
ЭКОНОМЕТРИКА

УЧЕБНИК

Под редакцией
члена-корреспондента РАН
И. И. Елисеевой

*Рекомендовано УМО по образованию в области статистики в качестве
учебника для студентов высших учебных заведений по специальности
080601 «Статистика» и другим междисциплинарным специальностям*



УДК 330.43(075.8)

ББК 65вбя73

Э40

Авторы:

И. И. Елисеева — чл.-корр. РАН, д-р экон. наук, проф. (введение, гл. 1, 2); С. В. Курышева — д-р экон. наук, проф. (гл. 6, 7, 8, 9.1—9.3, 9.4.1); Ю. В. Нерадовская — канд. экон. наук, доц. (гл. 3, 4, 9.4.2, 9.4.3, приложение); Д. К. Павелеску — ассистент (гл. 5); Ю. В. Лемешко — ассистент (гл. 10).

Рецензенты:

кафедра исследования операций в экономике им. Ю. А. Львова Санкт-Петербургского государственного инженерно-экономического университета (ИНЖЭКОН);

В. С. Мхитарян, д-р экон. наук, проф., зав. кафедрой математической статистики и эконометрики Московского государственного университета экономики, статистики и информатики (МЭСИ).

Под редакцией члена-корреспондента РАН И.И. Елисеевой.

Эконометрика: учеб. / под ред. И. И. Елисеевой. — М.
Э40 Проспект, 2009. — 288 с.

ISBN 978-5-392-00186-6

Учебник охватывает все основные разделы современного курса эконометрики для студентов экономических специальностей, включая исследование парной регрессии, множественной регрессии, использование фиктивных переменных, модели дискретного выбора. Детально рассматривается экономический анализ временных рядов: моделирование отдельного временного ряда, построение моделей на основе системы временных рядов, модели с лаговыми переменными. Излагаются системы одновременных уравнений, модели анализа панельных данных, ARCH и GARCH модели.

Для студентов экономических специальностей высших учебных заведений.

УДК 330.43(075.8)
ББК 65вбя73

Учебное издание

Елисеева Ирина Ильинична,
Курышева Светлана Владимировна,
Нерадовская Юлия Владимировна
и др.

ЭКОНОМЕТРИКА

Учебник

Подписано в печать 01.08.08. Формат 60 × 90^{1/16}.
Печать офсетная. Печ. л. 18.0. Тираж 3000 экз. Заказ № 9520.

ООО «Проспект»
111020, г. Москва, ул. Боровая, д. 7, стр. 4.

Отпечатано с готовых диапозитивов в ОАО ордена «Знак Почета»
«Смоленская областная типография им. В. И. Смирнова».
214000, г. Смоленск, проспект им. Ю. Гагарина, 2.



9 785392 001866

© Коллектив авторов, 2009
© ООО «Издательство Проспект», 2009

ОГЛАВЛЕНИЕ

Введение	5
Глава 1. Понятие эконометрики. Условия ее применения	6
1.1. Определение эконометрики. Предпосылки возникновения и этапы развития	6
1.2. Методы оценивания	11
1.3. Испытание гипотез и доверительные интервалы	16
Глава 2. Парная регрессия	21
2.1. Основные понятия	21
2.2. Оценка параметров регрессии методом наименьших квадратов	24
2.3. Свойства остатков	28
Глава 3. Множественная регрессия	35
3.1. Спецификация модели	35
3.2. Натуральная и стандартизованная форма модели множественной регрессии	41
3.3. Оценка параметров уравнения множественной регрессии	42
3.4. Показатели силы связи в модели множественной регрессии	48
3.5. Изучение тесноты связи на основе множественной регрессии	54
3.6. Оценка значимости модели множественной регрессии и ее параметров	66
3.7. Прогнозирование по модели множественной регрессии	75
3.8. Анализ случайных остатков в модели регрессии	76
3.9. Обобщенный метод наименьших квадратов	82
Глава 4. Фиктивные переменные	91
4.1. Особенности включения в модели регрессии неколичественных показателей	91
4.2. Спецификация моделей регрессии с фиктивными независимыми переменными	92
4.3. Модели регрессии с фиктивными переменными сдвига	92
4.4. Модели регрессии с фиктивными переменными наклона	95
4.5. Общий вид модели регрессии с фиктивными переменными	98
4.6. Исследование структурных изменений с помощью теста Чоу	103
Глава 5. Модели дискретного выбора	106
5.1. Модели бинарного выбора. Использование линейной регрессии	106
5.2. Модели множественного выбора	114

Глава 6. Моделирование изолированного динамического ряда	128
6.1. Компоненты динамического ряда	128
6.2. Автокорреляция уровней динамического ряда и характеристика его структуры	132
6.3. Модели тенденции развития	138
6.4. Моделирование периодических колебаний	159
Глава 7. Модели регрессии по временным рядам	183
7.1. Специфика изучения взаимосвязей по рядам динамики	183
7.2. Учет тенденции при построении модели регрессии	184
7.3. Обобщенный метод наименьших квадратов при построении модели регрессии по временным рядам	193
7.4. Учет сезонности при построении модели регрессии	200
Глава 8. Модели с лаговыми переменными	205
8.1. Общая характеристика	205
8.2. Модели с распределенными лагами	207
8.3. Модели авторегрессии	218
8.4. Авторегрессионные процессы и их моделирование	223
Глава 9. Система эконометрических уравнений	230
9.1. Общая характеристика системы эконометрических уравнений	230
9.2. Структурная и приведенная формы модели	233
9.3. Идентификация структурной модели	235
9.4. Оценивание параметров системы одновременных уравнений	241
Глава 10. Модели с авторегрессионной условной гетероскедастичностью	261
10.1. Оценка порядка интегрируемости. Тест Дикки—Фуллера	261
10.2. Модель ARCH	271
10.3. Модель GARCH	276
Список рекомендуемой литературы	281
Статистико-математические таблицы	282

ВВЕДЕНИЕ

С середины 1990-х гг. эконометрика вошла в учебные планы подготовки экономистов в высших учебных заведениях России. Это решение существенно приблизило российское экономическое образование к подготовке в зарубежных университетах. Несомненно, большое влияние на этот процесс оказало присоединение России к Болонским соглашениям.

Любая дисциплина имеет свои пререквизиты, т. е. те учебные дисциплины, на которые она опирается. Изучение эконометрики предполагает, что студенты знакомы с микро- и макроэкономикой, матричной алгеброй и математической статистикой.

Учебник по эконометрике посвящен прежде всего эконометрическим методам, поэтому изложение в первую очередь опирается на те знания, которые получены студентами при изучении статистики. Прежде всего это понятие случайной переменной (дискретной или непрерывной), связанные с этими понятиями функции вероятности распределения или плотности вероятности. Студент должен понимать такие статистические термины, как кумулятивное распределение, условная вероятность, совместная вероятность, уметь вычислять математические ожидания, понимать условие независимости случайных переменных, знать теорему Бейеса, распределение вероятностей непрерывных и дискретных распределений, включая распределение Бернулли, Пуассона, геометрическое, равномерное, нормальное, гамма-распределение, хи-квадрат-распределение, экспоненциальное, t- и F-распределения, а также знать экономические индексы.

Введение эконометрики в учебные планы студентов экономических специальностей потребовало создания учебной литературы: учебников, задачников, практикумов.

Настоящий учебник подготовлен группой преподавателей кафедры статистики и эконометрики Санкт-Петербургского государственного университета экономики и финансов. Коллектив авторов данного учебника имеет опыт создания учебника по эконометрике. Первый подготовленный нами учебник «Эконометрика» для вузов был издан в 1998 г. издательством «Финансы и статистика» и многократно допечатывался. В 2005 г. вышло второе издание этого учебника, переработанное и дополненное. Большой вклад во второе издание внесли преподаватели кафедры статистики и эконометрики Потсдамского университета (Германия): профессор доктор Г. Г. Штroe (H. G. Strohe) и доктор К. Бартелс (K. Bartels). Следует упомянуть и подготовленный тем же составом авторов «Практикум по эконометрике», который также выдержал два издания (1-е – в 1999 г., 2-е – в 2006 г., оба издания вышли в издательстве «Финансы и статистика»).

Данный учебник содержит достаточно подробное изложение необходимых сведений, касающихся методов построения парной и множественной регрессий, оценивания параметров регрессионных моделей, способов их тестирования, выявления гетероскедастичности остатков. В учебнике исследуются принципы и методы построения системы одновременных уравнений, подробно рассматриваются подходы к эконометрическому моделированию на основе временных рядов. Специальные главы посвящены моделям дискретного выбора и анализу панельных данных. Учебник включает изложение методов, используемых в экономическом анализе финансовых рынков: анализ коинтеграции, ARCH- и GARCH-моделей и др.

Таким образом, предлагаемый учебник охватывает все основные разделы курса «Эконометрика».

ГЛАВА 1. ПОНЯТИЕ ЭКОНОМЕТРИКИ. УСЛОВИЯ ЕЕ ПРИМЕНЕНИЯ

1.1. Определение эконометрики.

Предпосылки возникновения и этапы развития

В качестве самостоятельного научного направления эконометрика выделилась в конце 20-х – 30-е гг. XX в. Но уже до этого в 1910 г. польский бухгалтер Павел Цьомпа использовал термин «эконометрия» (*oekonometrie*), вкладывая в это понятие математическое описание рядов экономических данных и отображение их в геометрической или графической форме. Буквально слово «эконометрика» означает изменение в экономике.

Во второй половине 20-х гг. XX в. стало понятно, что смысл эконометрики — в моделировании экономических явлений. Именно поэтому ее развитие стало необходимым для продвижения экономической науки. Сам термин «эконометрика» был введен в научный оборот в 1926 г. норвежским экономистом Рагнаром Фришем (1895–1973). Развернутое определение эконометрики было дано им во вступительной статье первого номера журнала «Эконометрика» (*Econometrica*) в 1933 г.:

«Эконометрика — это ни в коем случае не то же самое, что экономическая статистика. Она отнюдь не идентична тому, что мы называем общей экономической теорией, хотя значительная доля этой теории носит определенно количественный характер. Также эконометрика не должна восприниматься как синоним применения математики в экономике. Опыт показывает, что и статистика, и экономическая теория, и математика, взятые по отдельности, являются необходимыми, но не достаточными для действительного понимания количественных отношений в современной экономической жизни. Именно объединение всех трех частей дает мощный эффект. И именно это объединение и составляет эконометрику»*.

В этом определении доминирует идея несводимости эконометрики к какой-то одной области знаний; подчеркивается ее междисциплинарный характер.

С современных позиций эконометрику можно определить как науку о моделировании экономических явлений, позволяющую объяснять и прогнозировать их развитие, выявлять и измерять определяющие факторы.

На основе экономической теории разрабатываются концепции развития изучаемых процессов; с помощью статистики эти процессы выражаются в статистических показателях; математико-статистические методы позволяют строить модели изучаемых процессов, оценивать их параметры, степень соответствия реальным данным и, наконец, прогнозировать развитие изучаемого явления.

* Цит. по Бернхт Э. Р. Практика эконометрики: классика и современность. М. 2005. С. 1.

Знание экономической теории необходимо как на начальном этапе моделирования при постановке решаемой задачи выявления объективно существующих законов и связей, так и на заключительном этапе исследования при интерпретации полученных результатов.

Экономическая статистика призвана дать информационное обеспечение построения эконометрической модели, позволить выбрать те показатели, содержание которых наилучшим образом отвечает поставленной задаче. Если требуемые показатели отсутствуют в системе официальной статистики, то знание статистики позволяет составить проект выборочного наблюдения и получить необходимые данные.

Математический инструментарий эконометрики в основном сводится к математико-статистическим методам регрессионного анализа, анализа временных рядов, решению систем одновременных уравнений, решению проблем спецификации и идентификации моделей, тестированию статистических гипотез.

Польский ученый Оскар Ланге (1904—1965) включал в число эконометрических методов и метод «затраты — выпуск» (статическую и динамическую модели).

Иногда в состав эконометрических методов включают и многомерные статистические методы, направленные на выявление структуры данных, латентных переменных (метод главных компонент, факторный анализ, дискриминантный анализ).

Предпосылки возникновения эконометрики. Среди предпосылок возникновения эконометрики можно назвать разработку количественных методов в экономических исследованиях, накопление учетно-статистических данных и, наконец, создание современной микромакроэкономики.

Зарождение количественного подхода к экономике относится ко второй половине XVII в. и связано со школой политических арифметиков. В трудах ее основателей Вильяма Петти (1623—1687) и Джона Граунта (1620—1674) была провозглашена особенность этого направления: говорить об экономике на языке мер, весов и чисел. К политическим арифметикам относятся англичане Г. Кинг (1648—1712), Ч. Давенант (1656—1714), Э. Галлей (1656—1742); немец И. Зюсмилх (1707—1767); французы А. Депарсье (1707—1768), С. Вобан (1633—1707); швед П. Варгентин (1717—1783) и др.

Именно представителям этого направления принадлежит честь первых попыток исчисления народного дохода, сравнения богатства разных стран, разработки подходов к выявлению законов воспроизводства населения, социальной структуры общества и социального неравенства.

Следующий этап математизации в изучении социальных процессов связан с именем Адольфа Кетле (1786—1874) — с созданием им «теории среднего человека» и «социальной физики». Он включил в методы познания кривые распределения, средние величины и отклонения от средних, убедительно показал могущество статистики в раскрытии социальных законов.

Дальнейший шаг состоял в попытке выявить на статистической основе закономерности формирования потребления: установить пропорциональность отношения полных расходов и расходов на питание (затем расходов на жилье и т. д.) к приросту доходов домашнего хозяйства. В исследование этой проблемы внесли вклад Э. Энгель (1821–1896), А. Швабе, Е. Слуцкий (1880–1948), А. Боули (1869–1957). Закономерности распределения доходов анализировались итальянским математиком-экономистом В. Парето (1848–1923) (кривая Парето).

Но особенно важным для развития количественного подхода был статистический анализ поведения цен на различные товары — отечественные и импортные. Появились исследования динамики цен на важнейшие товары, группы товаров, анализ региональных особенностей роста цен; уделялось внимание построению индексов цен, попыткам выявления цикличности в изменениях цен и связи с бизнес-циклами. Такого рода исследования возникли уже в XVIII в., но особенно они были популярны в конце XIX — первой четверти XX в. Заметных успехов в этой области добились американские ученые, работавшие в Национальном бюро экономических исследований: Уэсли Митчелл (1874–1948), Фредерик Миллс (1892–1951). Проблема бизнес-циклов разрабатывалась не только в связи с поведением цен, но и в связи с динамикой объемов производства, а также состоянием рынка труда (У. Мур, У. Митчелл и др.). К этому же этапу исследований относятся разного рода экономические барометры, в первую очередь *гарвардский барометр*, просуществовавший с 1917 до 1929 г.

Возникновение эконометрики. Великая депрессия 1929–1933 гг., потрясшая экономически развитые страны мира, показала необходимость рассмотрения экономики как системы,ключающей определенные элементы и взаимосвязи между ними. Именно в этот период пробудился интерес к методу В. В. Леонтьева «затраты — выпуск». Осознание сложности экономики, взаимозависимости экономик разных стран привело к идею создания эконометрики.

Исследование экономики с неизбежностью опирается на данные так называемого пассивного эксперимента, поскольку исследователь никак не может воздействовать на данные. Таковыми являются все реальные данные, которые нам предлагает официальная статистика, или учет, или специальное наблюдение. В этих данных заложено влияние множества факторов, в которых трудно разобраться. Новое знание в экономике может быть получено только посредством абстракции, т. е. с помощью модели. Модель отражает попытку реконструировать в упрощенной форме тот механизм, который скрыт за изучаемыми явлениями. Поэтому модель представляет собой результат абстракции. На основе модели формируется предварительное объяснение, которое затем должно пройти проверку путем статистико-математического тестирования.

В основе модели лежат постулаты экономической теории и правила логики. К этому всегда добавляется то, — что можно назвать интуицией. И если построение модели прошло все этапы (теоретическое

обоснование, наполнение конкретными данными, спецификацию, идентификацию, тестирование), то полученная на ее основе информация может быть весьма полезной для целей принятия решений, разработки прогнозов. Очевидно, что успехи в построении эконометрических моделей были обеспечены формированием современной экономической теории. Микроэкономика в основном была разработана к концу XIX в., а макроэкономика — к 30-м гг. XX в., что и подтолкнуло к выделению эконометрики.

В 1930 г. было образовано Эконометрическое общество, основателями которого выступили И. Фишер, Р. Фриш, О. Андерсон, О. Моргенштерн и др. С 1933 г. начал издаваться журнал «Эконометрика» (*Econometrica*). В 1928 г. Ч. Коббом и П. Дугласом была построена производственная функция, ставшая одной из самых популярных эконометрических моделей:

$$P = a \cdot L^{b_1} \cdot K^{b_2} e,$$

где P — объем производства;
 L — численность занятых;
 K — объем капитала;
 a, b_1, b_2 — параметры;
 e — случайная компонента.

В 1930—1939 гг. Я. Тинбергеном начато макроэконометрическое моделирование, которое опиралось на теорию Дж. М. Кейнса и разработку системы национальных счетов в США и других странах. В 1955 г. Л. Клейном и А. Голдбергером была построена одна из первых комплексных эконометрических моделей. Эта модель состояла из 15 регрессионных уравнений и 5 тождеств и охватывала 40 макроэкономических показателей по данным за 20 лет. Большой вклад в развитие эконометрики внесла разработка адаптивных моделей и методов, позволивших повысить точность прогнозов.

Разработка методов оценки параметров системы одновременных уравнений Т. Хаавельмо, Т. С. Купмансом, Т. В. Андерсоном, Л. Клейном и др. ознаменовали окончательное создание специальных эконометрических методов.

Таким образом, примерно за 50 лет — с конца 30-х по 80-е гг. были сделаны основополагающие разработки в области эконометрической теории и ее применения.

В 1986 г. начал издаваться «Журнал прикладной эконометрики» (*Journal of Applied Econometrics*), который выходит с периодичностью 2 раза в месяц. В журнале публикуются статьи по новым эконометрическим методам и сопряженным вопросам: измерения в экономике, оценка параметров, тестирование, предсказания и т. д. Акцент делается на тщательное и строгое применение эконометрических методов и соответствующую интерпретацию результатов.

Развитие финансовых рынков и инструментов дало новый толчок возникновению специальных эконометрических методов анализа временных рядов. Семидесятые годы XX в. были ознаменованы созданием

и плодотворным применением моделей линейной фильтрации. Следующим шагом стала разработка методов, направленных на выявление шоков, тестирование рядов на стационарность, исследование коинтегрированности рядов и т. д. В этой области следует отметить прежде всего заслуги Р. Энгла и К. Грэндженера (оба получили Нобелевскую премию по экономике в 2005 г.).

