению проблемы оценки  $\beta$  в случае сверхидентификации заключаются в следующем. Пусть  $A$  является симметричной положительно полуопределенной матрицей весов размерности  $(m + r + 1) \times (m + r + 1)$  и пусть  $\hat{\beta}_A^{IV}$  является оценкой, которая минимизирует:

$$\min_b (\mathbf{Y} - \mathbf{X}b)' \mathbf{Z} A \mathbf{Z}' (\mathbf{Y} - \mathbf{X}b). \quad (18.58)$$

Решение к этой задачи можно найти, взяв производную от целевой функции по  $b$ , и затем приравнять результат к нулю и перегруппировать. Делая так, получаем  $\hat{\beta}_A^{IV}$  – IV-оценку, основанную на матрице весов  $A$ :

$$\hat{\beta}_A^{IV} = (\mathbf{X}' \mathbf{Z} A \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} A \mathbf{Z}' \mathbf{Y}. \quad (18.59)$$

Сравнение выражений (18.59) и (18.48) показывает, что 2МНК – это IV-оценка с  $A = (\mathbf{Z}' \mathbf{Z})^{-1}$ . Другими словами, 2МНК-оценка и есть решение задачи минимизации из (18.58) при  $A = (\mathbf{Z}' \mathbf{Z})^{-1}$ .

Вычисления, которые ведут к выражениям (18.51) и (18.52), применительно к  $\hat{\beta}_A^{IV}$  показывают:

$$\sqrt{n}(\hat{\beta}_A^{IV} - \beta) \xrightarrow{d} N(0, \Sigma_A^{IV}),$$

где  $\Sigma_A^{IV} = (\mathbf{Q}_{XZ} A \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} A H A \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} A \mathbf{Q}_{ZX})^{-1}$ . (18.60)

Второй способ получить класс IV-оценок, которые используют линейную комбинацию  $Z$ , – это рассмотреть IV-оценки, в которых инструментами являются  $ZB$ , где  $B$  – это матрица  $(m+r+1) \times (k+r+1)$  с полным рангом системы строк. Тогда система из  $(k+r+1)$  уравнений,  $(Y - Xb)'ZB = 0$ , может быть решена единственным способом для  $(k+r+1)$  неизвестных элементов  $b$ . Решая эти уравнения относительно  $b$ , получаем  $\hat{b}^{IV} = (B'Z'X)^{-1}(B'Z'Y)$  и, подставляя  $B = AZX$  в это выражение, получаем выражение (18.59). Таким образом, два подхода определения IV-оценки, которые являются линейными комбинациями инструментов, дают одинаковые семейства IV-оценок. Традиционно используют первый способ, в котором IV-оценка решает задачу минимизации квадратичной целевой функции (18.58), и поэтому этот подход применяется и здесь.

**Асимптотическая эффективность 2МНК при наличии гомоскедастичности ошибок.** Если ошибки гомоскедастичны, то  $H = Q_{ZZ} \sigma_u^2$ , и выражение для  $\Sigma_A^{IV}$  в (18.60) принимает вид:

$$\Sigma_A^{IV} = (\mathbf{Q}_{XZ} A \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} A \mathbf{Q}_{ZZ} A \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} A \mathbf{Q}_{ZX})^{-1} \sigma_u^2. (18.61)$$

Чтобы показать, что 2МНК-оценка асимптотически эффективна в классе оценок, которые являются линейной комбинацией  $Z$  при гомоскедастичных ошибках, нам необходимо показать, что при наличии гомоскедастичности

$$c' \Sigma_A^{IV} c \geq c' \Sigma^{TSLS} c (18.62)$$

для любой положительно полуопределенной матрицы  $A$  и для любого вектора  $c$  размерности  $(k+r+1) \times 1$ , где  $\Sigma^{TSLS} = (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} \sigma_u^2$  [см. (18.55)]. Неравенство (18.62), которое доказывается в приложении 18.6, является тем же самым критерием эффективности, что используется в множественной теореме Гаусса–Маркова из вставки «Основные понятия 18.3». Следовательно, 2МНК является эффективной оценкой в классе оценок, в которых инструменты являются линейной комбинацией  $Z$ .

**J-статистика в случае гомоскедастичных ошибок.** J-статистика (вставка «Основные понятия 12.6») используется для проверки нулевой гипотезы о том, что все сверхидентифицируемые ограничения выполняются против альтернативной гипотезы, что некоторые или все из них не выполняются.

Идея построения J-статистики заключается в том, что если все сверхидентифицируемые ограничения выполняются, то  $u_i$  будут некоррелированы с инструментами и, таким образом, в регрессии  $U$  на  $Z$  все теоретические коэффициенты будут равны нулю. На практике  $U$  не наблюдается, но она может быть оценена при помощи 2МНК-остатков  $\hat{U}$ , так что регрессия  $\hat{U}$  на  $Z$  должна давать статистически незначимые коэффициенты. Соответственно, 2МНК J-статистики – это F-статистика для проверки гипотезы о том, что все коэффициенты при  $Z$  равны нулю, рассчитанная в предположении гомоскедастичности ошибок,

в регрессии  $\hat{U}$  на  $Z$ , и умноженная на  $(m+r+1)$ , так что  $F$ -статистика имеет свое хи-квадрат распределение.

Точную формулу для  $J$ -статистики можно получить, используя выражение (7.13) для  $F$ -статистики при гомоскедастичности ошибок. Модель регрессии без ограничений представляет собой регрессию  $\hat{U}$  на  $(m+r+1)$  регрессоров из  $Z$ , а модель с ограничениями – это модель без регрессоров. Таким образом, в обозначениях выражения (7.13)  $SSR_{unrestricted} = \hat{U}'M_Z\hat{U}$  и  $SSR_{restricted} = \hat{U}'\hat{U}$ , так что  $SSR_{unrestricted} = SSR_{restricted} = \hat{U}'\hat{U} - \hat{U}'M_Z\hat{U} = \hat{U}'P_Z\hat{U}$ , и  $J$ -статистика вычисляется так:

$$J = \frac{\hat{U}'P_Z\hat{U}}{\hat{U}'M_Z\hat{U}/(n-m-r-1)}. \quad (18.63)$$

Метод вычисления  $J$ -статистики, описанный во вставке «Основные понятия 12.6», подходит только для проверки гипотезы о том, что коэффициенты при исключенных инструментах равны нулю. Несмотря на то что эти два метода используют разные вычислительные этапы, они дают идентичные  $J$ -статистики (см. упражнение 18.14).

Как показано в приложении 18.6, что в условиях выполнения нулевой гипотезы, что  $E(u_iZ_i) = 0$ ,

$$J \xrightarrow{d} \chi^2_{m-k}. \quad (18.64)$$

### **Оценка обобщенного метода моментов для линейных моделей**

Если ошибки гетероскедастичны, то 2МНК-оценка не является более эффективной в классе IV-оценок при использовании линейной комбинации  $Z$  в качестве инструментов. Эффективной оценкой в данном случае является оценка обобщенным методом моментов (GMM). В дополнение к этому, если ошибки гетероскедастичны, то  $J$ -статистика, определяемая по формуле (18.63), не имеет больше распределения хи-квадрат. Однако альтернативная формула для  $J$ -статистики, сконструированная с использованием GMM-оценки, имеет распределения хи-квадрат с  $m-k$  степенями свободы.

Эти результаты аналогичны результатам, касающимся оценок обычной модели регрессии с экзогенными регрессорами и гетероскедастичными ошибками: если ошибки гетероскедастичны, то оценка МНК неэффективна среди оценок, которые линейны по  $Y$  (условия Гаусса–Маркова не выполняются), и  $F$ -статистика, рассчитанная для случая гомоскедастичности, не имеет больше  $F$ -распределения даже в больших выборках. В модели регрессии с экзогенными регрессорами и гетероскедастичностью ошибок эффективная оценка – это оценка взвешенного метода наименьших квадратов; в модели регрессии с инструментальными переменными при гетероскедастичности ошибок в эффективной оценке используется иная матрица весов, чем в 2МНК, и конечная оценка – это эффективная GMM-оценка.

**GMM-оценка.** Обобщенный метод моментов (GMM) – это общий метод для оценки параметров линейных или нелинейных моделей, в которых параметры

выбираются так, чтобы наилучшим образом подходить к множеству уравнений, каждое из которых приравнивает выборочные моменты к нулю. Эти уравнения в контексте GMM называются моментными условиями, обычно они не могут быть все удовлетворены одновременно.

В линейной IV-регрессии с экзогенными переменными класс GMM-оценок состоит из всех оценок, которые являются решением квадратичной задачи минимизации (18.58). Таким образом, класс GMM-оценок, основанный на полном наборе инструментов  $Z$  с различными весовыми матрицами  $A$ , является тем же самым, что и класс IV-оценок, в котором инструменты являются линейными комбинациями  $Z$ . В линейной регрессии GMM – это просто другое название для класса оценок, которые мы изучали, то есть оценки, которые являются решением (18.58)

**Асимптотическая эффективность GMM-оценки.** Среди класса GMM-оценок эффективная GMM-оценка – это такая GMM-оценка, которая имеет наименьшую асимптотическую матрицу дисперсий [где наименьшая матрица дисперсий определена в выражении (18.62)]. Таким образом, результат, приведенный в (18.62), может быть пересмотрен как говорящий о том, что 2МНК является эффективной GMM-оценкой в модели линейной регрессии при гомоскедастичности ошибок.

Чтобы вывести выражение для эффективной GMM-оценки в случае гетероскедастичности ошибок, вспомним, что когда ошибки гомоскедастичны,  $H$  [дисперсионная матрица  $Z_i u_i$ ; см. выражение (18.50)] равна  $Q_{zz} \sigma_u^2$ , и асимптотически эффективная матрица весов получается, если положить  $A = (Z'Z)^{-1}$ , что включает в себя 2МНК-оценку. В больших выборках использование матрицы весов  $A = (Z'Z)^{-1}$  эквивалентно использованию  $A = (Q_{zz} \sigma_u^2)^{-1} = H^{-1}$ . Такая интерпретация 2МНК-оценки предполагает, по аналогии, что эффективная IV-оценка при гетероскедастичных ошибках может быть получена, положив  $A = H^{-1}$  и решив

$$\min_b (Y - Xb)' Z H^{-1} Z' (Y - Xb). \quad (18.65)$$

Эта аналогия верна: решение задачи минимизации в (18.65) является эффективной GMM-оценкой. Пусть  $\tilde{\beta}^{Eff.GMM}$  обозначает решение задачи минимизации (18.65). По формуле (18.59) эта оценка равна:

$$\tilde{\beta}^{Eff.GMM} = (X' Z H^{-1} Z' X)^{-1} X' Z H^{-1} Z' Y. \quad (18.66)$$

Асимптотическое распределение  $\tilde{\beta}^{Eff.GMM}$  получается подстановкой  $A = H^{-1}$  в (18.60) и упрощением полученного выражения:

$$\sqrt{n} (\tilde{\beta}^{Eff.GMM} - \beta) \xrightarrow{d} N(0, \Sigma^{Eff.GMM}),$$

$$\text{где } \Sigma^{Eff.GMM} = (Q_{xz} H^{-1} Q_{zx})^{-1}. \quad (18.67)$$

Результат, говорящий о том, что  $\tilde{\beta}^{Eff.GMM}$  является эффективной GMM-оценкой, может быть показан путем доказательства того, что  $c' \Sigma_A^{IV} c \geq c' \Sigma^{Eff.GMM} c$  для любого вектора  $c$ , где  $\Sigma_A^{IV}$  дана в выражении (18.60). Доказательство этого утверждения приводится в приложении 18.6.

**Доступная эффективная GMM-оценка.** GMM-оценка, определенная в (18.66), является недоступной оценкой, так как она зависит от неизвестной дисперсионной матрицы  $\mathbf{H}$ . Однако доступная эффективная оценка может быть вычислена подстановкой состоятельной оценки  $\mathbf{H}$  в задачу минимизации (18.65) или, эквивалентно, подстановкой состоятельной оценки матрицы  $\mathbf{H}$  в формулу для  $\tilde{\beta}^{Eff.GMM}$  в выражении (18.66).

Эффективная GMM-оценка может быть вычислена в два шага. На первом шаге оценивается  $\beta$ , с использованием любой состоятельной оценки. Используйте эту оценку  $\beta$ , чтобы вычислить остатки в интересующем вас уравнении, а затем используйте эти остатки, чтобы вычислить оценку матрицы  $\mathbf{H}$ . На втором шаге используется полученная оценка  $\mathbf{H}$ , чтобы вычислить оптимальную матрицу весов  $\mathbf{H}^{-1}$  и эффективную GMM-оценку. Более точно, в модели линейной IV-регрессии обычно используется 2МНК-оценка на первом шаге и находятся 2МНК-остатки для оценки  $\mathbf{H}$ . Если 2МНК-оценка используется на первом шаге, то эффективная доступная GMM-оценка, вычисляемая на втором шаге, равна:

$$\tilde{\beta}^{Eff.GMM} = \left( \mathbf{X}' \mathbf{Z} \hat{\mathbf{H}}^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{Z} \hat{\mathbf{H}}^{-1} \mathbf{Z}' \mathbf{Y}, \quad (18.68)$$

где  $\hat{\mathbf{H}}$  приведена в выражении (18.54).

$$\begin{aligned} \text{Так как } \hat{\mathbf{H}} &\xrightarrow{p} \mathbf{H}, \sqrt{n} \left( \hat{\beta}^{Eff.GMM} - \tilde{\beta}^{Eff.GMM} \right) \xrightarrow{p} 0 \text{ (упражнение 18.12)} \\ \text{и } \sqrt{n} \left( \hat{\beta}^{Eff.GMM} - \beta \right) &\xrightarrow{d} N(0, \Sigma^{Eff.GMM}), \end{aligned} \quad (18.69)$$

где  $\Sigma^{Eff.GMM} = (\mathbf{Q}_{XZ} \mathbf{H}^{-1} \mathbf{Q}_{ZX})^{-1}$  [см. (18.67)]. То есть доступная двухшаговая оценка  $\tilde{\beta}^{Eff.GMM}$  в (18.68) является асимптотически эффективной GMM-оценкой.

**Устойчивая при гетероскедастичности J-статистика.** Устойчивая при гетероскедастичности ошибок J-статистика, известная как GMM J-статистика, в отличие от 2МНК J-статистики, вычисляется на основе эффективной GMM-оценки и функции весов. То есть J-статистика вычисляется так:

$$J^{GMM} = \left( \mathbf{Z}' \hat{\mathbf{U}}^{GMM} \right)' \hat{\mathbf{H}}^{-1} \left( \mathbf{Z}' \hat{\mathbf{U}}^{GMM} \right) / n, \quad (18.70)$$

где  $\hat{\mathbf{U}}^{GMM} = \mathbf{Y} - \mathbf{X} \hat{\beta}^{Eff.GMM}$  – остатки интересующего нас уравнения, оцененного (доступным) эффективным GMM, и  $\hat{\mathbf{H}}^{-1}$  – матрица весов, используемая для вычисления  $\hat{\beta}^{Eff.GMM}$ .

При выполнении нулевой гипотезы  $E(\mathbf{Z}_i u_i) = \mathbf{0}$  –  $J^{GMM} \xrightarrow{d} \chi^2_{m-k}$  (см. приложение 18.6).

**GMM для временных рядов.** Результаты, полученные в этом разделе, имеют место при предпосылках IV-регрессии для межъобъектных данных. Тем не менее во многих прикладных задачах результаты для IV-регрессии и GMM расширяются на случай временных рядов. Несмотря на то что формальный математический аппарат GMM применительно к временным рядам лежит за рамками данной книги (для этого см. Hayashi, 2000, гл. 6), мы все же приведем клю-

чевые идеи использования GMM для временных рядов. Это короткое описание напомнит материал из глав 14 и 15. Предположим, что все переменные стационарны.

Полезно различать два типа прикладных задач: задачу, в которой ошибки  $u_t$  серийно коррелированы, и задачу, в которой ошибки  $u_t$  не коррелированы. Если ошибки  $u_t$  серийно коррелированы, то асимптотическое распределение GMM-оценки продолжает быть нормальным, но формула для  $\mathbf{H}$  (18.50) не является корректной. Вместо этого верное выражение для  $\mathbf{H}$  зависит от автоковариаций  $Z_t u_t$  и аналогично формуле (15.14) для дисперсии МНК-оценки для случая серийно коррелированных ошибок. Эффективная GMM-оценка также вычисляется с помощью состоятельной оценки  $\mathbf{H}$ , тем не менее состоятельная оценка должна быть получена с использованием метода НАС, обсужденного в главе 15.

Если ошибки  $u_t$  серийно не коррелированы, тогда нет необходимости в НАС-оценке  $\mathbf{H}$ , и все формулы, представленные в этом разделе, применимы к GMM и для случая временных рядов. В современных прикладных исследованиях по финансам и макроэкономике обычно используются модели, в которых ошибки представлены неожиданными или непредсказуемыми шоками, в этом случае из модели вытекает, что  $u_t$  серийно не коррелированы. Например, рассмотрим модель с единственной включенной эндогенной переменной и без включенных экзогенных переменных, так что интересующее нас уравнение выглядит как  $Y_t = \beta_0 + \beta_1 X_t + u_t$ . Предположим, что из экономической теории следует, что  $u_t$  является непредсказуемой при помощи имеющейся информации о прошлом. Тогда из теории следует такое моментное условие:

$$E(u_t | Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \dots) = 0, \quad (18.71)$$

где  $Z_{t-1}$  – лаговое значение какой-то другой переменной. Моментное условие (18.71) означает, что все запаздывающие переменные  $Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \dots$  являются кандидатами для того, чтобы быть состоятельными инструментами (они удовлетворяют требованиям экзогенности). Более того, так как  $u_{t-1} = Y_{t-1} - \beta_0 - \beta_1 X_{t-1}$ , моментное условие (18.71) эквивалентно условию  $E(u_t | u_{t-1}, X_{t-1}, Z_{t-1}, u_{t-2}, X_{t-2}, Z_{t-2}, \dots) = 0$ . Так как  $u_t$  серийно не коррелирована, то нет необходимости в вычислении НАС-оценки  $\mathbf{H}$ . Теория GMM, представленная в данном разделе, включая эффективную GMM-оценку и GMM J-статистику, применяется напрямую к временным рядам с моментными условиями в форме, представленной в (18.71), в предположении, что моментные условия (18.71) в действительности верны.

## Выводы

- Линейная модель множественной регрессии в матричной форме имеет вид:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$ , где  $\mathbf{Y}$  – это вектор  $n \times 1$  наблюдений зависимой переменной,  $\mathbf{X}$  – это матрица размерности  $n \times (k + 1)$ , состоящая из  $n$  наблюдений  $(k + 1)$  регрессора (включая константу),  $\boldsymbol{\beta}$  – это вектор  $(k + 1)$  неизвестных параметров и  $\mathbf{U}$  – вектор ошибок размерности  $n \times 1$ .

2. МНК-оценка равна:  $\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$ . При выполнении первых четырех предположений МНК, изложенных во вставке «Основные понятия 18.1»,  $\hat{\beta}$  состоятельна и асимптотически нормально распределена. Если, в дополнение к этому, ошибки гомоскедастичны, то условная дисперсия  $\hat{\beta}$  равна:  $\text{var}(\hat{\beta} | \mathbf{X}) = \sigma_u^2 (\mathbf{X}' \mathbf{X})^{-1}$ .
3. Линейные ограничения на  $\beta$  могут быть записаны как  $q$  уравнений  $\mathbf{R}\beta = \mathbf{r}$ , и такая формулировка может быть использована для проверки совместных гипотез относительно нескольких коэффициентов или для построения доверительных интервалов для элементов вектора  $\beta$ .
4. Если ошибки регрессии нормально одинаково независимо распределены условно по  $\mathbf{X}$ , то  $\beta$  имеет точное нормальное распределение, а  $t$ -статистики и  $F$ -статистики при наличии гомоскедастичности ошибок имеют, соответственно, точные  $t_{n-k-1}$  и  $F_{q, n-k-1}$ -распределения.
5. Теорема Гаусса–Маркова показывает, что если ошибки гомоскедастичны и условно не коррелированы по наблюдениям и если  $E(u_i | \mathbf{X}) = 0$ , МНК-оценка является лучшей (т.е. самой эффективной) среди линейных условно несмещенных оценок (МНК является BLUE-оценкой).
6. Если ковариационная матрица ошибок  $\Omega$  не пропорциональна единичной матрице и если  $\Omega$  известна или может быть оценена, тогда ОМНК-оценка асимптотически более эффективна, чем МНК. Тем не менее для ОМНК требуется, чтобы  $u_i$  были некоррелированы со всеми наблюдениями регрессоров, а не только с  $\mathbf{X}$ , как это требуется в МНК, что является требованием, которое должно быть аккуратно проверено на практике.
7. 2МНК-оценка входит в класс GMM-оценок линейных моделей. В GMM коэффициенты оцениваются путем вычисления выборочной ковариации между ошибкой регрессии и экзогенными переменными настолько малой, насколько это возможно – или, более точно, решая проблему  $\min_b [(Y - Xb)' Z] A [Z'(Y - Xb)]$ , где  $A$  является матрицей весов. Асимптотическая эффективность GMM-оценки достигается при  $A = [E(Z_i Z'_i u_i^2)]^{-1}$ .

Если ошибки гомоскедастичны, то асимптотически эффективная GMM-оценка в линейной IV-регрессии – это 2МНК-оценка.

## Основные понятия

Условия Гаусса–Маркова для множественной регрессии (с. 743).

Теорема Гаусса–Маркова для множественной регрессии (с. 744).

Обобщенный метод наименьших квадратов (ОМНК, GLS) (с. 746).

Недоступный ОМНК (с. 749).

Доступный ОМНК (с. 749).

Обобщенный метод моментов (GMM) (с. 756).

Эффективный GMM (с. 757).

Устойчивая при гетероскедастичности ошибок  $J$ -статистика (с. 758).

GMM  $J$ -статистика (с. 758).

Вектор средних (с. 767).

Ковариационная матрица (с. 770).

### **Вопросы для повторения и закрепления основных понятий**

- 18.1. Исследователь изучает соотношение между доходами и полом индивида для группы рабочих, специфицировав регрессионную модель в виде  $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + u_i$ , где  $X_{1i}$  – фиктивная переменная, которая равна 1, если  $i$ -й индивид женщина, а  $X_{2i}$  – фиктивная переменная, которая равна 1, если  $i$ -й индивид мужчина. Запишите модель в матричной форме (18.2) для гипотетической выборки из  $n=5$  наблюдений. Покажите, что столбцы  $X$  линейно зависимы, так что  $X$  не имеет полного ранга. Объясните, как вы измените модель, чтобы исключить совершенную мультиколлинеарность?
- 18.2. Вы анализируете линейную регрессию с 500 наблюдениями и одним регрессором. Объясните, как вы будете строить доверительные интервалы для  $\beta$ , если:
  - a) Предположения 1–4 из вставки «Основные понятия 18.1» выполняются, но вы думаете, что предположения 5 и 6 могут не выполняться.
  - б) Предположения 1–5 из вставки «Основные понятия 18.1» выполняются, но вы думаете, что предположение 6 может не выполняться (постройте доверительный интервал двумя способами).
  - в) Предположения 1–6 выполняются.
- 18.3. Пусть предположения 1–5 из вставки «Основные понятия 18.1» выполняются, но предположение 6 – нет. Выполняется ли результат, представленный в (18.31)? Поясните ваш вывод.
- 18.4. Можете ли вы вычислить BLUE-оценку  $\beta$ , если предположение (18.41) выполняется и вы не знаете  $\Omega$ ? Что будет, если вы знаете  $\Omega$ ?
- 18.5. Постройте пример регрессии, которая удовлетворяет предпосылке  $E(u_i|X_i)=0$ , но для которой  $E(U|X)\neq 0_n$ .

### **Упражнения**

- 18.1. Рассмотрите теоретическую регрессию зависимости результатов тестов от доходов и квадратов доходов, описанную в уравнении (8.1).
  - а) Запишите регрессию из уравнения (8.1) в матричной форме (18.5). Определите  $Y, X, U$  и  $\beta$ .
  - б) Объясните, как проверить нулевую гипотезу о том, что зависимость между результатами тестов и доходом линейна против гипотезы, что она квадратична. Выпишите нулевую гипотезу аналогично (18.20). Что такое  $R, r$  и  $q$ ?
- 18.2. Предположим, что выборка из  $n=20$  домохозяйств имеет выборочные средние и выборочные ковариации для зависимой переменной и двух регрессоров, выписанные ниже:

	Выборочные средние	Выборочная ковариация		
		$Y$	$X_1$	$X_2$
$Y$	6,39	0,26	0,22	0,32
$X_1$	7,24		0,80	0,28
$X_2$	4,00			2,40

- a) Вычислите МНК-оценки  $\beta_0$ ,  $\beta_1$  и  $\beta_2$ . Вычислите  $s_u^2$ . Вычислите  $R^2$  регрессии.
- б) Предположим, что все шесть предпосылок из вставки «Основные понятия 18.1» выполняются. Проверьте гипотезу о том, что  $\beta_1 = 0$  на 5 %-м уровне значимости.
- 18.3. Пусть  $W - m \times 1$  вектор с ковариационной матрицей  $\Sigma_W$ , где  $\Sigma_W$  – конечная и положительно определенная матрица. Пусть  $c$  – неслучайный вектор размерности  $m \times 1$  и пусть  $Q = c' W$ .
- а) Покажите, что  $\text{var}(Q) = c' \Sigma_W c$ .
- б) Предположим  $c \neq \mathbf{0}_m$ . Покажите, что  $0 < \text{var}(Q) < \infty$ .
- 18.4. Рассмотрим модель регрессии из главы 4,  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , и предположим, что предположения из вставки «Основные понятия 4.3» выполняются.
- а) Запишите модель в матричной форме, приведенной в уравнениях (18.2) и (18.4).
- б) Покажите, что предположения 1–4 из вставки «Основные понятия 18.1» выполнены.
- в) Используя общую формулу для  $\hat{\beta}$  из уравнения (18.11), получите выражения для  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , данные во вставке «Основные понятия 4.2».
- г) Покажите, что элемент  $(1, 1)$  матрицы  $\Sigma_{\hat{\beta}}$  в уравнении (18.13) равен выражению для  $\sigma_{\hat{\beta}_1}^2$ , приведенному во вставке «Основные понятия 4.4».
- 18.5. Пусть  $P_X$  и  $M_X$  будут определены как в выражениях (18.24) и (18.25).
- а) Докажите, что  $P_X M_X = \mathbf{0}_{n \times n}$  и что  $P_X$  и  $M_X$  идемпотентны.
- б) Выведите выражения (18.27) и (18.28).
- 18.6. Рассмотрим регрессию в матричной форме,  $Y = X\beta + W + U$ , где  $X$  – матрица  $n \times k_1$  регрессоров и  $W$  – матрица  $n \times k_2$  регрессоров. Тогда, как показано в упражнении 18.17, МНК-оценка  $\hat{\beta}$  может быть выражена так:

$$\hat{\beta} = (X' M_W X)^{-1} (X' M_W Y).$$

Теперь пусть  $\hat{\beta}_1^{BV}$  будет оценкой фиксированных эффектов с использованием «бинарных переменных», вычисленной оцениванием уравнения (10.11) при помощи МНК, и пусть  $\hat{\beta}_1^{DM}$  будет «центрированной» оценкой с фиксированными эффектами, вычисленной путем оценивания выражения (10.14) при помощи МНК, в котором  $X$  и  $Y$  были отцентрированы при помощи среднего для объекта значения. Используйте выражение для  $\hat{\beta}$ , данное выше, чтобы доказать, что  $\hat{\beta}_1^{BV} = \hat{\beta}_1^{DM}$ . [Подсказка: выпишите вы-

ражение (10.11), используя полный набор фиксированных эффектов  $D1_i, D2_i, \dots, Dn_i$  и не включая константу. Включите все фиксированные эффекты в  $\mathbf{W}$ . Выпишите матрицу  $\mathbf{M}_w X$ .]

- 18.7. Рассмотрим регрессию  $Y_i = \beta_1 X_i + \beta_2 W_i + u_i$ , где для простоты константа пропущена и предполагается, что все переменные имеют нулевые средние. Предположим, что  $X_i$  распределена независимо от  $(W_i, u_i)$ , но  $W_i$  и  $u_i$  могут быть коррелированы, и пусть  $\hat{\beta}_1$  и  $\hat{\beta}_2$  будут МНК-оценками для этой модели. Покажите, что:

- Независимо от того, коррелированы ли  $W_i$  и  $u_i$  или нет,  $\hat{\beta}_1 \xrightarrow{p} \beta_1$ .
- Если  $W_i$  и  $u_i$  коррелированы, тогда  $\hat{\beta}_2$  несостоительна.
- Пусть  $\hat{\beta}_1$  будет МНК-оценкой из регрессии  $Y$  на  $X$  (регрессия с ограничениями, из которой исключена переменная  $W$ ). Выведите условия, при которых  $\hat{\beta}_1$  имеет асимптотическую дисперсию меньшую, чем у  $\hat{\beta}_1$ , допуская возможность, что  $W_i$  и  $u_i$  коррелированы.