Путь эконометрики в нашу страну был долгим и сложным. Первая попытка внедрить эконометрику (или *эконометрию*, как нередко говорили) в советскую науку принадлежит Василию Сергеевичу Немчинову (1894–1964). Эта попытка привела к выделению такого направления, как экономико-математические методы и экономическая кибернетика. Но по-настоящему эконометрика стала развиваться в России лишь с середины 1990-х гг. Первым учебником по эконометрике стал учебник Я. Р. Магнуса, П. К. Катышева, А. А. Персецкого «Эконометрика. Начальный курс» (М., 1997).

С 2006 г. в нашей стране выходит ежеквартальный журнал «Прикладная эконометрика» (главный редактор — доктор физико-математических наук С. А. Айвазян), что, несомненно, свидетельствует о признании значимости эконометрики как инструмента микро- и макроэкономического исследования.

Подтверждением значимости эконометрики является и то, что самая почетная премия, которая присуждается ученым-экономистам, — премия Банка Швеции в память Альфреда Нобеля (именно ее называют «нобелевской премией» для экономистов), который отметил (по поводу модели Тинбергена) — не раз присуждалась эконометристам. Первым звеном в этой цепочке награждений явилась Нобелевская премия 1969 г., которую разделили Ян Тинберген и Рагнар Фриш.

Победное шествие эконометрики не означает отсутствия критики в ее адрес. Первым и главным критиком был Дж. М. Кейнс (1883–1946), что в эконометрических построениях слишком много субъективизма, и даже высказал мнение, что эконометрика сродни алхимии.

Несмотря на это замечание, надо признать, что ценность эконометрических моделей перевешивает их недостатки и эконометрические исследования ведутся во всех значимых научных центрах, занимающихся проблемами экономики.

Применяемые в настоящее время эконометрические модели можно подразделить на:

- статические и динамические — по характеру используемых данных (первые основаны на одновременных данных по совокупности объектов, вторые — на временных рядах). Промежуточное положение занимают модели панельных данных, основанные на данных по одной и той же совокупности за ряд лет;
- комплексные или некомплексные (первые отличаются тем, что отражают связи между макроэкономическими показателями на всех стадиях процесса воспроизводства);

- аналитические, имитационные и прогностические. Это деление моделей производится по целям их применения.

Построение любой эконометрической модели проходит следующие этапы:

- первый — теоретический, в ходе которого формулируется цель исследования, определяется круг участвующих в модели экономических характеристик, создается априорное описание formalизованных связей между ними;
- второй — информационный, когда осуществляется поиск требуемых данных, проверяется их достоверность, сопоставимость, осуществляются необходимые пересчеты, используются пространственные и временные данные;
- третий — спецификация модели, когда устанавливаются экзогенные (внешние) и эндогенные (внутренние) переменные, выявляются связи и соотношения;
- четвертый — идентификация модели, т. е. выявление условий корректного оценивания параметров модели на основе соотношения количества переменных и связей между ними;
- пятый — оценка параметров модели;
- шестой — верификация модели, т. е. проверка адекватности модели, делается вывод о том, какова точность расчетов на основе модели, получаемых прогнозных оценок; производится анализ остатков (случайных компонент).

Важным фактором применения и развития эконометрического моделирования является наличие специальных пакетов прикладных программ. Нередко эконометристу даже в решении какой-либо одной задачи приходится пользоваться несколькими пакетами, например Econometric Views, Gauss, Statistica, SPSS и др.

Если модель удовлетворяет всем предъявляемым требованиям, то она может быть использована либо для прогнозирования, либо для объяснения (анализа) скрытых механизмов исследуемых процессов.

1.2. Методы оценивания

Начнем с краткого обзора основных положений статистики, которые необходимы для понимания эконометрики.

Рассмотрим нормальное распределение со средней μ и дисперсией σ^2 . Как известно, такое гауссовское распределение симметрично, имеет колоколообразную форму, полностью определяется центром распределения — средней (μ) и дисперсией, т. е. σ^2 . И μ , и σ^2 — это так называемые генеральные параметры. Если мы работаем с выборкой, извлеченной из нормальной генеральной совокупности, то выборочную среднюю будем

$$\text{обозначать } \bar{x}, \text{ а дисперсию } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

Если, например, μ — средний душевой доход, приходящийся на одного жителя Северо-Западного федерального округа, то \bar{x} — средний душевой доход, рассчитанный по выборке 1000 домохозяйств этого региона.

Оценка генеральных параметров может быть получена двумя методами: а) методом моментов; б) методом максимального правдоподобия.

Процедура метода моментов такова: определяются выборочные моменты (моменты эмпирического распределения) и в количестве, равном числу оцениваемых параметров, приравниваются к соответствующим моментам распределения, являющимся функциями от неизвестных параметров. Полученные уравнения решают относительно параметров и находят искомые оценки. Метод моментов позволяет найти оценки, которые при больших объемах выборки n асимптотически нормальны, имеют математическое ожидание, лишь величиной порядка $\frac{1}{n}$ отличающееся от истинного значения параметра, и стандартное отклонение, отличающееся от истинного лишь величиной порядка $\frac{1}{\sqrt{n}}$.

Метод моментов обеспечивает соответствие генеральных параметров выборочным. Например, плотность нормального распределения полностью идентифицируется двумя параметрами μ и σ^2 . Необходимо иметь в виду, что

$$\mu \approx \bar{x} \text{ и } \hat{\mu} + \hat{\sigma}^2 = \sum_{i=1}^n x_i^2 / n.$$

Подставим первое выражение во второе и получим оценку дисперсии:

$$\sigma^2 = \sum_{i=1}^n x_i^2 / n - \bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n.$$

В ряде случаев задача определения параметров эквивалентна задаче отыскания минимума некоторого функционала. Наиболее важным методом решения такой задачи является *метод наименьших квадратов* (МНК) (см. гл. 2 и 3).

Оценки, найденные методом моментов (или МНК), не являются наилучшими из возможных с точки зрения эффективности — их дисперсия не является наименьшей возможной дисперсией.

Другим методом оценивания параметров распределения по данным выборки является *метод максимального правдоподобия*, предложенный Р. Фишером. Рассмотрим выборку из значений случайной величины X объема n : x_1, x_2, \dots, x_n . Пусть $p(x, \theta)$ — плотность абсолютно непрерывной случайной величины x . Тогда $L(x_1, x_2, \dots, x_n, \theta) = p(x_1, \theta) \dots p(x_n, \theta)$ называется функцией правдоподобия.

Аналогичным образом можно определить функцию правдоподобия и для дискретной случайной величины, но поскольку в этом случае

значения выборки могут повторяться, функцию правдоподобия можно несколько преобразовать.

Пусть X — дискретная случайная величина с возможными значениями x_1, \dots, x_n , а m_1, \dots, m_n — частоты этих значений в выборке. Очевидно, что $\sum_{i=1}^n m_i = n$.

Пусть $p(x=y_i) = p_i(\theta)$. В этом случае функция правдоподобия определяется следующим соотношением:

$$L(x_1, x_2, \dots, x_n, \theta) = p_1^{m_1}(\theta) \dots, p_n^{m_n}(\theta).$$

Будем рассматривать L как функцию от θ , считая выборочные значения x_1, \dots, x_n заданными. Метод максимального правдоподобия заключается в том, что в качестве оценки для неизвестного параметра θ берется такое значение θ , при котором функция правдоподобия L достигает наибольшего значения.

Это значение является функцией от значений выборки x_1, \dots, x_n и называется оценкой максимального правдоподобия. Поскольку $\ln L$ достигает максимума при том же значении θ , что и L , то согласно правилам дифференциального исчисления необходимо решить уравнение $\frac{\sigma \ln L}{\sigma \theta} = 0$. В случае если параметр θ — многомерный, его оценка находится из системы:

$$\left\{ \begin{array}{l} \frac{\sigma \ln L}{\sigma \theta_1} = 0 \\ \frac{\sigma \ln L}{\sigma \theta_2} = 0 \\ \dots \\ \frac{\sigma \ln L}{\sigma \theta_k} = 0 \end{array} \right.$$

Пример 1

В качестве первого примера рассмотрим абсолютно непрерывную величину, имеющую нормальное распределение с параметрами μ и σ :

$$p(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Функция правдоподобия в данном случае имеет следующий вид:

$$\begin{aligned} L(x_1, \dots, x_n, \mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma_n} \cdot e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \dots \frac{1}{\sqrt{2\pi}\sigma_n} \cdot e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} = \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} \end{aligned}$$

Теперь прологарифмируем функцию правдоподобия:

$$\ln L = -n\left(\frac{1}{2}\ln 2\pi + \ln \sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Продифференцировав $\ln L$ по μ и σ , получим систему из двух уравнений:

$$\begin{cases} \frac{\sigma \ln L}{\sigma \mu} = -\frac{(-2)}{2\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\sigma \ln L}{\sigma \delta} = -\frac{n}{\sigma} - \frac{(-2)}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 \end{cases}$$

Решая эту систему, находим оценки для μ и σ :

$$\begin{aligned} \mu &= \bar{x} \\ \sigma^2 &= s^2, \end{aligned}$$

где \bar{x} и s^2 — выборочные средняя и дисперсия, соответственно

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

Пример 2

В качестве второго примера рассмотрим дискретную случайную величину x , имеющую распределение Пуассона с параметром:

$$pk(\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Данное распределение часто встречается, например, в страховых расчетах. В этом случае функция правдоподобия имеет следующий вид:

$$L = \prod_{k=0}^r \left(\frac{\lambda^k}{k!} e^{-\lambda} \right)^{m_k}.$$

Представим функцию правдоподобия в виде, пригодном для логарифмирования:

$$L = e^{-\lambda n} \prod_{k=0}^r e^{m_k(k \ln \lambda - \ln k!)}$$

$$\text{Тогда } \ln L = -\lambda n + \sum_{k=0}^r m_k (k \ln \lambda - \ln k!),$$

$$\frac{\sigma \ln L}{\sigma \lambda} = -n + \frac{1}{\lambda} \sum_{k=0}^r k \cdot m_k = 0 \text{ — уравнение метода максимального правдоподобия.}$$

Учитывая, что $\sum_{k=0}^r k \cdot m_k = \sum_{i=0}^n x_i$, получим:

$\lambda = \bar{x}$ — оценка параметра λ распределения Пуассона методом максимального правдоподобия.

Надо отметить, что не всегда уравнение метода максимального правдоподобия разрешимо. В том случае, когда явно выписать корни уравнения максимального правдоподобия не удается, необходимо прибегнуть к пошаговому перебору параметров на ЭВМ для максимизации логарифма функции правдоподобия.

Также отметим, что, несмотря на совпадение в двух рассмотренных примерах оценок параметров распределения методом максимального правдоподобия и методом моментов, в других случаях оценки, полученные этими методами, могут не совпадать. При этом метод максимального правдоподобия обладает по сравнению с другими методами важными достоинствами: он всегда приводит к состоятельным (хотя иногда и смещенным — как в случае, например, нормального распределения) оценкам, распределенным асимптотически нормально, имеющим наименьшую возможную дисперсию по сравнению с другими, также асимптотически нормальными оценками. Если для существует эффективная оценка, то это оценка максимального правдоподобия.

При оценке параметров число уравнений, которое приходится решать, зависит от числа параметров. Например, экспоненциальное распределение определяется одним параметром, значит, требуется только одно уравнение для оценки параметра, в то время как гамма-распределение имеет два параметра и требует двух уравнений. Отметим, что метод максимального правдоподобия трудно подбирать к форме оцениваемого распределения, зато у этого метода имеются, как было отмечено выше, свои привлекательные свойства.

Свойства оценок

Несмещенность

μ называется несмещенной оценкой $\hat{\mu}$ только тогда, когда $E(\bar{x}) = \mu$.

Так как $E(\bar{x}) = \sum_{i=1}^n E(x_i)$ $n = \mu$ т. е. \bar{x} — несмещенная оценка μ . Не-

смещенностю означает, что «в среднем» наша оценка соответствует параметру. Пусть, например, мы будем производить выборку объемом 1000 домохозяйств 200 раз из одной и той же генеральной совокупности. Мы получим 200 значений \bar{x} , осредняя которые, получим μ (точнее говоря, приблизимся к оценке μ). Ввиду того что фактически мы всегда производим только одну выборку, возможно, что наблюдаемое значение \bar{x} будет удалено от μ . При увеличении объема выборки n дисперсия \bar{x} уменьшается, поскольку $\text{var}(\bar{x}) = \sigma^2 / n$, и соответственно становится меньше правдоподобие того, что \bar{x} будет удалено от μ . Это подводит нас к понятию эффективности.

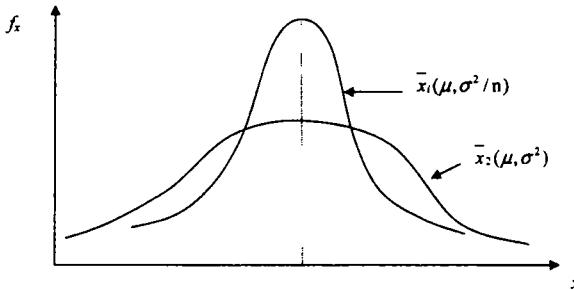


Рис. 1.1. Свойство эффективности оценок

Эффективность

Если мы будем сравнивать две исследуемые оценки, то мы обратимся к их эффективности, основанной на соотношении их дисперсий. Та из оценок, которая имеет меньшую дисперсию, является более эффективной. Это свойство показано на рис. 1.1.

Обе оценки \bar{x}_1 и \bar{x}_2 являются несмешенными оценками μ , однако оценка \bar{x}_1 является более эффективной оценкой: она имеет дисперсию меньше по сравнению с дисперсией для \bar{x}_2 .

При сравнении исследуемых оценок мы должны найти ту, которая имеет минимальную дисперсию. Такая оценка называется *MVUE* – *minimum variance unbiased estimator*.

Состоятельность

Еще одним асимптотическим свойством является состоятельность. Это свойство записывается следующим образом, так как $\epsilon = 0$ для любого положительного ϵ . Другими словами, \bar{x} не отличается от μ , когда $n \rightarrow \infty$.

Согласно неравенству Чебышева

$$Pr = \{|\bar{x} - \mu| > \sigma_{\bar{x}}\} \leq 1/k^2.$$

Если $\epsilon = k\sigma_{\bar{x}}$, то $1/k^2 = \sigma_{\bar{x}}^2/\epsilon^2 = \sigma^2/n\epsilon^2$ стремится к 0 при $n \rightarrow \infty$, поскольку σ^2 и ϵ – конечные положительные значения. Таким образом, достаточным условием того, чтобы оценка рассматривалась как состоятельная, выступает то, что асимптотически исследуемая оценка имеет дисперсию, приближающуюся к нулю при возрастании объема выборки, $n \rightarrow \infty$.

1.3. Испытание гипотез и доверительные интервалы

Статистической гипотезой называется предположение о свойстве генеральной совокупности, которое можно проверить, опираясь на данные выборки. Обозначается гипотеза буквой H (лат. *hypothesis*). Так, может быть выдвинута гипотеза о том, что средняя в генеральной совокупности равна некоторой величине a , т. е. $H: \mu = a$, или о том, что генеральная средняя больше некоторой величины b : $H: \mu > b$.

Различают *простые* и *сложные* гипотезы. Гипотеза называется простой, если она однозначно характеризует параметр распределения случайной величины. Например, $H: \mu = a$. Сложная гипотеза состоит из конечного или бесконечного числа простых гипотез, при этом указывается некоторая область вероятных значений параметра. Например, $H: \mu > b$. Эта гипотеза состоит из множества простых гипотез: $H: \mu = c$, где c — любое число, большее b .

Гипотезы о параметрах генеральной совокупности называются *параметрическими*; гипотезы о распределениях или структурных характеристиках (моде, меднане) называются *непараметрическими*.

Гипотеза о том, что две совокупности, сравниваемые по одному или нескольким признакам, не отличаются, называется *нулевой гипотезой* (или нуль-гипотезой). Она обозначается H_0 . При этом предполагается, что действительное различие сравниемых величин равно нулю, а выявленное по данным выборки отличие от нуля носит случайный характер. Например, $H_0: \mu = \mu_2$ и т. д.

Поясним обсуждаемый вопрос на примере. На кафедре статистики разработаны новые тесты для проведения письменного экзамена по системе национального счетоводства. Протестируем нулевую гипотезу, что 80% студентов сдадут экзамен. Альтернативная гипотеза состоит в утверждении, что только 50% прослушавших курс сдадут экзамен. Пусть была сделана выборка 20 студентов из потока. Заметим, что распределение в данном случае подчиняется закону Бернулли с вероятностью успешной сдачи 0, значит, можно записать $H_0: \theta_0 = 0,80$ и $H_1: \theta_1 = 0,5$; иначе говоря, ожидалось, в соответствии с H_0 , что сдадут $[E_{(x)} = n\theta_0] 16$ студентов против $(n\theta_1) 10$ студентов в соответствии с H_1 . Чтобы принять нулевую гипотезу, нужно, чтобы не менее 13 студентов сдали экзамен (13 — это средняя точка в интервале [10, 16]).

Нулевая гипотеза отвергается тогда, когда по выборке получается результат, который при истинности выдвинутой нулевой гипотезы маловероятен. Границей невозможного или маловероятного обычно считают $\alpha = 0,05$ (т. е. 5%), или 0,01 (т. е. 1%), или 0,001 (т. е. 0,1%). Если ориентироваться на «правило трех сигм», то вероятность ошибки должна быть равна $1 - 0,9973 = 0,0027$. Однако для этого уровня вероятности ошибки значения критериев редко табулируются: как правило, значения критериев в статистико-математических таблицах рассчитаны для вышеуказанных вероятностей ошибки: 0,05; 0,01.

Статистическим критерием называют определенное правило, устанавливающее условия, при которых проверяемую нулевую гипотезу следует либо отклонить, либо не отклонять. Критерий проверки статистической гипотезы определяет, противоречит ли выдвинутая гипотеза фактическим данным или нет.

Проверка статистических гипотез складывается из следующих этапов:

- формулируется в виде статистической гипотезы задача исследования;

- выбирается статистическая характеристика гипотезы;
- выбираются испытуемая и альтернативная гипотезы на основе анализа возможных ошибочных решений и их последствий;
- определяются область допустимых значений, критическая область, а также критическое значение статистического критерия (t , F , χ^2) по соответствующей таблице;
- вычисляется фактическое значение статистического критерия;
- проверяется испытуемая гипотеза на основе сравнения фактического и критического значений критерия и в зависимости от результатов проверки гипотеза либо отклоняется, либо не отклоняется.

При испытании гипотезы возможны две ошибки.

Первая — отклонение H_0 , когда фактически она верна. Такая ошибка называется ошибкой I рода, вероятность такой ошибки обозначается α . Вторая — принятие H_1 , когда фактически она ложна. Эта ошибка называется ошибкой II рода и ее вероятность обозначается β (табл. 1.1).

Таблица 1.1

Виды ошибок при испытании гипотез

Решение	Правильная гипотеза	
	H_0	H_1
Принять	Нет ошибки	Ошибка II рода
Отклонить	Ошибка I рода	Нет ошибки

Согласно теории Неймана—Пирсона испытание гипотезы производится при фиксированном α и минимизации β или максимизации $(1-\beta)$, называемой *мощностью критерия*. Чем она больше, тем меньше вероятность ошибки II рода.

Альтернативная гипотеза H_1 может быть сформулирована по-разному в зависимости от того, какие отклонения от гипотетической величины нас особенно беспокоят: положительные, отрицательные либо и те, и другие. Соответственно альтернативные гипотезы могут быть записаны как

$$H_1: \mu > a; H_1: \mu < a; H_1: \mu \neq a.$$

От того, как формулируется альтернативная гипотеза, зависят границы критической области и области допустимых значений (рис. 1.2).

Критической областью называется область, попадание значения статистического критерия в которую приводит к отклонению. Вероятность попадания значения критерия в эту область равна принятому уровню значимости.

Область допустимых значений дополняет критическую область. Если значение критерия попадает в область допустимых значений, это свидетельствует о том, что выдвинутая гипотеза не противоречит фактическим данным (H_0 не отклоняется).

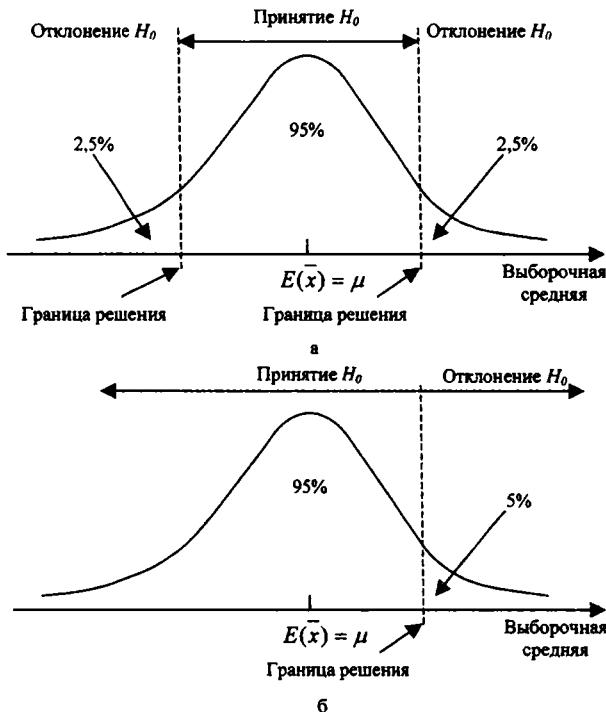


Рис. 1.2. Проверка нулевой гипотезы
на 5%-ном уровне значимости:
а – двусторонняя; б – односторонняя

Точки, разделяющие критическую область и область допустимых значений, называются *критическими точками* или *границами критической области*. В зависимости от формулировки альтернативной гипотезы критическая область может быть *двусторонняя* (рис. 1.2, а) или *односторонняя* (рис. 1.2, б) — либо левосторонняя, либо правосторонняя.

Если вычисленное значение критерия попадает в критическую область, нулевая гипотеза отклоняется (не принимается), поскольку она противоречит фактическим данным.

Метод тестирования приводит к точечной оценке параметра. Однако лучше давать оценку в форме интервала с некоторой, например, 95%-ной, доверительной вероятностью. По своей сути построение доверительного интервала является как бы другой стороной проверки гипотез. Покажем это на примере следующей записи:

$$Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1 - \alpha.$$

Для табличного значения $z_{\alpha/2} = 1,96$ получим

$$Pr[-1,96 \leq z \leq 1,96] = 0,95.$$

Зная, что $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$, запишем:

$$\Pr[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}] = 1 - \alpha.$$

Если α известна, то последнее выражение можно представить как

$$\Pr[\bar{x} - z_{\alpha/2}(\sigma / \sqrt{n}) \leq \mu \leq \bar{x} + z_{\alpha/2}(\sigma / \sqrt{n})] = 1 - \alpha,$$

т. е. получено выражение интервальной оценки генерального параметра μ .

С точки зрения проверки гипотез можно сказать, что мы не отклоняем H_0 , если μ находится в принятом доверительном интервале. Например, если $\mu = 2$ находится в 95%-ном доверительном интервале (в соответствии с нулевой гипотезой), эта гипотеза не может быть отклонена на 5%-ном уровне значимости по данным той же выборки. Так что значение, находящееся в 95%-ном доверительном интервале (в соответствии с проверяемой нулевой гипотезой), не может быть отклонено на 5%-ном уровне. Именно поэтому в статистике не принято говорит: «принимаем H_0 », — вместо этого предпочитают формулировку «не отклоняем H_0 ».

Контрольные вопросы

1. Сформулируйте задачи, решаемые эконометрикой.
2. Что явилось предпосылкой выделения эконометрики как самостоятельной науки?
3. Какие периодические издания посвящены проблемам эконометрики?
4. Какие методы оценивания вам известны?
5. Какие требования предъявляются к статистическим оценкам?
6. В чем состоят сходство и различие метода испытания гипотез и построения доверительных интервалов?
7. Охарактеризуйте основные типы эконометрических моделей.

ГЛАВА 2. ПАРНАЯ РЕГРЕССИЯ

2.1. Основные понятия

Парная регрессия — это уравнение, описывающее корреляционную связь между парой переменных: зависимой переменной (результатом) y и независимой переменной (фактором) x .

$$y = f(x).$$

Функция может быть как линейной, так и нелинейной, например, $\hat{y} = \alpha + \frac{\beta}{x} + \varepsilon$, или $y = \alpha x + \beta e$, или $\hat{y} = e^{\alpha+\beta x+\varepsilon}$ и т. д.

В этой главе мы будем прежде всего рассматривать парную регрессию, описывающую линейную связь между двумя переменными, которая представлена в следующей форме:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \dots, N, \quad (2.1)$$

где y_i — i -е значение зависимой переменной y ;
 x_i — i -е значение независимой переменной x ;
 α и β — генеральные параметры парной линейной регрессии;
 N — объем генеральной совокупности.

Это уравнение можно использовать для изучения зависимости потребления (y) от уровня доходов (x); инвестиций (y), от процентной ставки (x) и для многих других задач. Уравнение парной регрессии в виде (2.1) относится к генеральной совокупности. Практически регрессия строится по данным выборки и записывается в виде:

$$y_i = a + bx_i + e_i, \quad i = 1, 2, \dots, n,$$

где n — объем выборки;
 a и b — выборочные оценки параметров парной линейной регрессии.

Данные выборки могут быть получены по объектам: домохозяйствам, агропромышленным комплексам, банкам, регионам и проч. — на какой-то один и тот же для всех объектов момент времени или за какой-то один период (месяц, квартал, год). Это так называемые *пространственные выборки*. А могут быть получены данные во времени, образующие временные ряды. Этот тип данных обсуждается в специальных главах учебника (см. гл. 6–8).

Итак, задача состоит в том, чтобы по данным выборки оценить неизвестные параметры линейной парной регрессии: параметр a — пересечение (*intercept*) и параметр β — наклон линии регрессии (*slope*) (рис. 2.1).

Как показано на графике (см. рис. 2.1), параметр a соответствует отрезку прямой, отсекаемому линией регрессии при пересечении ею оси ординат. Параметр β определяет наклон линии регрессии к оси абсцисс.

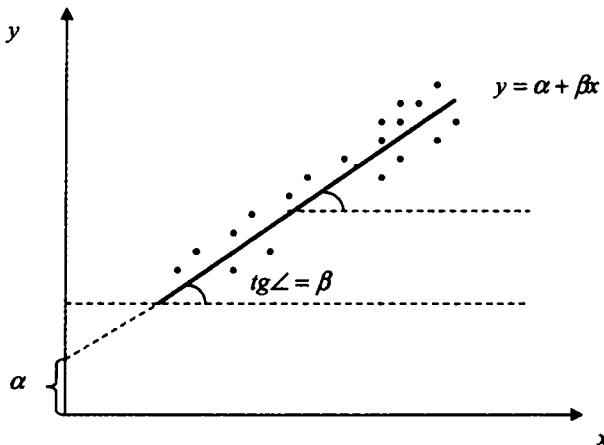


Рис. 2.1. Линия регрессии

Обратимся к примеру. Пусть y_i — объем выпуска продукции за месяц на i -м малом предприятии обрабатывающей промышленности (в тыс. руб.), x_i — стоимость основных фондов на i -м предприятии (в тыс. руб.). Выборка включает 30 малых предприятий обрабатывающей промышленности (условные данные) (табл. 2.1). Задача состоит в том, чтобы построить линейную парную регрессию $y = a + bx$.

Таблица 2.1

Исходные данные

Номер предприятия <i>a</i>	Выпуск, тыс. руб. <i>y</i>	Стоимость основных фондов, тыс. руб. <i>x</i>			
			<i>a</i>	<i>y</i>	<i>x</i>
1	2320	430			
2	2500	410			
3	2560	530			
4	2700	560			
5	2893	480			
6	2941	430			
7	3020	440			
8	3648	510			
9	3400	550			
10	3110	580			
11	3150	630			
12	3246	570			
13	3380	620			
14	3420	660			
15	3460	650			

Окончание табл. 2.1

Номер предприятия <i>a</i>	Выпуск, тыс. руб. <i>y</i>	Стоимость основных фондов, тыс. руб. <i>x</i>
16	3465	670
17	3480	720
18	3510	640
19	3555	670
20	3600	720
21	3662	770
22	3679	760
23	3700	800
24	3571	710
25	3500	650
26	3680	580
27	3710	690
28	3125	430
29	3176	520
30	3720	650
Среднее значение	3296,03	601

Параметр a называют свободным членом регрессии; параметр b – коэффициент регрессии, который измеряет, на сколько единиц в среднем изменится y при изменении x на одну единицу. Параметр $a \neq 0$, если вариация зависимой переменной (V_y) превосходит вариацию независимой переменной (V_x); в противном случае $a = 0$. Иногда, если данные выборки включают значения x_i , равные нулю, то свободный член интерпретируется как значение y при $x_i = 0$. Однако чаще свободный член не имеет экономического смысла. Параметр $b \neq 0$ в случае прямой связи между x и y , $b \neq 0$ в случае обратной связи между x и y .

Представим исходные данные на графике (рис. 2.2). Каждая точка графика соответствует конкретному предприятию; координаты точек определяются значениями переменных – объем выпуска (y_i) и стоимость основных фондов (x_i). Если бы параметры уравнения регрессии были известны, мы бы нанесли прямую линию $y_i = a + bx_i$ на график, где $a = \bar{y} - b\bar{x}$. Очевидно, что не все наблюдаемые точки будут лежать непосредственно на прямой $(a + bx_i)$. Отклонения обусловлены случайной ошибкой e_i . Эта ошибка может быть вызвана случайным искаложением значений переменных, ошибками в измерении переменных, наконец, неправильным выбором линейной формы регрессии, тогда как на самом деле связь между x и y нелинейна. Указанные причины по-разному влияют на распределение ошибки e_i .

Параметры a и b должны быть оценены так, чтобы линия $(a + bx_i)$ наилучшим образом подходила к данным. Поскольку по одним и тем

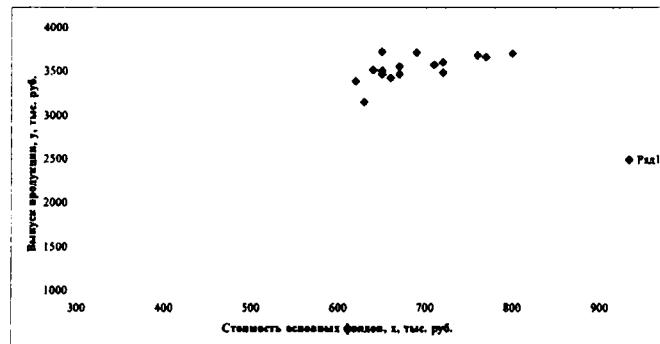


Рис. 2.2. Поле корреляции для зависимости выпуска продукции от основных фондов

же данным можно провести несколько линий, возникает необходимость понять, чем одна линия лучше другой. Пусть наилучшая линия имеет вид:

$$\hat{y}_i = a + bx_i,$$

где \hat{y} , a и b — выборочные оценки.

Тогда качество подобранный линии, т. е. то, насколько хорошо она соответствует фактическим данным, можно измерить как

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (2.2)$$

Это приводит к оценке параметров регрессии методом наименьших квадратов, который рассматривается в следующем параграфе.

2.2. Оценка параметров регрессии методом наименьших квадратов

Метод наименьших квадратов позволяет минимизировать сумму квадратов остатков:

$$e_i = y_i - a - bx_i, \quad i = 1, 2, \dots, n, \quad (2.3)$$

где a и b — оценки параметров регрессии α и β соответственно.

Обозначим сумму квадратов остатков как *RSS* (*residual sum of squares*):

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 \rightarrow \min.$$

Это выражение минимизируется при выполнении двух условий:

$$\frac{\partial a(\sum_{i=1}^n e_i^2)}{\partial a} = -2 \sum_{i=1}^n e_i = 0 \text{ или } \sum_{i=1}^n y_i + na + b \sum_{i=1}^n x_i = 0; \quad (2.4)$$

$$\frac{\partial a(\sum_{i=1}^n e_i^2)}{\partial b} = -2 \sum_{i=1}^n e_i x_i = 0 \text{ или } \sum_{i=1}^n y_i x_i + a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = 0. \quad (2.5)$$

Отсюда получим систему нормальных уравнений:

$$\begin{cases} an + b \sum_{(i)} x_i = \sum_{(i)} y_i \\ a \sum_{(i)} x_i + b \sum_{(i)} x_i^2 = \sum_{(i)} y_i x_i \end{cases}$$

Решая систему нормальных уравнений относительно a и b , получим

$$a = \bar{y} - b\bar{x}, \quad (2.6)$$

$$b = \left(\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y} \right) / \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \quad (2.7)$$

или

$$b = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - \bar{x}^2}$$

Остатки, оцененные таким образом, можно представить как

$$e_i = y_i - a - bx_i. \quad (2.8)$$

Они удовлетворяют обоим выдвинутым условиям (формулы (2.4) и (2.5)).

Сумма остатков равна нулю, т. е. $\sum_{i=1}^n e_i = 0$.

Из уравнения (2.6) следует, что

$$\bar{y} = a + b\bar{x}. \quad (2.9)$$

По исходным данным (см. табл. 2.1) с помощью пакета прикладных программ Econometric Views были произведены расчеты и получены значения параметров уравнения парной линейной регрессии (табл. 2.2).

Таблица 2.2

Результаты оценки параметров МНК

	Коэффициент	Стандартная t -статистика	P -значение	Нижнее 95%	Верхнее 95%
y -пересечение	1675,499	265,3978	6,313159	7,91E-07	1131,856 2219,141
x	2,696397	0,434528	6,205348	1,06E-06	1,806307 3,586487

Таблица 2.2 включает оценки параметров, их среднеквадратические ошибки, вероятности ошибочного решения (P -значение), нижние и верхние интервальные оценки параметров с вероятностью 95%. Согласно полученным значениям уравнение парной регрессии записывается как

$$\hat{y} = 1675,5 + 2,696x, \quad (2.9)$$

т. е. с увеличением объема основных факторов на 1 тыс. руб. выпуск продукции возрастает в среднем на 2,696 тыс. руб. Предприятие не может не иметь основных фондов, т. е. x_i не может быть равным нулю, поэтому свободный член не интерпретируется. Коэффициент эластичности показывает, на сколько процентов изменится значение y при изменении x на 1%:

$$\vartheta = b \frac{\bar{x}}{\bar{y}} = 2,696 \frac{601}{3296,03} = 0,49\%.$$

Полученное уравнение регрессии статистически значимо. Об этом свидетельствуют результаты дисперсионного анализа (табл. 2.3).

Таблица 2.3

Дисперсионный анализ

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	Значимость <i>F</i>
Регрессия	1	2 583 011	2 583 011	38,50634	1,06E-06
Остаток	28	1 878 244	67 080,14		
Итого	29	4 461 255			

В первой графе табл. 2.3 показаны источники вариации зависимой переменной; во второй — число степеней свободы; в третьей — суммы квадратов отклонений; в четвертой — суммы квадратов отклонений, приходящиеся на одну степень свободы; в пятой — значение *F*-критерия.

Для парной линейной регрессии число степеней свободы равно числу параметров (*p*) минус единица:

$$\begin{aligned} df_{\text{факт.}} &= p - 1 = m, \\ df_{\text{факт.}} &= 2 - 1 = 1. \end{aligned} \quad (2.10)$$

Число степеней свободы для остаточной вариации равно

$$df_{\text{ост.}} = n - p = n - m - 1, \quad (2.11)$$

где *m* — число независимых переменных.

Сумма квадратов отклонений для регрессии называется *объясненной* (или *факторной*) и определяется по формуле:

$$SS_{\text{факт.}} = \sum (\hat{y}_i - \bar{y})^2. \quad (2.12)$$

Остаточная сумма квадратов (отклонений) имеет вид

$$SS_{\text{ост.}} = \sum (y_i - \hat{y}_i)^2. \quad (2.13)$$

В соответствии с правилом сложения дисперсий сумма объясненной и остаточной вариации есть не что иное, как общая вариация зависимой переменной:

$$ESS_{\text{общ.}} = SS_{\text{факт.}} + SS_{\text{ост.}}. \quad (2.14)$$

$$\text{Тогда } MS_{\text{факт.}} = SS_{\text{факт.}} : df_{\text{факт.}}; \quad (2.15)$$

$$MS_{\text{ост.}} = SS_{\text{ост.}} : df_{\text{ост.}},$$

где MS — сумма квадратов отклонений в расчете на одну степень свободы.

Отсюда

$$F = MS_{\text{факт.}} : MS_{\text{ост.}} \quad (2.16)$$

В нашем примере $F = 38,5$. Полученное значение F -критерия необходимо сравнить с табличным значением, соответствующим $H_0: \sigma_{\text{факт.}}^2 = \sigma_{\text{ост.}}^2$. Распределение статистики F зависит от числа степеней свободы числителя ($df_{\text{факт.}}$) и знаменателя ($df_{\text{ост.}}$), а также от уровня значимости, т. е. вероятности ошибочного отклонения H_0 . На 5%-ном уровне значимости $df_{\text{факт.}} = 1$, $df_{\text{ост.}} = 28$, $F_{\text{факт.}} \pm = 4,23$. Поскольку $F_{\text{факт.}} > F_{\text{ост.}}$, H_0 не принимается.

В табл. 2.4 приведены значения зависимой переменной, рассчитанные по уравнению регрессии $\hat{y}_i = 1675,5 + 2,696x_i$, а также значение остатков $e_i = y_i - \hat{y}_i$.