- 18.8. Рассмотрите модель регрессии  $Y_i = \beta_0 + \beta_1 X_i + u_i$ , где  $u_1 = \tilde{u}_1$  и  $u_i = 0,5u_{i-1} + \tilde{u}_i$  для  $i = 2, 3, \dots, n$ . Предположим, что  $u_i$  независимо одинаково распределены с нулевым средним и единичной дисперсией и распределены независимо от  $X_j$  для всех  $i$  и  $j$ .

- Выполните выражение для  $E(\mathbf{U}\mathbf{U}') = \Omega$ .
- Объясните, как оценить модель методом ОМНК без точного вычисления обратной матрицы  $\Omega$ . (Подсказка: преобразуйте модель таким образом, чтобы ошибки стали  $\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_n$ .)

- 18.9. Это упражнение показывает, что МНК-оценка некоторого подмножества коэффициентов регрессии является состоятельной в предположении независимости условного среднего, введенного в приложении 7.2. Рассмотрите модель множественной регрессии в матричной форме  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \mathbf{u}$ , где  $\mathbf{X}$  и  $\mathbf{W}$  являются, соответственно,  $n \times k_1$  и  $n \times k_2$ -матрицами регрессоров. Пусть  $\mathbf{X}'_i$  и  $\mathbf{W}'_i$  будут  $i$ -ми строками матриц  $\mathbf{X}$  и  $\mathbf{W}$  [как в выражении (18.3)]. Предположим, что (i)  $E(u_i | \mathbf{X}_i, \mathbf{W}_i) = \mathbf{W}'_i \boldsymbol{\delta}$ , где  $\boldsymbol{\delta}$  является  $k_2 \times 1$  вектором неизвестных параметров; (ii)  $(\mathbf{X}_i, \mathbf{W}_i, Y_i)$  независимо одинаково распределены, (iii)  $(\mathbf{X}_i, \mathbf{W}_i, u_i)$  имеют четыре конечных ненулевых момента и (iv) нет совершенной мультиколлинеарности. Все это предпосылки № 1–4 из вставки «Основные понятия 18.1» с предположением о независимости условного среднего (i), заменяющего обычное предположение о нулевом условном среднем.

- Используйте выражение для  $\hat{\boldsymbol{\beta}}$ , полученное в упражнении (18.6), чтобы записать:  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (n^{-1}\mathbf{X}'\mathbf{M}_w\mathbf{X})^{-1}(n^{-1}\mathbf{X}'\mathbf{M}_w\mathbf{U})$ .
- Покажите, что  $n^{-1}\mathbf{X}'\mathbf{M}_w\mathbf{X} \xrightarrow{p} \Sigma_{XX} - \Sigma_{XW} \Sigma_{WW}^{-1} \Sigma_{WX}$ , где  $\Sigma_{XX} = E(\mathbf{X}_i \mathbf{X}'_i)$ ,  $\Sigma_{XW} = (\mathbf{X}_i \mathbf{W}'_i)$ , и так далее. [Матрица  $A_n \xrightarrow{p} A$ , если  $A_{n,ij} \xrightarrow{p} A_{ij}$  для всех  $i, j$ , где  $A_{n,ij}$  и  $A_{ij}$  – элементы  $(i, j)$  матриц  $A_n$  и  $A$ .]
- Покажите, что из предположений (i) и (ii) следует, что  $E(\mathbf{U} | \mathbf{X}, \mathbf{W}_i) = \mathbf{W}\boldsymbol{\delta}$ .
- Используйте пункт (в) и закон повторного математического ожидания, докажите, что  $n^{-1}\mathbf{X}'\mathbf{M}_w\mathbf{U} \xrightarrow{p} \mathbf{0}_{k_1 \times 1}$ .
- Используйте результаты пунктов (a) – (c), чтобы вывести, что в предположениях (i) – (iv)  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ .

18.10. Пусть  $C$  является симметричной идемпотентной матрицей.

a) Покажите, что собственные значения матрицы  $C$  равны или 1, или 0. (Подсказка: заметьте, что из  $Cq = \gamma q$  следует, что  $0 = Cq - \gamma q = CCq - \gamma q = \gamma Cq - \gamma q = \gamma^2 q - \gamma q$ , и решите уравнение относительно  $\gamma$ .)

б) Покажите, что  $\text{trace}(C) = \text{rank}(C)$ .

в) Пусть  $d$  будет вектором  $n \times 1$ . Покажите, что  $d' C q \geq 0$ .

18.11. Предположим, что  $C$  – симметричная и идемпотентная матрица  $n \times n$  с рангом  $r$ , и пусть  $V \sim N(\mathbf{0}_n, I_n)$ .

a) Покажите, что  $C = AA'$ , где  $A$  это  $n \times r$  и  $A'A = I_r$ . (Подсказка:  $C$  – положительно полуопределенная матрица и может быть записана в форме  $Q\Lambda Q'$ , как объяснено в приложении 18.1.)

б) Покажите, что  $A'V \sim N(\mathbf{0}_r, I_r)$ .

в) Покажите, что  $V'CV \sim \chi_r^2$ .

18.12. а) Покажите, что  $\tilde{\beta}^{Eff.GMM}$  является эффективной GMM-оценкой, то есть  $\tilde{\beta}^{Eff.GMM}$  в (18.66) является решением уравнения (18.65).

б) Покажите, что  $\sqrt{n}(\hat{\beta}^{Eff.GMM} - \tilde{\beta}^{Eff.GMM}) \xrightarrow{p} 0$ .

в) Покажите, что  $J^{GMM} \xrightarrow{d} \chi_{m-k}^2$ .

18.13. Рассмотрите проблему минимизации суммы квадратов остатков при ограничении, что  $Rb = r$ , где  $R$  это матрица  $q \times (k+1)$  с рангом  $q$ . Пусть  $\tilde{\beta}$  обозначает значение  $b$ , которое является решением задачи минимизации при ограничениях.

а) Покажите, что Лангранжиан для задачи минимизации – это  $L(b, \gamma) = -(Y - Xb)'(Y - Xb) + \gamma'(Rb - r)$ , где  $\gamma$  – это вектор  $q \times 1$  множителей Лагранжа.

б) Покажите, что  $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$ .

в) Покажите, что  $(Y - X\tilde{\beta})'(Y - X\tilde{\beta}) - (Y - X\hat{\beta})'(Y - X\hat{\beta}) = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$ .

г) Покажите, что  $\tilde{F}$  из (18.36) является эквивалентом  $F$ -статистики, рассчитанной при наличии гомоскедастичности ошибок, из (7.13).

18.14. Рассмотрите модель регрессии  $Y = X\beta + U$ . Разобьем  $X$  на  $[X_1 \quad X_2]$  и  $\beta$  на  $[\beta'_1 \quad \beta'_2]$ , где  $X_1$  содержит  $k_1$  столбцов и  $X_2$  содержит  $k_2$  столбцов. Предположим, что  $X_2'Y = \mathbf{0}_{k_2 \times 1}$ . Пусть  $R = [I_{k_1} \quad \mathbf{0}_{k_1 \times k_2}]$ .

а) Покажите, что  $\hat{\beta}'(X'X)\hat{\beta} = (R\hat{\beta})' [R(X'X)^{-1}R']^{-1}(R\hat{\beta})$ .

б) Рассмотрите регрессию, описанную в (12.17). Пусть  $W = [\mathbf{1} \quad W_1 \quad W_2 \dots W_r]$ , где  $\mathbf{1}$  – это  $n \times 1$  вектор, состоящий из единиц,  $W_i$  –  $n \times 1$  вектор с  $i$ -м элементом  $W_{1i}$  и так далее. Пусть  $\hat{U}^{TSL}$  – вектор остатков регрессии, оцененной двухшаговым методом наименьших квадратов.

(i) Покажите, что  $W'\hat{U}^{TSL} = 0$ .

(ii) Покажите, что метод для вычисления  $J$ -статистики, описанный во вставке «Основные понятия 12.6» (с использованием  $F$ -статистики, рассчитанной в предположении гомоскедастичности) и формула (18.63)

дают одинаковые значения  $J$ -статистики. [Подсказка: используйте результаты, полученные в пунктах (a), (b, i) и упражнении 18.13.]

- 18.15. (Состоятельность кластеризованных стандартных ошибок.) Рассмотрите модель регрессии панельных данных  $Y_{it} = \beta X_{it} + \alpha_i + u_{it}$ , где все переменные являются скалярами. Предположим, что требования № 1, № 2 и № 4 из вставки «Основные понятия 10.3» выполняются и усилим требование № 3, так что  $X_{it}$  и  $u_{it}$  имеют по восемь ненулевых конечных моментов. Пусть  $\mathbf{M} = \mathbf{I}_T - T^{-1}\mathbf{1}'$ , где  $\mathbf{1} = T \times 1$  вектор, состоящий из единиц. Также пусть  $\mathbf{Y}_i = (Y_{i1} \ Y_{i2} \dots \ Y_{iT})'$ ,  $\mathbf{X}_i = (X_{i1} \ X_{i2} \dots \ X_{iT})'$ ,  $\mathbf{u}_i = (u_{i1} \ u_{i2} \dots \ u_{iT})'$ ,  $\tilde{\mathbf{Y}}_i = \mathbf{M}\mathbf{Y}_i$ ,  $\tilde{\mathbf{X}}_i = \mathbf{M}\mathbf{X}_i$ , и  $\tilde{\mathbf{u}}_i = \mathbf{M}\mathbf{u}_i$ . Для асимптотических вычислений в данной задаче предположим, что  $T$  фиксировано, а  $n \rightarrow \infty$ .

a) Покажите, что оценка фиксированных эффектов коэффициентов  $\beta$  из

раздела 10.3 может быть записана как  $\hat{\beta} = \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i \right)^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}_i' \tilde{\mathbf{Y}}_i$

б) Покажите, что  $\hat{\beta} - \beta = \left( \sum_{i=1}^n \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i \right)^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}_i' \mathbf{u}_i$  (Подсказка:  $\mathbf{M}$  – идемпотентная матрица.)

в) Пусть  $\mathcal{Q}_{\tilde{\mathbf{X}}} = T^{-1}E(\tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i)$  и  $\hat{\mathcal{Q}}_{\tilde{\mathbf{X}}} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{X}_{it}^2$ . Покажите, что  $\hat{\mathcal{Q}}_{\tilde{\mathbf{X}}} \xrightarrow{p} \mathcal{Q}_{\tilde{\mathbf{X}}}$ .

г) Пусть  $\eta_i = \tilde{\mathbf{X}}_i' \mathbf{u}_i / \sqrt{T}$  и  $\sigma_\eta^2 = \text{var}(\eta_i)$ . Покажите, что  $\sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i^2} \xrightarrow{d} N(0, \sigma_\eta^2)$ .

д) Используя результаты пунктов (б) – (г), докажите (10.25), то есть покажите, что  $\sqrt{nT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma_\eta^2 / \mathcal{Q}_{\tilde{\mathbf{X}}}^2)$ .

е) Пусть  $\tilde{\sigma}_{\eta, \text{clustered}}^2$  – недоступная оценка кластеризованной дисперсии, вычисленная с использованием реальных ошибок вместо остатков, так что  $\tilde{\sigma}_{\eta, \text{clustered}}^2 = \frac{1}{nT} \sum_{i=1}^n (\tilde{\mathbf{X}}_i' \mathbf{u}_i)^2$ . Покажите, что  $\tilde{\sigma}_{\eta, \text{clustered}}^2 \xrightarrow{p} \sigma_\eta^2$ .

ж) Пусть  $\hat{\mathbf{u}}_i = \mathbf{Y}_i - \hat{\beta} \tilde{\mathbf{X}}_i$  и  $\hat{\sigma}_{\eta, \text{clustered}}^2 = \frac{n}{n-1} \frac{1}{nT} \sum_{i=1}^n (\tilde{\mathbf{X}}_i' \hat{\mathbf{u}}_i)^2$  [это есть формула (10.27) в матричной форме]. Покажите, что  $\hat{\sigma}_{\eta, \text{clustered}}^2 \xrightarrow{p} \sigma_\eta^2$  [Подсказка: используйте логику, аналогичную той, при помощи которой выводится (17.16), чтобы показать, что  $\sigma_{\eta, \text{clustered}}^2 \xrightarrow{p} \tilde{\sigma}_{\eta, \text{clustered}}^2 \xrightarrow{p} 0$ , затем используйте ваш ответ из пункта (e).]

- 18.16. В приведенном упражнении поднимается вопрос о пропущенных значениях в данных, обсуждаемый в разделе 9.2. Рассмотрите модель регрессии  $Y_i = X_i \beta + u_i$ ,  $i = 1, \dots, n$ , где все переменные скалярные и константа пропущена для удобства.

а) Предположим, что предпосылки МНК из вставки «Основные понятия 4.3» выполнены. Покажите, что оценка МНК коэффициента  $\beta$  не смешена и состоятельна.

б) Теперь предположим, что несколько наблюдений пропущены. Пусть  $I_i$  является бинарной случайной переменной, которая указывает

на непропущенные наблюдения, то есть  $I_i = 1$ , если наблюдение  $i$  не пропущено, и  $I_i = 0$ , если наблюдение  $i$  пропущено. Предположим, что  $\{I_i, X_i, u_i\}$  независимо одинаково распределены.

(i) Покажите, что МНК-оценка может быть записана так:

$$\hat{\beta} = \left( \sum_{i=1}^n I_i X_i X_i' \right)^{-1} \left( \sum_{i=1}^n I_i X_i Y_i \right) = \beta + \left( \sum_{i=1}^n I_i X_i X_i' \right)^{-1} \left( \sum_{i=1}^n I_i X_i u_i \right).$$

(ii) Предположим, что наблюдения пропущены совершенно случайно в том смысле, что  $\Pr(I_i = 1 | X_i, u_i) = p$ , где  $p$  является константой. Покажите, что  $\hat{\beta}$  – несмешенная и состоятельная оценка.

(iii) Предположим, что вероятность того, что  $i$ -ое наблюдение пропущено, зависит от  $X_i$ , но не от  $u_i$ , то есть  $\Pr(I_i = 1 | X_i, u_i) = p(X_i)$ . Покажите, что  $\hat{\beta}$  – несмешенная и состоятельная оценка.

(iv) Предположим, что вероятность того, что  $i$ -ое наблюдение пропущено, зависит и от  $X_i$ , и от  $u_i$ , то есть  $\Pr(I_i = 1 | X_i, u_i) = p(X_i, u_i)$ . Является ли  $\hat{\beta}$  несмешенной? Состоятельной? Объясните.

в) Предположим, что  $\beta = 1$  и что  $X_i$  и  $u_i$  – взаимно независимые стандартные нормальные случайные величины [так что  $X_i$  и  $u_i$  распределены как  $N(0, 1)$ ]. Предположим, что  $I_i = 1$  при  $Y_i \geq 0$  и  $I_i = 0$  при  $Y_i < 0$ . Является ли  $\hat{\beta}$  несмешенной? Является ли  $\hat{\beta}$  состоятельной? Объясните.

18.17. Рассмотрите модель регрессии в матричной форме  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \mathbf{U}$ , где  $\mathbf{X}$  и  $\mathbf{W}$  – матрицы регрессоров и  $\boldsymbol{\beta}$  и  $\boldsymbol{\gamma}$  – векторы неизвестных коэффициентов регрессии. Пусть  $\tilde{\mathbf{X}} = \mathbf{M}_w \mathbf{X}$  и  $\tilde{\mathbf{Y}} = \mathbf{M}_w \mathbf{Y}$ , где  $\mathbf{M}_w = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}$ .

а) Покажите, что МНК-оценка коэффициентов  $\boldsymbol{\beta}$  и  $\boldsymbol{\gamma}$  может быть записана так:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{W}'\mathbf{Y} \end{bmatrix}.$$

б) Покажите, что

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{X}'\mathbf{M}_w \mathbf{X})^{-1} \\ -(\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{X} (\mathbf{X}'\mathbf{M}_w \mathbf{X})^{-1} \\ -(\mathbf{X}'\mathbf{M}_w \mathbf{X})^{-1} \mathbf{X}'\mathbf{W} (\mathbf{W}'\mathbf{W})^{-1} \\ (\mathbf{W}'\mathbf{W})^{-1} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{X} (\mathbf{X}'\mathbf{M}_w \mathbf{X})^{-1} \mathbf{X}'\mathbf{W} (\mathbf{W}'\mathbf{W})^{-1} \end{bmatrix}.$$

(Подсказка: покажите, что произведение двух матриц равно единичной матрице.)

в) Покажите, что  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_w \mathbf{X})^{-1} \mathbf{X}'\mathbf{M}_w \mathbf{Y}$ .

г) Согласно теореме Фриша–Во (приложение 6.2),  $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$ . Используя результат из пункта (в), докажите теорему Фриша–Во.

## Приложения

### Приложение 18.1. Основы линейной алгебры

В данном приложении рассказывается о векторах, матрицах и элементах матричной алгебры, используемых в главе 18. Основная цель – привести основные идеи и определения из курса линейной алгебры, но не заменить этот курс.

#### *Определения вектора и матрицы*

*Вектором* называется набор из  $n$  чисел или элементов, собранных или в столбец (*вектор-столбец*), или в строку (*вектор-строка*). Вектор-столбец  $b$  размерности  $n$  и вектор-строка  $c$  размерности  $n$ :

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \text{ и } c = [c_1 \ c_2 \ \cdots \ c_n]$$

где  $b_1$  – это первый элемент  $b$ , и в целом  $b_i$  является  $i$ -м элементом  $b$ .

Повсюду символы, выделенные жирным шрифтом, означают, что это матрица или вектор.

*Матрицей* называется набор из чисел или элементов, в котором элементы находятся в столбцах и строках. Размерность матрицы равна  $n \times m$ , где  $n$  обозначает количество строк, а  $m$  – количество столбцов. Матрица  $A$  размерности  $n \times m$ :

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

где  $a_{ij}$  является  $(i, j)$  элементом  $A$ , то есть  $a_{ij}$  – это элемент, который находится в  $i$ -й строке и  $j$ -м столбце.

Чтобы отличать одномерные числа от векторов и матриц, одномерное число называется *скаляром*.

#### *Типы матриц*

**Диагональные, симметричные и квадратные матрицы.** Матрица называется *квадратной*, если количество ее строк равно количеству столбцов. Квадратная матрица называется *симметричной*, если ее элемент  $(i, j)$  равен элементу  $(j, i)$ . *Диагональная* матрица – это квадратная матрица, в которой все внедиагональные

элементы равны нулю, то есть если квадратная матрица  $A$  диагональная, тогда  $a_{ij}=0$  для  $i \neq j$ .

**Специальные матрицы.** Важной матрицей является *единичная матрица*,  $I_n$ , которая является диагональной матрицей  $n \times n$  с единицами на диагонали. *Нулевая матрица*  $0_{n \times m}$  является матрицей  $n \times m$ , где все элементы – нули.

**Транспонирование.** При транспонировании матриц меняются местами строки и столбцы. Другими словами, транспонирование матриц делает из матрицы  $A$   $n \times m$  матрицу  $m \times n$ , которая обозначается  $A'$ ; говоря иначе, при транспонировании  $A$  строки матрицы  $A$  становятся столбцами матрицы  $A'$ . Если  $a_{ij}$  – это  $(i, j)$ -элемент матрицы  $A$ , тогда  $A'$  (транспонированная матрица  $A$ ) равна:

$$A' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}.$$

Транспонирование вектора является частным случаем транспонирования матриц. Таким образом, при транспонировании вектора мы получаем из вектора-строки вектор-столбец, то есть если  $b$  – вектор-столбец  $n \times 1$ , тогда при его транспонировании мы получим вектор-строку  $1 \times n$ :

$$b' = [b_1 \ b_2 \ \cdots \ b_n].$$

Транспонирование вектора-строки дает вектор-столбец.

### Операции в линейной алгебре: сложение и умножение

**Матричное сложение.** Две матрицы, имеющие одинаковые размерности (обе  $n \times m$ ), можно сложить. Суммой матриц называется сумма их элементов, то есть если  $C = A + B$ , тогда  $c_{ij} = a_{ij} + b_{ij}$ . Частный случай матричного сложения – это векторное сложение: если  $a$  и  $b$  оба векторы-столбцы  $n \times 1$ , тогда их сумма  $c = a + b$  является поэлементной суммой, то есть  $c_i = a_i + b_i$ .

**Матричное и векторное умножение.** Пусть  $a$  и  $b$  – векторы-столбцы размерности  $n \times 1$ . Тогда произведение транспонированного вектора  $a$  (который является вектором-строкой) и вектора  $b$  – это  $a'b = \sum_{i=1}^n a_i b_i$ . Применяя это определение к ситуации, когда  $b=a$ , получаем  $a'a = \sum_{i=1}^n a_i^2$ .

Аналогично, матрицы  $A$  и  $B$  могут быть перемножены, если они являются конформными, то есть если количество столбцов матрицы  $A$  равно количеству строк матрицы  $B$ . В частности, предположим, что  $A$  имеет размерность  $n \times m$ , а  $B$   $n \times r$ . Тогда произведение матриц  $A$  и  $B$  – это матрица  $C$  размерности  $n \times r$ , то есть  $C = AB$ , где  $(i, j)$ -й элемент  $C$  равен  $c_{ij} = \sum_{k=1}^m a_{ik} b_{kj}$ . Говоря иначе,  $(i, j)$ -й элемент  $AB$  – это произведение, результат умножения вектора-строки, то есть  $i$ -й строки матрицы  $A$  на вектор-столбец, то есть  $j$ -й столбец матрицы  $B$ .

Произведение скаляра  $d$  и матрицы  $A$  дает  $(i, j)$  элемент, равный  $da_{ij}$ , то есть каждый элемент  $A$  умножается на скаляр  $d$ .

**Несколько полезных свойств матричного сложения и умножения.** Пусть  $A, B$  и  $C$  – матрицы, тогда:

$$A + B = B + A;$$

$$(A + B) + C = A + (B + C);$$

$$(A + B)' = A' + B';$$

Если  $A$  размерности  $n \times m$ , тогда:

$$AI_m = A \text{ и } I_n A = A;$$

$$A(BC) = (AB)C;$$

$$(A + B)C = AC + BC;$$

$$\text{и } (AB)' = B'A'.$$

В общем случае произведение матриц некоммутативно, то есть  $AB \neq BA$ , хотя есть несколько частных случаев, когда произведение матриц коммутативно, например, если обе  $A$  и  $B$  являются диагональными матрицами размерности  $n \times n$ , тогда  $AB = BA$ .

### Обратная матрица, квадратный корень из матрицы и так далее

**Обратная матрица.** Пусть  $A$  является квадратной матрицей. Обратной матрицей к  $A$  (если такая существует) называется матрица, для которой выполняется равенство  $A^{-1}A = I_n$ . Если в действительности  $A^{-1}$  существует, то говорят, что  $A$  обратима или невырождена. Если обе матрицы  $A$  и  $B$  обратимы, то  $(AB)^{-1} = B^{-1}A^{-1}$ .

**Положительно определенные матрицы, положительно полуопределенные матрицы.** Пусть  $V$  является квадратной матрицей  $n \times n$ . Тогда  $V$  положительно определена, если  $c'Vc > 0$  для любого ненулевого вектора  $c$  размера  $n \times 1$ . Аналогично,  $V$  положительно полуопределена, если  $c'Vc \geq 0$  для любого ненулевого вектора  $c$  размера  $n \times 1$ . Если матрица  $V$  положительно определена, тогда она обратима.

**Линейная независимость.** Два вектора размерности  $n \times 1$ , векторы  $a_1$  и  $a_2$ , называются линейно независимыми, если не существует ненулевых скалярных величин  $c_1$  и  $c_2$ , таких что  $c_1a_1 + c_2a_2 = \mathbf{0}_{n \times 1}$ . В общем случае набор из  $k$  векторов  $a_1, a_2, \dots, a_k$  линейно независим, если не существует ненулевых скалярных величин  $c_1, c_2, \dots, c_k$  таких что  $c_1a_1 + c_2a_2 + \dots + c_ka_k = \mathbf{0}_{n \times 1}$ .

**Ранг матрицы.** Рангом матрицы  $A$  размерности  $n \times m$  называется количество линейно независимых столбцов  $A$ . Ранг матрицы  $A$  обозначается как  $\text{rank}(A)$ . Если ранг матрицы  $A$  равен количеству столбцов  $A$ , тогда считается, что матрица  $A$  имеет полный ранг системы столбцов (строк). Если матрица  $A$  размера  $n \times m$  имеет полный ранг системы столбцов, то не существует ненулевого вектора  $c$  размера  $m \times 1$  такого, что  $Ac = \mathbf{0}_{n \times 1}$ . Если  $A$  имеет размерность  $n \times n$  и  $\text{rank}(A) = n$ , тогда  $A$  не вырождена. Если матрица  $A$  размерности  $n \times m$  имеет полный ранг системы столбцов, то  $A'A$  не вырождена.

**Квадратный корень из матрицы.** Пусть  $V$  является квадратной симметричной положительно определенной матрицей  $n \times n$ . Квадратным корнем из матрицы  $V$  называется матрица  $F$  размера  $n \times n$ , такая что  $F'F = V$ . Квадратный корень из положительно определенной матрицы всегда существует, но не

единствен. Свойство квадратного матричного корня  $\mathbf{F}\mathbf{V}^{-1}\mathbf{F}' = \mathbf{I}_n$ . Кроме того, квадратный корень из положительно определенной матрицы обратим, так что  $\mathbf{F}'^{-1}\mathbf{V}\mathbf{F}^{-1} = \mathbf{I}_n$ .

**Собственные значения и собственные векторы.** Пусть  $A$  является матрицей  $n \times n$ . Если  $n \times 1$  вектор  $\mathbf{q}$  и скаляр  $\lambda$  довлетворяют  $A\mathbf{q} = \lambda \mathbf{q}$ , где  $\mathbf{q}'\mathbf{q} = 1$ , тогда  $\lambda$  называется *собственным значением* матрицы  $A$  и  $\mathbf{q}$  – *собственным вектором* матрицы  $A$ , связанным с этим собственным значением. Матрица размера  $n \times n$  имеет  $n$  собственных значений, необязательно являющихся различными, и  $n$  собственных векторов.

Если  $V$  является симметричной положительно определенной матрицей  $n \times n$ , тогда все собственные значения  $V$  – положительные действительные числа и все собственные векторы  $V$  являются действительными. Также  $V$  может быть записана в терминах собственных значений и собственных векторов как  $V = Q \Lambda Q'$ , где  $\Lambda$  диагональная матрица  $n \times n$  с диагональными элементами, которые равны собственным значениям  $V$ , и  $Q$  – это матрица  $n \times n$ , состоящая из собственных векторов  $V$ , записанная так, что  $i$ -й столбец матрицы  $Q$  является собственным вектором, соответствующим определенному собственному значению, то есть  $i$ -му диагональному элементу  $\Lambda$ . Собственные векторы ортогональны, так что  $QQ' = I_n$ .

**Идемпотентные матрицы.** Матрица  $C$  называется *идемпотентной*, если  $C$  квадратная и  $C = CC$ . Если  $C$  идемпотентная матрица  $n \times n$  также симметрична, то  $C$  положительно полуопределена и  $C$  имеет  $r$  собственных значений, которые равны 1, и  $n - r$  собственных значений, которые равны 0, где  $r = \text{rank}(C)$  (упражнение 18.10).

## Приложение 18.2. Многомерные распределения

В данном приложении собраны разные определения и факты о распределениях векторов случайных величин. Мы начнем с определения математического ожидания и ковариационной матрицы вектора  $V$  случайных переменных размерности  $n$ . Далее представим многомерное нормальное распределение и обобщим несколько фактов о линейной и квадратичной функциях совместно нормально распределенных случайных величин.

### Обратная матрица, квадратный корень из матрицы и прочее

Первые и вторые моменты вектора  $m \times 1$  случайных величин  $V = (V_1 \ V_2 \dots V_m)'$  сводятся к его вектору средних и ковариационной матрице.

Так как  $V$  – вектор, то вектор его средних, то есть *вектор математического ожидания*, это  $E(V) = \mu_V$ .  $i$ -й элемент вектора средних значений – это математическое ожидание  $i$ -го элемента  $V$ .

*Ковариационная матрица*  $V$  – это матрица, состоящая из дисперсий  $\text{var}(V_i)$ ,  $i = 1, \dots, n$ , стоящих на главной диагонали, и  $\text{cov}(V_i, V_j)$  – на  $(i, j)$  вне диагональных элементах. В матричной форме ковариационная матрица  $\Sigma_V$  – это:

$$\Sigma_V = E[(V - \mu_V)(V - \mu_V)'] = \begin{bmatrix} \text{var}(V_1) & \dots & \text{cov}(V_1, V_m) \\ \vdots & \ddots & \vdots \\ \text{cov}(V_m, V_1) & \dots & \text{var}(V_m) \end{bmatrix}. \quad (18.72)$$

## Многомерное нормальное распределение

Вектор  $m \times 1$  случайных величин  $V$  имеет многомерное нормальное распределение с вектором средних значений  $\mu_V$  и ковариационной матрицей  $\Sigma_V$ , если совместная плотность распределения имеет вид:

$$f(V) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma_V)}} \exp\left[-\frac{1}{2}(V - \mu_V)' \Sigma_V^{-1} (V - \mu_V)\right], \quad (18.73)$$

где  $\det(\Sigma_V)$  – определитель матрицы  $\Sigma_V$ . Многомерное нормальное распределение обозначается  $N(\mu_V, \Sigma_V)$ .

Важным моментом, касающимся многомерного нормального распределения, является то, что если две совместно распределенные случайные величины не коррелированы (что эквивалентно тому, что они имеют диагональную ковариационную матрицу), тогда они независимо распределены. То есть пусть  $V_1$  и  $V_2$  совместно нормально распределены с соответственно размерностями  $m_1 \times 1$  и  $m_2 \times 1$ . Тогда, если  $\text{cov}(V_1, V_2) = E[(V_1 - \mu_{V_1})(V_2 - \mu_{V_2})'] = 0_{m_1 \times m_2}$ , то  $V_1$  и  $V_2$  независимы.

Если  $\{V_i\}$  является i.i.d.  $N(0, \sigma_V^2)$ , то  $\Sigma_V = \sigma_V^2 I_m$ , и многомерное нормальное распределение упрощается до произведения  $m$  одномерных нормальных плотностей.

## Распределение линейной комбинации и квадратичных форм нормально распределенных случайных величин

Линейные комбинации многомерной нормальной случайной величины сами по себе нормально распределены, и некоторые квадратичные формы многомерной нормальной случайной величины имеют распределение хи-квадрат. Пусть  $V$  является вектором  $m \times 1$  случайных величин, распределенных как  $N(\mu_V, \Sigma_V)$ , пусть  $A$  и  $B$  – неслучайные матрицы размерности  $a \times m$  и  $b \times m$  и пусть  $d$  – неслучайный вектор  $a \times 1$ . Тогда

$$d + AV \text{ распределен как } N(d + A\mu_V, A\Sigma_V A'); \quad (18.74)$$

$$\text{cov}(AV, BV) = A\Sigma_V B'. \quad (18.75)$$

$$\begin{aligned} \text{Если } A\Sigma_V B' = \mathbf{0}_{a \times b}, \text{ то } AV \text{ и } BV \text{ независимо} \\ \text{распределены и} \end{aligned} \quad (18.76)$$

$$(V - \mu_V)' \Sigma_V^{-1} (V - \mu_V) \text{ распределен как } \chi_m^2. \quad (18.77)$$

Пусть  $U$  является  $m$ -мерной многомерной стандартной нормальной случайной величиной с распределением  $N(\mathbf{0}, I_m)$ . Если  $C$  является симметричной и идемпотентной матрицей, тогда

$$U' C U \text{ имеет } \chi_r^2 \text{ распределение, где } r = \text{rank}(C). \quad (18.78)$$

Выражение (18.78) доказывается в упражнении (18.11).

### **Приложение 18.3. Вывод асимптотического распределения $\hat{\beta}$**

В данном приложении приводится доказательство асимптотической нормальности распределения  $\sqrt{n}(\hat{\beta} - \beta)$ , данного в (18.12). Из этого следует, что  $\hat{\beta} \xrightarrow{p} \beta$ .

Во-первых, рассмотрим матрицу в «знаменателе»  $(X'X / n) = \frac{1}{n} \sum_{i=1}^n X_i X_i'$  (18.15). Элемент  $(j, l)$  этой матрицы – это  $\frac{1}{n} \sum_{i=1}^n X_{ji} X_{li}$ . По второй предпосылке из вставки «Основные понятия 18.1»  $X_i$  является i.i.d., так что  $X_{ji} X_{li}$  тоже i.i.d. По третьей предпосылке из вставки «Основные понятия 18.1» каждый элемент из  $X_i$  имеет четвёртый момент, следовательно, по неравенству Коши–Шварца (приложение 17.2),  $X_{ji} X_{li}$  имеет два момента. Так как  $X_{ji} X_{li}$  является i.i.d. с двумя моментами,  $\frac{1}{n} \sum_{i=1}^n X_{ji} X_{li}$  удовлетворяет закону больших чисел, поэтому  $\frac{1}{n} \sum_{i=1}^n X_{ji} X_{li} \xrightarrow{p} E(X_{ji} X_{li})$ . Это справедливо для всех элементов из  $(X'X / n)$ , поэтому

$$\text{этому } \left( \frac{X'X}{n} \right) \xrightarrow{p} E(X_i X_i') = Q_X.$$

Далее рассмотрим матрицу «числитель» из (18.15),  $X'U / \sqrt{n} = \sqrt{\frac{1}{n} \sum_{i=1}^n V_i}$ ,

где  $V_i = X_i u_i$ . По первой предпосылке из вставки «Основные понятия 18.1» и по закону повторного математического ожидания  $E(V_i) = E[X_i E(u_i | X_i)] = 0_{k+1}$ . По второй предпосылке МНК  $V_i$  является i.i.d. Пусть  $c$  будет конечным вектором размерности  $k+1$ . По неравенству Коши–Шварца,  $E[(c'V_i)^2] = E[(c'X_i u_i)^2] = E[(c'X_i)^2 (u_i)^2] \leq \sqrt{E[(c'X_i)^4]} E(u_i^4)$ , что является конечным по третьей предпосылке МНК. Это верно для каждого такого вектора  $c$ , так что  $E(V_i V_i') = \Sigma_V$  конечна и, предположим, положительно определена. Таким образом, из многомерной центральной предельной теормы следует:  $\sqrt{\frac{1}{n} \sum_{i=1}^n V_i} = \frac{1}{\sqrt{n}} X'U$ , то есть

$$\frac{1}{\sqrt{n}} X'U \xrightarrow{d} N(0_{k+1}, \Sigma_V). \quad (18.79)$$

Результат из (18.12) следует из (18.15) и (18.79), состоятельности  $X'X / n$  четвертой предпосылки МНК (в которой утверждается, что  $X'X / n$  существует) и теоремы Слуцкого.

### **Приложение 18.4. Вывод точных распределений МНК-статистик при нормальных ошибках**

В данном приложении выводятся МНК-распределения  $t$ -статистики, рассчитанной в предположении гомоскедастичности ошибок, в условиях выполне-

ния нулевой гипотезы, представленной в (18.35), и  $F$ -статистики, рассчитанной в предположении гомоскедастичности ошибок, из (18.37), предполагая, что все шесть предпосылок из вставки «Основные понятия 18.1» выполняются.

### Доказательство (18.35)

Если (i)  $Z$  имеет стандартное нормальное распределение, (ii)  $W$  имеет  $\chi^2_m$ -распределение, (iii)  $Z$  и  $W$  независимо распределены, тогда случайная величина  $Z / \sqrt{W/m}$  имеет  $t$ -распределение с  $m$  степенями свободы (приложение 17.1). Подставляя  $\tilde{t}$  в эту запись, заметим, что  $\hat{\Sigma}_{\hat{\beta}} = (s_{\hat{u}}^2 / \sigma_u^2) \Sigma_{\hat{\beta}|X}$ . Затем перепишем (18.34) так:

$$\tilde{t} = \frac{(\hat{\beta}_j - \beta_{j,0}) / \sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}}{\sqrt{W/(n-k-1)}}, \quad (18.80)$$

где  $W = (n-k-1)(s_{\hat{u}}^2 / \sigma_u^2)$ , и пусть  $Z = (\hat{\beta}_j - \beta_{j,0}) / \sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}$  и  $m = n-k-1$ . В этих обозначениях  $\tilde{t} = Z / \sqrt{W/m}$ . Таким образом, чтобы доказать результат, представленный в (18.35), мы должны доказать пункты (i) – (iii) для  $Z$ ,  $W$  и  $m$ .

(i) Следствие выражения (18.30) заключается в том, что при выполнении нулевой гипотезы  $Z = (\hat{\beta}_j - \beta_{j,0}) / \sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}$  имеет точное стандартное нормальное распределение, которое показано в (i).

(ii) Из (18.31) следует, что  $W$  имеет распределение  $\chi^2_m$ .

(iii) Чтобы показать (iii), необходимо сначала показать, что  $\hat{\beta}_j$  и  $s_{\hat{u}}^2$  независимо распределены. Из (18.14) и (18.29) следует, что  $\hat{\beta} - \beta = (X'X)^{-1} X'U$  и  $s_{\hat{u}}^2 = (M_x U)' (M_x U) / (n-k-1)$ . Таким образом,  $\hat{\beta} - \beta$  и  $s_{\hat{u}}^2$  независимы, если  $(X'X)^{-1} X'U$  и  $M_x U$  независимы. Обе  $(X'X)^{-1} X'U$  и  $M_x U$  являются линейной комбинацией  $U$ , которая имеет условное  $N(\mathbf{0}_{(n-k+1)}, \sigma_u^2 I_{n-k+1})$  распределение относительно  $X$ . Но так как  $M_x X (X'X)^{-1} = \mathbf{0}_{n \times (k+1)}$  [выражение (18.26)], из этого следует, что  $(X'X)^{-1} X'U$  и  $M_x U$  независимо распределены [выражение (18.76)]. Следовательно, при выполнении всех шести предпосылок из вставки «Основные понятия 18.1»:

$$\hat{\beta} \text{ и } s_{\hat{u}}^2 \text{ независимо распределены,} \quad (18.81)$$

что доказывает (iii) и, следовательно, является доказательством выражения (18.35).

### Доказательство (18.37)

$F_{n_1, n_2}$ -распределение – это распределение  $(W_1 / n_1) / (W_2 / n_2)$ , где (i)  $W_1$  имеет распределение  $\chi^2_{n_1}$ , (ii)  $W_2$  имеет распределение  $\chi^2_{n_2}$  и (iii)  $W_1$  и  $W_2$  распределены независимо (приложение 17.1). Для того чтобы выразить  $\tilde{F}$  в этих обозначениях, положим  $W_1 = (R\hat{\beta} - r)' [R(X'X)^{-1} R' \sigma_u^2] (R\hat{\beta} - r)$  и  $W_2 = (n-k-1)s_{\hat{u}}^2 / \sigma_u^2$ . Подстановка этих выражений в (18.36) дает  $\tilde{F} = (W_1 / q) / [W_2 / (n-k-1)]$ .

Таким образом, по определению  $F$ -распределения,  $\tilde{F}$  имеет  $F_{q, n-k-1}$ -распределение, если выполняются условия (i) – (iii) с  $n_1 = q$  и  $n_2 = n - k - 1$ .

(i) В условиях нулевой гипотезы  $\mathbf{R}\hat{\beta} - \mathbf{r} = \mathbf{R}(\hat{\beta} - \beta)$ . Так как  $\hat{\beta}$  имеет условное нормальное распределение (18.30) и так как  $\mathbf{R}$  является неслучайной матрицей, то  $\mathbf{R}(\hat{\beta} - \beta)$  имеет распределение  $N(\mathbf{0}_{q \times 1}, \mathbf{R}(X'X)^{-1} \mathbf{R}'\sigma_u^2)$ , условное по  $X$ . Таким образом, по (18.77) из приложения 18.2 следует, что  $(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(X'X)^{-1} \mathbf{R}'\sigma_u^2] (\mathbf{R}\hat{\beta} - \mathbf{r})$  имеет распределение  $\chi_q^2$ , доказывая (i).

(ii) Требование (ii) показано в (18.31).

(iii) Как было показано,  $\hat{\beta} - \beta$  и  $s_u^2$  независимо распределены [выражение (18.81)]. Из этого следует, что  $\mathbf{R}\hat{\beta} - \mathbf{r}$  и  $s_u^2$  распределены независимо, из чего в свою очередь следует, что  $W_1$  и  $W_2$  распределены независимо, доказывая (iii) и завершая доказательство.

### Приложение 18.5. Вывод точных распределений МНК-статистик при нормальных ошибках

В этом приложении доказывается теорема Гаусса–Маркова для модели множественной регрессии (вставка «Основные понятия 18.1»). Пусть  $\tilde{\beta}$  является линейной условно несмешенной оценкой  $\beta$ , при этом  $\tilde{\beta} = A'Y$  и  $E(\tilde{\beta} | X) = \beta$ , где  $A$  – матрица размерности  $n \times (k+1)$ , которая может зависеть от  $X$  и неслучайной константы. Мы покажем, что  $\text{var}(\tilde{\beta}) \leq \text{var}(\tilde{\beta})$  для всех векторов  $c$  размерности  $(k+1)$ , где неравенство выполняется как равенство, только если  $\tilde{\beta} = \tilde{\beta}$ .

Так как  $\tilde{\beta}$  линейна, она может быть записана как  $\tilde{\beta} = A'Y = A'(X\beta + U) = (A'X)\beta + A'U$ . По первому условию теоремы Гаусса–Маркова  $E(U|X) = \mathbf{0}_{n \times 1}$ , поэтому  $E(\tilde{\beta} | X) = (A'X)\beta$ , но так как  $\tilde{\beta}$  является условно несмешенной оценкой,  $E(\tilde{\beta} | X) = \beta = (A'X)\beta$ , из чего следует, что  $A'X = I_{k+1}$ . Таким образом,  $\tilde{\beta} = \beta + A'U$ , поэтому  $\text{var}(\tilde{\beta} | X) = \text{var}(A'U | X) = E(A'UU'A | X) = A'E(UU'|X)A = \sigma_u^2 A'A$ , где третье равенство является следствием того, что  $A$  может зависеть от  $X$ , но не от  $U$ , и последнее равенство следует из второго условия теоремы Гаусса–Маркова. Иначе говоря, если  $\tilde{\beta}$  линейная и несмешенная, тогда при выполнении условий теоремы Гаусса–Маркова:

$$A'X = I_{k+1} \text{ и } \text{var}(\tilde{\beta} | X) = \sigma_u^2 A'A. \quad (18.82)$$

Результат (18.82) также применим к  $\hat{\beta}$  при  $A = X(X'X)^{-1}$ , где  $(X'X)^{-1}$  существует по третьему условию теоремы Гаусса–Маркова.

Теперь положим  $A = \hat{A} + D$ , где  $D$  является разностью между матрицей весов  $A$  и  $\hat{A}$ . Заметим, что  $\hat{A}'A = (X'X)^{-1}X'A = (X'X)^{-1}$  [по (18.82)] и  $\hat{A}'\hat{A} = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$ , поэтому  $\hat{A}'D = \hat{A}'(A - \hat{A}) = \hat{A}'A - \hat{A}'\hat{A} = \mathbf{0}_{(k+1) \times (k+1)}$ . Подставляя  $A = \hat{A} + D$  в формулу для условной дисперсии в (18.82), получаем:

$$\begin{aligned} \text{var}(\tilde{\beta} | X) &= \sigma_u^2 (\hat{A} + D)' (\hat{A} + D) = \\ &= \sigma_u^2 [\hat{A}'\hat{A} + \hat{A}'D + D'\hat{A} + D'D] = \sigma_u^2 (X'X)^{-1} + \sigma_u^2 D'D, \end{aligned} \quad (18.83)$$

где в финальном равенстве используется тот факт, что  $\hat{\mathbf{A}}'\hat{\mathbf{A}} = (\mathbf{X}'\mathbf{X})^{-1}$  и  $\hat{\mathbf{A}}'\mathbf{D}' = \mathbf{0}_{(k+1) \times (k+1)}$ .

Так как  $\text{var}(\hat{\beta} | X) = \sigma_u^2 (\mathbf{X}'\mathbf{X})^{-1}$ , то из (18.82) и (18.83) следует, что  $\text{var}(\tilde{\beta} | X) - \text{var}(\hat{\beta} | X) = \sigma_u^2 \mathbf{D}'\mathbf{D}$ . Разность между дисперсиями двух оценок линейной комбинации  $\mathbf{c}'\beta$  равна:

$$\text{var}(\mathbf{c}'\tilde{\beta} | X) - \text{var}(\mathbf{c}'\hat{\beta} | X) = \sigma_u^2 \mathbf{c}'\mathbf{D}'\mathbf{D}\mathbf{c} \geq 0. \quad (18.84)$$

Неравенство (18.84) выполняется для всех линейных комбинаций  $\mathbf{c}'\beta$ , и неравенство выполняется как равенство для любого ненулевого  $\mathbf{c}$ , только если  $\mathbf{D} = \mathbf{0}_{n \times (k+1)}$ , то есть  $\mathbf{A} = \hat{\mathbf{A}}$ , или, что эквивалентно,  $\tilde{\beta} = \hat{\beta}$ . Таким образом,  $\mathbf{c}'\hat{\beta}$  имеет наименьшую дисперсию среди всех линейно несмещенных оценок  $\mathbf{c}'\beta$ , то есть оценка МНК является BLUE.

### Приложение 18.6. Вывод некоторых результатов для IV- и GMM-оценок

**Эффективность 2МНК-оценки при наличии гомоскедастичности ошибок [доказательство (18.62)]**

Если ошибки  $u_i$  гомоскедастичны, разность между  $\Sigma_A^{IV}$  (18.61) и  $\Sigma^{\text{TSLS}}$  (18.85)] равна:

$$\begin{aligned} \Sigma_A^{IV} - \Sigma^{\text{TSLS}} &= (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZZ} \mathbf{A} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \sigma_u^2 - \\ &\quad - (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} \sigma_u^2 = (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{A} \left[ \mathbf{Q}_{ZZ} - \right. \\ &\quad \left. - \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \right] \mathbf{A} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \sigma_u^2, \end{aligned} \quad (18.85)$$

где второй член в скобках во втором равенстве следует из равенства  $(\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX} = \mathbf{I}_{(k+r+1)}$ . Пусть  $\mathbf{F}$  обозначает матрицу, являющуюся квадратным корнем из  $\mathbf{Q}_{ZZ}$ , так что  $\mathbf{Q}_{ZZ} = \mathbf{F}'\mathbf{F}$  и  $\mathbf{Q}_{ZZ}^{-1} = \mathbf{F}^{-1}\mathbf{F}^{-1'}$  [последнее равенство следует из замечания о том, что  $(\mathbf{F}'\mathbf{F})^{-1} = \mathbf{F}^{-1}\mathbf{F}^{-1'} \mathbf{F}^{-1} = \mathbf{F}^{-1'}$ ]. Последнее выражение в (18.85) может быть переписано, чтобы получить:

$$\begin{aligned} \Sigma_A^{IV} - \Sigma^{\text{TSLS}} &= (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{A} \mathbf{F}' \left[ \mathbf{I} - \right. \\ &\quad \left. - \mathbf{F}^{-1'} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{F}^{-1} \mathbf{F}^{-1'} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{F}^{-1} \right] \times \\ &\quad \times \mathbf{F} \mathbf{A} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \sigma_u^2, \end{aligned} \quad (18.86)$$

где второе выражение в скобках использует  $\mathbf{F}'\mathbf{F}^{-1} = \mathbf{I}$ . Таким образом,

$$\mathbf{c}' (\Sigma_A^{IV} - \Sigma^{\text{TSLS}}) \mathbf{c} = \mathbf{d}' \left[ \mathbf{I} - \mathbf{D} (\mathbf{D}'\mathbf{D})^{-1} \mathbf{D}' \right] \mathbf{d} \sigma_u^2, \quad (18.87)$$

где  $\mathbf{d} = \mathbf{F} \mathbf{A} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{A} \mathbf{Q}_{ZX})^{-1} \mathbf{c}$   $\mathbf{D} = \mathbf{F}^{-1'} \mathbf{Q}_{ZX}$ . Теперь  $\mathbf{I} - \mathbf{D} (\mathbf{D}'\mathbf{D})^{-1} \mathbf{D}'$  – симметричная идемпотентная матрица (упражнение 18.5). В результате  $\mathbf{I} - \mathbf{D} (\mathbf{D}'\mathbf{D})^{-1} \mathbf{D}'$

имеет собственные значения, которые равны 0 или 1 и  $\mathbf{d}'(\mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}')\mathbf{d} \geq 0$  (упражнение 18.10). Таким образом,  $\mathbf{c}'(\Sigma_A^{IV} - \Sigma^{\text{TSLS}})\mathbf{c} \geq 0$ , что доказывает, что 2МНК-оценка эффективней при гомоскедастичности.

### *Асимптотическое распределение J-статистики при наличии гомоскедастичности ошибок*

J-статистика определена в (18.63). Для начала заметим, что

$$\begin{aligned}\hat{\mathbf{U}} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{TSLS}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{Y} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{U}) - \\ &- \mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}) = \mathbf{U} - \\ &- \mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{U} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z]\mathbf{U}.\end{aligned}\quad (18.88)$$

Таким образом,

$$\begin{aligned}\hat{\mathbf{U}}'\mathbf{P}_Z\hat{\mathbf{U}} &= \mathbf{U}'[\mathbf{I} - \mathbf{P}_Z\mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}']\mathbf{P}_Z[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z]\mathbf{U} = \\ &= \mathbf{U}'[\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z]\mathbf{U},\end{aligned}\quad (18.89)$$

где второе равенство получается путем упрощения предыдущего выражения. Так как  $\mathbf{Z}'\mathbf{Z}$  – симметричная и положительно определенная матрица, она может быть записана через собственные квадратные корни,  $\mathbf{Z}'\mathbf{Z} = (\mathbf{Z}'\mathbf{Z})^{1/2}(\mathbf{Z}'\mathbf{Z})^{1/2}$ , и этот матричный квадратный корень обратим, то есть  $(\mathbf{Z}'\mathbf{Z})^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1/2}(\mathbf{Z}'\mathbf{Z})^{-1/2}$ , где  $(\mathbf{Z}'\mathbf{Z})^{-1/2} = [(\mathbf{Z}'\mathbf{Z})^{1/2}]^{-1}$ . Таким образом,  $\mathbf{P}_Z$  может быть записана как  $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{B}\mathbf{B}'$ , где  $\mathbf{B} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$ . Подставляя это выражение для  $\mathbf{P}_Z$  в последнее выражение в (18.89), получаем:

$$\begin{aligned}\hat{\mathbf{U}}'\mathbf{P}_Z\hat{\mathbf{U}} &= \mathbf{U}'[\mathbf{B}\mathbf{B}' - \mathbf{B}\mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}\mathbf{B}']\mathbf{U} = \\ &= \mathbf{U}'\mathbf{B}[\mathbf{I} - \mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}]\mathbf{B}'\mathbf{U} = \mathbf{U}'\mathbf{B}\mathbf{M}_{\mathbf{B}'\mathbf{X}}\mathbf{B}'\mathbf{U},\end{aligned}\quad (18.90)$$

где  $\mathbf{M}_{\mathbf{B}'\mathbf{X}} = \mathbf{I} - \mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}$  – симметричная и идемпотентная матрица.

Асимптотическое распределение  $\hat{\mathbf{U}}'\mathbf{P}_Z\hat{\mathbf{U}}$  тестовой статистики в условиях нулевой гипотезы может быть найдено с помощью умножения пределов по вероятности и по распределению различных членов в последнем выражении равенства (18.90). В условиях нулевой гипотезы, что  $E(\mathbf{Z}_i u_i) = 0$ ,  $\mathbf{Z}'\mathbf{U} / \sqrt{n}$  имеет нулевое среднее и, применяя центральную предельную теорему, получаем, что  $\mathbf{Z}'\mathbf{U} / \sqrt{n} \xrightarrow{d} N(0, \mathbf{Q}_{ZZ} \sigma_u^2)$ . В дополнение к этому  $\mathbf{Z}'\mathbf{Z} / n \xrightarrow{p} \mathbf{Q}_{ZZ}$  и  $\mathbf{X}'\mathbf{Z} / n \xrightarrow{p} \mathbf{Q}_{XZ}$ . Таким образом,  $\mathbf{B}'\mathbf{U} = (\mathbf{Z}'\mathbf{Z})^{-1/2} \mathbf{Z}'\mathbf{U} = (\mathbf{Z}'\mathbf{Z} / n)^{-1/2} (\mathbf{Z}'\mathbf{U} / \sqrt{n}) \xrightarrow{d} \sigma_u z$ , где  $z$  распределена как  $N(\mathbf{0}_{m+r+1}, \mathbf{I}_{m+r+1})$ . В дополнение имеем  $\mathbf{B}'\mathbf{X} / \sqrt{n} = (\mathbf{Z}'\mathbf{Z} / n)^{-1/2} (\mathbf{Z}'\mathbf{X} / n) \xrightarrow{p} \mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZX}$ , так что  $\mathbf{M}_{\mathbf{B}'\mathbf{X}} \xrightarrow{p} \mathbf{I} - \mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZX} (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1/2})$ .

$$\mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1/2} = \mathbf{M}_{\mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZX}}. \text{ Таким образом,}$$

$$\hat{U}' \mathbf{P}_z \hat{U} \xrightarrow{d} \left( z' M_{Q_{zz}^{-1/2} Q_{zx}} z \right) \sigma_u^2. \quad (18.91)$$

В условиях нулевой гипотезы 2МНК-оценка состоятельна и коэффициент в регрессии  $\hat{U}$  на  $Z$  сходится по вероятности к нулю [следствие (18.91)], так что знаменатель в определении  $J$ -статистики — состоятельная оценка  $\sigma_u^2$ .

$$\hat{U}' M_z \hat{U} / (n - m - r - 1) \xrightarrow{p} \sigma_u^2. \quad (18.92)$$

Из определения  $J$ -статистики и (18.91) и (18.92) следует:

$$J = \frac{\hat{U}' \mathbf{P}_z \hat{U}}{\hat{U}' M_z \hat{U} / (n - m - r - 1)} \xrightarrow{d} z' M_{Q_{zz}^{-1/2} Q_{zx}} z. \quad (18.93)$$

Так как  $z$  — стандартный нормальный случайный вектор и  $M_{Q_{zz}^{-1/2} Q_{zx}}$  — симметричная и идемпотентная матрица,  $J$  имеет распределение хи-квадрат со степенью свободы, равной рангу  $M_{Q_{zz}^{-1/2} Q_{zx}}$  [(18.78)]. Так как  $Q_{zz}^{-1/2} Q_{zx}$  имеет размерность  $(m+r+1) \times (k+r+1)$ ,  $m > k$ , ранг  $M_{Q_{zz}^{-1/2} Q_{zx}}$  равен  $m-k$  [упражнение (18.5)].

Так что  $J \xrightarrow{d} \chi_{m-k}^2$ , что и сформулировано в (18.64).

### Эффективность эффективной GMM-оценки

Недоступная эффективная GMM-оценка  $\tilde{\beta}^{Eff.GMM}$  приведена в (18.66). Доказательство того, что  $\tilde{\beta}^{Eff.GMM}$  эффективна, заключается в том, чтобы показать, что  $c' (\Sigma_A^{IV} - \Sigma^{Eff.GMM}) c \geq 0$  для всех векторов  $c$ . Доказательство аналогично доказательству для 2МНК-оценки в первом разделе данного приложения с одной лишь модификацией — заменой  $Q_{zz} \sigma_u^2$  на  $H^{-1}$  в (18.85) и в дальнейшем.

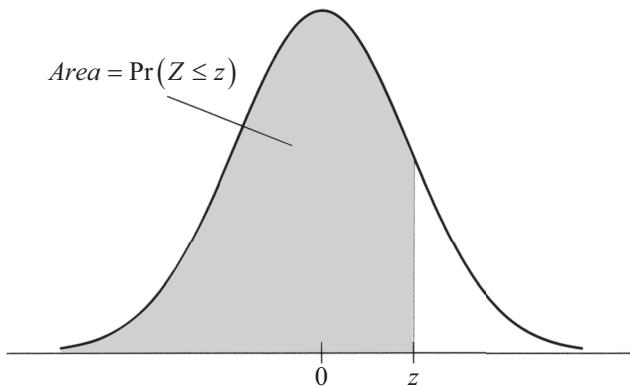
### Распределение GMM J-статистики

GMM  $J$ -статистика приведена в (18.70). Доказательство того, что в условиях нулевой гипотезы  $J^{GMM} \xrightarrow{d} \chi_{m-k}^2$ , аналогично соответствующему доказательству для 2МНК  $J$ -статистики, рассчитанной при наличии гомоскедастичности.

# Приложения

Таблица 1

**Функция стандартного  
нормального распределения  $\Phi(z) = \Pr(Z \leq z)$**



Второй знак после запятой у $z$											
$z$	0	1	2	3	4	5	6	7	8	9	
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019	
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036	
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048	
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064	
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084	
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110	
-2,1	0,0179	0,0174	0,170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143	
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183	
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233	
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294	
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367	
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455	
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559	
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681	

## Продолжение таблицы 1

<i>z</i>	Второй знак после запятой у <i>z</i>									
	0	1	2	3	4	5	6	7	8	9
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
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	0,6029	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	0,9699	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

Окончание таблицы 1

<b>Второй знак после запятой у <math>z</math></b>										
<b><math>z</math></b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
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

**Примечание.** Эта таблица может быть использована для вычисления  $\Pr(Z \leq z)$ , где  $Z$  – стандартная нормальная случайная величина. Например, если  $z=1,17$ , то вероятность равна 0,8790, и ее можно найти в таблице на пересечении строки 1,1 и столбца 7.