Таблица 2.4
Вывод остатков

Наблюдение	Расчетное значение выпуска продукции, тыс. руб.	Остаток, тыс. руб.
	\hat{y}_i	e_i
1	2834,949	-514,949
2	2781,021	-281,021
3	3104,589	-544,589
4	3185,481	-485,481
5	2969,769	-76,7693
6	2834,949	106,0506
7	2861,913	158,0866
8	3050,661	597,3388
9	3158,517	241,4829
10	3239,409	-129,409
11	3374,229	-224,229
12	3212,445	33,55498
13	3347,265	32,73512
14	3455,121	-35,1208
15	3428,157	31,84321
16	3482,085	-17,0847
17	3616,905	-136,905
18	3401,193	108,8072

Окончание табл. 2.4

Наблюдение	Расчетное значение выпуска продукции, тыс. руб.	Остаток, тыс. руб. e_i
	\hat{y}_i	
19	3482,085	72,91527
20	3616,905	-16,9046
21	3751,724	-89,7244
22	3724,76	-45,7605
23	3832,616	-132,616
24	3589,941	-18,9406
25	3428,157	71,84321
26	3239,409	440,591
27	3536,013	173,9873
28	2843,949	290,0506
29	3077,625	98,37483
30	3428,157	291,8432

2.3. Свойства остатков

Первое свойство остатков следует из уравнения (2.3), которое показывает, что

$$\sum_{i=1}^n e_i x_i = 0, \quad (2.17)$$

т. е. остатки и объясняющая переменная не коррелированы.

Второе свойство, которым обладают оценки, полученные МНК, состоит в следующем:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

или же

$$\sum_{i=1}^n e_i \hat{y}_i = 0, \quad (2.18)$$

т. е. остатки и предсказанные значения \hat{y}_i не коррелированы.

Третье свойство остатков (e_i):

$$E(e_i) = 0, \quad (2.19)$$

математическое ожидание остатков равно нулю. В выборке $\bar{e} = 0$.

Четвертое свойство остатков: остатки имеют постоянную дисперсию, т. е. $\text{var}(e_i) = \sigma^2$ для всех $i = 1, 2, \dots, N$. Дисперсия остатков

$$\sigma_e^2 = 1 \text{ (в выборке } s_e^2 = 1\text{).} \quad (2.20)$$

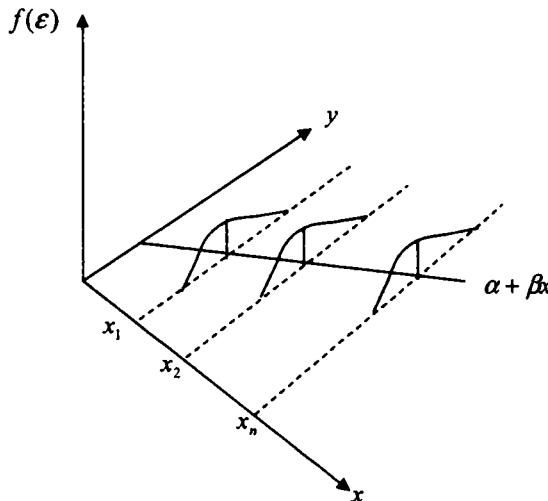


Рис. 2.3. Распределение остатков при фиксированных значениях независимой переменной x

Пятое свойство: остатки не коррелированы между собой:

$$E(\epsilon_i \epsilon_j) = 0 \quad (2.21)$$

для $i \neq j, ij = 1, 2, \dots, n$.

Зная остаток для i -го наблюдения, мы ничего не можем сказать об остатке для j -го наблюдения ($i \neq j$). Заметим, что это свойство остатков исчезает при построении регрессии по временным рядам, в которых наблюдение каждого последующего года (месяца, квартала) зависит от наблюдения предыдущего года (месяца, квартала).

Четвертое и пятое свойства остатков можно проиллюстрировать графически (рис. 2.3).

Постоянство дисперсии остатков называют *гомоскедастичностью* остатков. Если же дисперсия остатков непостоянна, то имеет место *гетероскедастичность* остатков.

Дисперсия остатков регрессии σ_e^2 неизвестна и должна быть оценена. При этом нужно принять во внимание, что дисперсии оцененных параметров s_b^2 и s_a^2 зависят от дисперсии остатков s_e^2 . МНК-оценки дисперсии остатков имеют вид:

$$s_e^2 = \sum_{i=1}^n e_i^2 / (n-2) \quad (2.22)$$

или же

$$E(s^2) = E\left(\sum_{i=1}^n e_i^2 / (n-2)\right) = \sigma_e^2 \quad (2.23)$$

Если распределение остатков ненормально, то наилучшим методом их оценки будет не МНК, а метод максимального правдоподобия.

Измерение ошибки аппроксимации, аналитическая и предсказательная силы уравнения регрессии определяются величиной коэффициента детерминации R^2 :

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (2.24)$$

Эта величина принимает значения от 0 до 1, R^2 может быть представлен как

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n y_i^2}, \quad (2.25)$$

где e_i — остаток (величина отклонения от линии регрессии, которая минимизируется МНК).

Если $\sum_{i=1}^n e_i^2$ велика, то регрессия не объясняет вариацию y . В этом случае коэффициент детерминации R^2 будет невелик (близок к нулю). Если многие точки фактических наблюдений лежат на линии регрессии или располагаются вблизи от нее, $\sum_{i=1}^n e_i^2$ мала и соответственно ко-

эффициент детерминации R^2 будет близок к 1. Если все наблюдения располагаются на линии регрессии, то $y_i = \hat{y}_i$, $e_i = 0$, а значит, и $\sum_{i=1}^n e_i^2 = 0$, тогда $R^2 = 1$. Если же точки, соответствующие наблюдени-

ям, не лежат на линии регрессии, то $\sum_{i=1}^n y_i^2 = 0$, а значит, регрессия не объясняет вариацию y_i . В случае если $\sum_{i=1}^n \hat{y}_i^2 = \sum_{i=1}^n e_i^2$, $R^2 = 0$. Тогда наи-

лучшей аппроксимацией данных будет линия $y_i = \bar{y}$ для всех i , это горизонтальная линия, параллельная оси абсцисс и проходящая через точку $y_i = \bar{y}$, что соответствует случаю полной независимости переменных x и y .

Степень аппроксимации данных выборки полученной регрессией $y = a + bx$ оценивается с помощью средней ошибки аппроксимации:

$$\bar{e} = \frac{\sum |y_i - \bar{y}|}{n}. \quad (2.26)$$

Большей информативностью обладает средняя относительная ошибка аппроксимации:

$$\bar{e}^* = \left[\sum_{i=1}^n |y_i - \bar{y}| / \sum_{i=1}^n y_i \right] \cdot 100. \quad (2.27)$$

Значения средней относительной ошибки аппроксимации, не превышающие 10%, свидетельствуют о хорошем соответствии линии регрессии исходным данным.

Коэффициент детерминации, R^2 , имеет две трактовки:

1) это квадрат коэффициента парной корреляции между фактическими и расчетными значениями зависимой переменной, т. е. $R^2 = r_{yx}^2$;

2) это квадрат коэффициента парной корреляции между y и x , т. е.: $R^2 = r_{xy}^2$:

$$r_{yx}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.28)$$

Коэффициент парной корреляции — это мера тесноты линейной связи:

$$r_{yx} = \frac{\sum_{(i)} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}; \quad -1 \leq r_{yx} \leq 1.$$

Можно представить коэффициент парной корреляции r_{yx} через коэффициент регрессии b , определяющий наклон регрессии к оси абсцисс:

$$r_{yx} = b \cdot \frac{s_x}{s_y}. \quad (2.29)$$

По данным рассматриваемого примера коэффициент парной корреляции равен $r_{yx} = 0,761$, а коэффициент детерминации $r_{yx}^2 = 0,579$, т. е. на 57,9% вариация выпуска продукции зависит от основных фондов, а на 42,1% — от других факторов.

Подчеркнем, что коэффициент парной корреляции представляет собой меру линейной связи между y и x . Если, например, существует полная квадратическая связь между y и x , которая описывается уравнением регрессии $y = a + bx + cx^2$, то, значит, и r_{yx} может иметь значения, далекие от 1. В этом случае нужно провести преобразование переменных, с тем чтобы линеаризировать уравнение регрессии. Примем $x^2 = z$, тогда параболическое уравнение регрессии будет иметь вид $\hat{y} = a + bx + cz$, т. е. это уравнение адекватно линейному уравнению множественной регрессии (см. гл. 3).

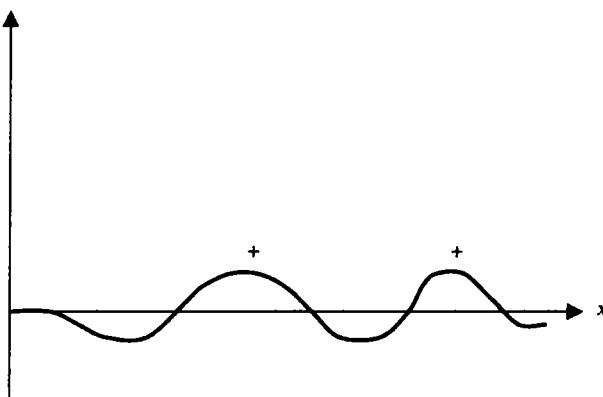


Рис. 2.4. График остатков (случай гомоскедастичности)

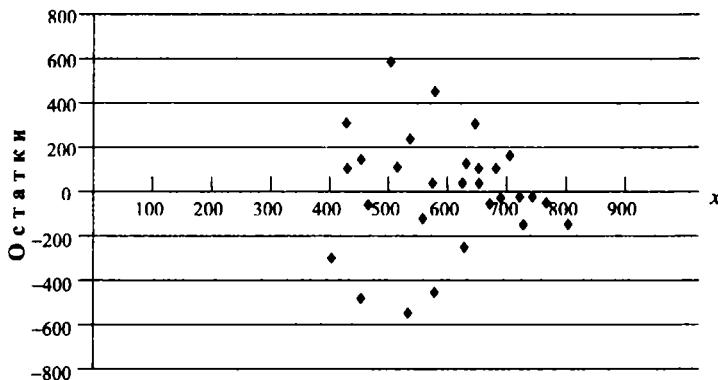


Рис. 2.5. График остатков по данным примера

Сформулированные свойства остатков проверяются после нахождения параметров уравнения регрессии. По уравнению регрессии находятся расчетные (предсказанные) значения зависимой переменной (\hat{y}). После этого производится расчет остатков $e_i = y_i - \hat{y}_i$. Затем строится график остатков (*residual plot*). В случае гомоскедастичности остатков положительные и отрицательные значения остатков чередуются и находятся в области, параллельной оси абсцисс (рис. 2.4).

График остатков по данным нашего примера о зависимости выпуска продукции от стоимости основных фондов представлен на рис. 2.5. Мы видим, что число точек с положительными и отрицательными значениями отклонений оказалось равным (по 15 единиц). Сумма остатков равна нулю. Однако расположение «облака» остатков позволяет предположить наличие гетероскедастичности.

Кроме визуального анализа остатков, существует ряд специальных тестов, позволяющих выявить гетероскедастичность остатков: тесты Гольдфельда—Квандта, Парка, Глейзера, Уайта, ранговой корреляции Спирмена и др. Названные тесты рассмотрены в гл. 3.

Наличие гетероскедастичности сказывается на точности предсказания значения зависимой переменной на основе регрессии. Ошибка предсказания может быть представлена следующим образом:

$$y_0 - \hat{y}_0 = [y_0 - E(y_0)] + [E(y_0) - \hat{y}_0] = e_0 + [E(y_0) - \hat{y}_0]$$

где y_0 — фактическое значение для объекта 0;
 \hat{y}_0 — предсказанное значение для объекта 0;
 $[y_0 - E(y_0)]$ — отклонение фактического значения от «истинного» в генеральной совокупности;
 $[E(y_0) - \hat{y}_0]$ — отклонение «истинного» значения от предсказанного по регрессии.

Тогда дисперсия ошибки предсказания имеет вид

$$\text{var}(e_0) + \text{var}[E(y_0) - \hat{y}_0] + 2\text{cov}[e_0, E(y_0) - \hat{y}_0] = \\ = \sigma_{y_{\text{oct}}}^2 \left[1 + (1 - n) \right] + (x_0 - \bar{x})^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2.$$

Таким образом, можно утверждать, что ошибка предсказания значения y на основе уравнения регрессии зависит от остаточной дисперсии y (дисперсии остатков), от объема выборки n , от того, насколько значение x_0 (для объекта предсказания) отличается от среднего значения по наблюдаемым данным. Чем меньше остаточная дисперсия, чем больше объем выборки n , и чем меньше вариация x , т. е. $(\sum(x_i - \bar{x})^2)$, а также чем ближе x_0 к \bar{x} , тем меньше ошибка предсказания.

Как уже отмечалось, на основе уравнения регрессии могут быть получены точечные прогнозные значения y_i и интервальные оценки. Например, 95%-ный доверительный интервал значения y_0 при x_0 имеет вид

$$(a + bx_0) \pm t_{0,025, n-2} \left\{ s_{\text{oct}} \cdot \left[1 + (1 - n) + (x_0 - \bar{x})^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2} \right\},$$

где s_{oct} — выборочная оценка остаточной дисперсии y (заменияет σ_{oct});

$t_{0,025, df=n-2}$ — критическое 2,5%-ное значение, полученное по таблице t -распределения с $n-2$ степенями свободы.

Как показано на рис. 2.6, границы доверительного интервала представляют собой гиперболу. Самое «узкое» значение интервала — в точке \bar{x} и ее окрестностях; чем дальше удалено значение x_i от \bar{x} , тем шире становится доверительный интервал.

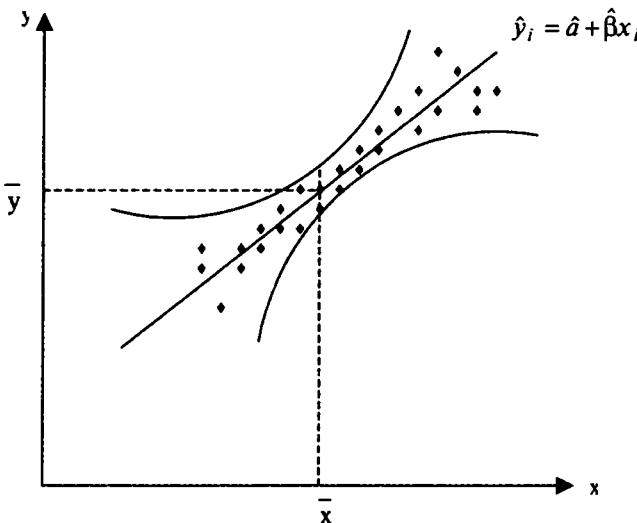


Рис. 2.6. 95%-ный доверительный интервал

Устранение из выборки предприятий, для которых значение остатка оказалось большим (в нашем примере это предприятия с номерами 1, 3, 8), позволило бы существенно улучшить свойства регрессии и повысить точность предсказания объема выпуска.

В заключение отметим, что парная регрессия довольно редко выступает в качестве эконометрической модели, поскольку исследуемые экономические явления формируются под влиянием не одного, а нескольких факторов. Этим обстоятельством определяется гораздо большая распространенность множественной регрессии в эконометрическом моделировании.

Контрольные вопросы

1. Дайте определение парной регрессии.
2. Что такое линия регрессии?
3. Поясните экономическую сущность параметров уравнения парной регрессии.
4. Как производится оценка параметров уравнения парной регрессии?
5. Как производится оценка качества уравнения в целом?
6. Каковы свойства остатков модели парной регрессии?
7. Как измеряется ошибка аппроксимации?
8. Какие трактовки коэффициента детерминации вам известны?
9. Как осуществляется линеаризация модели?
10. Как анализируется график остатков модели регрессии?
11. От чего зависит точность предсказания значения зависимой переменной на основе уравнения парной регрессии?

ГЛАВА 3. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

3.1. Спецификация модели

Модель множественной регрессии – это уравнение, отражающее корреляционную связь между результатом и несколькими факторами. В общем виде оно может быть записано как

$$y = f(x_1, x_2, \dots, x_p, e) \quad (3.1)$$

где y – зависимая переменная (результат);

x_1, x_2, \dots, x_p – независимые переменные (факторы);

e – случайный остаток;

f – некая математическая функция.

Необходимо различать модель множественной регрессии для генеральной совокупности и для выборки: W_0 . В качестве функций множественной регрессии часто выбирают наиболее простые – линейную, показательную и степенную функции:

$y = a + b_1 x_1 + \dots + b_p x_p + e$ – линейная функция;

$y = ax_1^{b_1} \cdot \dots \cdot x_p^{b_p} e$ – степенная функция;

$y = ab_1^{x_1} \cdot \dots \cdot b_p^{x_p} e$ – показательная функция.

где a, b_1, \dots, b_p – параметры функций.

Эти функции могут быть использованы и при формировании «смешанных» моделей. Например, можно построить следующее уравнение множественной регрессии с тремя факторами:

$$y = ab_1^{x_1} x_2^{b_2} b_3^{x_3} e$$

В данном уравнении использованы две функции: показательная (для учета влияния факторов x_1 и x_3) и степенная (для учета влияния фактора x_2).

При проведении корреляционно-регрессионного анализа предполагается, что наблюдения, на основе которых он проводится, были получены по однородной совокупности единиц. То есть механизм воздействия факторов на результат должен быть примерно одинаков на разных единицах совокупности. Для обеспечения статистической достоверности модели количество наблюдений должно быть в 8–10 раз больше количества параметров, не считая «свободного члена»*.

Результат и факторы – это количественные показатели. В прошлом случае считают, что для них нет границ изменения, т. е. они принадлежат интервалу $(-\infty; +\infty)$, и что факторы неслучайны. При построении эконометрической модели предполагается, что факторы оказывают влияние на результат, причем влияние отдельного фактора не зависит от влияния других факторов. В противном случае изменение значения какого-либо фактора окажет на результат как прямое воздей-

* Термин «свободный член» относится к линейной функции. В то же время так удобно называть также параметр, не относящийся непосредственно к какому-либо фактору. В наших моделях он обозначается буквой a .

ствие, так и опосредованное — через другие факторы. Это, в свою очередь, приведет к ошибкам в интерпретации результатов корреляционно-регрессионного анализа.

Корреляционная связь может существовать как между двумя факторами (*интеркорреляция*), так и между несколькими факторами (*мультиколлинеарность*). Существование корреляционной связи между факторами может быть выявлено с помощью показателей корреляции между ними, в частности с помощью парных коэффициентов корреляции, которые можно записать в виде матрицы

$$r_{xx} = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_p} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_p} \\ r_{x_px_1} & r_{x_px_2} & r_{x_px_p} \end{pmatrix}. \quad (3.2)$$

Коэффициент корреляции фактора с самим собой равен ($r_{x_ix_i} = 1$), а коэффициент корреляции фактора i с фактором j равен коэффициенту корреляции фактора j с фактором i ($r_{x_ix_j} = r_{x_jx_i}$). Следовательно, данная матрица является симметрической, поэтому в ней указывают только главную диагональ и элементы под ней:

$$r_{xx} = \begin{pmatrix} 1 & & \\ r_{x_2x_1} & 1 & \\ r_{x_px_1} & r_{x_px_2} & 1 \end{pmatrix}. \quad (3.3)$$

Наличие мультиколлинеарности можно подтвердить, найдя определитель матрицы (3.2). Если связь между факторами полностью отсутствует, то недиагональные элементы будут равны нулю, а определитель матрицы — единице. Если связь между факторами близка к функциональной (то есть является очень тесной), то определитель матрицы r_{xx} будет близок к нулю.

Пусть, например, изучается влияние на объем выпуска продукции (y , тыс. руб.) различных факторов: количества занятых (x_1 , чел.), стоимости основных фондов (x_2 , тыс. руб.), средней заработной платы на предприятии (x_3 , руб.) (табл. 3.1, данные условные).

Таблица 3.1

Показатели деятельности малых предприятий

Номер предприятия	Объем выпуска продукции (y , тыс. руб.)	Количество занятых (x_1), чел.	Стоимость основных фондов (x_2), тыс. руб.	Средняя заработка на предприятии (x_3), руб.
1	2320	40	430	5200
2	2500	40	410	5610
3	2560	42	530	6800

Окончание табл. 3.1

Номер предприятия	Объем выпуска продукции (y), тыс. руб.	Количество занятых (x_2), чел.	Стоимость основных фондов (x_2), тыс. руб.	Средняя заработка на предприятии (x_2), руб.
4	2700	45	560	6300
5	2893	47	480	6400
6	2941	49	430	7000
7	3020	52	440	6900
8	3648	55	510	6950
9	3400	59	550	7200
10	3110	44	580	8000
11	3150	41	630	7830
12	3246	51	570	7500
13	3380	60	620	7600
14	3420	46	660	7760
15	3460	53	650	7800
16	3465	57	670	7800
17	3480	44	720	7900
18	3510	48	640	8040
19	3555	56	670	8600
20	3600	54	720	8900
21	3662	51	770	9600
22	3679	46	760	8400
23	3700	56	800	9200
24	3571	57	710	8360
25	3500	48	650	8100
26	3680	59	580	7900
27	3710	60	690	7220
28	3125	55	430	6500
29	3176	45	520	5630
30	3720	60	650	6840

Матрица парных коэффициентов корреляции между факторами x_1 , x_2 , x_3 равна

$$r_{xx} = \begin{pmatrix} 1 & & \\ 0,305 & 1 & \\ 0,353 & 0,840 & 1 \end{pmatrix}.$$

Следовательно, связь между факторами x_1 и x_2 ; x_1 и x_3 слабая ($r_{x_1x_2} = 0,305$; $r_{x_1x_3} = 0,353$), а между факторами x_2 и x_3 — тесная ($r_{x_2x_3} = 0,840$). То есть вариация количества занятых не связана с вариа-

цией стоимости основных фондов и вариацией средней заработной платы. В то же время вариация стоимости основных фондов связана с вариацией средней заработной платы, причем связь прямая (коэффициент корреляции больше нуля). Это можно объяснить, например, тем, что на более дорогом оборудовании должны работать более квалифицированные рабочие, которым выплачивается более высокая заработная плата.

Определитель данной матрицы равен 0,258. Он близок к нулю, что является следствием наличия связи между факторами x_1 , x_2 и x_3 .

Рассмотрим влияние каждого из факторов на результат. Парные коэффициенты корреляции результата с каждым из факторов равны

$$r_{yx_1} = 0,682; \quad r_{yx_2} = 0,761; \quad r_{yx_3} = 0,737.$$

Все факторы оказывают заметное влияние на результат. Наибольшее влияние оказывает фактор x_2 , наименьшее — фактор x_1 , однако различия в тесноте связи невелики.

В модель регрессии должны быть включены факторы, тесно связанные с результатом и слабо связанные друг с другом. В данной задаче этим требованиям удовлетворяют две модели регрессии:

$$\hat{y} = f(x_1, x_3) \quad \text{и} \quad \hat{y} = f(x_1, x_2).$$

Вывод о слабой связи между факторами можно сделать как рассматривая парные коэффициенты корреляции (для этих пар факторов они невелики), так и найдя определители соответствующих матриц. Например, для модели матрица $\hat{y} = f(x_1, x_3)$ парных коэффициентов межфакторной корреляции равна

$$r_{xx} = \begin{pmatrix} 1 & \\ 0,353 & 1 \end{pmatrix}$$

Ее определитель примерно равен 0,87. Это значение близко к единице, следовательно, связь между факторами слабая.

Аналогично для модели $\hat{y} = f(x_1, x_2)$ определитель матрицы коэффициентов корреляции между факторами равен 0,91, что подтверждает обоснованность включения обоих факторов в модель.

Выбор конкретной модели зависит от целей исследования. В данном примере первая модель содержит независимые переменные, характеризующие один и тот же фактор производства — труд. Таким образом, эта модель может быть использована для решения задачи изучения влияния данного фактора на объем продукции. Вторая модель учитывает уже два фактора производства — труд и капитал. Область ее применения — анализ влияния на объем продукции разных факторов производства.

Часто на результат оказывает влияние достаточно большое количество факторов и отсутствует возможность выделить из них наиболее значимые, подлежащие включению в модель регрессии. В этом случае рассматривают несколько моделей с разным составом факторов. Наилучшей является модель, имеющая значимые параметры и максимальный показатель тесноты связи.

Существует несколько алгоритмов (методов) перебора моделей, например метод включения факторов, метод исключения факторов, шаговый регрессионный анализ, ступенчатый регрессионный анализ.

Метод последовательного включения факторов предполагает, что сначала будет построена модель с фактором, наиболее тесно связанным с результатом. Затем поочередно добавляются другие факторы. Каждый раз оценивается целесообразность включения нового фактора с точки зрения сокращения остаточной дисперсии.

При использовании *метода исключения факторов* сначала строится модель с максимально большим количеством факторов, из которой затем поочередно исключаются незначимые факторы до тех пор, пока модель не будет иметь только значимые параметры при факторах.

Шаговый регрессионный анализ можно рассматривать как развитие метода включения факторов. Построение модели начинается с расчета парной регрессии с фактором, наиболее тесно связанным с результатом. Добавление каждого нового фактора сопровождается не только оценкой значимости включения данного фактора, но и проверкой значимости влияния на результат прочих факторов, уже включенных в модель. Выявленные незначимые факторы исключаются из модели. Процесс завершается, если добавление нового фактора не приводит к заметному улучшению качества модели.

С построения парного уравнения регрессии с наиболее значимым (по степени влияния на результат) фактором начинается и *ступенчатый регрессионный анализ*. Затем по полученной модели находят случайные остатки ϵ . Так как эти остатки отражают влияние факторов, не включенных в уравнение регрессии, далее строится уравнение зависимости ϵ от следующего по степени влияния на результат фактора. По этому уравнению также находят случайные остатки, которые рассматриваются как результат влияния третьего фактора. Процедура повторяется до тех пор, пока вновь полученное уравнение регрессии значимо. Уравнение множественной регрессии, построенное на основе этого метода, должно включать в себя все факторы, содержащиеся в рассмотренных значимых уравнениях парной регрессии. Этот метод более прост с точки зрения расчетов, однако недостаточно точен, так как не учитывает взаимосвязь факторов.

Показатели, выбранные в качестве результата и факторов, могут оказаться неколичественными переменными. Модели, в которых есть факторы, являющиеся неколичественными переменными, рассматриваются в главе «Фиктивные переменные». Модели, в которых результат является неколичественной переменной, рассматриваются в главе «Модели бинарного и множественного выбора».

Достаточно распространенной является ситуация, когда результат является количественной переменной, но его значения ограничены определенным интервалом. Различают два типа моделей: с усеченными данными и с цензурированными данными.

Если наблюдение проводится не над всей статистической совокупностью, а над ее частью, для которой свойственно попадание значений результативного признака в определенный числовой интервал, то полученная выборка называется *усеченной*. Например, изучаются доходы только малообеспеченного населения. Тогда верхней границей результата «среднедушевой доход» будет величина прожиточного минимума или какая-либо другая, аналогичная ей.

Наблюдения могут проводиться над всей совокупностью исследуемых объектов, однако в силу каких-либо причин значениям результата, меньшим или большим определенной числовой границы, присваивается значение, равное этой границе. В этом случае говорят, что полученные данные представляют собой *цензуринную* выборку. Например, на основе объема продаж какого-либо товара можно оценить спрос на него. Эта оценка будет верна только в том случае, если предложение превышает спрос. Если предложение не обеспечивает спрос, то полученные данные будут занижать его величину, ограничивая ее величиной предложения. В данном случае имеет место так называемое цензирование сверху. Если обозначить наблюдавшее значение результата как y , а ненаблюдавшее — как y^ϕ , объем предложения — как Q , то цензирование сверху можно записать как

$$y_i = \begin{cases} y_i^\phi, & \text{если результат (спрос) меньше предложения: } y_i^\phi < Q \\ Q, & \text{если результат (спрос) больше предложения: } y_i^\phi \geq Q \end{cases}$$

Частным случаем модели с цензуринными данными является *tobit-модель*, названная так в честь ее создателя Джеймса Тобина. Он исследовал расходы отдельных семей на автомобиль. Эти расходы ограничены снизу нулем:

$$y_i = \begin{cases} y_i^\phi, & \text{если } y_i^\phi > 0 \\ 0, & \text{если } y_i^\phi \leq 0 \end{cases} \quad (3.4)$$

Таким образом, отсутствие расходов можно рассматривать как отрицательные расходы, т. с. доходы. Дж. Тобин показал, что неучет этого обстоятельства ведет к некорректной оценке параметров модели регрессии.

В зависимости от типа исходных данных могут применяться различные методы оценки параметров регрессионных моделей.

Как и в парной регрессии, в множественной регрессии различают *фактическое и теоретическое (выровненное) значение* результата. Фактическое значение — это наблюдавшее значение результата. В модели регрессии оно обозначается как y . Фактическое значение результата можно представить в виде двух составляющих: теоретического (выровненного) значения и случайного остатка e . Теоретическое значение (\hat{y}) равно значению функции регрессии при определенных значениях факторов и отсутствующем случайном остатке:

$$\hat{y} = f(x_1, x_2, \dots, x_p).$$

Например, для линейного уравнения множественной регрессии с двумя факторами вида

$$y = a + b_1 x_1 + b_2 x_2 + e \quad (3.5)$$

теоретическое значение равно

$$\hat{y} = a + b_1 x_1 + b_2 x_2. \quad (3.6)$$

Таким образом, для линейной модели можно записать:

$$y = \hat{y} + e. \quad (3.7)$$

Фактическое значение результата может быть связано со случайными остатками как аддитивно (как в формуле для множественной линейной регрессии), так и мультипликативно — например, для двухфакторной степенной регрессии вида

$$y = ax_1^{b_1} x_2^{b_2} e \quad (3.8)$$

фактическое значение результата связано с выровненным значением и случайным остатком мультипликативной зависимостью

$$y = \hat{y} \cdot e. \quad (3.9)$$

При построении регрессии анализу остатков уделяется большое внимание, так как их величина и характер распределения определяют качество модели.

3.2. Натуральная и стандартизованная форма модели множественной регрессии

Уравнение регрессии, построенное по исходным данным, называется моделью в *натуральной форме* или в *натуральном масштабе*. Например, множественная линейная регрессия в натуральном масштабе имеет вид

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + \varepsilon. \quad (3.10)$$

Если провести стандартизацию переменных, входящих в модель, т. е. выполнить следующие преобразования:

$$t_y = \frac{y - \bar{y}}{\sigma_y}; \quad t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}} \text{ для всех } i, \quad (3.11)$$

а затем построить по новым переменным модель множественной линейной регрессии

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \dots + \beta_p t_{x_p} + u, \quad (3.12)$$

то такая модель будет называться моделью в стандартизованной форме или в стандартизованном масштабе. Стандартизованные переменные имеют среднее, равное нулю, и среднее квадратическое отклонение, равное единице. Например, для результата y имеем:

$$\bar{t}_y = \frac{\sum t_y}{n} = \frac{\sum (y - \bar{y})}{n \cdot \sigma_y} = \frac{0}{n \cdot \sigma_y} = 0; \quad (3.13)$$

$$\sigma_{t_y} = \sqrt{\frac{\sum (t_y - \bar{t}_y)^2}{n}} = \sqrt{\frac{\sum t_y^2}{n}} = \sqrt{\frac{\sum (y - \bar{y})^2}{n \cdot \sigma_y^2}} = \sqrt{\frac{\sigma_y^2}{\sigma_y^2}} = 1. \quad (3.14)$$

Коэффициенты модели в стандартизованной форме отличаются от коэффициентов исходной модели и поэтому обозначены другими символами — β , а случайный остаток — u . Свободный член в этой модели отсутствует, что следует из свойств стандартизированной переменной и результатов применения метода наименьших квадратов (см. п. 3.3).

3.3. Оценка параметров уравнения множественной регрессии

Данные, используемые в корреляционно-регрессионном анализе, рассматриваются как *выборочные*, неполные. Поэтому количественные характеристики связи между показателями, полученные на основе этих данных, также являются выборочными, т. е. содержащими некую ошибку, отличающимися от объективно существующих, но неизвестных «подлинных» характеристик.

Параметры уравнения регрессии, найденные на основе имеющихся у исследователя данных, называют *оценками параметров*, подчеркивая то, что они рассчитаны по выборочным данным. Оценки параметров могут меняться от выборки к выборке, поэтому они рассматриваются как *случайные величины*.

Так как найденные параметры являются лишь выборочными оценками неизвестных параметров по генеральной совокупности, то возникает вопрос об их качестве. Считается, что оценками параметров можно пользоваться для дальнейшего анализа и прогноза, если эти оценки являются несмещенными, эффективными и состоятельными.

Оценка параметра является *несмещенной*, если ее математическое ожидание равно оцениваемому параметру. Например, математическое ожидание оценки коэффициента регрессии b_j равно его значению в генеральной совокупности $b_j^{\text{ген}}$:

$$M b_j = b_j^{\text{ген}} \quad (3.15)$$

Оценка параметра является *эффективной*, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок данного параметра по выборкам одного и того же объема:

$$M(b_j - b_j^{\text{ген}})^2 = \sigma_{b_j}^2 = \min \sigma_{b_j}^2 \quad (3.16)$$

Оценка параметра является *состоятельной*, если с увеличением числа наблюдений оценка параметра стремится к его значению в генеральной совокупности:

$$b_j \xrightarrow{n \rightarrow \infty} b_j^{\text{ген}} \quad (3.17)$$

Наиболее простым методом, применяемым для оценки параметров множественной регрессии, является *метод наименьших квадратов* (МНК). Для того чтобы МНК-оценки параметров были несшенными, эффективными и состоятельными, необходимо выполнение определенных требований, называемых *предпосылками МНК*. Эти требования касаются статистических свойств исходных данных:

- факторы являются неслучайными величинами, не связанными между собой;

- результат является случайной величиной, не ограниченной сверху или снизу;
- для каждого конкретного значения фактора (или факторов) результат рассматривается как отдельная случайная величина Y_i . Ее распределение описывается нормальным законом с математическим ожиданием, равным выровненному значению результата:

$$MY_i = f(x_1, x_2, \dots, x_p) = \hat{y}_i, \quad (3.18)$$

где $f(x_1, x_2, \dots, x_p)$ – модель регрессии, линейная по параметрам. Вторая характеристика нормального распределения – среднее квадратическое отклонение (σ) – может быть любым, однако оно должно быть одинаковым для всех распределений результата;

- различные случайные величины Y_i и Y_j независимы друг от друга:

$$r_{Y_i Y_j} = 0, \quad i \neq j. \quad (3.19)$$

Линейная модель регрессии, для которой выполняются эти условия, называется *классической нормальной линейной моделью*.

Метод наименьших квадратов может быть применен к моделям регрессии, линейным по параметрам. Если функция регрессии нелинейна по параметрам, необходима ее предварительная линеаризация.

В частности, для степенной, показательной функций и их комбинаций в качестве приема линеаризации может быть использовано логарифмирование правой и левой частей уравнения регрессии. Например, двухфакторная функция вида

$$y = ax_1^{b_1} b_2^{x_2} e \quad (3.20)$$

после линеаризации путем логарифмирования принимает вид

$$\ln y = \ln a + b_1 \ln x_1 + x_2 \ln b_2 + \ln e. \quad (3.21)$$

Введя новые обозначения для логарифмированных величин:

$$\ln y \equiv Y \quad \ln a \equiv A \quad \ln x_1 \equiv X_1 \quad \ln b_2 \equiv B_2 \quad \ln e \equiv e_1,$$

получим:

$$Y = A + b_1 X_1 + x_2 B_2 + e_1. \quad (3.22)$$

Для обратной функции вида

$$y = \frac{1}{a + b_1 x_1 + \dots + b_p x_p + e} \quad (3.23)$$

линеаризация заключается в следующих преобразованиях:

$$\frac{1}{y} = a + b_1 x_1 + \dots + b_p x_p + e;$$

$$\frac{1}{y} \equiv Y;$$

$$Y = a + b_1 x_1 + \dots + b_p x_p + e.$$

Применение МНК к функции множественной линейной регрессии вида $y = a + b_1x_1 + b_2x_2 + \dots + b_px_p + e$ приводит к следующей системе уравнений:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + \dots + b_p \sum x_p \\ \sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \dots + b_p \sum x_1 x_p \\ \sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 + \dots + b_p \sum x_2 x_p \\ \vdots \\ \sum x_p y = a \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p + \dots + b_p \sum x_p^2 \end{cases} \quad (3.24)$$

Из первого уравнения этой системы можно выразить свободный член как

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \dots - b_p \bar{x}_p. \quad (3.25)$$

Отсюда следует, что для стандартизованной формы модели свободный член тоже равен нулю (так как средние всех переменных, входящих в нее, равны нулю). Применение МНК к стандартизованной форме модели дает следующий результат:

$$\begin{cases} \sum t_{x_1} t_y = \beta_1 \sum t_{x_1}^2 + \beta_2 \sum t_{x_1} t_{x_2} + \dots + \beta_p \sum t_{x_1} t_{x_p} \\ \sum t_{x_2} t_y = \beta_1 \sum t_{x_1} t_{x_2} + \beta_2 \sum t_{x_2}^2 + \dots + \beta_p \sum t_{x_2} t_{x_p} \\ \vdots \\ \sum t_{x_p} t_y = \beta_1 \sum t_{x_1} t_{x_p} + \beta_2 \sum t_{x_2} t_{x_p} + \dots + \beta_p \sum t_{x_p}^2 \end{cases} \quad (3.26)$$

Эту систему можно упростить. Так как

$$\sum t_{x_i} t_{x_j} = \sum \frac{(x_i - \bar{x}_i) \cdot (x_j - \bar{x}_j)}{\sigma_{x_i} \sigma_{x_j}} = r_{x_i x_j}, \quad \sum t_y t_{x_i} = r_{yx_i}, \quad \sum t_{x_i}^2 = r_{x_i x_i} = 1,$$

то

$$\begin{cases} \beta_1 + \beta_2 r_{x_1 x_2} + \dots + \beta_p r_{x_1 x_p} \\ r_{x_2 y} = \beta_1 r_{x_1 x_2} + \beta_2 + \dots + \beta_p r_{x_2 x_p} \\ \vdots \\ r_{x_p y} = \beta_1 r_{x_1 x_p} + \beta_2 r_{x_2 x_p} + \dots + \beta_p \end{cases} \quad (3.27)$$

Отсюда для модели с двумя факторами имеем:

$$\beta_1 = \frac{r_{x_1 y} - r_{x_2 y} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}; \quad (3.28)$$

$$\beta_2 = \frac{r_{x_2 y} - r_{x_1 y} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}.$$

Развитие вычислительной техники позволяет сократить трудоемкость расчетов, сконцентрировавшись на содержательных вопросах моделирования. Исследователю достаточно задать лишь общий алгоритм решения. Этот алгоритм в компактной форме записывается в виде операций над матрицами.