Таблица 2

**Критические значения для двустороннего и одностороннего тестов  
с использованием  $t$ -распределения Стьюдента**

<b>Уровень значимости</b>					
<b>Число степеней свободы</b>	<b>20 % (двустороннее) 10 % (одностороннее)</b>	<b>10 % (двустороннее) 5 % (одностороннее)</b>	<b>5 % (двустороннее) 2,5 % (одностороннее)</b>	<b>2 % (двустороннее) 1 % (одностороннее)</b>	<b>1 % (двустороннее) 0,5 % (одностороннее)</b>
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,36	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,41	1,89	2,36	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,11
12	1,36	1,78	2,18	2,68	3,05
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,14	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,72	2,09	2,53	2,85
21	1,32	1,72	2,08	2,52	2,83
22	1,32	1,72	2,07	2,51	2,82
23	1,32	1,71	2,07	2,50	2,81
24	1,32	1,71	2,06	2,49	2,80

## Окончание таблицы 2

Число степеней свободы	Уровень значимости				
	20 % (двустороннее) 10 % (одностороннее)	10 % (двустороннее) 5 % (одностороннее)	5 % (двустороннее) 2,5 % (одностороннее)	2 % (двустороннее) 1 % (одностороннее)	1 % (двустороннее) 0,5 % (одностороннее)
25	1,32	1,71	2,06	2,49	2,79
26	1,32	1,71	2,06	2,48	2,78
27	1,31	1,70	2,05	2,47	2,77
28	1,31	1,70	2,05	2,47	2,76
29	1,31	1,70	2,05	2,46	2,76
30	1,31	1,70	2,04	2,46	2,75
60	1,30	1,67	2,00	2,39	2,66
90	1,29	1,66	1,99	2,37	2,63
120	1,29	1,66	1,98	2,36	2,62
$\infty$	1,28	1,64	1,96	2,33	2,58

Примечание. Критические значения приведены для двусторонней ( $\neq$ ) и односторонней ( $>$ ) альтернативных гипотез. Критические значения для одностороннего ( $<$ ) теста равны критическому значению для одностороннего ( $>$ ) теста из таблицы, взято му со знаком минус. Например, 2,13 – это двустороннее критическое значение на уровне значимости 5 %, полученное с использованием  $t$ -распределения Стьюдента с 15 степенями свободы.

Таблица 3

Критические значения для  $\chi^2$ -распределения

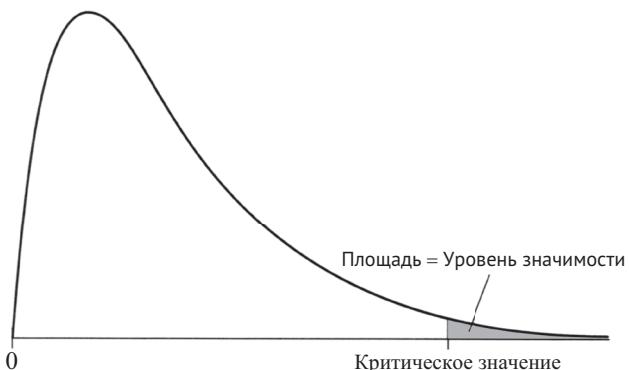
Число степеней свободы	Уровень значимости		
	10 %	5 %	1 %
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,81
19	27,20	30,14	36,19
20	28,41	31,41	37,57

Окончание таблицы 3

Число степеней свободы	Уровень значимости		
	10 %	5 %	1 %
21	29,62	32,67	38,93
22	30,81	33,92	40,29
23	32,01	35,17	41,64
24	33,20	36,41	42,98
25	34,38	37,65	44,31
26	35,56	38,89	45,64
27	36,74	40,11	46,96
28	37,92	41,34	48,28
29	39,09	42,56	49,59
30	40,26	43,77	50,89

Примечание. Эта таблица содержит 90, 95 и 99-е квартили распределения  $\chi^2$ . Что позволяет получить критические значения для проверки гипотез на уровнях значимости 10, 5 и 1%.

Таблица 4

**Критические значения для  $F_{m, \infty}$ -распределения**


Число степеней свободы	Уровень значимости		
	10 %	5 %	1 %
1	2,71	3,84	6,63
2	2,30	3,00	4,61
3	2,08	2,60	3,78
4	1,94	2,37	3,32
5	1,85	2,21	3,02
6	1,77	2,10	2,80
7	1,72	2,01	2,64
8	1,67	1,94	2,51
9	1,63	1,88	2,41
10	1,60	1,83	2,32
11	1,57	1,79	2,25

## Окончание таблицы 4

Уровень значимости			
Число степеней свободы	10 %	5 %	1 %
12	1,55	1,75	2,18
13	1,52	1,72	2,13
14	1,50	1,69	2,08
15	1,49	1,67	2,04
16	1,47	1,64	2,00
17	1,46	1,62	1,97
18	1,44	1,60	1,93
19	1,43	1,59	1,90
20	1,42	1,57	1,88
21	1,41	1,56	1,85
22	1,40	1,54	1,83
23	1,39	1,53	1,81
24	1,38	1,52	1,79
25	1,38	1,51	1,77
26	1,37	1,50	1,76
27	1,36	1,49	1,74
28	1,35	1,48	1,72
29	1,35	1,47	1,71
30	1,34	1,46	1,70

Примечание. Эта таблица содержит 90-е, 95-е и 99-е перцентили распределения  $F_{m, \infty}$ -distribution. Они служат критическими значениями для тестов с уровнем значимости в 10, 5 и 1 %.

Таблица 5А

Критические значения для  $F_{n_1, n_2}$ -распределения – 10 %-й уровень значимости

Число степеней свободы знаменателя ( $n_2$ )	Число степеней свободы числителя ( $n_1$ )									
	1	2	3	4	5	6	7	8	9	10
1	39,86	49,50	53,59	55,83	57,24	58,20	58,90	59,44	59,86	60,20
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25

Окончание таблицы 5A

Число степеней свободы числителя ( $n_1$ )										
Число степеней свободы знаменателя ( $n_2$ )	1	2	3	4	5	6	7	8	9	10
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71
90	2,76	2,36	2,15	2,01	1,91	1,84	1,78	1,74	1,70	1,67
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65
$\infty$	<b>2,71</b>	<b>2,30</b>	<b>2,08</b>	<b>1,94</b>	<b>1,85</b>	<b>1,77</b>	<b>1,72</b>	<b>1,67</b>	<b>1,63</b>	<b>1,60</b>

Примечание. Таблица содержит 90-е квантили – распределения  $F_{n_1, n_2}$ , которые позволяют получить критические значения для проверки гипотез на уровне значимости 10 %.

Таблица 5B

#### Критические значения для $F_{n_1, n_2}$ -распределения – 5 %-й уровень значимости

Число степеней свободы числителя ( $n_1$ )										
Число степеней свободы знаменателя ( $n_2$ )	1	2	3	4	5	6	7	8	9	10
1	161,40	199,50	215,70	224,60	230,20	234,00	236,80	238,90	240,50	241,90
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64

## Окончание таблицы 5В

Число степеней свободы числителя ( $n_1$ )										
Число степеней свободы знаменателя ( $n_2$ )	1	2	3	4	5	6	7	8	9	10
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
$\infty$	<b>3,84</b>	<b>3,00</b>	<b>2,60</b>	<b>2,37</b>	<b>2,21</b>	<b>2,10</b>	<b>2,01</b>	<b>1,94</b>	<b>1,88</b>	<b>1,83</b>

Примечание. Таблица содержит 95-е квантили – распределения  $F_{n_1, n_2}$ , которые позволяют получить критические значения для проверки гипотез на уровне значимости 5 %.

Таблица 5С

Критические значения для  $F_{n_1, n_2}$ -распределения – 1 %-й уровень значимости

Число степеней свободы числителя ( $n_1$ )										
Число степеней свободы знаменателя ( $n_2$ )	1	2	3	4	5	6	7	8	9	10
1	4052,00	4999,00	5403,00	5624,00	5763,00	5859,00	5928,00	5981,00	6022,00	6055,00
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40

Окончание таблицы 5С

Число степеней свободы знаменателя ( $n_2$ )	Число степеней свободы числителя ( $n_1$ )									
	1	2	3	4	5	6	7	8	9	10
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
$\infty$	<b>6,63</b>	<b>4,61</b>	<b>3,78</b>	<b>3,32</b>	<b>3,02</b>	<b>2,80</b>	<b>2,64</b>	<b>2,51</b>	<b>2,41</b>	<b>2,32</b>

Примечание. Таблица содержит 99-е квантили – распределения, которые позволяют получить критические значения для проверки гипотез на уровне значимости 1 %.

## Список литературы

- Adda, Jerome, and Francesca Cornaglia.* 2006. "Taxes, Cigarette Consumption, and Smoking Intensity". *American Economic Review* 96 (4): 1013–1028.
- Aggarwal, Rajesh K., and Philippe Jorion.* 2010. "The Performance of Emerging Hedge Funds and Managers". *Journal of Financial Economics* 96: 238–256.
- Anderson, Theodore W., and Herman Rubin.* 1950. "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations". *Annals of Mathematical Statistics* 21: 570–582.
- Andrews, Donald W.K.* 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation". *Econometrica* 59 (3): 817–858.
- Andrews, Donald W.K.* 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point". *Econometrica* 61 (4): 821–856.
- Andrews, Donald W.K.* 2003. "Tests For Parameter Instability and Structural Change with Unknown Change Point: A Corrigendum". *Econometrica* 71: 395–397.
- Angrist, Joshua.* 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records". *American Economic Review* 80 (3): 313–336.
- Angrist, Joshua, and William Evans.* 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size". *American Economic Review* 88 (3): 450–477.
- Angrist, Joshua, Kathryn Graddy, and Guido Imbens.* 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish". *Review of Economic Studies* 67 (232): 499–527.
- Angrist, Joshua, and Alan Krueger.* 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (4): 979–1014.
- Angrist, Joshua, and Alan B. Krueger.* 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments". *Journal of Economic Perspectives* 15 (4), Fall: 69–85.
- Arellano, Manuel.* 2003. *Panel Data Econometrics*. Oxford: Oxford University Press.
- Ayres, Ian, and John Donohue.* 2003. "Shooting Down the 'More Guns Less Crime' Hypothesis". *Stanford Law Review* 55: 1193–1312.
- Barendregt, Jan J.* 1997. "The Health Care Costs of Smoking". *New England Journal of Medicine* 337 (15): 1052–1057.
- Beck, Thorsten, Ross Levine, and Norman Loayza.* 2000. "Finance and the Sources of Growth". *Journal of Financial Economics* 58: 261–300.
- Benartzi, Shlomo, and Richard H. Thaler.* 2007. "Heuristics and Biases in Retirement Savings Behavior". *Journal of Economic Perspectives* 21 (3): 81–104.
- Bergstrom, Theodore A.* 2001. "Free Labor for Costly Journals?" *Journal of Economic Perspectives* 15 (4), Fall: 183–198.

- Bertrand, Marianne, and Kevin Hallock.* 2001. "The Gender Gap in Top Corporate Jobs". *Industrial and Labor Relations Review* 55 (1): 3–21.
- Bertrand, Marianne, and Sendhil Mullainathan.* 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". *American Economic Review* 94 (4): 991–1013.
- Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian.* 2008. "The Importance of Default Options for Retirement Saving Outcomes: Evidence from the United States", in *Lessons from Pension Reform in the Americas*, edited by Stephen J. Kay and Tapen Sinha. Oxford: Oxford University Press, 59–87.
- Bollersev, Tim.* 1986. "Generalized Autoregressive Conditional Heteroskedasticity". *Journal of Econometrics* 31 (3): 307–327.
- Bound, John, David A. Jaeger, and Regina M. Baker.* 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instrument and the Endogenous Explanatory Variable Is Weak". *Journal of the American Statistical Association* 90 (430): 443–450.
- Campbell, John Y.* 2003. "Consumption-Based Asset Pricing". Ch. 13 in *Handbook of the Economics of Finance*, edited by Milton Harris and Rene Stulz. Amsterdam: Elsevier.
- Campbell, John Y., and Motohiro Yogo.* 2005. "Efficient Tests of Stock Return Predictability". *Journal of Financial Economics* 81 (1): 27–60.
- Card, David.* 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market". *Industrial and Labor Relations Review* 43 (2): 245–257.
- Card, David.* 1999. "The Causal Effect of Education on Earnings". Ch. 30 in *The Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier.
- Card, David, and Alan B. Krueger.* 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry". *American Economic Review* 84 (4): 772–793.
- Carhart, Mark M.* 1997. "On Persistence in Mutual Fund Performance". *Journal of Finance* 52 (1): 57–82.
- Carpenter, Christopher, and Philip J. Cook.* 2008. "Cigarette Taxes and Youth Smoking: New Evidence from National, State, and Local Youth Risk Behavior Surveys", *Journal of Health Economics* 27: 287–299.
- Chaloupka, Frank J., Michael Grossman, and Henry Saffer.* 2002. "The Effect of Price on Alcohol Consumption and Alcohol-Related Problems". *Alcohol Research & Health* 26: 22–34.
- Chaloupka, Frank J., and Kenneth E. Warner.* 2000. "The Economics of Smoking". Ch. 29 in *The Handbook of Health Economics*, edited by Joseph P. Newhouse and Anthony J. Cuyler. New York: North Holland.
- Chow, Gregory.* 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions". *Econometrica* 28 (3): 591–605.
- Clements, Michael P.* 2004. "Evaluating the Bank of England Density Forecasts of Inflation". *Economic Journal* 114: 844–866.
- Cochrane, D., and Guy Orcutt.* 1949. "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms". *Journal of the American Statistical Association* 44 (245): 32–61.
- Cohen, Alma, and Liran Einav.* 2003. "The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities". *The Review of Economics and Statistics* 85 (4): 828–843.

- Cook, Philip J., and Michael J. Moore.* 2000. "Alcohol". Ch. 30 in *The Handbook of Health Economics*, edited by Joseph P. Newhouse and Anthony J. Cuyler. New York: North Holland.
- Cooper, Harris, and Larry V. Hedges.* 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.
- Dang, Jennifer N.* 2008. "Statistical Analysis of Alcohol- Related Driving Trends, 1982–2005". Technical Report DOT HS 810 942. Washington, D.C.: U.S. National Highway Traffic Safety Administration.
- Davidson, James E.H., David F. Hendry, Frank Srba, and Stephen Yeo.* 1978. "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom". *Economic Journal* 88: 661–692.
- Dickey, David A., and Wayne A. Fuller.* 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". *Journal of the American Statistical Association* 74 (366): 427–431.
- Diebold, Francis X.* 2007. *Elements of Forecasting* (fourth edition). Cincinnati: South-Western.
- Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and J. Douglas Willms.* 2001a. "Class Size and Student Achievement". *Psychological Science in the Public Interest* 2 (1): 1–30.
- Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and J. Douglas Willms.* 2001b. "Does Class Size Matter?" *Scientific American* 285 (5): 80–85.
- Eicker, F.* 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1,59–82. Berkeley: University of California Press.
- Elliott, Graham, Thomas J. Rothenberg, and James H. Stock.* 1996. "Efficient Tests for an Autoregressive Unit Root". *Econometrica* 64 (4): 813–836.
- Enders, Walter.* 1995. *Applied Econometric Time Series*. New York: Wiley.
- Engle, Robert F.* 1982. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation". *Econometrica* 50 (4): 987–1007.
- Engle, Robert F., and Clive W.J. Granger.* 1987. "Cointegration and Error Correction: Representation, Estimation and Testing". *Econometrica* 55 (2): 251–276.
- Evans, William, Matthew Farrelly, and Edward Montgomery.* 1999. "Do Workplace Smoking Bans Reduce Smoking?" *American Economic Review* 89 (4): 728–747.
- Foster, Donald.* 1996. "Primary Culprit: An Analysis of a Novel of Politics". *New York Magazine* 29 (8), February 26.
- Fuller, Wayne A.* 1976. *Introduction to Statistical Time Series*. New York: Wiley.
- Garvey, Gerald T., and Gordon Hanka.* 1999. "Capital Structure and Corporate Control: The Effect of Antitakeover Statutes on Firm Leverage". *Journal of Finance* 54 (2): 519–546.
- Gillespie, Richard.* 1991. *Manufacturing Knowledge: A History of the Hawthorne Experiments*. New York: Cambridge University Press.
- Goering, John, and Ron Wienk, eds.* 1996. *Mortgage Lending, Racial Discrimination, and Federal Policy*. Washington, DC: Urban Institute Press.
- Goyal, Amit, and Ivo Welch.* 2003. "Predicting the Equity Premium with Dividend Ratios". *Management Science* 49 (5): 639–654.
- Granger, Clive W.J.* 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods". *Econometrica* 37 (3): 424–438.

- Granger, Clive W.J., and A.A. Weiss.* 1983. "Time Series Analysis of Error-Correction Models". Pp. 255–278 in *Studies in Econometrics: Time Series and Multivariate Statistics*, edited by S. Karlin, T. Amemiya, and L.A. Goodman. New York: Academic Press.
- Greene, William H.* 2000. *Econometric Analysis* (fourth edition). Upper Saddle River, NJ: Prentice Hall.
- Gruber, Jonathan.* 2001. "Tobacco at the Crossroads: The Past and Future of Smoking Regulation in the United States". *Journal of Economic Perspectives* 15 (2): 193–212.
- Haldrup, Niels, and Michael Jansson.* 2006. "Improving Size and Power in Unit Root Testing". Pp. 252–277 in *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, edited by Terrence Mills and Kerry Patterson. Basingstoke U.K.: Palgrave MacMillan.
- Hamermesh, Daniel, and Amy Parker.* 2005. "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity". *Economics of Education Review* 24 (4): 369–376.
- Hamilton, James D.* 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Hansen, Bruce.* 1992. "Efficient Estimation and Testing of Cointegrating Vectors in the Presence of Deterministic Trends". *Journal of Econometrics* 53 (1–3): 86–121.
- Hansen, Bruce.* 2001. "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity". *Journal of Economic Perspectives* 15 (4), Fall: 117–128.
- Hanushek, Eric.* 1999a. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects". *Educational Evaluation and Policy Analysis* 21: 143–164.
- Hanushek, Eric.* 1999b. "The Evidence on Class Size". Ch. 7 in *Earning and Learning: How Schools Matter*, edited by S. Mayer and P. Peterson. Washington, DC: Brookings Institution Press.
- Hayashi, Fumio.* 2000. *Econometrics*. Princeton, NJ: Princeton University Press.
- Heckman, James J.* 1974. "Shadow Prices, Market Wages, and Labor Supply", *Econometrica*, 42: 679–694.
- Heckman, James J.* 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture". *Journal of Political Economy* 109 (4): 673–748.
- Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith.* 1999. "The Economics and Econometrics of Active Labor Market Programs". Ch. 31 in *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier.
- Hedges, Larry V., and Ingram Olkin.* 1985. *Statistical Methods for Meta-analysis*. San Diego: Academic Press.
- Hetland, Lois.* 2000. "Listening to Music Enhances Spatial-Temporal Reasoning: Evidence for the 'Mozart Effect'." *Journal of Aesthetic Education* 34 (3–4): 179–238.
- Hoxby, Caroline M.* 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation". *Quarterly Journal of Economics* 115 (4): 1239–1285.
- Huber, P.J.* 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–233. Berkeley: University of California Press.
- Imbens, Guido W., and Joshua D. Angrist.* 1994. "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62: 467–476.

- Johansen, Søren. 1988. "Statistical Analysis of Cointegrating Vectors". *Journal of Economic Dynamics and Control* 12: 231–254.
- Jones, Stephen R.G. 1992. "Was There a Hawthorne Effect?" *American Journal of Sociology* 98 (3): 451–468.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn", *The Review of Economics and Statistics* 91: 437–456.
- Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions". *Quarterly Journal of Economics* 14 (2): 497–562.
- Ladd, Helen. 1998. "Evidence on Discrimination in Mortgage Lending". *Journal of Economic Perspectives* 12 (2), Spring: 41–62.
- Levitt, Steven D. 1996. "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation". *Quarterly Journal of Economics* 111 (2): 319–351.
- Levitt, Steven D., and Jack Porter. 2001. "How Dangerous Are Drinking Drivers?" *Journal of Political Economy* 109 (6): 1198–1237.
- List, John. 2003. "Does Market Experience Eliminate Market Anomalies". *Quarterly Journal of Economics* 118 (1): 41–71.
- Maddala, G. S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
- Maddala, G. S., and In-Moo Kim. 1998. Unit Roots, Cointegration, and Structural Change. Cambridge: Cambridge University Press.
- Madrian, Brigitte C.\* and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior". *Quarterly Journal of Economics* 116 (4): 1149–1187.
- Malkiel, Burton G. 2003. A Random Walk Down Wall Street. New York: W.W. Norton.
- Manning, Willard G., et al. 1989. "The Taxes of Sin: Do Smokers and Drinkers Pay Their Way?" *Journal of the American Medical Association* 261 (11): 1604–1609.
- Matsudaira, Jordan D. 2008. "Mandatory Summer School and Student Achievement". *Journal of Econometrics* 142: 829–850.
- McClellan, Mark, Barbara J. McNeil, and Joseph P. Newhouse. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association* 272 (11): 859–866.
- Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics". *Journal of Business and Economic Statistics* 13 (2): 151–161.
- Meyer, Bruce D., W. Kip Viscusi, and David L. Durbin. 1995. "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment". *American Economic Review* 85 (3): 322–340.
- Moreira, M. J. 2003. "A Conditional Likelihood Ratio Test for Structural Models". *Econometrica* 71: 1027–1048.
- Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades". *The Future of Children: Critical Issues for Children and Youths* 5 (2), Summer/Fall: 113–127.
- Mosteller, Frederick, Richard Light, and Jason Sachs. 1996. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size". *Harvard Educational Review* 66 (4), Winter: 631–676.

- Mosteller, Frederick, and David L. Wallace.* 1963. "Inference in an Authorship Problem". Journal of the American Statistical Association 58: 275–309.
- Munnell, Alicia H., Geoffrey M.B. Tootell, Lynne E. Browne, and James McEneaney.* 1996. "Mortgage Lending in Boston: Interpreting HMDA Data". American Economic Review 86 (1): 25–53.
- Neumark, David, and William Wascher.* 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment". American Economic Review 90 (5): 1362–1396.
- Newey, Whitney, and Kenneth West.* 1987. "A Simple Positive Semi-definite, Heteroskedastic and Autocorrelation Consistent Covariance Matrix". Econometrica 55 (3): 703–708.
- Newhouse, Joseph P. et al.* 1993. Free for All? Lessons from the Rand Health Insurance Experiment. Cambridge, MA: Harvard University Press.
- Perry, Craig, and Harvey S. Rosen.* 2004. "The Self-Employed Are Less Likely Than Wage-Earners to Have Health Insurance. So What?" Pp. 23–58 in Public Policy and the Economics of Entrepreneurship, edited by Douglas Holtz-Eakin and Harvey S. Rosen. Boston: MIT Press.
- Phillips, Peter C.B., and Sam Ouliaris.* 1990. "Asymptotic Properties of Residual Based Tests for Cointegration". Econometrica 58 (1): 165–194.
- Porter, Robert.* 1983. "A Study of Cartel Stability: Tbl Joint Executive Committee, 1880–1886". Bell Journal of Economics 14 (2): 301–314.
- Quandt, Richard.* 1960. "Tests of the Hypothesis Tha: a Linear Regression Systemobeys Two Separate Regimes". Journal of the American Statistical Association 55 (290): 324–330.
- Rauscher, Frances, Gordon L. Shaw, and Katherine 4 Ky.* 1993. "Music and Spatial Task Performance". Nature 365 (6447): 611.
- Roll, Richard.* 1984. "Orange Juice and Weather."\* American Economic Review 74 (5): 861–880.
- Rosenzweig, Mark R., and Kenneth I. Wolpin.* 2000 "Natural 'Natural Experiments' in Economics". Journal of Economic Literature 38 (4): 827–874.
- Rouse, Cecilia.* 1995. "Democratization or Diversions? The Effect of Community Colleges on Educational Attainment". Journal of Business and Economic Statistics 12 (2): 217–224.
- Ruhm, Christopher J.* 1996. "Alcohol Policies and Highway Vehicle Fatalities". Journal of Health Economics 15 (4): 435–454.
- Ruud, Paul.* 2000. An Introduction to Classical Econometric Theory. New York: Oxford University Press.
- Shadish, William R., Thomas D. Cook, and Donald X Campbell.* 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin.
- Shiller, Robert J.* 2005. Irrational Exuberance (second edition). Princeton, NJ: Princeton University Press.
- Sims, Christopher A.* 1980. "Macroeconomics and Reality". Econometrica 48 (1): 1–48.
- Stock, James H.* 1994. "Unit Roots, Structural Breaks, and Trends". Ch. 46 in Handbook of Econometrics, volume IV, edited by Robert Engle and Daniel McFadden. Amsterdam: Elsevier.

- Stock, James H., and Francesco Trebbi.* 2003. "Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17: 177–194.
- Stock, James H., and Mark W. Watson.* 1988. "Variable Trends in Economic Time Series". *Journal of Economic Perspectives* 2 (3): 147–174.
- Stock, James H., and Mark W. Watson.* 1993. "A Simple Estimator of Cointegrating Vectors in Higher-Order Integrated Systems". *Econometrica* 61 (4): 783–820.
- Stock, James H., and Mark W. Watson.* 2001. "Vector Autoregressions". *Journal of Economic Perspectives* 15 (4), Fall: 101–115.
- Stock, James H., and Motohiro Yogo.* 2005. "Testing for Weak Instruments in Linear IV Regression". Ch. 5 in *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, edited by Donald W.K. Andrews and James H. Stock. Cambridge: Cambridge University Press.
- Tobin, James.* 1958. "Estimation of Relationships for Limited Dependent Variables". *Econometrica* 26 (1): 24–36.
- Wagenaar, Alexander C., Matthew J. Salois, and Kelli A. Komro.* 2009. "Effects of Beverage Alcohol Price and Tax Levels on Drinking: A Meta-Analysis of 1003 Estimates from 112 Studies". *Addiction* 104: 179–190.
- Watson, Mark W.* 1994. "Vector Autoregressions and Cointegration". Ch. 47 in *Handbook of Econometrics*, volume IV, edited by Robert Engle and Daniel McFadden. Amsterdam: Elsevier.
- White, Halbert.* 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". *Econometrica*, 48, 827–838.
- Winner, Ellen, and Monica Cooper.* 2000. "Mute Those Claims: No Evidence (Yet) for a Causal Link Between Arts Study and Academic Achievement". *Journal of Aesthetic Education* 34 (3–4): 11–76.
- Wooldridge, Jeffrey.* 2002. *Economic Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- Wright, Philip G.* 1915. "Moore's Economic Cycles". *Quarterly Journal of Economics* 29: 631–641.
- Wright, Philip G.* 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.
- Young, Douglas J., and Agnieszka Bielinska-Kwapisz.* 2006. "Alcohol Prices, Consumption, and Traffic Fatalities." *Southern Economic Journal* 72: 690–703.

# Глоссарий

**95 %-я доверительная область (95 % confidence set):** Доверительная область с 95 %-м уровнем доверия. См. *Доверительный интервал*.

**ADL ( $p, q$ ):** См. *Авторегрессионная модель с распределенными лагами (ADL)*.

**AIC:** См. *Информационный критерий*.

**ARCH:** См. *Авторегрессионная модель с условной гетероскедастичностью*.

**AR ( $p$ ):** См. *Авторегрессия*.

**BIC:** См. *Информационный критерий*.

**BLUE:** См. *Наилучшая линейная несмещенная оценка (BLUE)*.

**$F_{m,n}$ -распределение (distribution):** Распределение отношения независимых случайных величин, где числитель является случайной величиной хи-квадрат с  $m$  степенями свободы, деленной на  $m$ , а знаменатель – случайной величиной хи-квадрат с  $n$  степенями свободы, деленной на  $n$ .

**$F_{m,\infty}$ -распределение distribution):** Распределение случайной величины с распределением хи-квадрат с  $m$  степенями свободы, деленной на  $m$ .

**F-статистика (F-statistic):** Статистика, используемая для проверки совместной гипотезы, рассматривающей более одного коэффициента регрессии.

**F-статистика, рассчитанная при условии гомоскедастичности ошибок регрессии (Homoskedasticity-only F-statistic):** F-статистика, которая может быть использована только когда остаточный член регрессии гомоскедастичен.

**GARCH:** См. *Обобщенный метод наименьших квадратов*.

**GMM:** См. *Обобщенный метод моментов*.

**HAC-стандартные ошибки (HAC standard errors):** См. *Устойчивые к гетероскедастичности и автокорреляции (HAC) стандартные ошибки*.

**I(0), I(1), and I(2):** См. *Порядок интегрированности*.

**J-статистика (J-statistic):** Статистика, используемая для тестирования сверхидентифицирующих ограничений в модели регрессии с инструментальными переменными.