Так, уравнение множественной линейной регрессии в матричной форме будет иметь вид:

$$Y = XB + E, \quad (3.29)$$

где X – матрица значений факторов вида

$$X = \begin{pmatrix} 1 & x_{11} & x_{p1} \\ 1 & x_{12} & x_{p2} \\ \dots & & \\ 1 & x_{1n} & x_{pn} \end{pmatrix}. \quad (3.30)$$

Первый столбец этой матрицы состоит из единиц, которые рассматриваются как значения дополнительной переменной — сомножителя свободного члена.

Y, B, E – соответственно матрицы-столбцы значений результата, параметров регрессии и случайных остатков:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; B = \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_p \end{pmatrix}; E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Тогда выражение для нахождения параметров уравнения регрессии с помощью МНК в матричной форме будет иметь вид

$$B = (X^T X)^{-1} X^T Y, \quad (3.31)$$

где X^T – транспонированная матрица, X^{-1} – обратная матрица.

Средства нахождения вектора коэффициентов B имеются в различных пакетах прикладных программ, например в *MS Excel*.

Если исходные данные не удовлетворяют перечисленным в начале этого параграфа предпосылкам МНК, то для нахождения параметров модели регрессии можно использовать другой метод – *метод максимального правдоподобия (ММП)*.

Для его применения необходимо знать закон распределения результативного признака. Предполагается, что для каждого отдельного значения фактора (в парной регрессии) или совокупности значений факторов (в множественной регрессии) результат подчиняется распределению одного вида, но с разными параметрами или даже разным видом распределений.

Сущность метода состоит в построении *функции правдоподобия*, которая представляет собой закон совместного распределения величин $Y_i : P(Y_1 Y_2 \dots Y_p)$. В зависимости от вида функции правдоподобия ее в ряде случаев для упрощения решения можно прологарифмировать, полу-

чив логарифмическую функцию правдоподобия. Дальнейшее применение ММП связано с нахождением максимума функции правдоподобия (или ее логарифма) путем ее дифференцирования по неизвестным параметрам, приравнивания полученных выражений нулю и решения полученной системы уравнений.

Например, если соблюдаются предпосылки МНК, то функция правдоподобия будет иметь вид:

$$L(Y_i, MY_i, \sigma) = \prod_{i=1}^n f_N(Y_i, MY_i, \sigma), \quad (3.32)$$

где $L(Y_i, MY_i, \sigma)$ — функция правдоподобия для случайных величин Y_i , распределенных с неизвестными параметрами MY_i и σ ;

$f_N(Y_i, MY_i, \sigma)$ — плотность нормального распределения для случайной величины Y_i ,

$$f(Y_i, MY_i, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}} = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_p x_{pi})^2}{2\sigma^2}}$$

Множитель $\frac{1}{\sqrt{2\pi}}$ не влияет на достижение максимума функции правдоподобия, поэтому его при дальнейших преобразованиях опускают.

Логарифмическая функция правдоподобия будет тогда иметь вид:

$$l(Y_i, MY_i, \sigma) = \sum_{i=1}^n \left(\frac{-(y_i - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_p x_{pi})^2}{2\sigma^2} - \ln \sigma \right). \quad (3.33)$$

Несложно убедиться, что дифференцирование функции (3.33) по неизвестным параметрам a, b_1, b_2, \dots, b_p приведет к системе (3.24), полученной по методу наименьших квадратов. Дифференцирование по неизвестному среднему квадратическому отклонению приведет к формуле

$$\sigma^2 = \frac{\sum_{i=1}^n (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_p x_{pi})^2}{n} = \frac{\sum_{i=1}^n \varepsilon^2}{n}. \quad (3.34)$$

Если рассматриваются модели регрессии, для которых не соблюдаются предпосылки МНК, то применение ММП даст уже другой результат. Рассмотрим, например, модель множественной линейной регрессии (3.10) как *tobit*-модель (3.4). Для нее обычно предполагается, что случайные величины Y_i^ϕ распределены по нормальному закону распределения. Если $Y_i^\phi > 0$, то наблюдаемые и ненаблюдаемые значения результата совпадут и функция распределения каждой такой величины определяется плотностью распределения $f_N(Y_i, MY_i, \sigma)$.

Если $Y_i^\phi \leq 0$, то наблюдаемые значения результата будут равны нулю. Для каждой такой величины Y_i^ϕ можно определить вероятность ее

попадания в интервал $(-\infty; 0]$. Можно доказать, что эта вероятность составит

$$P(Y_i^\phi \leq 0) = F_i(0) = 1 - \Phi\left(\frac{\hat{y}_i}{\sigma}\right),$$

где $\Phi(\cdot)$ — нормированная функция нормального распределения.

Таким образом, функция правдоподобия будет равна

$$L = \prod_{y_i > 0} f(y_i) \prod_{y_i \leq 0} (1 - \Phi\left(\frac{\hat{y}_i}{\sigma}\right)).$$

После логарифмирования этой функции и ее дифференцирования по неизвестным параметрам получим следующую систему уравнений:

$$\left\{ \begin{array}{l} \sum_{y_i > 0} \left(\frac{y_i - a - b_1 x_1 - \dots - b_p x_p}{\sigma^2} \right) + \sum_{y_i = 0} \frac{-\Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right) \frac{1}{\sigma}}{1 - \Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right)} = 0 \\ \sum_{y_i > 0} \left(\frac{(y_i - a - b_1 x_1 - \dots - b_p x_p)x_1}{\sigma^2} \right) + \sum_{y_i = 0} \frac{-\Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right) x_1}{1 - \Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right)} = 0 \\ \sum_{y_i > 0} \left(\frac{(y_i - a - b_1 x_1 - \dots - b_p x_p)x_p}{\sigma^2} \right) + \sum_{y_i = 0} \frac{-\Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right) x_p}{1 - \Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right)} = 0 \\ \sum_{y_i > 0} \left(\frac{(y_i - a - b_1 x_1 - \dots - b_p x_p)^2}{\sigma^3} - \frac{1}{\sigma} \right) + \sum_{y_i = 0} \frac{\Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right) (a + b_1 x_1 + \dots + b_p x_p)}{\left(1 - \Phi\left(\frac{a + b_1 x_1 + \dots + b_p x_p}{\sigma}\right)\right)} = 0 \end{array} \right.$$

В данной системе $\varphi(\cdot)$ — плотность нормированного нормального распределения $\Phi(\cdot)$:

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}};$$

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{z^2}{2}} dz.$$

Полученная для нахождения параметров *tobit*-модели система уравнений существенно отличается от системы уравнений, получен-

ной с применением ММП для классической нормальной линейной модели. Она содержит нелинейные выражения, что не позволяет решить ее путем простых алгебраических преобразований. Процедура решения реализована в специальных статистических программных пакетах, например в *EViews*.

3.4. Показатели силы связи в модели множественной регрессии

В классической множественной линейной регрессии абсолютным показателем силы связи результата с фактором x_j является коэффициент b_j при этом факторе. Действительно, для линейной модели (3.10)

$$\frac{dy}{dx_j} = b_j. \quad (3.35)$$

Коэффициенты b_j линейной модели множественной регрессии называют коэффициентами условно-чистой регрессии. Они показывают, на сколько единиц в среднем изменится результат при изменении фактора x_j на 1 при фиксированном уровне других факторов, включенных в модель.

Вернемся к примеру, в котором рассматривалась зависимость объема продукции от количества занятых, стоимости основных фондов и средней заработной платы (п. 3.1). Как уже отмечалось, в данном случае можно построить две модели регрессии. Предположим, что для каждой из моделей были выбраны линейные функции. После применения МНК для первой модели имеем

$$\hat{y} = 225,68 + 29,21x_1 + 0,21x_3.$$

Коэффициенты при факторах интерпретируются следующим образом. Коэффициент при x_1 равен 29,21, следовательно, при изменении количества занятых на предприятии на одного человека объем произведенной продукции в среднем изменится в ту же сторону на 29,21 тыс. руб. при фиксированном уровне средней заработной платы. Коэффициент при x_3 равен 0,21, следовательно, при изменении средней заработной платы на 1 руб. объем произведенной продукции изменится в ту же сторону в среднем на 0,21 тыс. руб. при фиксированном количестве занятых.

Вторая модель после нахождения параметров будет иметь вид

$$\hat{y} = 474,09 + 30,08x_1 + 2,16x_2.$$

То есть при изменении количества занятых на одного человека объем произведенной продукции в среднем изменится в ту же сторону на 30,08 тыс. руб. при фиксированной стоимости основных фондов. При изменении стоимости основных фондов на 1 тыс. руб. объем произведенной продукции в среднем изменится в ту же сторону на 2,16 тыс. руб. при фиксированной численности занятых.

Сравнивая две полученные модели, можно отметить, что коэффициенты при одном и том же факторе количество занятых (x_1) различны. Это связано с тем, что на их величину оказывают влияние другие фак-

торы, не учтенные в модели, и перечень неучтенных факторов для разных моделей разный.

Исследуемые связи могут носить как линейный, так и нелинейный характер. В частности, в нашем примере зависимость объема производства от количества занятых и стоимости основных фондов может рассматриваться в форме степенной функции, которая для данной спецификации модели является разновидностью *производственной функции*:

$$P = a \cdot L^{b_1} \cdot K^{b_2} \cdot e, \quad (3.36)$$

где P — объем продукции;
 L — затраты труда;
 K — величина капитала;
 a, b_1, b_2 — неизвестные параметры;
 e — случайный остаток.

Если принять за затраты труда количество занятых (x_1), а за величину капитала — стоимость основных фондов (x_2), то получим

$$\hat{y} = a \cdot x_1^{b_1} \cdot x_2^{b_2}. \quad (3.37)$$

Для нахождения неизвестных параметров производственной функции с помощью метода наименьших квадратов данную функцию необходимо предварительно линеаризовать, произведя логарифмирование по одному и тому же основанию правой и левой части равенства:

$$\ln y = \ln a + b_1 \ln x_1 + b_2 \ln x_2. \quad (3.38)$$

Применив к данной функции МНК, были получены следующие оценки параметров модели:

$$\ln y = 3,62 + 0,5 \ln x_1 + 0,4 \ln x_2. \quad (3.39)$$

Произведя потенцирование, получим функцию регрессии в исходной форме:

$$\hat{y} = e^{3,62} \cdot x_1^{0,5} \cdot x_2^{0,4},$$

$$\hat{y} = 37,46 \cdot x_1^{0,5} \cdot x_2^{0,4}$$

Величины абсолютных показателей силы связи определяются единицами измерения факторов и поэтому не сравнимы между собой. Для сопоставления факторов по силе влияния используют относительные показатели силы связи — *коэффициенты эластичности*.

Общая формула коэффициента эластичности по фактору x_j имеет вид

$$\Theta_{x_j} = \frac{dy}{dx_j} \cdot \frac{x_j}{\hat{y}}, \quad (3.40)$$

где dy/dx_j — частная производная функции регрессии по фактору x_j ;
 \hat{y} — выровненное значение результата при заданном значении фактора x_j .

Например, для множественной линейной регрессии коэффициент эластичности по фактору x_j будет равен

$$\mathcal{E}_{x_j} = b_j \cdot \frac{x_j}{a + b_1 x_1 + b_2 x_2 + \dots + b_j x_j + \dots + b_p x_p}. \quad (3.41)$$

Из формулы видно, что коэффициент эластичности для линейной функции зависит от конкретных значений факторов, включенных в модель. Так как коэффициент эластичности измеряет влияние фактора x_j на результат, то значения остальных факторов принято фиксировать на их среднем уровне:

$$\mathcal{E}_{x_j} = b_j \cdot \frac{x_j}{a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_j \bar{x}_j + \dots + b_p \bar{x}_p}. \quad (3.42)$$

Таким образом, при фиксированных значениях других факторов существует целый ряд коэффициентов эластичности по фактору x_j , определяемый областью значений этого фактора. Они называются частными коэффициентами эластичности. Если зафиксировать значение x_j на среднем уровне, получим средний коэффициент эластичности (или коэффициент эластичности для среднего значения x_j):

$$\bar{\mathcal{E}}_{x_j} = b_j \cdot \frac{\bar{x}_j}{a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_j \bar{x}_j + \dots + b_p \bar{x}_p}. \quad (3.43)$$

Как следует из МНК для линейной регрессии, выражение в знаменателе равно среднему значению результата, что позволяет упростить формулу:

$$\bar{\mathcal{E}}_{x_j} = b_j \cdot \frac{\bar{x}_j}{\bar{y}}. \quad (3.44)$$

Коэффициенты эластичности показывают, на сколько процентов в среднем изменится результат при изменении фактора x_j на 1% и значениях других факторов, фиксированных на средних уровнях. Так как величина коэффициента эластичности зависит от выбранного значения фактора, при формулировке вывода следует указывать заданное значение фактора x_j и соответствующее ему выровненное значение результата.

Рассмотрим средние коэффициенты эластичности для рассчитанных нами моделей регрессии. Для линейной зависимости объема производства от количества занятых и средней заработной платы имеем

$$\bar{y} = 3296; \bar{x}_1 = 50,7; \bar{x}_3 = 7461,3.$$

Тогда

$$\mathcal{E}_1 = b_1 \frac{\bar{x}_1}{\bar{y}} = 29,21 \frac{50,7}{3296} = 0,45;$$

$$\mathcal{E}_3 = b_3 \frac{\bar{x}_3}{\bar{y}} = 0,21 \frac{7461,3}{3296} = 0,48.$$

То есть при изменении количества занятых на 1% от среднего уровня объем производства в среднем изменится в ту же сторону на 0,45% от своего среднего уровня при фиксированной средней заработной

плате. Коэффициент эластичности при факторе x_3 показывает, что при изменении средней заработной платы на 1% от своего среднего уровня объем производства в среднем изменится в ту же сторону на 0,48% от своего среднего уровня при фиксированном количестве занятых.

Коэффициенты условно-чистой регрессии в данной модели являются абсолютными показателями силы связи и поэтому несопоставимы. Сопоставимые показатели — коэффициенты эластичности — практически равны, что указывает на примерно одинаковую силу влияния факторов на результат.

Для второй линейной модели (зависимости объема производства от количества занятых и стоимости основных фондов) средние значения равны

$$\bar{y} = 3296; \bar{x}_1 = 50,7; \bar{x}_2 = 601.$$

Следовательно, коэффициенты эластичности составят

$$\mathcal{E}_1 = 30,08 \frac{50,7}{3296} = 0,46,$$

$$\mathcal{E}_2 = 2,16 \frac{601}{3296} = 0,39.$$

Сравнивая эти коэффициенты эластичности между собой, можно сделать вывод о том, что количество занятых более сильно влияет на объем производства, чем стоимость основных производственных фондов.

Третья из рассмотренных моделей регрессии по своей математической форме была степенной функцией. Как в парной регрессии, так и в множественной коэффициенты b_j степенной функции являются ее коэффициентами эластичности

$$\mathcal{E}_j = \frac{dy}{dx_j} \cdot \frac{x_j}{\hat{y}} = b_j a x_1^{b_1} x_2^{b_2} \cdots x_j^{b_j-1} \cdots x_k^{b_k} \cdot \frac{x_j}{a x_1^{b_1} x_2^{b_2} \cdots x_j^{b_j} \cdots x_k^{b_k}} = b_j. \quad (3.45)$$

Тогда для нашей модели степенной регрессии коэффициенты эластичности составят:

$$\mathcal{E}_1 = 0,5; \mathcal{E}_2 = 0,4.$$

Эти показатели являются константами степенной функции и, следовательно, не зависят от конкретного значения фактора. Поэтому в выводе не оказывается, от какого уровня изменились результат и фактор. В частности, для нашего примера:

- при изменении количества занятых на 1% объем производства в среднем изменится в ту же сторону на 0,5% при фиксированной стоимости основных фондов;
- при изменении стоимости основных фондов на 1% объем производства в среднем изменится в ту же сторону на 0,4% при фиксированном количестве занятых.

Анализируя полученные по линейной и степенной функции коэффициенты эластичности, можно отметить, что количество занятых

оказывает на объем производства более сильное влияние, чем стоимость основных фондов. Выводы о степени влияния факторов на результат, которые можно сделать по степенной и линейной функциям, одинаковы: количество занятых оказывает более сильное влияние, чем стоимость основных фондов.

В множественной линейной регрессии относительными показателями силы связи являются также стандартизованные коэффициенты регрессии. Как и коэффициенты эластичности, они сопоставимы между собой по силе влияния факторов на результат. Стандартизованные коэффициенты регрессии и коэффициенты условно-чистой регрессии характеризуют одну и ту же модель и в силу этого взаимосвязаны между собой. Изучим эту связь.

Коэффициент условно-чистой регрессии показывает, на сколько в среднем изменится результат при изменении соответствующего фактора на единицу и фиксированном уровне других факторов. То есть если Δ_{x_j} — абсолютное изменение переменной x_j , $\Delta_{x_j} = 1$ и $\Delta_{x_k} = 0$ (для $k \neq j$), то $\Delta_y = b_j$.

Стандартизованная переменная-фактор в этом случае изменится следующим образом:

$$\Delta_{t_{x_j}} = \frac{x_j + 1 - \bar{x}_j}{\sigma_{x_j}} - \frac{x_j - \bar{x}_j}{\sigma_{x_j}} = \frac{1}{\sigma_{x_j}}.$$

Как следует из стандартизованной формы уравнения регрессии, изменение стандартизированной переменной-результата составит:

$$\Delta_{t_y} = \beta_j \frac{1}{\sigma_{x_j}}.$$

С другой стороны, зная величину изменения значения результата y , можно найти величину изменения стандартизированного результата

$$\Delta_{t_y} = \frac{y + b_j - \bar{y}}{\sigma_y} - \frac{y - \bar{y}}{\sigma_y} = \frac{b_j}{\sigma_y}.$$

Отсюда

$$\frac{b_j}{\sigma_y} = \frac{\beta_j}{\sigma_{x_j}} \Rightarrow \beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y} \quad (3.46)$$

Стандартизованные коэффициенты регрессии (β_j) показывают, на сколько своих среднеквадратических отклонений (σ_y) в среднем изменится результат при изменении фактора x_j на одно свое среднеквадратическое отклонение (σ_{x_j}) при фиксированном уровне других факторов, включенных в модель регрессии.

Покажем справедливость такого вывода. Пусть переменная t_{x_j} в стандартизованном уравнении регрессии изменится на 1. Тогда переменная x_j изменится с некоторого значения $x_j^{(1)}$ на значение $x_j^{(2)}$

Имеем:

$$\Delta_{t_{x_j}} = (t_{x_j} + 1) - t_{x_j} = \frac{x_j^{(2)} - \bar{x}_j}{\sigma_{x_j}} - \frac{x_j^{(1)} - \bar{x}_j}{\sigma_{x_j}};$$

$$1 = \frac{x_j^{(2)} - \bar{x}_j - x_j^{(1)} + \bar{x}_j}{\sigma_{x_j}} = \frac{x_j^{(2)} - x_j^{(1)}}{\sigma_{x_j}}$$

откуда

$$x_j^{(2)} - x_j^{(1)} = \sigma_{x_j},$$

т. е. изменение фактора x_j составит σ_{x_j} .

Изменение результата:

$$\Delta_y = b_j \sigma_{x_j} = \beta_j \frac{\sigma_y}{\sigma_{x_j}} \sigma_{x_j} = \beta_j \sigma_y,$$

что и требовалось доказать.

Найдем стандартизованные коэффициенты регрессии для нашего примера. Для этого предварительно найдем средние квадратические отклонения результата и каждого из факторов:

$$\sigma_y = 385,63, \quad \sigma_{x_1} = 6,37, \quad \sigma_{x_2} = 108,82, \quad \sigma_{x_3} = 1025,97.$$

Для линейной модели регрессии $\hat{y} = f(x_1, x_3)$ стандартизованные коэффициенты регрессии будут равны:

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 29,21 \frac{6,37}{385,63} = 0,48; \quad \beta_3 = b_3 \frac{\sigma_{x_3}}{\sigma_y} = 0,21 \frac{1025,97}{385,63} = 0,56.$$

Следовательно, при изменении количества занятых на $1\sigma_{x_1}$ объем производства в среднем изменится в ту же сторону на $0,48\sigma_y$ при фиксированном уровне средней заработной платы. С другой стороны, изменение средней заработной платы на $1\sigma_{x_3}$ приводит к изменению объема производства в ту же сторону в среднем на $0,56$ при фиксированном количестве занятых. Фактор «средняя заработная плата» оказывает большее влияние на результат, чем фактор «количество занятых», что согласуется с выводами по коэффициентам эластичности.

Стандартизованные коэффициенты регрессии для линейной модели $\hat{y} = f(x_1, x_2)$ равны

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 30,08 \frac{6,37}{385,63} = 0,50; \quad \beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 2,16 \frac{108,82}{385,63} = 0,61.$$

Выводы по данным коэффициентам формулируются аналогично выводам по предыдущей модели. Сравнивая данные коэффициенты между собой, можно отметить, что второй коэффициент выше, т. е. фактор «стоимость основных фондов» оказывает более сильное влияние на объем производства, чем фактор «количество занятых». Это противоречит выводу о соотношении влияния этих факторов на результат, полученному по коэффициентам эластичности. Причина та-

кого расхождения — в разном уровне вариации фактора и результата. Коэффициенты эластичности и стандартизованные коэффициенты регрессии соотносятся между собой как коэффициенты вариации фактора и результата:

$$\frac{\beta_j}{\vartheta_j} = b_j \frac{\sigma_{x_j}}{\sigma_y} : b_j \frac{\bar{x}_j}{\bar{y}} = \frac{\sigma_{x_j}}{\bar{x}_j} : \frac{\sigma_y}{\bar{y}} = v_{x_j} : v_y. \quad (3.47)$$

Следовательно, ранжирование факторов по коэффициентам эластичности и стандартизованным коэффициентам регрессии не изменится, если коэффициенты вариации факторов будут примерно равны. В нашем случае $v_{x_1} = 12,57\%$, а $v_{x_2} = 18,11\%$, что привело к изменению ранжирования факторов по силе их влияния на результат.

3.5. Изучение тесноты связи на основе множественной регрессии

В модели множественной регрессии, как и в парной регрессии, измерителем тесноты связи являются показатели, в основе которых лежит величина суммы квадратов остатков:

$$\sum(y - \hat{y})^2 = \sum e^2. \quad (3.48)$$

В парной линейной регрессии такими показателями были коэффициенты корреляции и детерминации (r и r^2), в парной нелинейной регрессии — индексы корреляции и детерминации (R и R^2). Термин «коэффициент» относится к линейной регрессии, «индекс» — к нелинейной.

В множественной регрессии показателями тесноты связи являются коэффициенты (индексы) множественной корреляции и детерминации. Они обозначаются R (коэффициент (индекс) множественной корреляции) и R^2 (коэффициент (индекс) множественной детерминации) как для линейной, так и для нелинейной функции.

Формулы для их расчета совпадают с формулами для индексов корреляции и детерминации в парной регрессии. В частности, коэффициент (индекс) множественной детерминации равен

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{SS_{\text{ост}}}{SS_{\text{общ}}}, \quad (3.49)$$

или с учетом правила сложения дисперсий:

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{SS_{\text{факт}}}{SS_{\text{общ}}}, \quad (3.50)$$

где $SS_{\text{ост}}$ — остаточная сумма квадратов: $SS_{\text{ост}} = \sum(y - \hat{y})^2$;

$SS_{\text{факт}}$ — факторная сумма квадратов: $SS_{\text{факт}} = \sum(\hat{y} - \bar{y})^2$;

$SS_{\text{общ}}$ — общая сумма квадратов: $SS_{\text{общ}} = \sum(y - \bar{y})^2$.

Коэффициент (индекс) множественной корреляции есть корень из коэффициента (индекса) множественной детерминации:

$$R = \sqrt{R^2}. \quad (3.51)$$

Показатели множественной корреляции изменяются в диапазоне [0; 1]. Чем ближе они к единице, тем связь теснее.

Найдем показатели тесноты связи по первой модели нашего примера ($\hat{y} = 225,68 + 29,21x_1 + 0,21x_3$). Остаточная и общая суммы квадратов равны

$$\begin{aligned} SS_{\text{ост}} &= \sum(y - \hat{y})^2 = 1126725; \\ SS_{\text{общ}} &= \sum(y - \bar{y})^2 = 4461255. \end{aligned}$$

Отсюда коэффициент множественной детерминации равен

$$R^2 = 1 - \frac{1126725}{4461255} \approx 0,747.$$

Можно сказать, что вариация количества занятых и средней зарплатной платы на 74,7% объясняет вариацию объема выпуска.

Коэффициент множественной корреляции равен

$$R = \sqrt{R^2} = \sqrt{0,747} \approx 0,865.$$

Полученная величина близка к единице, следовательно, связь между результатом и факторами тесная.

Показатели тесноты связи для второй модели примера $\hat{y} = 225,68 + 29,21x_1 + 0,21x_3$ равны

$$R^2 = 0,803; R = 0,896.$$

Очевидно, что связь между числом занятых, стоимостью основных фондов и объемом выпуска продукции тесная.

Рассмотренные функции были *линейными*. При изучении связи по *нелинейным* функциям мы рассматриваем фактически не одно, а два уравнения регрессии: нелинейное и линеаризованное. Поэтому возникает вопрос: совпадают ли показатели тесноты связи, найденные по этим уравнениям, и если не совпадают, то каким отдать предпочтение?

В формулах для расчета показателей тесноты связи (3.49), (3.50) используются только значения результата: исходное (y), среднее (\bar{y}), выровненное (\hat{y}). Если нелинейная функция *линейна по параметрам*, например:

$$y = a + b_{11}x_1 + b_{12}x_1^2 + b_{21} \frac{1}{x_2} + e, \quad (3.52)$$

то ее приведение к линейному виду будет связано только с преобразованием переменных-факторов (x_1, x_2):

$$x_1^2 \equiv x_3 \quad \frac{1}{x_2} \equiv x_4,$$

тогда

$$y = a + b_{11}x_1 + b_{12}x_3 + b_{21}x_4 + e. \quad (3.53)$$

Переменная-результат не была преобразована, поэтому показатели тесноты связи совпадают для нелинейной (3.52) и линеаризованной (3.53) форм уравнения.

Если для нахождения параметров регрессии применяется линеаризация, предполагающая преобразование результативного признака, то отношения дисперсий для линеаризованной и исходной функции не равны друг другу:

$$\frac{D_{\text{ост}}(y=f(x_1 \dots x_p))}{D_{\text{общ}}(y=f(x_1 \dots x_p))} \neq \frac{D_{\text{ост}}(Y=f_1(x_1 \dots x_p))}{D_{\text{общ}}(Y=f_1(x_1 \dots x_p))} \text{ и}$$

$$\frac{D_{\text{факт}}(y=f(x_1 \dots x_p))}{D_{\text{общ}}(y=f(x_1 \dots x_p))} \neq \frac{D_{\text{факт}}(Y=f_1(x_1 \dots x_p))}{D_{\text{общ}}(Y=f_1(x_1 \dots x_p))},$$

где $D_{\text{ост}}(\cdot)$, $D_{\text{факт}}(\cdot)$, $D_{\text{общ}}(\cdot)$ — остаточная, факторная и общая дисперсии результативного признака, найденные для уравнения, указанного в скобках;

$y = f(x_1 \dots x_p)$ — исходное нелинейное уравнение регрессии;

$Y = f_1(x_1 \dots x_p)$ — линеаризованное уравнение регрессии с преобразованным результативным признаком.

Поэтому показатели тесноты связи для исходной и линеаризованной формы не равны:

$$R^2_{y=f(x_1 \dots x_p)} \neq R^2_{Y=f_1(x_1 \dots x_p)}, \quad (3.54)$$

где $R^2_{y=f(x_1 \dots x_p)}$ — индекс множественной детерминации для исходной функции;

$R^2_{Y=f_1(x_1 \dots x_p)}$ — коэффициент множественной детерминации для линеаризованной функции с преобразованным результатом (Y).

Примерами таких функций являются степенная и показательная, которые линеаризуются путем логарифмирования правой и левой части уравнения регрессии, а также обратная функция.

Для исходной нелинейной функции регрессии не выполняется также правило сложения дисперсий:

$$\frac{\sum(y - \bar{y})^2}{n} \neq \frac{\sum(\hat{y} - \bar{y})^2}{n} + \frac{\sum(y - \hat{y})^2}{n} \quad (3.55)$$

Поэтому индексы множественной детерминации, рассчитанные по формулам (3.49) и (3.50), имеют разные значения.

Из-за несоблюдения правила сложения дисперсий индексы детерминации и корреляции, рассчитанные по исходной нелинейной функции, не могут быть однозначно определены. Их нельзя проинтерпретировать как долю объясненной дисперсии в общей. Исходя из этого для оценки тесноты связи применяют коэффициенты детерминации

и корреляции, рассчитанные по линеаризованной функции ($R^2_{Y=f_1(x_1 \dots x_p)}$).

Для линеаризованной функции правило сложения дисперсий соблюдается — общая дисперсия равна сумме факторной и остаточной:

$$D_{\text{общ}} = D_{\text{факт}} + D_{\text{ост}}, \quad (3.56)$$

$$\frac{\sum(Y - \bar{Y})^2}{n} = \frac{\sum(\hat{Y} - \bar{Y})^2}{n} + \frac{\sum(Y - \hat{Y})^2}{n}, \quad (3.57)$$

где Y — преобразованная переменная-результат.

Следовательно, для расчета коэффициента множественной детерминации по линеаризованной функции можно использовать как формулу (3.49), так и формулу (3.50):

$$R^2_{Y=f_1(x_1 \dots x_p)} = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}; \quad (3.58)$$

$$R^2_{Y=f_1(x_1 \dots x_p)} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}. \quad (3.59)$$

Расчеты по этим формулам дадут одинаковый результат.

Индексы множественной детерминации по исходной нелинейной модели используют в ряде случаев как характеристику степени приближения модели к реальным данным, т. е. величины остаточной дисперсии по отношению к общей. В силу невыполнения правила сложения дисперсий (3.56) для этого типа моделей применяется только формула (3.49). Такие индексы иногда называют *квази-коэффициенты детерминации и корреляции* (*квази-R²* и *квази-R*).

Вернемся к третьей модели нашего примера ($\hat{y} = 37,46 \cdot x_1^{0,5} \cdot x_2^{0,4}$).

Она является нелинейной функцией, приводимой к линейному виду логарифмированием правой и левой части. После применения метода наименьших квадратов уравнение в логарифмах имело вид (3.39):

$$\ln \hat{y} = \ln 37,46 + 0,5 \ln x_1 + 0,4 \ln x_2.$$

Найдем показатели тесноты связи по линеаризованной модели. Остаточная, факторная и общая суммы квадратов по ней равны соответственно:

$$SS_{\text{ост}} = \sum(\ln y - \ln \hat{y})^2 = 0,09442;$$

$$SS_{\text{факт}} = \sum(\ln \hat{y} - \ln \bar{y})^2 = 0,380639;$$

$$SS_{\text{общ}} = \sum(\ln y - \ln \bar{y})^2 = 0,475059.$$

Коэффициент множественной детерминации по формуле (3.49) составит

$$R^2 = 1 - \frac{0,09442}{0,475059} \approx 0,801.$$

То же значение получим, применив формулу (3.50):

$$R^2 = \frac{0,380639}{0,475059} \approx 0,801.$$

Следовательно, по степенной функции вариация результата (объема выпуска) на 80,1% объясняется вариацией факторов (числом занятых и стоимостью основных фондов).

Коэффициент множественной корреляции равен

$$R = \sqrt{R^2} = \sqrt{0,801} \approx 0,895,$$

что свидетельствует о высокой тесноте связи.

Найдем теперь показатели тесноты связи для исходной нелинейной модели. Суммы квадратов для нее равны

$$SS_{\text{ост}} = \sum(y - \hat{y})^2 = 938556,9;$$

$$SS_{\text{факт}} = \sum(\hat{y} - \bar{y})^2 = 3910191,0;$$

$$SS_{\text{общ}} = \sum(y - \bar{y})^2 = 4461255,0.$$

Отметим, что в данном случае не выполняется правило сложения дисперсий и соответственно сложения их числителей:

$$SS_{\text{общ}} < (SS_{\text{факт}} + SS_{\text{ост}});$$

$$4461255,0 < (3910191,0 + 938556,9).$$

Применив для расчета индекса множественной детерминации формулу (3.49), получим

$$R^2 = 1 - \frac{938556,9}{4461255,0} \approx 0,7896.$$

Если применить формулу (3.50), то ответ будет другим:

$$R^2 = \frac{3910191,0}{4461255,0} \approx 0,8765.$$

Как отмечалось ранее, величины R^2 различаются, и довольно значительно. Они не совпадают и со значением коэффициента множественной детерминации, полученным по линеаризованной функции.

Сравним показатели тесноты связи рассмотренных уравнений регрессии. Каждый из них близок к единице, следовательно, все функции достаточно хорошо описывают связь результата и факторов. В зависимости от целей исследования каждая функция может быть применена для дальнейшего анализа и прогноза.

Формально максимальный показатель тесноты связи имеет линейная модель с факторами x_1, x_2 ($R^2=0,803$). Он, однако, мало отличается от показателя тесноты связи по нелинейной функции ($R^2=0,801$). Если нет каких-либо теоретических ограничений, то предпочтение следует отдать линейной функции. Если изучается конкретная теоретическая модель — производственная функция, то можно утверждать, что эта модель подтверждается эмпирическими данными и может быть использована для дальнейшего анализа.

Коэффициент множественной детерминации (т. е. показатель тесноты связи для линейной множественной регрессии) можно рассчитать, используя определители матриц парных коэффициентов корреляции: полной, включающей коэффициенты корреляции факторов с результатом и между собой (r_{yx}), и матрицы коэффициентов корреляции между факторами (r_{xx}):

$$r_{yx} = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_p} \\ r_{x_1y} & 1 & r_{x_1x_p} \\ \dots & & \\ r_{x_py} & r_{x_px_1} & 1 \end{pmatrix}, \quad (3.60)$$

$$r_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1x_p} \\ r_{x_2x_1} & 1 & r_{x_2x_p} \\ r_{x_px_1} & r_{x_px_2} & 1 \end{pmatrix}. \quad (3.61)$$

Можно показать, что коэффициент множественной детерминации равен

$$R^2_{yx_1\dots x_p} = 1 - \frac{\det r_{yx}}{\det r_{xx}}, \quad (3.62)$$

где $\det r_{yx}$ и $\det r_{xx}$ — определители матриц r_{yx} и r_{xx} . Для множественной линейной регрессии с двумя факторами эта формула примет вид:

$$R^2_{yx_1x_2} = \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}. \quad (3.63)$$

В частности, для наилучшей линейной модели нашего примера $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ имеем:

$$R^2_{yx_1x_2} = \frac{0,682^2 + 0,761^2 - 2 \cdot 0,682 \cdot 0,761 \cdot 0,305}{1 - 0,305^2} \approx 0,802^*.$$

Используя стандартизованную форму линейного уравнения множественной регрессии, можно получить еще одну формулу для расчета коэффициента множественной детерминации:

$$R^2_{yx_1x_2\dots x_p} = \sum \beta_i r_{yx_i}. \quad (3.64)$$

Для модели $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ имеем

$$R^2_{yx_1x_2} = 0,682 \cdot 0,50 + 0,761 \cdot 0,61 \approx 0,805.$$

*Значения парных коэффициентов корреляции были приведены в п. 3.1.

Результаты расчетов по формулам (3.63) и (3.64) незначительно отличаются от полученного ранее (0,803) из-за ошибок округления.

При измерении тесноты связи в множественной регрессии возникают проблемы корректности ее оценки и сопоставимости показателей тесноты связи для различных функций регрессии. Рассмотрим эти проблемы более подробно.