**p-значение (вероятность значимости) (p-value (significance probability)):** Вероятность получения статистики, при которой нулевая гипотеза отвергается, в предположении, что она на самом деле верна. Также называемое предельным значением вероятности, *p*-значение является наименьшим уровнем значимости, при котором нулевая гипотеза может быть отвергнута.

**$R^2$ :** В регрессии – доля выборочной дисперсии зависимой переменной, которая объясняется регрессорами.

**$\bar{R}^2$ :** См. *Скорректированный R<sup>2</sup>*.

**TSLS:** См. *Двухшаговый метод наименьших квадратов*.

**t-отношение (t-ratio):** См. *t-статистика*.

***t*-распределение (*t-distribution*):** См. *t*-распределение Стьюдента.

***t*-распределение Стьюдента (Student *t-distribution*):** *t*-распределение Стьюдента с  $m$  степенями свободы является распределением отношения стандартной нормальной случайной величины к квадратному корню независимо распределенной случайной величины хи-квадрат с  $m$  степенями свободы, деленному на  $m$ . Когда  $m$  становится большим, *t*-распределение Стьюдента сходится к стандартному нормальному распределению.

***t*-статистика (*t-statistic*):** Статистика, используемая для тестирования гипотез. См. вставку «Основные понятия 5.1».

**VAR:** См. *Векторная авторегрессия*.

**Автоковариация (Autocovariance):** Ковариация между временным рядом и его запаздывающим значением.  $j$ -я автоковариация  $Y$  является ковариацией между  $Y_t$  и  $Y_{t-j}$ .

**Автокорреляция (Autocorrelation):** Корреляция между временным рядом и его запаздывающим значением.  $j$ -я автокорреляция  $Y$  является корреляцией между  $Y_t$  и  $Y_{t-j}$ .

**Авторегрессионная модель с распределенными лагами (Autoregressive distributed lag (ADL) model):** Модель линейной регрессии, в которой временной ряд  $Y_t$  выражается как функция от запаздываний  $Y_t$  и некоторой переменной  $X_t$ . Модель обозначается ADL ( $p, q$ ), где  $p$  обозначает число запаздываний  $Y_t$  и  $q$  обозначает число запаздываний  $X_t$ .

**Авторегрессионная модель с условной гетероскедастичностью (Autoregressive conditional heteroskedasticity (ARCH)):** Модель временного ряда, позволяющая моделировать условную гетероскедастичность.

**Авторегрессия (Autoregression):** Модель линейной регрессии, которая связывает временной ряд с его прошлыми (т.е. запаздывающими или лагированными) значениями. Авторегрессия с  $p$  запаздывающими значениями, включенными в регрессию, обозначается AR ( $p$ ).

**Альтернативная гипотеза (Alternative hypothesis):** Гипотеза, которая предполагается верной, если нулевая гипотеза неверна. Альтернативная гипотеза обычно обозначается  $H_1$ .

**Асимметрии (Skewness):** Мера асимметрии функции плотности вероятности.

**Асимптотическое распределение (Asymptotic distribution):** Приближение выборочного распределения случайной величины, вычисленное с использованием большой выборки. Например, асимптотическое распределение выборочного среднего является нормальным.

**Асимптотическое нормальное распределение (Asymptotic normal distribution):** Нормальное распределение, которое приближает выборочное распределение вычисленной статистики, используя большую выборку.

**Базовая спецификация (Base specification):** Базовая или исходная спецификация регрессии, включающая некий набор регрессоров, отобранных с использованием комбинации экспертного мнения, экономической теории и знаний о том, как данные были собраны.

**Байесовский информационный критерий (Bayes information criterion (BIC)):** См. Информационный критерий.

**Безусловное распределение вероятностей (Marginal probability distribution):** Другое название для функции распределения случайной величины  $Y$ , которое характеризует распределение  $Y$  отдельно (безусловное распределение) от совместного распределения  $Y$  и других случайных величин.

**Бернуlliевская случайная величина (Bernoulli random variable):** Случайная величина, принимающая значения либо 0, либо 1.

**Бинарная переменная (Binary variable):** Переменная, которая принимает значения либо 0, либо 1. Бинарная переменная используется для обозначения бинарного исхода. Например,  $X$  является бинарной (или индикаторной, или дамми, или фиктивной) переменной, если она обозначает пол человека,  $X = 1$ , если индивид является женщиной, и  $X = 0$ , если индивид является мужчиной.

**Векторная авторегрессия (Vector autoregression):** Модель  $k$  временных рядов, содержащая  $k$  уравнений – по одному для каждой переменной, в которой во всех уравнениях регрессорами являются запаздывающие значения всех переменных.

**Вероятность (Probability):** Доля, с которой элементарное событие (или событие) принимает то или иное значение в долгосрочной перспективе.

**Взвешенный метод наименьших квадратов (ВМНК) (Weighted least squares (WLS)): Альтернатива методу наименьших квадратов, которая может быть использована, когда ошибка регрессии является гетероскедастичной и форма гетероскедастичности известна или может быть оценена.**

**Включенные экзогенные переменные (Included exogenous variables):** Регрессоры, которые не коррелированы с остаточным членом (обычно в контексте регрессии с инструментальными переменными).

**Включенные эндогенные переменные (Included endogenous variables):** Регрессоры, которые коррелированы с остаточным членом (обычно в контексте регрессии с инструментальными переменными).

**Внешняя обоснованность (External validity):** Выводы и заключения, сделанные на основе статистического исследования, являются внешне обоснованными, если они могут быть обобщены с изучаемой генеральной совокупностью и заданных условий на другие генеральные совокупности и желаемые условия.

**Внутренняя обоснованность (Internal validity):** Если статистическая проверка о причинных эффектах обоснована для изучаемой генеральной совокупности.

**Временные ряды (Time series data):** Данные для одного и того же объекта в разные моменты времени.

**Временные фиксированные эффекты (Time fixed effects):** См. *Временные эффекты*.

**Временные эффекты (Time effects):** Бинарные переменные, характеризующие временной период в регрессии панельных данных.

**Выборочная ковариация (Sample covariance):** Оценка ковариации между двумя случайными величинами.

**Выборочная дисперсия (Sample variance):** Оценка дисперсии случайной величины.

**Выборочное распределение (Sampling distribution):** Распределение статистики во всех возможных выборках; распределение, получаемое повторяющейся оценкой статистик с использованием серий случайно отобранных выборок из одной и той же генеральной совокупности.

**Выборочное стандартное отклонение (Sample standard deviation):** Оценка стандартного отклонения случайной величины.

**Выброс (Outlier):** Исключительно большие или маленькие значения случайной величины.

**Генеральная совокупность (Population):** Группа изучаемых объектов – таких как люди, компании или школьные округа.

**Гетероскедастичность (Heteroskedasticity):** Ситуация, в которой условная относительно регрессоров дисперсия ошибки регрессии  $u_i$  не является постоянной.

**Гомоскедастичность (Homoskedasticity):** Условная относительно регрессоров дисперсия ошибки регрессии  $u_i$  является константой.

**Дамми-переменная или фиктивная переменная (Dummy variable):** См. Бинарная переменная.

**Дата структурного сдвига (Break date):** Дата дискретного изменения в коэффициентах теоретической регрессии временных рядов.

**Двухмерное нормальное распределение (Bivariate normal distribution):** Обобщение нормального распределения, используемое для описания совместного распределения двух случайных величин.

**Двухсторонняя альтернативная гипотеза (Two-sided alternative hypothesis):** Альтернативная гипотеза, в которой оцениваемый параметр не равен значению, равенство которому тестируется в рамках нулевой гипотезы.

**Двухшаговый метод наименьших квадратов (2МНК) (Two stage least squares (TSLS):** Оценка метода инструментальных переменных, описанная во вставке «Основные понятия 12.2».

**Детерминированный тренд (Deterministic trend):** Устойчивое долгосрочное движение переменной во времени, которое может быть представлено как неслучайная функция времени.

**Диаграмма рассеяния (Scatterplot):** График  $n$  наблюдений  $X_i$  и  $Y_i$ , на которой каждое отношение представлено точкой  $(X_i, Y_i)$ .

**Динамический мультиплликатор (Dynamic multiplier):** Динамический мультиплликатор  $h$ -го периода характеризует влияние единичного изменения временного ряда  $X_t$  на  $X_{t+h}$ .

**Динамическое причинное влияние (Dynamic causal effect):** Влияние одной переменной на текущее и будущие значения другой переменной.

**Дискретная случайная величина (Discrete random variable):** Случайная величина, которая принимает дискретные значения.

**Дисперсия (Variance):** Математическое ожидание квадрата разности между случайной величиной и ее средним значением; дисперсия  $Y$  обозначается  $\sigma_y^2$ .

**Доверительный интервал (доверительная область) (Confidence interval (confidence set)): Интервал (или область), содержащий истинные значения теоретических параметров с заранее определенной вероятностью при вычислении в повторяющихся выборках.**

**Долгосрочный совокупный динамический мультипликатор (Long-run cumulative dynamic multiplier):** Накопленное долгосрочное влияние изменения  $X$  на временной ряд  $Y$ .

**Доступная или реализуемая ОМНК-оценка (Feasible GLS estimator):** Вариант оценки обобщенного метода наименьших квадратов (ОМНК), в котором используется оценка условной дисперсии ошибок регрессии и ковариации между ошибками регрессии для различных наблюдений.

**Доступный или реализуемый ВМНК (Feasible WLS):** Вариант оценки взвешенного метода наименьших квадратов (ВМНК), в котором используется оценка условной дисперсии ошибок регрессии.

**Единичный корень (Unit root):** Относится к авторегрессии с наибольшим корнем, равным единице.

**Естественный эксперимент (Natural experiment):** См. Квазиэксперимент.

**Зависимая переменная (Dependent variable):** Переменная, которую нужно объяснить в регрессии или при помощи любой другой статистической модели; эта переменная стоит в левой части регрессии.

**Закон больших чисел (Law of large numbers):** Согласно этому результату из теории вероятности, при выполнении общих предположений в больших выборках выборочное среднее будет с высокой вероятностью близко к среднему в генеральной совокупности.

**Закон повторного математического ожидания (Law of iterated expectations):** Результат из теории вероятности, говорящий о том, что математическое ожидание  $Y$  равно ожидаемому значению его условного математического ожидания относительно  $X$ , то есть  $E(Y) = E[E(Y|X)]$ .

**Запаздывания или лаги (Lags):** Значение временного ряда в предыдущий момент времени.  $j$ -е запаздывание  $Y_t$  равно  $Y_{t-j}$ .

**Импульсный эффект (Impact effect):** Одновременное или немедленное влияние единичного изменения временного ряда  $X_t$  на  $Y_t$ .

**Индикаторная переменная (Indicator variable):** См. Бинарная переменная.

**Инструмент (Instrument):** См. Инструментальная переменная.

**Инструментальная переменная (Instrumental variable):** Переменная, коррелированная с эндогенным регрессором (релевантность инструмента) и некоррелированная с ошибкой регрессии (экзогенность инструмента).

**Интегральная функция распределения (Cumulative distribution function (c.d.f.)): См. Интегральное распределение вероятностей.**

**Интегральное распределение вероятностей (Cumulative probability distribution):** Функция, показывающая вероятность того, что случайная величина меньше либо равна данному числу.

**Интервальный прогноз (Forecast interval):** Интервал, содержащий будущее значение временного ряда с определенной вероятностью.

**Информационный критерий (Information criterion):** Статистика, используемая для оценки числа запаздывающих значений переменной, которое необходимо включить в авторегрессию или модель с распределенными лагами. Примерами

являются информационный критерий Акаике (AIC) и байесовский информационный критерий (BIC).

**Информационный критерий Акаике (Akaike information criterion (AIC)):** См. Информационный критерий.

**Исследуемая (экспериментальная) группа (Treatment group):** Группа, объекты которой подвергаются экспериментальному воздействию.

**Истощение выборки (Attrition):** Потеря изучаемых объектов после их распределения между контрольной и экспериментальной группами.

**Квадратный корень из среднеквадратичной ошибки прогнозирования (Root mean squared forecast error (RMSFE)): Квадратный корень из среднеквадратической ошибки прогноза.**

**Квазиэксперимент (Quasi-experiment):** Условие, при котором случайность вводится при помощи изменений индивидуальных условий, что можно рассматривать, как если бы воздействие было оказано случайно.

**Кластеризованная волатильность (Volatility clustering):** Ситуация, в которой временной ряд имеет периоды (кластеры) с высокой дисперсией и периоды (кластеры) с дисперсией.

**Кластеризованные стандартные ошибки (Clustered standard errors):** Метод вычисления стандартных ошибок в панельной регрессии.

**Критическое значение (Critical value):** Значение тестовой статистики, для которого тест отвергает нулевую гипотезу на заданном уровне значимости.

**Ковариационная матрица (Covariance matrix):** Матрица, составленная из дисперсий и ковариаций вектора случайных величин.

**Ковариация (Covariance):** Величина, увеличение которой означает, что две случайные величины движутся или изменяются вместе. Ковариация между  $X$  и  $Y$  равна математическому ожиданию  $E[(X - \mu_x)(Y - \mu_y)]$  и обозначается  $\text{cov}(X, Y)$  или  $\sigma_{XY}^2$ .

**Кointеграция (Cointegration):** Ситуация, когда два или более временных ряда имеют общий стохастический тренд.

**Компонента взаимодействия (Interaction term):** Регрессор, который получается перемножением двух других регрессоров, например  $X_{11} \times X_{21}$ .

**Контрольная группа (Control group):** Группа, которая не получает воздействия в эксперименте.

**Контрольная переменная (Control variable):** Регрессор, который контролирует пропущенные факторы, определяющие зависимую переменную.

**Корреляция (Correlation):** Безмерная величина, увеличение которой означает, что две случайные величины движутся или изменяются вместе. Корреляция (или коэффициент корреляции) между  $X$  и  $Y$  равна  $\sigma_{XY}/\sigma_x\sigma_y$  и обозначается  $\text{corr}(X, Y)$ .

**Коэффициент выборочной корреляции (выборочная корреляция) (Sample correlation coefficient (sample correlation)): Оценка корреляции между двумя случайными величинами.**

**Коэффициент детерминации (Coefficient of determination):** См.  $R^2$ .

**Коэффициент корреляции (Correlation coefficient):** См. Корреляция.

**Линейная в логарифмах модель (Log-log model):** Функция нелинейной регрессии, в которой зависимой переменной является  $\ln(Y)$ , а объясняющей —  $\ln(X)$ .

**Линейная вероятностная модель (Linear probability model):** Модель регрессии, в которой  $Y$  является бинарной переменной.

**Линейно-логарифмическая модель (Linear-log model):** Функция нелинейной регрессии, в которой зависимой переменной является  $Y$ , а объясняющей —  $\ln(X)$ .

**Линия МНК-регрессии (OLS regression line):** Линия регрессии с теоретическими коэффициентами, замененными на оценки, полученные методом наименьших квадратов.

**Линия регрессии генеральной совокупности или теоретическая линия регрессии (Population regression line):** В парной регрессии теоретическая линия регрессии — это  $\beta_0 + \beta_1 X_{1i}$ , во множественной регрессии —  $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ .

**Ловушка фиктивных переменных (Dummy variable trap):** Проблема, которая возникает при включении в регрессию полного набора бинарных переменных вместе с константой, что приводит к совершенной мультиколлинеарности.

**Логарифм (Logarithm):** Математическая функция, определенная для положительного аргумента; ее наклон всегда положителен, но стремится к нулю. Натуральный логарифм является функцией, обратной к экспоненциальной функции, то есть  $X = \ln(e^X)$ .

**Логарифмически-линейная модель (Log-linear model):** Функция нелинейной регрессии, в которой зависимой переменной является  $\ln(Y)$ , а объясняющей —  $X$ .

**Логит-регрессия (Logit regression):** Модель нелинейной регрессии для бинарной зависимой переменной, в которой теоретическая функция регрессии моделируется с помощью кумулятивной функции логистического распределения.

**Локальный средний эффект в эксперименте (Local average treatment effect):** Взвешенное среднее эффектов воздействия, оцененное, например, при помощи двухшагового метода наименьших квадратов.

**Лонгитюдные данные (Longitudinal data):** См. Панельные данные.

**Математическое ожидание (Expected value):** Долгосрочное среднее значение случайной переменной, рассчитанное для большого числа испытаний. Это взвешенное по вероятности среднее значение всех возможных значений, которые может принимать случайная величина. Математическое ожидание случайной величины  $Y$  обозначается  $E(Y)$  и также называется ожиданием  $Y$ .

**Межобъектные данные (Cross-sectional data):** Данные, собранные для различных объектов в один период времени.

**МНК-остатки (OLS residual):** Разность между  $Y_i$  и линией МНК-регрессии, обозначаемая в учебнике как  $u_i$ .

**МНК-оценка (OLS estimator):** См. Оценка метода наименьших квадратов.

**Модель квадратичной регрессии (Quadratic regression model):** Функция нелинейной регрессии, включающая  $X$  и  $X^2$  в качестве регрессоров.

**Модель кубической регрессии (Cubic regression model):** Функция нелинейной регрессии, которая включает  $X$ ,  $X^2$  и  $X^3$  в качестве регрессоров.

**Модель множественной регрессии (Multiple regression model):** Расширение модели парной регрессии, в которой  $Y$  зависит от  $k$  регрессоров.

**Модель полиномиальной регрессии (Polynomial regression model):** Функция нелинейной регрессии, включающая  $X, X^2$  и  $X^r$  в регрессию, где  $r$  – целое число.

**Модель регрессии с временными и индивидуальными фиксированными эффектами (Time and entity fixed effects regression model):** Регрессия панельных данных, включающая и индивидуальные фиксированные эффекты, и временные фиксированные эффекты.

**Модель регрессии с фиксированными эффектами (Fixed effects regression model):** Регрессия панельных данных, включающая индивидуальные фиксированные эффекты.

**Модель с распределенными лагами (Distributed lag model):** Модель регрессии, в которой регрессорами являются текущие и запаздывающие значения  $X$ .

**Моменты распределения (Moments of a distribution):** Математическое ожидание случайной величины, возведенной в различные степени.  $r$ -м моментом случайной величины  $Y$  является  $E(Y^r)$ .

**Мощность (Power):** Вероятность, с которой тест правильно отвергает нулевую гипотезу, когда альтернативная гипотеза верна.

**Мультиколлинеарность (Multicollinearity):** См. Совершенная мультиколлинеарность и несовершенная мультиколлинеарность.

**Наблюдаемые данные (Observational data):** Данные о фактическом поведении вне эксперимента, полученные при помощи наблюдения или измерения.

**Наилучшая линейная несмешенная оценка (BLUE) (Best linear unbiased estimator):** Оценка, имеющая наименьшую дисперсию по сравнению с другими оценками, являющимися линейными функциями от выборочных значений  $Y$ . При выполнении условий Гаусса–Маркова оценка метода наименьших квадратов является наилучшей линейной несмешенной оценкой коэффициентов регрессии условной относительно значений регрессоров.

**Натуральный логарифм (Natural logarithm):** См. Логарифм.

**Недоопределенность (Underidentification):** Ситуация, когда число инструментальных переменных меньше, чем число эндогенных переменных.

**Независимость (Independence):** Когда знание значения одной случайной величины не дает информации о значении другой случайной величине. Две случайные величины независимы, если их совместное распределение является произведением их безусловных распределений.

**Независимость условного среднего (Conditional mean independence):** Условное математическое ожидание ошибки регрессии  $\mu$ , полученное при данных регрессорах, зависит от каких-либо переменных, за исключением всех этих регрессоров.

**Независимо и одинаково распределенные (Independently and identically distributed (i.i.d.)):** Когда две или более независимые случайные величины имеют одинаковое распределение.

**Некоррелированные (Uncorrelated):** Две случайные величины являются некоррелированными, если их корреляция равна нулю.

**Нелинейный метод наименьших квадратов (Nonlinear least squares):** Аналог метода наименьших квадратов, применяемый когда функция регрессии является нелинейной по параметрам.

**Неправильно специфицированная функциональная форма (Functional form misspecification):** Когда форма оцененной функции регрессии не соответствует форме теоретической функции регрессии, например, когда использована линейная спецификация, но истинной является квадратичная теоретическая функция регрессии.

**Непрерывная случайная величина (Continuous random variable):** Случайная величина, которая может принимать континуум значений.

**Несбалансированная панель (Unbalanced panel):** Множество панельных данных, в которых пропущены некоторые наблюдения.

**Несмешенная оценка (Unbiased estimator):** Оценка, смещение которой равно нулю.

**Несовершенная мультиколлинеарность (Imperfect multicollinearity):** Когда два или более регрессоров сильно коррелированы.

**Нестационарность (Nonstationary):** Ситуация, когда совместное распределение временного ряда и его запаздываний изменяется во времени.

**Номер наблюдения (Observation number):** Уникальный номер, присваиваемый каждому объекту в базе данных.

**Нормальное распределение (Normal distribution):** Широко используемое колоколообразное распределение непрерывной случайной величины.

**Нулевая гипотеза (Null hypothesis):** Гипотеза, которая тестируется при проверке гипотез и часто обозначается  $H_0$ .

**Область принятия (Acceptance region):** Множество значений тестовой статистики, для которых нулевая гипотеза не отвергается.

**Область отвержения (Rejection region):** Множество значений тестовой статистики, для которых отвергается нулевая гипотеза.

**Обобщенная авторегрессионная модель с условной гетероскедастичностью (GARCH) (Generalized autoregressive conditional heteroskedasticity):** Модель временных рядов для условной гетероскедастичности.

**Обобщенный метод моментов (Generalized method of moments):** Метод оценки параметров подгонкой выборочных моментов к теоретическим моментам, которые являются функциями от неизвестных параметров. Оценки инструментальных переменных являются важным частным случаем.

**Обобщенный метод наименьших квадратов (ОМНК) (Generalized least squares (GLS)): Обобщение метода наименьших квадратов, которое можно использовать, когда ошибки регрессии имеют известную форму гетероскедастичности (в этом случае ОМНК называют взвешенным методом наименьших квадратов, ВМНК) или известную форму серийной корреляции.**

**Общий тренд (Common trend):** Тренд, который содержится в двух или более временных рядах.

**Объясненная сумма квадратов (Explained sum of squares (ESS)): Сумма квадратов отклонений предсказанных значений  $\bar{Y}_i$  от их среднего значения; см. уравнение (4.14).**

**Объясняемая переменная (Regressand):** См. Зависимая переменная.

**Объясняющая переменная (Explanatory variable):** См. Регрессор.

**Ограниченнная зависимая переменная (Limited dependent variable):** Зависимая переменная, принимающая лишь ограниченное множество значений. Например, переменная может быть бинарной и принимать значения 0 либо 1. Или возникать из моделей регрессии, описанных в приложении 11.3.

**Одноково распределенные (Identically distributed):** Когда две или более случайные величины имеют одинаковое распределение.

**Одновременная причинность (Simultaneous causality):** Ситуация, когда помимо интересующего нас влияния  $X$  на  $Y$  существует еще и обратное влияние  $Y$  на  $X$ . Одновременна причинность приводит к корреляции  $X$  и ошибке теоретической регрессии.

**Одновременные уравнения (Simultaneous equations):** См. Одновременная причинность.

**Односторонняя альтернативная гипотеза (One-sided alternative hypothesis):** Интересующий нас параметр находится с одной стороны от значения, тестируемого в условиях нулевой гипотезы.

**Ожидание (Expectation):** См. Математическое ожидание.

**Остаточный член или ошибка (Error term):** Разность между значением  $Y$  и теоретической функцией регрессии, обозначаемая в учебнике *и*.

**Оценка (Estimate):** Численное значение оценки, вычисленное на основе данных имеющейся выборки на компьютере.

**Оценка метода максимального правдоподобия (ММП) (Maximum likelihood estimator (MLE)): Оценка неизвестных параметров, которая получена максимизацией функции правдоподобия; см. приложение 11.2.**

**Оценка метода наименьших квадратов (Least squares estimator):** Оценка, полученная при помощи метода наименьших квадратов.

**Оценка нелинейного метода наименьших квадратов (Nonlinear least squares estimator):** Оценка, полученная минимизацией суммы квадратов остатков, когда функция регрессии является нелинейной по параметрам.

**Оценка программных документов (Program evaluation):** Область исследования, связанная с оценкой эффектов от реализации программ, политик или любой другой интервенции или «воздействия».

**Оценка разностей (Differences estimator):** Оценка причинного влияния, построенная как разность выборочных средних исходов в исследуемой и контрольной группах.

**Оценка «разности разностей» (Differences-in-differences estimator):** Среднее изменение  $Y$ , принадлежащих исследуемой группе, минус среднее изменение  $Y$ , принадлежащих контрольной группе.

**Оценки (Estimators):** Функция выборки данных, случайным образом извлеченных из генеральной совокупности. Оценка – это процедура, в которой выборка данных используется для того, чтобы понять, чему равно значение параметров генеральной совокупности, таких как среднее значение.

**Оценки метода наименьших квадратов (Ordinary least squares estimators):**

Оценки коэффициентов регрессии, которые минимизируют сумму квадратов остатков.

**Ошибка I рода (Type I error):** При проверке гипотез; ошибка, означающая неверное отвержение нулевой гипотезы, когда нулевая гипотеза верна.

**Ошибка II рода (Type II error):** При проверке гипотез; ошибка, означающая неверное неотвержение нулевой гипотезы, когда нулевая гипотеза неверна.

**Ошибка прогноза (Forecast error):** Разность между фактическим и прогнозным значениями переменной.

**Панельные данные (Panel data):** Данные для нескольких объектов, которые наблюдаются в течение двух или более периодов времени.

**Параметры (Parameters):** Константы, которые определяют характеристики распределения вероятности или теоретической функции регрессии.

**Первая разность (First difference):** Первая разность временного ряда  $Y_t$ , равная  $Y_t - Y_{t-1}$  и обозначается  $\Delta Y_t$ .

**Переопределенность или сверхидентифицируемость (Overidentification):**

Ситуация, когда число инструментов превышает число эндогенных регрессоров.

**Повторяющиеся межобъектные данные (Repeated cross-sectional data):** Набор множеств межобъектных данных, где каждое множество межобъектных данных относится к определенному периоду времени.

**Подобранное значение (Fitted value):** См. Предсказанное значение.

**Полная сумма квадратов (Total sum of squares (TSS)):** Сумма квадратов отклонений  $Y_i$  от их среднего.

**Порядок интегрированности (Order of integration):** Число разностей, которые необходимо взять, чтобы временной ряд стал стационарным. Если временной ряд является интегрированным порядка  $d$ , то необходимо взять  $d$  его разностей. Он обозначается как  $I(d)$ .

**Постоянный регрессор (Constant regressor):** Регрессор, который ставится в соответствие константе в регрессии; этот регрессор всегда равен единице.

**Постоянный член (Constant term):** Константа в регрессии.

**Потенциальный исход (Potential outcomes):** Множество исходов, которые могут произойти с индивидом (объектом воздействия) после получения или не получения экспериментального воздействия.

**Предположения метода наименьших квадратов (Least squares assumptions):** Предположения для модели линейной регрессии, сформулированные во вставках «Основные понятия 4.3» (для случая парной регрессии) и «Основные понятия 6.4» (для модели множественной регрессии)  $\bar{Y}_i$ .

**Предсказанное значение (Predicted value):** Значение  $Y_i$ , которое предсказывается линией МНК-регрессии, обозначаемое в учебнике как  $\hat{Y}_i$ .

**Причинный эффект или причинное влияние (Causal effect):** Ожидаемый эффект влияния от некоторого воздействия, который измерен в идеальном случайном контролируемом эксперименте.

**Пробит-регрессия (Probit regression):** Модель нелинейной регрессии для бинарной зависимой переменной, в которой теоретическая функция регрессии

моделируется с использованием функции кумулятивного стандартного нормального распределения.

**Простой случайный выбор (Simple random sampling):** Выбор, при котором объекты выбираются из генеральной совокупности с использованием метода, гарантирующего, что объекты имеют одинаковую вероятность быть выбранными.

**Псевдовневыборочный прогноз (Pseudo out-of-sample forecast):** Прогноз, вычисленный для части выборки с использованием процедуры, при которой данные выборки как бы нереализованы.

**Размер теста (Size of a test):** Вероятность, с которой тест неправильно отвергает нулевую гипотезу, когда она верна.

**Распределение Бернулли (Bernoulli distribution):** Функция плотности вероятности Бернульиевской случайной величины.

**Распределение вероятности (Probability distribution):** Для дискретной случайной величины – это перечень всех значений, которые она может принимать, и соответствующие им вероятности.

**Распределение хи-квадрат (Chi-squared distribution):** Распределение суммы  $m$  квадратов стандартных нормальных случайных величин. Параметр  $m$  называют числом степеней свободы распределения хи-квадрат.

**Расширенный тест (Augmented Dickey–Fuller (ADF) test):** Основанный на оценке регрессии тест для проверки гипотезы о наличии единичного корня в модели AR ( $p$ ).

**Регрессия с инструментальными переменными (Instrumental variables (IV) regression):** Способ получения состоятельных оценок неизвестных коэффициентов теоретической функции регрессии, если регрессор  $X$  коррелирован с ошибкой  $u$ .

**Регрессия, оцененная на первом шаге (First-stage regression):** Регрессия зависимости эндогенных переменных от экзогенных и инструментальных переменных, оцениваемая на первом шаге двухшагового метода наименьших квадратов.

**Регрессия с ограничениями (Restricted regression):** Регрессия, на коэффициенты которой наложены ограничения, удовлетворяющие некоторым условиям. Например, при вычислении  $F$ -статистики для случая гомоскедастичных ошибок такой регрессией является регрессия с ограничениями на коэффициенты, удовлетворяющими нулевой гипотезе.

**Регрессия без ограничений (Unrestricted regression):** При вычислении  $F$ -статистики в предположении гомоскедастичности – регрессия, предполагаемая в условиях альтернативной гипотезы, на коэффициенты которой не накладываются ограничения, чтобы удовлетворять нулевой гипотезе.

**Регрессор или объясняющая переменная (Regressor):** Переменная, стоящая с правой стороны регрессии; независимая переменная в регрессии.

**Сбалансированная панель (Balanced panel):** Множество панельных данных без пропусков наблюдений, то есть в котором данные наблюдаются для каждого объекта в каждый момент времени.

**Свободный член или константа (Intercept):** Значение  $\beta_0$  в модели линейной регрессии.

**Серийная корреляция (Serial correlation):** См. Автокорреляция.

**Серийно некоррелированные (Serially uncorrelated):** Временные ряды, у которых все автокорреляции равны нулю.

**Скорректированный  $R^2$  ( $\bar{R}^2$ ) (Adjusted  $R^2$ ):** Скорректированная версия  $R^2$ , которая необязательно возрастает при включении в регрессию новых объясняющих переменных.

**Слабые инструменты (Weak instruments):** Инstrumentальные переменные, которые имеют низкую корреляцию с эндогенными регрессорами.

**Случайное блуждание (Random walk):** Случайный процесс (временной ряд), значение которого равно его значению в предыдущий момент времени плюс непредсказуемая случайная ошибка.

**Случайное блуждание с дрейфом (сносом) (Random walk with drift):** Обобщение процесса случайного блуждания, изменение величины которого имеет ненулевое среднее, но продолжает оставаться непредсказуемым.

**Случайный управляемый (контролируемый) эксперимент (Randomized controlled experiment):** Эксперимент, в котором участники случайным образом причисляются к контрольной группе, которая не испытывает воздействия, или к исследуемой группе, которая испытывает воздействие.

**Смещение (Bias):** Математическое ожидание разности между параметром, который оценивается, и его оценкой. Если  $\hat{\mu}_Y$  является оценкой  $\mu_Y$ , то смещение равно:  $E(\hat{\mu}_Y - \mu_Y)$ .

**Смещение из-за отбора наблюдений (Sample selection bias):** Смещение оценки коэффициента регрессии, возникающее из-за процесса отбора наблюдений, на который влияет доступность данных, и этот процесс связаны с зависимой переменной. Это смещение возникает из-за корреляции между одним или несколькими регрессорами и ошибкой регрессии.

**Смещение из-за ошибок в переменных (Errors-in-variables bias):** Смещение оценок коэффициентов регрессии, которое возникает из-за ошибок измерения регрессоров.

**Смещение из-за пропущенных переменных (Omitted variables bias):** Смещение оценки, возникающее из-за того, что переменная, влияющая на  $Y$  и коррелированная с регрессором, пропущена в регрессии.

**Совершенная мультиколлинеарность (Perfect multicollinearity):** Случается, когда один из регрессоров является точной линейной комбинацией других регрессоров.

**Совместная гипотеза (Joint hypothesis):** Гипотеза, состоящая из двух или более индивидуальных гипотез, то есть содержащая более одного ограничения на параметры модели.

**Совместное распределение вероятностей (Joint probability distribution):** Распределение вероятностей, определяющее вероятности исходов двух или более случайных величин.

**Совокупный динамический мультипликатор (Cumulative dynamic multiplier):** Накопленное влияние единичного изменения временного ряда  $X$  на  $Y$ . Совокупный динамический мультипликатор на  $h$  шагов вперед равен влиянию единичного изменения  $X_t$  на  $Y_t + Y_{t+1} + \dots + Y_{t+h}$ .

**Состоятельная оценка (Consistent estimator):** Оценка, которая сходится по вероятности к оценивавшемуся параметру.

**Состоятельность (Consistency):** Означает, что оценка является состоятельной. См. *Состоятельная оценка*.

**Спецификация регрессии (Regression specification):** Описание регрессии, включающее набор регрессоров и все применяющиеся нелинейные преобразования.

**Среднее (Mean):** Математическое ожидание случайной величины. Среднее значение случайной величины  $Y$  обозначается  $\mu_Y$ .

**Средний причинный эффект (Average causal effect):** Теоретическое среднее значение индивидуального причинного эффекта в гетерогенной генеральной совокупности. Также называется средним эффектом воздействия в эксперименте.

**Стандартизация случайной величины (Standardizing a random variable):** Операция, заключающаяся в извлечении среднего и делении на стандартное отклонение, которая дает случайную величину с нулевым средним и единичным стандартным отклонением. Стандартизированное значение  $Y$  равно  $(Y - \mu_Y)/\sigma_Y$ .

**Стандартная ошибка оценки (Standard error of an estimator):** Оценка стандартного отклонения оценки.

**Стандартная ошибка регрессии (Standard error of the regression (SER)): Оценка стандартного отклонения ошибки регрессии  $u$ .**

**Стандартное нормальное распределение (Standard normal distribution):** Нормальное распределение с нулевым средним и единичной дисперсией, обозначаемое  $N(0, 1)$ .

**Стандартное отклонение (Standard deviation):** Квадратный корень из дисперсии. Стандартное отклонение случайной величины  $Y$ , обозначаемое  $\sigma_Y$ , имеет такие же единицы измерения, как  $Y$ , и измеряет разброс  $Y$  около своего среднего.

**Статистическая значимость (Statistically significant):** Нулевая гипотеза (обычно о том, что коэффициент регрессии равен нулю) отвергнута на данном уровне значимости.

**Статистическая незначимость (Statistically insignificant):** Нулевая гипотеза (как правило о том, что коэффициент регрессии равен нулю) не может быть отвергнута на данном уровне значимости.

**Стационарность (Stationarity):** Ситуация, когда совместное распределение временного ряда и запаздываний не изменяется во времени.

**Стохастический тренд (Stochastic trend):** Устойчивое, но случайное долгосрочное движение переменной во времени.

**Строгая экзогенность (Strict exogeneity):** Требование, согласно которому ошибка регрессии имеет нулевое условное среднее относительно текущего, прошлых и будущих значений регрессора с модели регрессии с распределенными лагами.

**Стандартные ошибки в предположении гомоскедастичности ошибок регрессии (Homoskedasticity-only standard errors):** Стандартные ошибки МНК-оценки,

которые можно использовать только тогда, когда остаточный член регрессии гомоскедастичен.

**Сумма квадратов остатков (Sum of squared residuals (SSR)):** Сумма квадратов МНК-остатков.

**Сходимость по вероятности (Convergence in probability):** Когда последовательность случайных величин сходится к определенному значению, например, когда выборочное среднее становится близким к среднему в генеральной совокупности при возрастании размера выборки; см. вставку «Основные понятия 2.6» и раздел 17.2.

**Сходимость по распределению (Convergence in distribution):** Когда последовательность распределений сходится к пределу; точное определение приведено в разделе 17.2.

**Теорема Гаусса–Маркова (Gauss–Markov theorem):** Математический результат, утверждающий, что при определенных условиях оценка метода наименьших квадратов является наилучшей условной линейной несмещенной оценкой коэффициентов регрессии относительно значений регрессоров.

**Теоретическая модель множественной регрессии (Population multiple regression model):** Модель множественной регрессии. См. вставку «Основные понятия 6.2».

**Теоретические коэффициенты (Population coefficients):** См. *Теоретические свободный член и угловой коэффициент*.

**Теоретические свободный член и угловой коэффициент (Population intercept and slope):** Истинные или теоретические значения в генеральной совокупности значения  $\beta_0$  (свободный член, константа) и  $\beta_1$  (коэффициент наклона) с парной регрессии. Во множественной регрессии коэффициентов наклона несколько ( $\beta_1, \beta_2, \dots, \beta_k$ ) – по одному для каждого регрессора.

**Тест Грейнджа на причинность (Granger causality test):** Процедура проверки того, помогают ли текущее и прошлые значения одного временного ряда предсказывать значения другого временного ряда.

**Тест Дики–Фуллера (Dickey–Fuller test):** Метод, используемый для проверки наличия единичного корня в авторегрессии первого порядка [AR(1)].

**Тест для разности между двумя средними (Test for the difference between two means):** Процедура тестирования гипотезы о том, равны ли средние значения двух генеральных совокупностей.

**Тест Чоу (Chow test):** Тест на наличие структурного сдвига в регрессии временных рядов в известный момент времени.

**Тестирование гипотезы (Hypothesis test):** Процедура, в которой выборочные характеристики позволяют определить, являются ли истинными или ложными специфические предположения о генеральной совокупности.

**Точная идентификация (Exact identification):** Ситуация, когда число инструментов равно числу эндогенных регрессоров.

**Точное распределение или распределение в конечных выборках (Exact (finite-sample) distribution):** Точное распределение вероятности случайной величины.

**Уровень доверия или доверительная вероятность (Confidence level):** Заранее определенная вероятность (или область), с которой доверительный интервал (или область) содержит истинное значение параметра.

**Уровень значимости (Significance level):** Заранее определенное значение вероятности отвержения нулевой гипотезы при тестировании статистической гипотезы в предположении, что нулевая гипотеза верна.

**Условная гетероскедастичность (Conditional heteroskedasticity):** Дисперсия (как правило, остаточного члена) зависит от других переменных.

**Условная дисперсия (Conditional variance):** Дисперсия условного распределения.

**Условное математическое ожидание (Conditional expectation):** Ожидаемое значение одной случайной величины, полученное при условии, что другая случайная величина принимает определенное значение.

**Условное распределение (Conditional distribution):** Функция плотности одной случайной величины, полученная при условии, что другая случайная величина принимает определенное значение.

**Условное среднее (Conditional mean):** Среднее значение условного распределения. См. Условное математическое ожидание.

**Устойчивая к гетероскедастичности *t*-статистика (Heteroskedasticity-robust *t*-statistic):** *t*-статистика, построенная с использованием устойчивых к гетероскедастичности стандартных ошибок.

**Устойчивые к гетероскедастичности и автокорреляции (HAC) стандартные ошибки (Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors):** Стандартные ошибки МНК-оценок, которые являются состоятельными независимо от того, являются или нет ошибки регрессии гетероскедастичными и автокоррелированными.

**Устойчивые к гетероскедастичности стандартные ошибки (Heteroskedasticity-robust standard error):** Стандартные ошибки МНК-оценок, которые являются состоятельными независимо от того, являются ли ошибки регрессии гомоскедастичными или гетероскедастичными.

**Фиксированные эффекты (Fixed effects):** Бинарная переменная, являющаяся характеристикой объекта или периода времени в панельной регрессии.

**Функция линейной регрессии (Linear regression function):** Функция регрессии с постоянным коэффициентом наклона.

**Функция нелинейной регрессии (Nonlinear regression function):** Функция регрессии с непостоянным наклоном.

**Функция плотности вероятности (Probability density function (p.d.f.)):**  Для непрерывной случайной величины площадь под графиком плотности вероятности между любыми двумя точками является вероятностью, с которой случайная величина попадает в интервал между этими двумя точками.

**Центральная предельная теорема (Central limit theorem):** Результат из математической статистики, который говорит о том, что при выполнении некоторых общих условий выборочное распределение стандартизированного выборочно-го среднего хорошо приближается стандартным нормальным распределением в выборках большого размера.

**Частичное соответствие (Partial compliance):** Случается, когда некоторые участники не следуют схеме эксперимента в случайному эксперименте.

**Частный эффект (Partial effect):** Влияние изменения одного из регрессоров на  $Y$  в предположении постоянства остальных регрессоров.

**Экзогенная переменная (Exogenous variable):** Переменная, которая является не коррелированной с остаточным членом регрессии.

**Экспериментальные данные (Experimental data):** Данные, полученные в эксперименте, разработаны для того, чтобы оценить влияние какого-либо воздействия или политики или изучить причинное влияние.

**Эксцесс (Kurtosis):** Мера того, насколько тяжелыми являются хвосты функции плотности вероятности.

**Эластичность спроса по цене (Price elasticity of demand):** Процентное изменение в объеме спроса, вызываемое увеличением цены на 1 %.

**Эндогенная переменная (Endogenous variable):** Переменная, которая коррелирована с ошибкой регрессии.

**Эффект воздействия в эксперименте (Treatment effect):** Причинное влияние в эксперименте или квазиэксперименте. См. *Причинный эффект или причинное влияние*.

# Указатель

2МНК-оценка. См. Оценка двухшаговым методом наименьших квадратов (2МНК)  
95 %-е доверительное множество, 233–234

## А

ADF-статистика. См. Расширенная статистика Дики–Фуллера

ADF-тест. См. Расширенный тест Дики–Фуллера

ADF-тест Энгла–Грейнджа, 684, 686

ADL-модель, 565, 596–598

в записи с использованием оператора запаздывания, 559–663

МНК-оценка, 635, 641–642

ОМНК-оценка, 641–642

оценка коэффициентов в, 637–638, 640–642

приложения, 648–650

ADL( $p, q$ )-модель, 565

AIC: состоятельности оценки глубины запаздывания, 614

AR (1)-модель

в teste Дики–Фуллера, 582–583

стационарность в, 610–611

AR(1)-ошибки, модель регрессии с распределенными лагами и, 634–637

AR ( $p$ )-модель, 560–563

ARCH-модель, 685, 686

кластеризованная волатильность и, 693

ARMA-модель, 612–613

## Б

BIC: состоятельности оценки глубины запаздывания, 613–614

BLUE (Наилучшая линейная несмещенная оценка), 71

МНК-оценка, 165–166

ОМНК-оценка, 639

## С

CLR-тест, 489–490

## Д

DF-статистика. См. Статистика Дики–Фуллера

DF-тест. См. Тест Дики–Фуллера

DF–GLS-тест, 677–680

и тест Дики–Фуллера, 678, 679

критические значения, 678, 679

приложение, 679

DOLS-оценка, 687–688

для нескольких переменных, 688

Dow Jones Industrial Average, 41–42

## Е

EG–ADF-тест, 685, 686

для нескольких переменных, 688

приложения, 689–690

## Ф

F-распределение, 44, 728

F-статистика, 226–228

$p$ -значения, 227–228

QLR-статистика и, 589–592, 593

Вальда, 743

выбор глубины запаздывания и, 573–574, 575–576

вычисленная на первом шаге, 460

для проверки совместных гипотез, 226–228, 739

для проверки значимости регрессии, 228

доверительная область для нескольких коэффициентов и, 233–234, 739–740

определение, 226

порядок авторегрессии и, 573–574

приложения, 228–229

при условии гомоскедастичности, 229–232, 743

распределение, 743, 773–774

устойчивая к гетероскедастичности, 227

F-статистика, вычисленная на первом шаге, 460

- F*-статистика, рассчитанная при условии гомоскедастичности, 229–232, 743
- G**
- GARCH-модель, 693–695
- GMM *J*-статистика, 758  
распределение, 777
- H**
- HAC-оценка, 631–633  
веса для, 633
- HAC-стандартная ошибка, 378–379, 619, 631–633, 634  
в прямых многошаговых прогнозах, 672–673
- I**
- $I(0)$ ,  $I(1)$  и  $I(2)$ , 675
- i.i.d., 46, 239–241, 745
- IV-регрессия. См. Модель регрессии с инструментальными переменными
- J**
- J*-статистика, 465
- GMM, 758  
асимптотическое распределение, 755, 775–777  
приложения, 468, 528  
при наличии гомоскедастичности, 755–756, 775–777  
устойчивая к гетероскедастичности, 758
- j*-я автоковариация, 554
- j*-е запаздывание, 551
- P**
- p*-значение, 74–76, 150
- F*-статистики, 228  
вычисление, 75–76, 78–79, 82  
для МНК-оценок, 150  
определение, 74
- R**
- $R^2$ . См.  $R^2$  регрессии
- $R^2$  регрессии
- бинарная зависимая переменная и, 402  
во множественной регрессии, 197–199, 239–241  
псевдо- $R^2$  и, 416, 435  
скорректированный, 197–199, 239–241
- T**
- t*-отношение. См. *t*-статистика
- t*-распределение, 44
- t*-распределение Стьюдента, 44  
в проверке гипотез, 89–93  
на практике, 93
- t*-статистика и, 167–168
- t*-статистика, 78–79, 82, 149–150.
- См. также Тестовая статистика
- t*-распределение Стьюдента и, 167–168  
асимптотическая, 713, 738  
асимптотическое распределение, 78–79  
в предположении гомоскедастичности, 167–168  
в регрессиях временных рядов, 581  
для МНК-оценки, 149  
для тестирования гипотезы о среднем, 89  
ненормальное распределение, 580  
общий вид, 148  
объединенная (для «пула»), 91–93  
определение, 78  
основанная на выборочном среднем, 710–711  
при проверке совместных гипотез, 225–226  
распределение, 743, 772–774  
распределение в конечных выборках, 89–93, 167–168
- V**
- VAR -модель, 665–669

**А**

Автоковариация, 554–555  
 Автокоррелированные ошибки, распределение МНК-оценки и, 629–631  
 Автокорреляция, 377–378, 554–555  
 Авторегрессионная модель с условной гетероскедастичностью (ARCH) 685, 686–687  
     кластеризованная волатильность и, 691–693  
 Авторегрессионная модель с распределенными лагами, 565–566, 595–597  
 Авторегрессионная модель фондового рынка, 562–563, 595–597  
 Авторегрессионные ошибки и модель с распределенными лагами, 640  
 Авторегрессия, 557–563. См. также Регрессия временных рядов  
     векторная, 664–669, 670–671. См. также Векторная авторегрессия  
         глубина запаздывания, 560  
         определение, 557  
         ошибка прогноза и, 558  
         первого порядка, 557–560  
         порядок, 573–576  
         порядка  $p$ , 560–563  
         смещение в, 580  
 Авторегрессия порядка  $p$ , 560–563  
 Авторегрессия первого порядка, 557–560  
 Асимптотическая нормальность, 736–737  
 Асимптотические распределения, 50–55  
     закон больших чисел и, 50–55  
     центральная предельная теорема и, 52–55  
 Асимптотическое распределение, 50, 706–713  
     2МНК-оценки, 752–753  
      $t$ -статистики, 713  
     МНК-оценки, 711, 735–737  
     нормальное, 55  
     определение, 709  
     сходимость и, 706–709  
     состоятельность и, 706–709, 711–713  
     теорема о непрерывном отображении и, 710  
     теорема Слуцкого и, 710  
     центральная предельная теорема и, 709–710  
 Асимптотическое нормальное распределение, 131–134

нестационарность и, 679–681  
 Асимптотическое распределение, 2МНК-оценки, 448–449, 484–488  
 Альтернативная гипотеза, 73  
     двухсторонняя, 74, 81  
     односторонняя, 82

**Б**

Банк Англии, 571–572  
 Байесовский информационный критерий (BIC), 573, 574, 613–614  
 База данных по результатам тестов в Калифорнии, 142  
 Баунд, Джон, 463  
 Безработица и инфляция, 550–551, 553, 563–565, 668–669  
 Безусловное распределение вероятностей, 29, 30  
 «Бета» на фондовом рынке, 121  
 Бинарные переменные  
      $R^2$  регрессии и, 402–403  
     взаимодействия между, 282  
     взаимодействия с непрерывной переменной, 284–288  
     зависимая переменная в регрессии, 398–437  
     критерии качества приближения данных моделью и, 415–416  
     линейная вероятностная модель и, 402–405  
     логит-модель и, 409–416  
     обзор, 399–401  
     оценка максимального правдоподобия и, 413–414, 433–434  
     оценка нелинейного метода наименьших квадратов, 412  
     приложения, 416–424  
     пробит-модель и, 404–409, 411–416  
 Бюро статистики труда, 73

**В**

Валовый внутренний продукт (GDP), Японии, квартальный, 556–557  
 Вальд, Абрахам, 743  
 Вектор математического ожидания, 770  
 Вектор средних, 770  
 Векторная авторегрессия, 664–669  
     глубина запаздывания в, 667  
     для анализа причинности, 667–668  
     для прогнозирования, 664–668

- итеративного, 669–671  
приложения, 668–669  
число коэффициентов и, 666–667
- Векторная модель коррекции ошибками, 681–683, 685, 690–691
- Вероятность попадания, 84
- Вероятность, 17–56
- значимость. См. *p*-значение  
определение, 18  
попадания, 84  
события, 20  
сходимость по, 50, 707–709  
элементарного события, 18
- Вест, Кеннет, 632
- Включенные экзогенные переменные, 451–453
- ВМНК-оценка. См. Оценка взвешенного метода наименьших квадратов
- Внешняя обоснованность, 320–321  
в квазиэкспериментах, 524  
в прогнозировании, 337–339  
изучаемая генеральная совокупность и, 320  
определение, 320  
оценка, 323  
схема эксперимента и, 322–323  
угрозы, 321–323, 502–503  
целевая генеральная совокупность и, 320
- Внутренняя обоснованность, 320–349  
в квазиэкспериментах, 521–524  
в прогнозировании, 337–339  
гетероскедастичность и, 335–336, 348  
изучаемая генеральная совокупность и, 320  
неправильная спецификация функциональной формы и, 326–327  
определение, 320  
ошибки измерения и, 327–330  
пропуски в данных и, 330–332  
смещение из-за отбора наблюдений, 332  
смещение из-за ошибок в переменных и, 330–332  
смещение из-за пропущенных переменных и, 323–325  
угрозы для, 321, 323–337, 337, 497–502  
целевая генеральная совокупность и, 320
- Вождение в нетрезвом виде  
база данных, 392  
модель регрессии с фиксированными эффектами и, 362–364, 371–372, 375  
панельные данные для, 362–364, 371
- Временные ряды, 11–13, 377–378, 550–557, 609–610, 746  
автоковариация и, 554–555  
автокорреляция и, 377–378  
запаздывания и, 551–553  
первые разности и, 551–553  
приложения, 549–553  
причинное влияние и, 620–621  
примеры, 555–557  
темпы роста и, 551–553  
логарифмы и, 551–553
- Вторая разность, 675
- Выбросы, 28  
в модели регрессии с фиксированными эффектами, 377  
закон больших чисел и, 51–52  
МНК-регрессия и, 129–130
- Выборка  
не i.i.d., 128–129  
случайная, 45–47
- Выборочная дисперсия, 76–78  
состоительность, 77–78, 107–108
- Выбор глубины запаздывания  
*F*-статистика для, 573–574, 575–576  
в регрессии временных рядов со множественными регрессорами, 575–576  
информационные критерии для, 573–574, 576–577
- Выборочное распределение, 45–49. См. также Распределение (я)  
2МНК-оценки, 447–448, 456, 484–488  
асимптотическое, 49–50  
в конечных выборках, 50  
в регрессии с инструментальными переменными, 447–448  
в регрессии с фиксированными эффектами, 370  
выборочного среднего, 47–49  
закон больших чисел и, 49–52  
МНК-оценки, 131–134, 142–146  
приближение в больших выборках, 50  
релевантный инструмент и, 459  
точное, 50  
центральная предельная теорема и, 52–54

- Выборочное стандартное отклонение, 77–78 нулевая. См. Нулевая гипотеза односторонняя
- Выборочное среднее значение альтернативная, 81–82
- выборочное распределение, 46–49 для углового коэффициента, 151–152
- как оценка, 69–72 совместная, критерий Бонферрони, 253–256
- определение, 46
- среднее и дисперсия, 47–48
- Выборочная ковариация, 94–97, 447–448 Глубина запаздывания (длина лага)
- состоятельность, 96 в авторегрессии, 560–561
- Выборочная корреляция, 94–97 выбор, 573–576
- состоятельность, 96 в векторной авторегрессии, 666–667
- Выпускники колледжей, заработка пла- Гомоскедастичность, 158–164, 176–177, 192
- та, 35–36 2МНК-оценка и, 753–756
- отдача от образования и, 290–291, 462–463 J-статистика при условии, 755, 775–777
- пол и, 35–37, 88–89, 162–163, 281–283, 290–291 во множественной регрессии, 192, 754–756
- математические следствия, 160
- определение, 158
- Г**
- Гендерный разрыв. См. Выпускники колледжей, заработные платы
- Генеральная совокупность, случайная выборка из 45–47
- Гладкий тренд, 675–676
- Грейнджер, Клайв, 684–685
- Гетерогенные генеральные совокупности, оценки в экспериментах/квазиэкспериментах в, 524–530, 541–542
- Гетероскедастичность, 158–164, 176, 193
- взвешенный метод наименьших квадратов и, 716–721
- во множественной регрессии, 158–164, 176, 193
- известной функциональной формы, 717–720
- корреляция ошибки и, 336–337, 348
- математические следствия, 160
- несостоятельные стандартные ошибки и, 335
- определение, 158
- условная авторегрессионная, 685, 686, 693–694
- Гипотеза эффективности финансовых рынков, 563
- Гипотеза
- альтернативная, 73
- двухсторонняя
- альтернативная, 73, 81
- для углового коэффициента, 148–151
- односторонняя
- альтернативная, 81–82
- для углового коэффициента, 151–152
- совместная, критерий Бонферрони, 253–256
- Глубина запаздывания (длина лага)
- в авторегрессии, 560–561
- выбор, 573–576
- в векторной авторегрессии, 666–667
- Гомоскедастичность, 158–164, 176–177, 192
- 2МНК-оценка и, 753–756
- J-статистика при условии, 755, 775–777
- во множественной регрессии, 192, 754–756
- математические следствия, 160
- определение, 158
- Д**
- Дамми-переменная, 156
- Данные
- временные ряды, 12–13
- дискретного выбора, 438
- источники и типы, 9–14
- межобъектные, 11–12
- наблюдения, 9–10
- панельные (лонгитюдные). См. Панельные данные
- экспериментальные, 10
- Данные дискретного выбора, 438
- Данные множественного выбора, 437–438
- Дата структурного сдвига, 588–592
- известная, 588–589
- неизвестная, 589–592
- Двойные слепые эксперименты, 501
- Двумерное нормальное распределение, 40–43, 727
- Двухсторонняя гипотеза
- альтернативная, 74, 149
- для коэффициента наклона, 146–151
- Двухшаговый метод наименьших квадратов (2МНК)
- асимптотическое распределение, 752–753
- в матричной форме, 752–756
- выборочное распределение, 447–448, 456, 485

- вычисление стандартных ошибок, 457, 752–753  
гомоскедастичность и, 752–756  
допустимость инструментов и, 449–451  
локальный средний эффект (воздействия) в эксперименте, 526–530  
оценка, 441–442, 453–458  
предположения, 455–456  
приложения, 443–447, 449–451, 465–470  
регрессия на втором шаге, 454  
регрессия на первом шаге и, 454  
с контрольными переменными, 491–492  
с несколькими эндогенными регрессорами, 453–454  
с одним эндогенным регрессором, 453  
слабые инструменты и, 449–461, 462–463, 488–491  
статистические выводы с использованием, 456–457  
формула для, 447, 484–485  
эффективность в предположении гомоскедастичности, 775–776  
Детерминированный тренд, 577  
Диаграмма рассеяния, 94, 95  
Диверсификация, 48–49  
Дики, Дэвид, 582–583  
Динамические мультипликаторы, 627–628  
долгосрочные совокупные, 628  
нулевого периода (одновременные), 627  
определение, 627  
приложения, 641–650  
совокупные, 627–628  
устойчивость, 647, 649–650  
Динамическая МНК-оценка, 687–688  
Динамическое причинное влияние, 615–663  
НАС-стандартные ошибки и, 628–633  
измерение, 622–623  
модель регрессии с распределенными лагами и, 619–620, 621–627. См. также Модель регрессии с распределенными лагами  
ОМНК-оценка, 633–634, 637–639, 641–642  
оценка с экзогенными регрессорами, 625–628  
оценка со строго экзогенными регрессорами, 633–642  
предположение об экзогенности и, 650–653  
приложения, 642–650  
экзогенность и, 623–625  
Дискретная случайная величина, 18  
распределение вероятности, 18–19  
Дисперсия, 24–25  
выборочная, 77–79  
выборочного среднего, 48–49  
определение, 24–25  
оценок, 69  
случайной величины, 24–26  
стандартное отклонение и, 24–25  
суммы случайных величин, 35–38  
условная, 33  
Доверительное множество, 82, 489–490  
Андерсона–Рубина тест, 489–490  
для нескольких коэффициентов, 233–234, 739–740  
для одного коэффициента, 220–221  
слабые инструменты и, 488–489  
Доверительный интервал, 82–84  
во множественной регрессии, 233–235  
для коэффициента регрессии, 154–156  
для МНК-оценки, 154–156  
для предсказанных эффектов, 737–738  
для одного коэффициента, 222  
для разностей средних, 86  
для регрессии с бинарной объясняющей переменной, 157–158  
для углового коэффициента, 154–156  
и интервальный прогноз, 570–571  
Доверительный эллипс, 234  
Доклад об инфляции, 571  
Долгосрочный совокупный динамический мультипликатор, 628  
Допустимость инструмента, 441, 449–450, 455  
в квазиэкспериментах, 523–524  
источники, 471–477  
оценка, 458–465  
Доступная GMM-оценка, 756–759  
Доступная ВМНК-оценка, 718  
Доступная ОМНК-оценка, 638, 749  
Доходы в округе и результаты тестов, 261–264, 266–267, 280, 315–316  
Дрейф, случайное блуждание с, 578–579

**Е**

Европейский центральный банк, 6  
 Единичный корень, 579  
     нулевая гипотеза о наличии, 586–587,  
     679–680  
 Естественные эксперименты. См. Квази-  
 эксперименты

**Ж**

Журналы, экономические, спрос на, 293–  
 295

**З**

Закон больших чисел, 50–52  
     асимптотическое распределение  
     и, 707–709, 710  
     доказательство, 707–708  
     неравенство Чебышева и, 707–708  
     состоительность и, 50–52, 707–709  
 Закон повторного математического ожи-  
 дания, 32–33  
 Зависимая переменная, 113  
 Запаздывания (лаги), 551–553  
     в ADL-модели, 565  
 Заморозки и цены на апельсиновый сок,  
 616–620  
     база данных, 660  
     регрессия временных рядов и, 642–  
     650  
 Значение вероятности. См. *p*-значение  
 Значение, ожидаемое. См. Математиче-  
 ское ожидание

**И**

Идеальный случайный контролируемый  
 эксперимент, 8  
 Изучаемая генеральная совокупность,  
 319–350  
 Иммиграция и рынок труда, 513, 524  
 Индекс потребительских цен, 553  
 Индивидуальные и временные фиксиро-  
 ванные эффекты, 374–375  
 Индикаторная переменная, 629–630  
 Инструменты. См. Инstrumentальные пе-  
 ременные  
 Инstrumentальные переменные, 751–757  
     в обобщенной модели IV-регрессии,  
     441, 450, 454–455  
     в оценке эффекта воздействия в экс-  
     перименте, 499–500  
     допустимость, 449, 454–455

источники, 471–477  
 релевантность, 441–442, 450, 454–  
 455  
 слабые, 459–461, 462–463, 488–491  
 случайно определенные, 523–524  
 экзогенность, 441, 450, 454–455  
 Инstrumentальные переменные (IV)  
     в квазиэкспериментах, 520  
     в матричной форме, 751–752  
     используя линейные комбинации  $Z$ ,  
     753–755  
     оценка, 751–757, 753–755  
     с гетерогенным причинным эффектом,  
     526–529  
 Интервальный прогноз, 570–572  
 Интегрированный порядка  $d$ , 675  
 Информационный критерий Акаике (AIC),  
 574–575  
 Информационный критерий Шварца  
 (SIC), 573, 574  
 Информационный критерий, 573–576  
     Акаике, 574–575  
     Байесовский, 573, 574  
     вычисление, 575  
     для авторегрессии, 573–576  
     для векторной авторегрессии, 667  
     Шварца, 573, 574  
 Ипотечное кредитование и расовая при-  
 надлежность, 5, 399–401, 416–424  
     база данных, 432  
     линейная вероятностная модель  
     и, 402–403  
     логит-модель и, 410  
     пробит-модель и, 407–410  
 Исследуемая (экспериментальная)  
 группа, 8  
 Истощение выборки  
     в квазиэкспериментах, 522–523  
     в случайном управляемом (контроли-  
     руемом) эксперименте, 500  
 Итеративная оценка Кохрейна–Оркэтта,  
 639–640  
 Итеративный AR (1) прогноз, 669–670  
 Итеративный многошаговый AR-прогноз,  
 671  
 Итеративный многошаговый прогноз,  
 669–671, 671  
     AR, 671  
     VAR, 671  
     и прямой многошаговый прогноз,  
     673–674

**Й**

Йохансен, Сорен, 687, 688

**К**

Казначейские векселя, ставка процента, по, 700  
Казначейские векселя США, ставка процента по, 700  
Кард, Дэвид, 516–517  
Катетеризация сердца, 474–477, 515, 529–530  
Квадратичная регрессия, 262–264  
Квадратный корень из среднеквадратичной ошибки прогнозирования, 559, 570  
псевдовневыборочное прогнозирование и, 593–594  
Квазидифференцированная модель, 635–637  
нулевое условное среднее в, 636–637  
Квазик эксперименты, 87–89, 513–530  
внешняя обоснованность, 524  
внутренняя обоснованность, 522–524  
гетерогенные генеральные совокупности и, 524–530, 540–542  
допустимость инструментов в, 523–524  
истощение выборки в, 522–523  
определение, 87, 513  
отсутствие случайности выбора, 522  
оценка «разности разностей», 516–519  
оценка метода инструментальных переменных и, 519–520  
оценка разрывных регрессий и, 520–521  
примеры, 88–89, 513–515  
потенциальные проблемы с, 521–524  
частичное соответствие в, 521–522  
экспериментальные эффекты и, 523  
Квартальный ВВП Японии, 556–557  
Классическая модель с ошибками измерения, 328  
Кластеризованная волатильность, 691–694, 695  
ARCH-модель для, 685, 693  
GARCH-модель для, 685, 693–694, 695  
Кластеризованные стандартные ошибки, 378–379  
Клейн, Джозеф, 444

Ковариационная матрица, 770

Ковариация, 34–37

выборочная, 95–96, 95–98, 447–448

определение, 34

состоительность, 96

Кointеграция, 681–691

для нескольких переменных, 688–689

коррекция ошибками, 681–683, 689

определение, 681, 683

приложение, 689–691

тестирование, 683–686

Кointегрирующий коэффициент, 683

оценка, 687–688

Компонента взаимодействия, 282

Константа, 113, 114

во множественной регрессии, 191

тестирование гипотезы для, 153–154

Константа во множественной регрессии, 192

Контрольная группа, 8

Контрольные переменные

в регрессии с инструментальными переменными, 452

во множественной регрессии, 190–191, 236–239

Корректирующий член, 682–683

Корреляция, 34

выборочная, 95–96

Коэффициент асимметрии, 26–27

Коэффициент эксцесса, 27, 28, 39

Коэффициент(ы), 147–154

ADL, 637–638

в обобщенной модели IV-регрессии, 451

в модели линейной регрессии, 113–114, 116–118

в нелинейной регрессии, 267, 272

в пробит-модели, 408–409

интерпретации в линейной регрессии, 630–632

коинтегрирующий, 683, 687–688

несколько, 191

доверительное множество для, 233–235

тестирование одного ограничения для, 232–233

с бинарными переменными, 283

тестирование совместных гипотез для, 224–232

Коэффициенты при распределенных лагах

ОМНК-оценка, 638–640, 641–642

- Коэффициенты регрессии. См. Коэффициент(ы)
- Кривая, логистическая, 261–262
- Кривая Филлипса, 7, 653  
стабильность, 592, 595–599
- Критерии качества приближения данных моделью, 123–126  
в логит- и пробит-моделях, 415–416
- Критическое значение, 80, 82
- Крюгер, Алан, 462–463, 516–517
- Кумулятивное распределение, 20. См. также Функция распределения  
дискретной случайной величины, 18–20  
непрерывной случайной величины, 20, 21–22  
нормальное, 39–40  
определение, 18–19
- Курение. См. Налоги на сигареты
- Л**
- Лаговый полином, 611–612
- Линейная вероятностная модель, 402–404  
и логит- и пробит-модели, 411–412  
ограничения, 404, 411–412  
определение, 402  
приложения, 402–404
- Линейная регрессия. См. также Регрессии  
выбросы в, 129–130  
коэффициенты в, 114, 116–118  
критерии качества приближения данных моделью, 123–126  
множественная, 183–208. См. также Множественная регрессия  
независимое и одинаковое распределение и, 129–130  
оценка метода наименьших квадратов и, 118–123  
парная, 111–169, 703–729. См. также Парная линейная регрессия  
предположения метода наименьших квадратов для, 126–131  
смещение из-за пропущенных переменных и, 183–190, 196  
терминология, 113, 114
- Линейная условно несмещенная оценка, 165–166, 744
- Линейная функция от случайной величины, среднее и дисперсия, 25–26
- Линейно-логарифмическая модель, 274–275, 277–278, 316  
и кубическая модель, 280–281
- Линия выборочной регрессии, 118–119
- Линия регрессии  
метода наименьших квадратов. См. Метод наименьших квадратов  
условное распределение и, 126–127  
теоретическая. См. Линия теоретической регрессии
- Линия регрессии метода наименьших квадратов, 118  
в нелинейной регрессии, 287–288  
во множественной регрессии, 194, 195  
выбросы в, 129–130
- Линия теоретической регрессии, 114, 115, 191, 193  
в нелинейной регрессии, 287–288  
во множественной регрессии, 191  
функция условной плотности вероятности, 126–127
- Ловушка фиктивных переменных, 205–206
- Логарифмически-линейная модель, 275–276, 277–279
- Логарифмическая регрессия, 274–281  
выбор, 278–279  
линейно-логарифмическая, 274–275, 278–279, 316  
логарифмически-линейная, 275–276, 277, 278–279  
линейная в логарифмах, 276–279, 277
- Логистическая (логит) регрессия, 312–313, 409–412  
и линейная вероятностная модель, 412  
и пробит-модель, 409–410  
критерии качества приближения данных моделью, 415–416  
множественная, 438  
определение, 409  
оценка и проверка статистических гипотез в, 411–416  
оценка метода максимального правдоподобия, 413–414, 434–435  
оценка нелинейного метода наименьших квадратов, 412–413  
приложения, 411
- Ложная регрессия, 581–582
- Локальный средний эффект (воздействия) в эксперименте, 526–530

Лонгитюдные данные. См. Панельные данные  
Лэндон, Альф, 72, 331

**М**

Макфадден, Дэниел, 425  
Математическое ожидание,  
повторное, закон, 32–33  
случайной величины, 20–21  
Математическое ожидание  
непрерывной случайной величины, 24  
определение, 20–21  
случайной величины Бернулли, 25  
случайной величины, 20–23, 28  
Математическое ожидание  
выборочное, 47–49  
выборочного среднего, 48–49  
генеральной совокупности. См. Теоретическое среднее генеральной совокупности, 20, 25–26, 28, 30–31, 35–38  
определение, 20  
условное, 31–32, 34–35, 127, 376, 566–567  
Матричная алгебра, 767–770  
Матричная запись  
2МНК-оценки, 753–756  
i.i.d. бернуlliевских случайных величин, 433  
IV-оценки, 751–752  
логит-модели, 414, 434–435  
МНК-оценки, 734–735, 740–741  
модели множественной регрессии, 731–732  
оценки метода максимального правдоподобия, 414, 433–435  
пробит-модели, 414, 434  
совместных гипотез, 738–739  
стандартной ошибки, 741  
Метод наименьших квадратов (МНК)  
ADL-модели и, 637–638  
*p*-значение для, 150  
*t*-статистика для, 149  
алгебраические факты о, 145–146  
асимптотическое распределение, 735–737  
взвешенный метод наименьших квадратов и, 716–721  
в матричной форме, 733–734, 740–741  
выборочное распределение  
в модели парной регрессии, 131–134, 143–146, 714–716  
во множественной регрессии, 202–203  
вывод формул, 142–143  
динамический, 687–688  
дисперсия, 206–207  
доверительный интервал для, 154  
и ОМНК-оценка, 641–642  
как BLUE-оценка, 165  
множественная регрессия и, 194–207  
модель регрессии с распределенными лагами и, 633–634  
нелинейный, 314–315  
несовершенная мультиколлинеарность и, 206–207  
обозначения и терминология, 118–119  
определение, 118  
оценка, 118–123, 705–706  
предположения метода наименьших квадратов и, 126–131, 704–706  
распределение, 202–203, 216–217, 629–631  
регрессия с фиксированными эффектами, 369–370  
с гетерогенными причинными эффектами, 525–526  
смещение в, 183–190, 333–334, 744.  
См. также Смещение  
смещение из-за пропущенных переменных в, 183–190  
состоительность, 132–133, 712  
стандартная ошибка, 149, 157–1641, 176–177, 219–220. См. также Гетероскедастичность; Гомоскедастичность  
условия Гаусса–Маркова и, 743–745  
теорема Гаусса–Маркова и, 165–167  
теоретические основы, 164–167  
эффективность, 160–161, 164  
Межобъектные данные, 10–12  
повторяющиеся, 519  
Методология потенциальных исходов, 542–543  
Минимальная заработная плата и уровень безработицы, 516–517  
МНК-оценка, 641–642. См. Оценка метода наименьших квадратов (МНК)  
для временных рядов, 620–622  
Многомерное распределение, 770–771  
нормальное, 40–43

- Многошаговые прогнозы, 669–674  
 выбор метода, 673–674  
 итеративные, 669–671, 671, 673–674  
 прямые, 669–674
- Модели регрессии с ограниченными зависимыми переменными, 436–438  
 бинарная зависимая переменная, 398–436  
 данные дискретного выбора и, 438  
 логит-модель, 409–416  
 модель формирования выборки, 437  
 пробит-модель, 404–409, 411–416  
 счетные данные и, 437  
 упорядоченные данные и, 438  
 усеченные зависимые переменные, 436  
 цензурированная зависимая переменная, 436
- Модели формирования выборки, 437
- Модель авторегрессии с ошибками в форме скользящего среднего (ARMA), 612–613
- Модель коррекции ошибками  
 векторная, 681–683, 684–685, 690–691  
 коинтеграция и, 681–683
- Модель кубической регрессии, 270–271, 280–281
- Модель множественной регрессии, 7, 183–208  
 $R^2$  регрессии в, 197–199, 240–241  
 в матричной форме, 731–732  
 в прогнозировании, 337–339  
 генеральная совокупность, 191–193  
 гетероскедастичность/гомоскедастичность и, 192, 753–756  
 единицы измерения переменных и, 243–244  
 константа в, 192  
 контролируя  $X$  в, 191  
 контрольные переменные в, 191, 235–239  
 коэффициенты в, 190–191  
 одно ограничение для нескольких, 232  
 критерии качества приближения данных моделью в, 197–199  
 линия регрессии генеральной совокупности в, 191, 193  
 МНК-оценка в. См. Оценка метода наименьших квадратов (МНК)
- мультиколлинеарность и  
 несовершенная, 206–207  
 совершенная, 100, 203–206  
 нелинейная, 268–269. См. Нелинейная регрессия  
 ограничения, 225, 229–232
- ОМНК-оценка и, 745–751  
 определение, 183, 193  
 панельные данные в, 359–385. См. также Панельные данные  
 постоянный член в, 192  
 предположения метода наименьших квадратов, 200–202, 236–238  
 предсказанное значение и, 194  
 представление результатов в таблице и, 242–246  
 приложения, 195–196, 199, 241–246  
 проверка гипотез и  
 для нескольких коэффициентов, 224–232  
 для одного коэффициента, 222  
 расширенные предпосылки метода наименьших квадратов для, 732–734  
 смещение из-за пропущенных переменных в, 183–190, 235–239, 324–326  
 совместные гипотезы и, 224–232  
 спецификации регрессии и, 235–241  
 альтернативная, 239, 241–242  
 базовая, 239, 241–242  
 стандартная ошибка регрессии и, 197  
 теорема Гаусса–Маркова для, 164–167, 178–181, 744–745, 774–775  
 угловые коэффициенты в, 191, 193  
 угрозы внутренней обоснованности и, 323–337  
 условия Гаусса–Маркова для, 743–745  
 частный эффект, 192
- Модель оценки финансовых активов (CAPM), 122
- Модель разрывной регрессии  
 с нечетким разрывом, 521  
 с четким разрывом, 520–521
- Модель разрывной регрессии с нечетким разрывом, 521
- Модель разрывной регрессии с четким разрывом, 520–521
- Модель регрессии с индивидуальными фиксированными эффектами, 367, 373
- Модель регрессии с усеченными переменными, 436

- Модель полиномиальной регрессии, 269–271, 280  
Модель регрессии с распределенными лагами, 621–627  
автокорреляция и, 626–627  
выводы и, 626–627  
динамическое причинное влияние и, 621–622  
обобщенный метод наименьших квадратов, 633–634  
предположения, 625–627  
приложение, 642–650  
расширение на случай множественной регрессии, 633  
с авторегрессионными ошибками, 640  
с дополнительными запаздываниями и AR(p)-ошибками, 640–642  
стандартная ошибка и, 626–627  
экзогенность и, 623–625  
Модель регрессии с распределенными лагами  
метода наименьших квадратов и, 633–634  
с ошибками в виде AR(1), 634–637  
Модель регрессии с фиксированными эффектами, 367–372  
автокорреляция в, 377–378  
большие выбросы и, 377  
временными, 372–375  
выборочное распределение и, 370  
индивидуальными, 367–368, 372–373  
индивидуальными и временными, 373–374  
МНК-оценка и, 369–370  
мультиколлинеарность и, 377  
определение, 369  
оценка и статистические выводы, 369–371  
предположения, 375–381  
серийная корреляция, 377–378  
сравнение «до и после» и, 364–367  
стандартные ошибки в, 371, 378–379, 393–397  
условное среднее и, 376  
Моменты распределения, 25–27  
Мостеллер, Фредерик, 444  
Мощность теста, 80  
Мультиколлинеарность  
в модели регрессии с фиксированными эффектами, 377  
ловушка фиктивной переменной и, 205  
несовершенная, 206–207  
решение проблемы, 206  
совершенная, 100, 204–206  
Мультиномиальные логит- и пробит-модели, 438  
Мэдриан, Бриджит, 92
- Н**
- Наблюдение, 7–8. См. также Данные в квазиэкспериментах, 524  
и экспериментальные данные, 510–512  
Наилучшая линейная несмещенная оценка. См. BLUE  
Наклон линии регрессии, 114, 115  
двухсторонняя гипотеза, 148–151  
доверительный интервал, 154–156  
нелинейной функции регрессии, 316–318  
односторонняя гипотеза, 151–152  
Налоги. См. Налоги на алкоголь; Налоги на сигареты  
Налоги на алкоголь и вождение в нетрезвом виде, 361–364  
Налоги на сигареты и курение  
инструментальные переменные и, 449–451, 457–458, 465–470  
панельные данные, 14, 484  
Натуральные логарифмы, 272–273  
в нелинейной регрессии, 272–280  
в регрессиях временных рядов, 551–553  
проценты и, 273–274  
Натуральные логарифмы, экспоненциальная функция и, 272–273  
Недоопределенные коэффициенты, 451  
Недоступная ВМНК-оценка, 717  
Недоступная ОМНК-оценка, 638, 749  
Независимая и одинаково распределенная (i.i.d.), 46, 239–241  
Независимость условного среднего, 238, 256–257  
Независимые случайные величины, 33, 114  
Некоррелированные случайные величины, 34  
Нелинейный метод наименьших квадратов, 314–315

- Нелинейная регрессия, 259–302  
 взаимодействие между двумя переменными, 281–293  
 для результатов тестов, 296–300  
 коэффициенты в, 268  
 кубическая, 270–271, 280–281  
 логарифмическая, 272–280  
 логит, 409–416  
 множественная регрессия и, 268–269  
 натуральные логарифмы в, 272  
 общая стратегия моделирования, 261–269  
 полиномиальная, 269–271  
 предсказанное значение и, 266–267, 279  
 пробит-, 404–409, 411–416  
 эффект влияния изменения  $X$  на  $Y$ , 264–268
- Неопределенность, прогноза, 569–570
- Неопределенность прогнозирования, 569–570
- Неправильная спецификация функциональной формы, 326–327
- Непрерывная случайная величина, 18  
 вероятность и моменты, 726  
 взаимодействие между, 288–293  
 взаимодействие с бинарными переменными, 284–288  
 математическое ожидание, 24  
 нормальная, 727–728  
     двумерная, 727  
 распределение, 725–728  
 условная, 772
- Неравенство Коши–Шварца, 729
- Неравенство Чебышева, 708, 728
- Несбалансированная панель, 361
- Несколько коэффициентов  
 доверительные множества для, 233–234  
 одно ограничение для, 232–233, 232–234
- Несовершенная мультиколлинеарность, 206–207
- Нестационарность  
 асимптотическое нормальное распределение и, 576, 680–681  
 структурные сдвиги и, 587–600  
 тренды и, 576–587
- Нобелевская премия по экономике, 425–426, 684–685
- Номер наблюдения, 12
- Нормальное распределение, 38–43  
 асимптотическое, 55  
 в больших выборках, 132–134, 680–681  
 двумерное, 40–43, 727  
 многомерное, 40–43  
 приближение, 49–50  
 условное, 727
- Нулевая гипотеза, 73, 149  
 $J$ -статистика и, 464–465, 468, 528, 755–756  
 в критерии Бонферрони, 253–255  
 неотвержение, 585–586  
 о наличии единичного корня, 585–586, 680  
 совместная, 224–225
- Ньюи, Уитни, 632
- О**
- Область принятия, 80
- Область отверждения, 80, 82
- Обменный курс фунта к доллару, 555–557
- Обобщенная IV-регрессия, 453–454  
 2МНК-оценка, 453–454  
 коэффициенты регрессии в, 451  
 определение, 452  
 терминология в, 452  
 экзогенность и релевантность инструментов в, 454–455
- Обобщенная ARCH-модель, 693–695
- Обобщенный метод наименьших квадратов (ОМНК)  
 в обозначениях оператора запаздывания, 559–663  
 доступный, 638–639, 749  
 интерпретация как оценки нелинейного метода наименьших квадратов, 639  
 и МНК-оценка, 641–642  
 когда  $\Omega$  содержит неизвестные параметры, 749  
 когда  $\Omega$  известна, 748–749  
 метод Кохрейна–Оркэтта, 640  
 недоступный, 638, 749  
 оценка, 633–634, 638–640, 745–751  
 преимущества и недостатки, 641–642  
 приложения, 648–650  
 предположения, 746–748

- предположение о нулевом условном среднем, 749–751  
эффективность, 639
- Обоснованность. См. Внешняя обоснованность; Допустимость инструментов; Внутренняя обоснованность
- Общий стохастический тренд, 582, 681–682
- Объясненная сумма квадратов, 123–124
- Ограниченнная зависимая переменная, 399
- Однократное распределение, 46
- Одно ограничение для нескольких коэффициентов, 232–233
- Одновременная причинность, 333–334, 347–348
- Одновременный динамический мультиплексор, 627–628
- Односторонняя гипотеза  
альтернативная, 81–82  
для коэффициента наклона, 151–153
- ОМНК-оценка. См. Оценка обобщенного метода наименьших квадратов
- Оператор запаздывания, обозначения, 611–612, 660–663
- Опросы о политических предпочтениях, 72, 331–332
- Остатки метода наименьших квадратов, 119  
во множественной регрессии, 194, 195
- Остатки  
МНК  
во множественной регрессии, 194, 195  
в модели парной регрессии, 119
- Сумма квадратов, 124
- Островершинное распределение, 28
- Отрицательный экспоненциальный рост, 313–314
- Отсутствие случайности выбора  
в квазиэкспериментах, 521–522  
в случайных контролируемых экспериментах, 497–498
- Оценка(и)  
AIC, 613, 614  
BIC, 613–614  
DOLS, 687, 688  
HAC, 631–633  
LIML, 490  
взвешенная, 166–167
- в квазиэкспериментах, 516–519  
выбор глубины запаздывания по AIC, 614  
выборочное среднее как, 70–72  
глубина запаздывания и, 613–614
- дисперсии, 69
- дисперсии Ньюи–Веста, 632–633
- наилучшая линейная несмещенная, 71, 165–166, 639
- Кохрейна–Оркэтта, 639
- линейная условно несмещенная, 165–166, 744–745
- метода инструментальных переменных. См. Оценка метода инструментальных переменных (IV)
- метода максимального правдоподобия, 413–414
- метода минимальных абсолютных отклонений, 166–167
- метода наименьших квадратов. См. Оценка метода наименьших квадратов
- модели разрывной регрессии, 520–521
- нелинейного метода наименьших квадратов, 314–315, 412–413
- несмещенная, 69, 165–166
- обобщенного метода наименьших квадратов. См. Оценка обобщенного метода наименьших квадратов (ОМНК)
- определение, 69
- разностей, 496
- «разности разностей», 516–519
- свойства, 68–70
- смещение из-за пропущенных переменных в, 183–190, 323–326
- состоительная, 69, 707–708
- стандартная ошибка как, 77
- стандартная ошибка регрессии как, 124
- эффективность, 69, 70, 71, 160–161, 164, 165–166, 744–745, 775–777
- Оценка взвешенного метода наименьших квадратов, 716–721  
доступная (реализуемая), 718  
и устойчивые к гетероскедастичности  
стандартные ошибки, 720–721  
недоступная (нереализуемая), 718
- определение, 717
- Оценка взвешенной регрессии, 165–166

- Оценка глубины запаздывания, 613–614  
 Оценка дисперсии Ньюи–Веста, 632–633  
 Оценка Кохрейна–Оркэтта, 639  
 Оценка максимального правдоподобия при ограниченной информации (LIML), 490  
 Оценка метода наименьших квадратов, 71–72, 107. См. также Оценка обобщенного метода наименьших квадратов; Оценка метода наименьших квадратов взвешенного, 716–721  
     нелинейного, 314–315, 412–413  
 Оценка минимальных абсолютных отклонений, 166–167  
 Оценка нелинейного метода наименьших квадратов, 314–315, 412–413  
 Оценка обобщенного метода наименьших квадратов (GMM), 756–758  
     эффективность, 777  
 Оценка объединенной дисперсии, 91–93  
 Оценка программных документов, 493  
 Оценка разностей, 496  
     с дополнительными регрессорами, 496  
 Оценка «разности разностей», 516–519  
     для повторяющихся межобъектных данных, 519  
     с дополнительными регрессорами, 518  
 Оценки регрессии. См. Оценка(и)  
 Оценки (численное значение), определение, 69  
 Ошибка I рода, 80  
 Ошибка II рода, 80  
 Ошибка (остаточный член), 114, 115  
     в AR( $p$ )-модели, 559–560  
     гетероскедастичная/гомоскедастичная, 158–164, 193  
     корреляция между наблюдениями, 336, 348  
     серийно коррелированная, 633  
 Ошибка прогнозирования (прогноза), 558  
     в AR ( $p$ ) -модели, 559–560  
     квадратный корень из среднеквадратичной, 560, 570  
     псевдовневыборочный прогноз и, 594–595  
     псевдо, 592–599  
 Ошибка регрессии. См. Ошибка (остаточный член)
- Ошибки измерения, 327–330  
     в  $Y$  vs.  $X$ , 328  
     классическая модель с, 328
- П**
- Паевые фонды, 72, 331–332  
 Панельные данные, 13–14, 359–386  
     анализ «до и после» и, 364–367, 370  
     вождение в нетрезвом виде и, 361–364  
     в регрессии с фиксированными эффектами, 376–372, 370  
     выборочное распределение и, 370  
     обозначения, 360  
     определение, 360  
     сбалансированные, 361  
 Параметр усечения НАС, 632, 646  
 Параметр усечения для НАС-стандартных ошибок, 632, 646  
 Параметры модели линейной регрессии, 113–114  
 Парная линейная регрессия, 111–135, 703–729  
     асимптотическое распределение и, 706–714. См. также Асимптотическое распределение  
     доверительные интервалы и, 147–182  
      проверка гипотез и, 147–182  
     распределение ошибок и, 714–716  
     расширенные предположения метода наименьших квадратов и, 704–706  
      $t$ -статистики в предположении гомоскедастичности в, 716  
 Пенсионные накопления, 92–93  
 Первое запаздывание (лаг), 551–552  
 Первые разности, во временных рядах, 551–553  
 Переопределенные коэффициенты, 451  
 Переменные. См. также Регрессоры  
     бинарные зависимые, 398–438  
     дамми (фиктивные), 156  
     дискретного выбора, 438  
     зависимые, 113–114  
     индикаторные, 156  
     инструментальные. См. Инструментальные переменные  
     интересующие, и контрольные переменные, 235–239  
     контрольные, 191, 235–239, 451–452  
     множественного выбора, 438

- независимые, 113–114  
взаимодействие между, 281–293  
непрерывные случайные. См. Непрерывная случайная величина  
ограниченные зависимые, 399. См. также Модель регрессии с ограниченной зависимой переменной  
эндогенные, 440–441, 453–454  
экзогенные, 440–441, 450, 451–453, 461–465  
Плотность, 21, 21–22  
Поведенческая экономика, 92–93  
Повторяющиеся межобъектные данные, 519  
Полная сумма квадратов, 123–124  
Порядок интегрированности, 674–677  
Постоянный член во множественной регрессии, 192  
Потенциальный исход, 494–495  
Правильно предсказанная доля наблюдений, 415  
Предположение о нулевом условном среднем, ОМНК-оценка и, 749–751  
Предположения метода наименьших квадратов, 236–237  
во множественной регрессии, 200–202  
в модели парной регрессии, 126–131  
выбросы маловероятны, 128–129, 200–201  
отсутствие совершенной мультиколлинеарности, 201  
расширенные, 704–705, 732–734  
регressоры независимо и одинаково распределены, 128–129, 200  
условное распределение имеет нулевое среднее, 126–128, 200  
Предположениями модели нормальной линейной регрессии с гомоскедастичными ошибками, 167  
Предсказанное при помощи метода наименьших квадратов значение, 118–119  
в нелинейной регрессии, 266–267, 279  
во множественной регрессии, 194, 195 и прогноз, 558  
Предсказанное значение, МНК  
в нелинейной регрессии, 266–267, 279  
во множественной регрессии, 194, 195 и прогноз, 558  
Приближение, большая выборка. См. Приближение в больших выборках  
Приведенная форма уравнения, 453  
Причинность по Грейндджеру, 568–569  
Причинный эффект. См. также Эффект воздействия (условий эксперимента)  
временные ряды и, 620–621  
гетерогенный, 524–530, 540–542  
динамический, 615–663. См. также Динамическое причинное влияние  
одновременная причинность и, 333–334  
определение, 8, 87, 524, 620  
оценка, 8–9, 86–87, 497  
разности между двумя средними, 86–87  
средний, 495  
Пробит-модель, 404–409, 411–416  
и линейная вероятностная модель, 411–412  
и логит-модель, 409–410, 411–412  
критерии качества приближения данных моделью в, 415–416  
мультиномиальная, 437  
определения, 407  
оценка коэффициентов и, 409  
оценка метода максимального правдоподобия и, 413–424, 434  
оценка нелинейного метода наименьших квадратов и, 412–413  
приложения, 408–409  
с несколькими regressорами, 406–407  
упорядоченная, 437  
эффект влияния изменения  $X$ , 406–407  
Прогноз  
инерционный, 561–563  
и предсказанное значение, 558  
многошаговый, 669–674  
Прогнозирование  
в векторных авторегрессиях, 664–669  
временные ряды и, 550–557  
в модели авторегрессии, 557–563  
внутренняя/внешняя обоснованность и, 337–339  
в моделях регрессии, 337–339, 548–550  
доходностей, 561–563  
инфляции, 6–7, 559–560, 560, 563–566

- ошибка прогнозирования и, 558  
причинность и, 9  
псевдовневыборочное, 592–599
- Прогнозирование инфляции, 6–7, 559–560, 560, 563–566, 571–572  
безработица и, 550–551, 553, 563–566, 668–669  
кривая Филлипса и, 7, 592, 595–599  
монетарная политика и, 652–653  
уровень цен и, 676, 677  
цены на нефть и, 652
- Проект STAR, база данных, 540
- Простая случайная выборка, 46
- Пространство элементарных событий, 18
- Процентные ставки  
векторная модель коррекции ошибками и, 684–685, 690–691  
временная структура, теория ожиданий, 683–684  
тесты на наличие единичных корней и, 689–691  
тест на отсутствие коинтеграции и, 689–691
- Проценты, логарифмы и, 273–274
- Прямой многошаговый прогноз, 672–674  
и итеративный многошаговый прогноз, 674
- Псевдовневыборочное прогнозирование, 592–599
- Псевдо- $R^2$ , 416, 435
- P**
- Размер класса. См. Соотношение учеников и учителей
- Размер выборки, 502
- Разность средних  
в оценке причинных эффектов, 86–87  
доверительный интервал, 86  
проверка гипотезы, 84–86  
 $t$ -статистика для проверки гипотезы, 91, 93
- Размер теста, 80
- Райт, Сьюэлл, 442, 443–444
- Райт, Филипп, 442, 443–444, 449
- Расовая принадлежность и ипотечное кредитование. См. Ипотечное кредитование и расовая принадлежность
- Распределение(я)  
 $F$ -статистики, 743, 773–774  
GMM J-статистики, 777  
 $t$ -статистики, 78, 713, 742–743, 772–773  
асимптотическое. См. Асимптотическое распределение  
Бернулли, 20  
асимптотически нормальное, 131–134  
безусловное распределение вероятностей, 29, 30  
вероятности. См. Функция плотности в конечных выборках, 50  
выборочное. См. Выборочное распределение  
коэффициент асимметрии, 26–27  
кумулятивное. См. Кумулятивное распределение  
линия регрессии и, 126–127  
моменты, 26–27  
МНК-оценки, 202–203  
многомерное, 40–43, 770–771  
независимо, 33, 46–47  
независимо и одинаково (i.i.d.), 46–47, 240–241  
непрерывное случайное, 20–21, 725–728  
нормальное, 38–43  
асимптотическое, 55  
в больших выборках, 131–134, 680–681  
двуухмерное, 40–43, 727  
приближение, 50  
условное, 727  
одинаковое, 6  
островершинное, 28  
условное, 30–31  
совместное, 28–30  
статистик регрессии с нормальными ошибками  
Стьюдента  $t$ , 44–45, 728  
сходимость по, 709  
теоретической дисперсии,  $p$ -значение и, 75–76  
точное, 50  
функция правдоподобия и, 414  
хи-квадрат, 43–44, 728  
эксцесс (коэффициент эксцесса), 27, 27–28
- Распределение Бернулли, 20
- Распределение в конечных выборках, 50
- Распределение вероятности. См. также Распределение(я)

- выборочного среднего, 47–49  
дискретной случайной величины, 18–20  
кумулятивное, 18–22  
непрерывной случайной величины, 20  
нормальной случайной величины, 38–43  
определение, 18
- Распределение хи-квадрат, 43, 728
- Расширенная статистика Дики–Фуллера, 583–586  
    критические значения, 585, 586
- Расширенные предположения метода наименьших квадратов, 704–705, 732
- Расширенный тест Дики–Фуллера, 584–586, 677–679  
    Энгл–Грейндженер, 584–586
- Регрессионные модели с цензурированными переменными, 436
- Регрессия  
    без ограничений, 229–230  
    в квазиразностях, 635  
    в прогнозировании, 337–339, 548–549. См. также Прогнозирование временных рядов, 547–614  
    Дики–Фуллера, 582–586  
    кажущаяся, 581  
    квадратичная, 262–264  
    линейная. См. Линейная регрессия  
    логарифмическая, 274–280  
    логистическая, 313, 409–416  
    множественная. См. Множественная регрессия  
    нелинейная, 259–302, 265, 312–318.  
        См. также Нелинейная регрессия  
    не первом шаге, 454  
    одно ограничение для нескольких коэффициентов и, 232–233  
    отрицательного экспоненциального роста, 313–314, 316  
    с бинарной объясняющей переменной, 156–158  
    с временными фиксированными эффектами, 372–375  
    с индивидуальными фиксированными эффектами, 367, 373  
    с инструментальными переменными, 439–492  
    с компонентой взаимодействия, 282, 287  
    с ограничениями, 229–230  
    с усеченными переменными, 436  
    с цензурированными переменными, 436  
    с фиксированными эффектами, 367–371  
    стандартная ошибка, 124–125  
    тобит-, 436  
    формирования выборки, 436
- Регрессия без ограничений, 229–230
- Регрессия временных рядов, 547–614  
    авторегрессия первого порядка и, 557–560  
    авторегрессия порядка  $p$  и, 557–560  
    авторегрессионные модели с условной гетероскедастичностью и, 684, 693–694  
    авторегрессионные модели с распределенными лагами, 565–566  
    в прогнозировании. См. Прогнозирование  
    векторная авторегрессия и, 664–669  
    данные для. См. Временные ряды  
    кластеризованная волатильность и, 691–692  
    коинтеграция и, 681–691  
    порядок интегрированности и, 674–677  
    предположения, 566–568  
    приложения, 642–650  
    причинность по Грейндженеру и, 566  
    с несколькими объясняющими переменными, 566–569  
    слабая зависимость и, 567–568  
    стационарность и, 566–567  
    структурные сдвиги в, 587–600  
    тренды в, 576–587  
    тесты на наличие единичных корней и, 677–681  
    условное среднее и, 566  
    AR( $p$ )-модель и, 557–560
- Регрессия Дики–Фуллера и ненормальное распределение, 679–681
- Регрессия, оцененная на втором шаге, 454
- Регрессия, оцененная на первом шаге, 454
- Регрессия с инструментальными переменными (IV), 439–492  
    2МНК-оценка в, 441–442, 453–458. См. также Оценка двухшагового метода наименьших квадратов (2МНК)

- для эффекта воздействия в эксперименте, 499  
 доверительное множество для, 489–490  
 допустимость инструмента и, 441, 450, 455–465  
 обобщенная модель для. См. также Тестирование гипотез в обобщенной IV-регрессии, 489–490  
 переменные в, 451–458  
 приложения, 442–447, 449–451, 465–470, 526–530  
 предположения, 440–441, 455–456  
 причинный эффект с гетерогенной генеральной совокупности и, 540–542  
 развитие метода, 443–444  
 релевантность инструмента и, 441, 450, 455, 459–461  
 слабые инструменты и, 459–461, 462–463, 488–491  
 с одним регрессором и одним инструментом, 440–451  
 экзогенность инструмента и, 441, 450, 455  
 эндогенные и экзогенные переменные и, 440–441  
 Регрессия с ограничениями, 229–232  
 Регрессия с фиксированными временными эффектами, 372–375  
 Регрессор взаимодействия, 282  
 Регрессоры, 113. См. также Постоянный член, 192  
     единицы измерения переменных, 243  
 Результаты лечения сердечных приступов, 474–477, 515, 529–530  
 Результаты тестов  
     соотношение учеников и учителей и. См. Соотношение числа учеников и учителей и результаты тестов  
 Река крови, 571–572  
 Ролл, Ричард, 651–652  
 Рузвельт, Франклин, 72
- С**  
 Сбалансированная панель, 360  
 Сводный фондовый индекс Нью-Йоркской фондовой биржи (NYSE), 556–557  
 Серийная корреляция, 377–378, 554, 555  
 Стабильность коэффициентов, QRL-тест для, 589–591  
 Стандартизированная случайная величина, 38–39  
 Стандартная ошибка  $\hat{\beta}_1$ , 149  
 Стандартное нормальное распределение, 38, 40f  
 Стандартная ошибка регрессии во множественной регрессии, 197  
     в парной регрессии, 124–126  
 Стандартное отклонение, 24–25  
     выборочное, 77  
     дисперсия и, 24–25  
     определение, 24  
 Стандартные ошибки, 78–79  
     в модели с распределенными лагами, 626–627  
     в матричной форме, 741  
     в нелинейной регрессии, 267–268  
     в парной линейной регрессии, 714–715  
     в предположении гомоскедастичности, 161, 164, 177–178, 742–743  
     в прямых многошаговых прогнозах, 672–673  
     в регрессии с фиксированными эффектами, 370, 378–379, 393–397  
     для МНК-оценки, 149, 158–164, 176–177, 219–220, 335–336, 629–631  
     для 2МНК-оценки, 457, 752–753  
     для предсказанных эффектов, 737  
     для предсказанных вероятностей, 435  
     кластеризованные, 378–379  
     НАС, 378–379, 619, 631–633, 634, 673  
     несостоительные, 335–336  
     определение, 77  
     при наличии автокорреляции, 629–631  
     состоительные при наличии гетероскедастичности, 629, 631–633  
     устойчивые при наличии гетероскедастичности, 176, 335, 336, 712–713, 720–721, 737  
     AR( $p$ ), 640–642  
 Стационарность, 566–568  
     в AR(1)-модели, 610–611  
 Скорректированный  $R^2$ , 197–199, 240–241  
 Слабая зависимость, 568–569  
 Слабые инструменты, 459–461, 462–463, 488–491  
 Служба в армии и уровень заработных плат гражданских лиц, 513–514, 524

- Случайная величина Бернулли, 20  
дисперсия, 25  
оценка максимального правдоподобия, 433
- Случайная выборка, 45–46  
в оценках, 72  
из генеральной совокупности, 45–46  
простая, 46  
i.i.d., 46
- Случайное блуждание, 578, 586, 675  
с дрейфом, 578
- Случайные величины, 18  
Бернулли. См. Случайная величина Бернулли  
дискретные, 18, 18–19  
корреляция между, 34–35  
ковариация между, 34  
линейная функция от, среднее и дисперсия, 25–26  
математическое ожидание, 22–23  
моменты, 26–27  
непрерывные, 18  
независимые, 34  
плотность распределения, 18–22  
стандартное отклонение, 24–25  
среднее значение, 22, 31–32  
сумма, среднее и дисперсия, 35–38  
условная дисперсия, 33  
условное математическое ожидание, 31–32  
 $r$ -й момент, 28
- Случайные контролируемые эксперименты, 8–9  
анализ данных в, 496–497, 505–510, 512, 542–543  
внешняя обоснованность, 502–503  
внутренняя обоснованность, 497–502  
гетерогенные генеральные совокупности и, 524–530, 540–542  
двойной слепой, 501  
для уменьшение размера класса, 510–512  
истощение выборки и, 500  
наблюдаемые и экспериментальные оценки, 510–512  
нерепрезентативные программы/политики в, 502–503  
нерепрезентативные выборки в, 502  
отсутствие случайности выбора, 497–499
- потенциальные исходы и, 494–496  
пример, 503–512  
размер выборки в, 502  
случайный выбор, зависящий от наблюдаемых переменных, и, 497  
сравнение оценок, на экспериментальных и наблюдаемых данных, 510–512  
средний причинный эффект и, 495  
частичное соответствие и, 499–500  
эффекты общего равновесия в, 503  
эффект Хоторна в, 500–501
- Случайный выбор, зависящий от независимых переменных, 497
- Смертность в ДТП. См. Вождение в нетрезвом виде
- Смещение  
в авторегрессии, 580  
в МНК-оценках, 183–190, 333–334, 744  
в оценках, 69, 70, 72  
вследствие выживаемости, 332–333  
из-за ошибок в переменных, 327–330, 347  
из-за пропущенных переменных, 346–347, 646–647  
из-за отбора наблюдений. См. Смещение из-за отбора наблюдений  
одновременная причинность и, 332–334, 347–348  
одновременные уравнения, 335–336  
состоительность и, 69–70, 707–709
- Смещение вследствие выживаемости, 332–333
- Смещение из-за отбора наблюдений, 331, 347  
в опросах, 72, 331  
паевые фонды и, 72, 331
- Смещение из-за ошибок в переменных, 327–330, 347
- Смещение из-за пропущенных переменных, 183–190, 196, 323–326, 346–347, 646–647  
внутренняя обоснованность и, 323–326  
во множественной регрессии, 183–190, 235–239, 241–246  
количество переменных и, 326  
определение, 184, 323  
предположения метода наименьших квадратов и, 186

- приложения, 187–188, 241–246  
 примеры, 184–185  
 решение проблемы, 323–326  
 формула, 186–188, 216  
 эффект Моцарта и, 187–188  
 Смещение одновременных уравнений, 335–336  
 Событие  
     вероятность, 18–19  
     определение, 18  
 Совершенная мультиколлинеарность, 100, 204–206  
     ловушка фиктивной переменной и, 205  
     решение проблемы, 206  
 Совместные гипотезы  
     в матричной форме, 738–739  
     для нескольких коэффициентов, 224–226, 232–233  
     для одного ограничения для нескольких переменных, 232–233  
     критерий Бонферонни, 253–256  
     нулевая, 224–225  
     одновременное тестирование гипотез для отдельных коэффициентов, 225–226  
     определение, 225  
     приложения, 228–229  
     при помощи *t*-статистики, 225–226  
     с помощью *F*-статистики, 226–228, 739  
     совместная нулевая гипотеза и, 224–225  
     тестирование, 224–225, 224–233, 739–740  
 Совместное распределение вероятности, 28–30  
     функция правдоподобия и, 413–414  
 Совокупный динамический мультипликатор, 627–628, 643–646  
     долгосрочный, 628  
 Соответствие, частичное, 499–500, 522–523  
 Соотношение числа учеников и учителей и результаты тестов, 4–5, 9–10  
     внешняя обоснованность и, 322, 339–346  
     внутренняя обоснованность и, 346–348  
     данные по Калифорнии, 11–12, 184  
     данные по штату Массачусетс, 355–356  
     данные по штату Теннесси, 10  
     доходы в округе и, 261–264  
     линейная регрессия и, 119–120, 125–130  
     МНК-оценки, 119–120, 195–196  
     не говорящие по-английски и, 280, 283–284, 292–293  
     нелинейные эффекты, 293–301  
     регрессионный анализ, 168–169, 195–196, 199, 241–246  
     смещение из-за пропущенных переменных и, 183–190, 241–246  
     экспериментальные оценки, 503–510  
     эксперимент в штате Теннесси, 503–513  
 Состоятельность, 51, 707–709  
     закон больших чисел и, 51–52, 707–709  
     оценки, 69, 70  
     смещение и, 69–70, 707–709  
     устойчивых к гетероскедастичности стандартных ошибок, 711–712  
     МНК-оценок, 132–133, 711  
     выборочной дисперсии, 77–78, 107–108  
 Состоятельная при наличии гетероскедастичности и автокорреляции. См. НАС-оценка  
 Состоятельная при наличии гетероскедастичности и автокорреляции стандартная ошибка. См. НАС-стандартная ошибка  
 Состоятельная при наличии гетероскедастичности стандартная ошибка, 628, 631–633  
 Состоятельные оценки, 707, 708–709  
 Спецификация модели, 235–241  
     альтернативная, 239, 241–242  
     базовая, 239, 241–242  
 Спецификация регрессии, 235–241  
     альтернативная, 239, 241–242  
     базовая, 239, 241–242  
 Сравнение «до и после», 364–367, 370  
 Сравнение теоретических средних, 84–89. См. также Разность средних  
     доверительные интервалы, 82–84  
     оценка, 68–73  
     проверка гипотез, 73–82, 148–149  
 Средний причинный эффект, 495  
 Средний эффект воздействия (условий эксперимента), 495, 527–530

- Ставка процента по федеральным фондам США, 555, 556–557  
Стандартная ошибка для «пула», 91–93  
Стандартная ошибка, рассчитанная при условии гомоскедастичности, 161, 164, 176–178, 742  
Статистика Вальда, 743–744  
Статистика Дики–Фуллера, 582  
расширенная, 582–586  
Статистика отношения правдоподобия Куандта (QLR), 589–592  
Стилометрия, 443–444  
Стохастические тренды, 577–578  
авторегрессия и, 578–579  
единичный корень и, 579  
коинтеграция и, 681–691  
ложная регрессия и, 581–582  
ненормальное распределение  $t$ -статистики и, 580–581  
общие, 582, 681  
обнаружение, 582–586  
порядок интегрированности и, 674–677  
проблемы, вызванные, 579–582, 586–587  
тест Дики–Фуллера, 582–583, 584  
Строгая экзогенность, 623–625  
Структурные VAR-модели, 666–668  
Структурный сдвиг, 588–592  
определение, 587–588  
проблемы, вызываемые, 588  
тестирование, 588–592  
Сумма квадратов  
объясненная, 123–124  
полная, 123–124  
Сумма квадратов остатков, 124  
Супремум-статистика Вальда, 589–592  
Сходимость  
по распределению, 709  
по вероятности, 50, 707–709  
Счетные данные, 437–438
- Т**  
Талеб, Насим, 42  
Текущее обследование населения, 73–74  
Текущее обследование населения США, 73–74, 106  
Телефонные опросы, 72, 331  
Темпы роста, регрессия временных рядов и, 551–553  
Теорема Гаусса–Маркова, 164–167, 178–181, 744–745  
доказательство, 774–775  
Теорема о непрерывном отображении, 710  
Теорема Слуцкого, 710  
Теорема Фриша–Во, 217–218  
Теоретическая модель множественной регрессии, 190–193  
Теоретические коэффициенты, 114, 116–118  
дисперсия,  $p$ -значение и, 75–76  
Теоретический свободный член, 114, 115  
Теоретическая функция регрессии, 114, 115  
структурные сдвиги и, 587  
оценка, 265–266  
Теория ожиданий временной структуры процентных ставок, 683–685  
Тест Андересона–Рубина, 489–490  
Тест Бонферонни, 253–256  
Тест Дики–Фуллера, 582–583, 678. См. также DF-GLS-тест  
и DF-GLS-тест, 678, 679  
расширенный, 584–586, 678  
Тест на сверхидентифицирующие ограничения, 464–465  
приложения, 468, 529  
Тест на случайность распределения между экспериментальной и контрольной группами, 498–499  
Тест отношения условного правдоподобия, 489–490  
Тест Чоу, 589–592  
Тестирование гипотез, 74–82  
вероятность попадания и, 84  
в регрессии с бинарными объясняющими переменными, 156–157  
выборочное стандартное отклонение и, 77  
выборочная дисперсия и, 77–78  
для генерального среднего, 148–149  
для нескольких коэффициентов, 224–232  
для одного коэффициента, 219–221  
для регрессии с инструментальными переменными, 489–491  
для свободного члена, 153–154  
для совместных гипотез, 224–233, 738–740

- доверительные интервалы в, 86  
 критические значения в, 80  
 мощность теста в, 80  
 на определенном уровне значимости, 79–81  
 область принятия в, 80  
 область отвержения в, 80, 82  
 ошибки I/II рода в, 80  
 размер критерия в, 80  
 разности средних, 84–86  
 слабые инструменты и, 489–491  
 стандартная ошибка и, 76–78  
 терминология, 80  
 уровень значимости и, 80–81  
 число степеней свободы и, 76  
 $p$ -значение и, 74–77, 79  
 $t$ -статистика в, 78–79, 80, 89–93
- Тестовая статистика, 78, 80  
 критическое значение, 80
- Тесты на наличие единичных корней  
 DF-GLS-тест, 677–680  
 Дики–Фуллера (DF), 582–586, 677, 679, 680  
 ненормальное распределение и, 680–681  
 приложения, 689–691  
 расширенный тест Дики–Фуллера (ADF), 584–586, 677–678
- Тобин, Джеймс, 436  
 Тобит-регрессия, 436  
 Точно определенные коэффициенты, 451  
 Точное распределение, 49–50
- Тренд, 576–587  
 гладкий, 675  
 детерминированный, 576  
 модель случайного блуждания и, 578, 586, 675  
 общий, 581, 681  
 определение, 576  
 порядок интегрированности и, 675–677  
 стохастический. См. Стохастические тренды
- Тюремное заключение и снижение преступности, 471–473
- у**  
 Угловой коэффициент во множественной регрессии, 191  
 Уоллес, Дэвид, 444  
 Упорядоченная пробит-модель, 438
- Упорядоченные данные, 438  
 Уровень безработицы и минимальная заработная плата, 516–517  
 Уровень доверия, 82, 154  
 Уровень значимости, 80  
 Уровень цен и инфляция, 676, 677  
 Уровень преступности и тюремные заключения, 471–473  
 Условие релевантности инструмента, 441, 450, 455, 459–461  
 Условие экзогенности инструмента, 441, 450, 455, 461–465  
 Условия Гаусса–Маркова, 743–745  
 Условная дисперсия, 33  
 Условное математическое ожидание, 31–32  
 разностей, 87  
 Условное распределение, 30–31  
 линия регрессии и, 126–127
- Условное среднее, 31–32, 127  
 в модели регрессии с фиксированными эффектами, 375–376  
 в моделях временных рядов, 566–567  
 корреляция, 34
- Устойчивая при гетероскедастичности  $F$ -статистика, 227  
 Устойчивая при гетероскедастичности  $J$ -статистика, 758  
 Устойчивая при гетероскедастичности стандартная ошибка, 176–177, 335, 336  
 асимптотическое распределение и, 711–712  
 во множественной регрессии, 737  
 и оценка взвешенным методом наименьших квадратов, 720–721  
 состоятельность, 711–712
- Ф**  
 Федеральный резервный банк Бостона, 5  
 Федеральная резервная система, 6  
 Финансовая диверсификация, 48–49  
 Фиксированные эффекты  
 временные, 372–373  
 индивидуальные, 367–368  
 Фиксированные временные эффекты, 372–373  
 Фондовый рынок США. См. Фондовый рынок  
 Фондовый рынок

- авторегрессионная модель, 561–563, 595–597  
«бета»-акции и, 121–122  
диверсификация и, 48–49  
кластеризованная волатильность и, 691–695  
модель оценки финансовых активов и, 121–122  
прогнозирование доходностей и, 561–563  
процентное изменение стоимости, 41–42  
смещение вследствие выживаемости паевых фондов, 72, 331  
сводный индекс NYSE, 556–557  
Формула дисперсии НАС, 631–633  
Формула дисперсии при гомоскедастичных ошибках, 161, 177–178  
Формулы оценок метода наименьших квадратов, 118–119, 121–122  
Фуллер, Уэйн, 582  
Функция выборочной регрессии, 119  
Функция нелинейной регрессии, 265  
для одной независимой переменной (парная нелинейная регрессия), 269–281  
как линейная функция от неизвестных параметров, 312–316  
многочлены в, 269–271  
наклон и эластичность, 316–318  
экспоненциальная, 272–273  
Функция плотности, 21, 22  
Функция плотности вероятности нормального распределения, 726  
двумерного, 727  
Функция плотности вероятности (p.d.f.), 21, 22  
двумерного нормального распределения, 727  
нормального распределения, 38–43, 727  
Функция правдоподобия, 413  
Функция распределения (c.d.f.), 20
- Х**  
Хекман, Джеймс, 425–426
- Ц**  
Целевая генеральная совокупность, 320  
Цены на апельсиновый сок и заморозки, 616–620, 651–652
- база данных, 660  
модель регрессии временных рядов и, 642–650  
Центральная предельная теорема, 52–55, 132–133, 709–710, 710  
многомерная, 735–736  
Центральная предельная теорема для многомерной случайной величины, 735–736
- Ч**  
Частичное соответствие, 499–500  
в квазиэкспериментах, 522–523  
Частичное соответствие, 499–500  
случайность «назначения воздействия», 499  
Черный лебедь (Талеб), 42  
Число степеней свободы, 77, 716
- Ш**  
Ши, Деннис, 92–93
- Э**  
Экзогенность, 623–625  
правдоподобность, 650–652  
строгая, 623–625  
Экзогенные инструменты, слабые, 459–461, 462–463, 488–491  
Экзогенные регрессоры, в оценках динамического причинного влияния, 625–628  
Экзогенные переменные, 440–441, 450, 451–453, 461–465  
включенные, 451–453  
в IV-регрессии, 440–441, 450, 451–453, 461–465  
определение, 440  
тест на сверхидентифицирующие ограничения и, 464–465  
Экономика, поведенческая, 92–93  
Экономические временные ряды. См. Временные ряды  
Экономические журналы, спрос на, 293–296  
Экспериментальные данные, 10. См. также Данные  
Эксперименты. См. Квазиэксперименты; Случайные управляемые эксперименты  
Экспоненциальный рост, 313–314, 316  
Экспоненциальная функция, натуральный логарифм и, 272–273  
Эндогенные переменные, 440–441

- в обобщенной модели регрессии с инструментальными переменными, 453–454  
определение, 440  
слабые инструменты и, 459–461, 462–463, 488–491
- Эластичность  
в нелинейных регрессионных моделях, 316–318  
предложения, 443–445  
спроса, 5–6, 272, 443–445, 449–450
- Эластичность спроса  
определение, 443  
цена и, 5–6, 272, 443–445, 449–450
- Эластичность спроса по цене, 5–6, 272
- Элементы математической статистики, 67–98
- Эластичность предложения, 443–445
- Элементарные события, 18
- вероятность, 18. См. также Вероятность  
определенение, 18  
Энгл, Роберт, 684–685  
Энгрист, Джошуа, 462, 514, 523, 524  
Эффект воздействия (импульсный эффект), 628  
Эффект воздействия в эксперименте, 87. См. также Причинный эффект  
локальный средний, 527–530  
оценка метода инструментальных переменных, 499  
средний, 527–530  
Эффект Моцарта, 187–188  
Эффект Хоторна, 500–501  
Эффективные оценки, 69, 70, 71, 160–161, 164, 744–745, 775–777  
Эффективная GMM-оценка, 756–759

*Учебная литература*

Джеймс Сток, Марк Уотсон

**Введение в эконометрику**

Серия «Академический учебник»

Выпускающий редактор *Е. В. Попова*

Редактор *М. А. Уварова*

Корректор *Г. А. Лакеева*

Художник *Е. Н. Спасская*

Оригинал-макет *О. З. Элоев*

Верстка *Ю. А. Тумакова*

Подписано в печать 19.06.2015. Формат 70x108 1/16

Гарнитура PT Sans. Усл. печ. л. 75,6.

Тираж 2000 экз. Изд № 1052. Заказ №

Издательский дом «Дело» РАНХиГС  
119571, Москва, пр-т Вернадского, 82

Коммерческий отдел – тел. (495) 433-2510, (495) 433 2502

[delo@rane.ru](mailto:delo@rane.ru)

[www.ranepa.ru](http://www.ranepa.ru)