Как уже отмечалось выше, достоверность результатов регрессионного анализа будет тем выше, чем большее число наблюдений было использовано для его проведения. Более того, если наблюдений слишком мало, может сложиться ситуация, когда параметры регрессии вообще невозможно найти. Например, если необходимо найти параметры уравнения парной линейной регрессии, одного наблюдения недостаточно: через одну точку может пройти бесконечное множество прямых.

Для двух наблюдений можно найти единственную прямую, причем она пройдет точно через эти точки, т. е. случайных отклонений не будет и коэффициент корреляции будет равен вне зависимости от того, есть ли связь между данными показателями или нет.

Соотношение количества наблюдений и количества оцениваемых параметров уравнения регрессии учитывается в так называемом *скорректированном (исправленном) коэффициенте детерминации*. В отличие от «обычного» коэффициента детерминации для его расчета используются дисперсии, рассчитанные на одну степень свободы

$$R_{\text{скорр}}^2 = 1 - \frac{MS_{\text{ост}}}{MS_{\text{общ}}}, \quad (3.65)$$

где MS — дисперсия на одну степень свободы ($MS_{\text{ост}}$ — остаточная, $MS_{\text{общ}}$ — общая). MS вычисляется по формуле

$$MS = \frac{SS}{df}, \quad (3.66)$$

где SS — соответствующая сумма квадратов;

df — число степеней свободы данной дисперсии.

Можно доказать, что остаточная дисперсия имеет $(n-m-1)$ степеней свободы, общая — $(n-1)$ степеней свободы, где m — число параметров уравнения регрессии без учета свободного члена. Тогда остаточная и общая дисперсии на одну степень свободы равны соответственно

$$MS_{\text{ост}} = \frac{SS_{\text{ост}}}{n-m-1} = \frac{\sum(y - \hat{y})^2}{n-m-1}, \quad (3.67)$$

$$MS_{\text{общ}} = \frac{SS_{\text{общ}}}{n-1} = \frac{\sum(y - \bar{y})^2}{n-1}. \quad (3.68)$$

Следовательно, скорректированный коэффициент детерминации равен

$$R_{\text{скорр}}^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \cdot \frac{n-1}{n-m-1}. \quad (3.69)$$

Эту формулу можно преобразовать:

$$R_{\text{скорр}}^2 = 1 - (1 - (1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2})) \cdot \frac{n-1}{n-m-1} = 1 - (1 - R^2) \cdot \frac{n-1}{n-m-1}. \quad (3.70)$$

Полученное выражение связывает два показателя тесноты связи — скорректированный и нескорректированный коэффициенты детерминации.

Скорректированный коэффициент детерминации применяется для решения двух задач: оценки реальной тесноты связи между результатом и факторами и сравнения моделей с разным числом параметров. В первом случае обращают внимание на близость скорректированного и нескорректированного коэффициентов детерминации. Если эти показатели велики и различаются незначительно, модель считается хорошей.

При сравнении разных моделей предпочтение при прочих равных условиях отдается той, у которой больше скорректированный коэффициент детерминации.

Следует отметить, что область применения скорректированного коэффициента детерминации ограничивается только этими задачами. Его нельзя использовать в формулах, где применяется обычный коэффициент детерминации*. Скорректированный коэффициент детерминации нельзя интерпретировать как долю вариации результата, объясненную вариацией факторов, включенных в модель регрессии.

Найдем скорректированные показатели детерминации для рассмотренных моделей регрессии (табл. 3.2). Отметим, что все модели имеют равное число параметров и поэтому коэффициенты детерминации сопоставимы между собой.

Таблица 3.2
Скорректированные коэффициенты детерминации

Модель	R^2	$R_{\text{скорр}}^2$
$\hat{y} = 225,68 + 29,21x_1 + 0,21x_3$	0,747	$1 - (1 - 0,747) \cdot \frac{30-1}{30-2-1} \approx 0,728$
$\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$	0,803	$1 - (1 - 0,803) \cdot \frac{30-1}{30-2-1} \approx 0,788$
$\hat{y} = 37,46 \cdot x_1^{0,5} \cdot x_2^{0,4}$	0,801**	$1 - (1 - 0,801) \cdot \frac{30-1}{30-2-1} \approx 0,786$

* Например, для оценки значимости уравнения регрессии по F-критерию.

** Для линеаризованной формы модели регрессии.

Скорректированные коэффициенты детерминации близки к не-скорректированным (исходным), что свидетельствует о хорошем качестве рассматриваемых моделей.

При отборе факторов в модель регрессии путем включения в нее дополнительных факторов возникает необходимость оценить целесообразность включения каждого из них. Для этого могут применяться *коэффициенты частной корреляции и детерминации*. Изучим их на примере линейной модели.

Предположим, что мы построили модель с наиболее значимым фактором x_1 :

$$y = a_1 + b_{11}x_1 + e_1.$$

Характеристикой близости наблюдаемых значений результата и полученных по данной модели является остаточная сумма квадратов:

$$SS_{\text{ост}} = \sum (y - \hat{y})^2.$$

Чем эта сумма меньше, тем лучше модель. Обозначим остаточную сумму квадратов по данной модели как $SS_{\text{ост}}^{(1)}$.

Далее строим новую модель уже с двумя факторами — x_1 и x_2 :

$$y = a_2 + b_{21}x_1 + b_{22}x_2 + e_2.$$

Включая в модель новый фактор, мы предполагаем, что он внесет существенный вклад в объяснение результата, т. е. остаточная дисперсия (и остаточная сумма квадратов) должна уменьшиться. Чем значительно это уменьшение, тем статистически обоснованнее включение фактора x_2 . Обозначим остаточную сумму квадратов по новой модели как $SS_{\text{ост}}^{(2)}$. Ее сокращение составит

$$\Delta SS_{\text{ост}} = SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)}. \quad (3.71)$$

Эта величина зависит от единиц измерения и размаха вариации исследуемого результативного признака, что затрудняет оценку величины уменьшения остаточной суммы квадратов. Однако если разделить ее на исходную остаточную сумму квадратов, получим относительный показатель:

$$\frac{SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)}}{SS_{\text{ост}}^{(1)}} \equiv r_{yx_2, x_1}^2 \quad (3.72)$$

Полученная величина называется *коэффициентом частной детерминации*. Он изменяется от 0 до 1. В обозначении в нижнем индексе указываются результат (y), фактор, дополнительно включаемый в модель (x_2), далее ставится точка и перечисляются все остальные факторы, которые были в модели и до, и после включения в нее нового фактора. В нашем случае это единственный фактор x_1 .

Чем больше сокращение остаточной дисперсии при включении в модель дополнительного фактора, тем ближе его величина к единице. Наоборот, если дополнительный фактор практически не изменяет величину остаточной дисперсии, значение коэффициента частной детерминации будет близко к нулю. Поэтому коэффициент частной де-

терминации интерпретируют как долю сокращения остаточной дисперсии по отношению к исходной остаточной дисперсии.

Рассмотрим коэффициенты частной детерминации на нашем примере. Наиболее значимым фактором был фактор x_2 (стоимость основных фондов, $r_{yx_2} = 0,761$). Уравнение парной регрессии с этим фактором будет иметь вид

$$\hat{y} = 1675,5 + 2,7x_2$$

Остаточная сумма квадратов равна

$$SS_{\text{ост}}^{(1)} = 1878244,0.$$

После включения второго фактора x_1 (число занятых) уравнение регрессии имеет вид

$$\hat{y} = 474,09 + 30,08x_1 + 2,16x_2.$$

Остаточная сумма квадратов равна

$$SS_{\text{ост}}^{(2)} = 879819,7.$$

Сокращение остаточной суммы квадратов составило:

- в абсолютных величинах:

$$SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)} = 1878244,0 - 879819,7 = 998424,4;$$

- в относительных величинах:

$$r^2_{yx_1, x_2} = \frac{SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)}}{SS_{\text{ост}}^{(1)}} = \frac{998424,4}{1878244,0} \approx 0,532.$$

То есть сокращение остаточной дисперсии при включении в модель фактора x_1 составило 53,2% от исходной величины. Можно утверждать, что этот фактор оказывает значимое влияние на результат и его следует использовать при построении уравнения регрессии.

Рассчитывая частные коэффициенты детерминации по формуле (3.72), не обязательно находить параметры первого уравнения регрессии (не включающего исследуемый фактор). Так как парные коэффициенты корреляции уже известны, достаточно найти общую сумму квадратов результативного признака:

$$SS_{\text{общ}} = \sum (y - \bar{y})^2,$$

а затем найти остаточную сумму квадратов, используя парный коэффициент корреляции:

$$SS_{\text{ост}} = SS_{\text{общ}} (1 - r_{yx_j}),$$

где x_j — фактор, включенный в модель регрессии ранее.

Ниже будут рассмотрены другие, так называемые рекуррентные формулы расчета частных показателей тесноты связи, которые чаще применяются в статистических расчетах.

Найдем коэффициенты частной детерминации каждого из рассмотренных факторов в двухфакторной модели. Результаты расчетов приведены в табл. 3.3.

Таблица 3.3

Парные и частные коэффициенты детерминации

Включаемый фактор x_j	Парные коэффициенты детерминации $r_{yx_j}^2$	Коэффициенты частной детерминации r_{yx_j, x_k}^2 для фактора x_j , включаемого после фактора x_k		
		x_1	x_2	x_3
x_1	0,465	—	0,532	0,466
x_2	0,579	0,631	—	0,149
x_3	0,543	0,527	0,078	—

По данным табл. 3.3 видно, что величина коэффициента частной детерминации зависит не только от тесноты связи фактора и результата, но и от того, какие другие факторы уже включены в модель и как они связаны с данным фактором. Коэффициенты частной детерминации, как и парные коэффициенты корреляции, используются для отбора факторов в модель регрессии. Данные, содержащиеся в таблице, подтверждают выводы, сделанные в п. 3.1 при спецификации модели регрессии. В частности, коэффициенты частной детерминации показывают, что нецелесообразно включать в модель фактор x_3 , после фактора x_2 ($r_{yx_3, x_2}^2 = 0,078$) и, наоборот, фактор x_2 после фактора x_3 ($r_{yx_2, x_3}^2 = 0,149$).

В общем случае рассматривается включение в модель фактора x_j после остальных факторов $x_1 \dots x_p$. Коэффициент частной детерминации показывает в этом случае долю сокращения остаточной дисперсии в модели со всеми факторами по сравнению с моделью без фактора x_j :

$$r_{yx_j, x_1 x_2 \dots}^2 = \frac{SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)}}{SS_{\text{ост}}^{(1)}}, \quad (3.73)$$

где y — результативный признак;
 x_j — фактор, дополнительно включаемый в модель;
 x_1, x_2, \dots, x_p — факторы, включенные в модель до фактора x_j ;
 $SS_{\text{ост}}^{(1)}$ — остаточная дисперсия для модели без фактора x_j ;
 $SS_{\text{ост}}^{(2)}$ — остаточная дисперсия для модели с фактором x_j .

Корень из коэффициента частной детерминации называется *коэффициентом частной корреляции*. Для фактора x_j при исключении влияния факторов $x_1 \dots x_p$ он обозначается как $r_{yx_j, x_1 x_2 \dots x_{j-1} x_{j+1} \dots x_p}$ и изменяется в границах от -1 до $+1$.

Преобразовав формулу для коэффициента частной детерминации:

$$r_{yx_j, x_1 x_2 \dots x_p}^2 = \frac{SS_{\text{ост}}^{(1)} - SS_{\text{ост}}^{(2)}}{SS_{\text{ост}}^{(1)}} = \frac{1 - \frac{SS_{\text{ост}}^{(1)}}{SS_{\text{общ}}} - \frac{SS_{\text{ост}}^{(2)}}{SS_{\text{общ}}}}{1 - \frac{SS_{\text{ост}}^{(1)}}{SS_{\text{общ}}}}, \quad (3.74)$$

получим

$$r_{yx_j \dots x_1 x_2 \dots x_p}^2 = \frac{R_{yx_1 x_2 \dots x_j \dots x_p}^2 - R_{yx_1 x_2 \dots x_{j-1} x_{j+1} \dots x_p}^2}{1 - R_{yx_1 x_2 \dots x_p}^2}. \quad (3.75)$$

Для фактора x_1 в двухфакторной модели формула коэффициента частной детерминации будет иметь вид

$$r_{yx_1 \dots x_2}^2 = \frac{R_{yx_1 x_2}^2 - r_{yx_2}^2}{1 - r_{yx_2}^2}. \quad (3.76)$$

Эту формулу можно преобразовать следующим образом:

$$r_{yx_1 \dots x_2}^2 = \frac{(r_{yx_1} - r_{yx_2} r_{x_1 x_2})^2}{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}. \quad (3.77)$$

Тогда коэффициент частной корреляции для фактора x_1 будет равен

$$r_{yx_1 \dots x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}}. \quad (3.78)$$

Формула коэффициента частной корреляции (3.78) называется рекуррентной, так как она получена на основе парных коэффициентов корреляции. Из этой формулы следует, что знак коэффициента частной детерминации зависит от знаков коэффициентов парной корреляции и от соотношения их абсолютных величин.

Коэффициент частной корреляции можно найти как корень из коэффициента частной детерминации, однако тогда он будет всегда положительной величиной. Рекуррентная формула позволяет определить как абсолютное значение, так и знак коэффициента частной корреляции.

Используем рекуррентную формулу для расчета коэффициента частной корреляции в нашем примере. Для фактора x_1 , включаемого в модель после фактора x_2 , коэффициент частной корреляции равен

$$r_{yx_1 \dots x_2} = \frac{0,682 - 0,761 \cdot 0,305}{\sqrt{(1 - 0,761^2) \cdot (1 - 0,305^2)}} \approx 0,728.$$

Это значение практически совпадает с величиной, полученной как корень из коэффициента частной детерминации:

$$r_{yx_1 \dots x_2} = \sqrt{0,532} \approx 0,729.$$

Формулы расчета показателей частной детерминации и корреляции для фактора x_2 в двухфакторной модели будут иметь вид

$$r_{yx_2 \dots x_1}^2 = \frac{R_{yx_1 x_2}^2 - r_{yx_1}^2}{1 - r_{yx_1}^2}; \quad (3.79)$$

$$r_{yx_2}^2 = \frac{(r_{yx_2} - r_{yx_1} r_{x_1 x_2})^2}{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}; \quad (3.80)$$

$$r_{yx_2} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}}. \quad (3.81)$$

В общем случае, когда фактор x_j включается в модель после остальных факторов $x_1 \dots x_p$, рекуррентная формула для коэффициента частной корреляции имеет вид:

$$r_{yx_j \cdot x_1 x_2 \dots x_p} = \frac{r_{yx_j \cdot x_1 x_2 \dots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \dots x_{p-1}} r_{x_j x_p \cdot x_1 x_2 \dots x_{p-1}}}{\sqrt{(1 - r_{yx_p \cdot x_1 x_2 \dots x_{p-1}}^2)(1 - r_{x_j x_p \cdot x_1 x_2 \dots x_{p-1}}^2)}}. \quad (3.82)$$

Коэффициент частной корреляции интерпретируется как мера тесноты связи фактора и результата при фиксированном влиянии других факторов, включенных в модель.

3.6. Оценка значимости модели множественной регрессии и ее параметров

В качестве исходных данных для расчета параметров уравнения регрессии используются данные статистических наблюдений над признаками-факторами и признаком-результатом. Эти данные рассматриваются как выборочные. Следовательно, любые числовые значения параметров уравнения регрессии, которые были получены, являются выборочными оценками объективно существующих, но неизвестных параметров. Как любые выборочные оценки, они содержат определенную ошибку и не совпадают со значениями параметров в генеральной совокупности.

Так как любой параметр уравнения регрессии определенным образом характеризует взаимосвязь факторов и результата, это влияние действительно имеет место, если ожидаемое значение того же параметра в генеральной совокупности с большой долей вероятности не равно нулю. В этом случае говорят, что параметр *значим*. Наоборот, если в области возможных значений генерального параметра входит нуль, говорят, что параметр *незначим*.

Для оценки значимости параметров уравнения множественной регрессии используют *критерий Стьюдента (t-критерий)*. Процедура оценки такая же, как и в парной регрессии. Формула для расчета ошибки конкретного параметра зависит от формы функции регрессии (линейная, параболическая и т. п.) и от числа факторов (один или более). Как и в парной регрессии, значимость параметра можно оценивать двумя способами: с помощью сравнения фактического и табличного значения *t*-критерия и с помощью доверительных интервалов.

Первый способ. Находят стандартную ошибку оцениваемого параметра (*se*) и рассчитывают фактическое значение *t*-критерия как соотношение значения самого параметра и его стандартной ошибки.

В линейной функции стандартную ошибку $se(b_j)$ любого ее параметра b_j * удобнее записать в матричной форме:

$$se(b_j) = \sqrt{\frac{SS_{\text{ост}}}{n-m-1}} [(X^T X)^{-1}]_{jj}, \quad (3.83)$$

где $[(X^T X)^{-1}]_{jj}$ — элемент (j, j) матрицы $(X^T X)^{-1}$.

Свободный член обозначим как b_0 . После проведения ряда преобразований формула стандартной ошибки $se(b_j)$ коэффициента условно-чистой регрессии b_j :

$$se(b_j) = \frac{\sigma_y}{\sigma_{x_j}} \sqrt{\frac{1 - R_{yx_1 \dots x_p}^2}{(1 - R_{x_j x_1 \dots x_p}^2)(n-m-1)}}, \quad (3.84)$$

- где σ_y — среднее квадратическое отклонение признака y ;
 σ_{x_j} — среднее квадратическое отклонение признака x_j ;
 $R_{yx_1 \dots x_p}^2$ — коэффициент детерминации, найденный для уравнения зависимости результата y от факторов $x_1 \dots x_p$, включая фактор x_j ;
 $R_{x_j x_1 \dots x_p}^2$ — коэффициент детерминации, найденный для уравнения зависимости переменной x_j от других факторов $x_1 \dots x_p$, входящих в рассматриваемую модель множественной регрессии.

Фактическое значение t -критерия для параметра b_j рассчитывается с помощью следующего выражения:

$$t_{b_j} = \frac{b_j}{se(b_j)}. \quad (3.85)$$

Полученную величину необходимо сравнить с табличным значением, которое находят для уровня значимости α и $(n - m - 1)$ степеней свободы**. Уровень значимости, как правило, задают равным 0,05; же 0,01 или 0,1.

Параметр считается значимым, если фактическое значение t -критерия по модулю больше его табличного значения.

Оценим значимость коэффициентов линейной множественной регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$.

Все данные для расчета стандартных ошибок нам уже известны:

$$R_{yx_1 x_2}^2 = 0,803; \quad r_{x_1 x_2}^2 = r_{x_2 x_1}^2 = 0,305^2 = 0,093; \quad n = 30; \quad m = 2; \\ \sigma_y = 385,63; \quad \sigma_{x_1} = 6,37; \quad \sigma_{x_2} = 108,82.$$

Отсюда стандартные ошибки коэффициентов регрессии b_1 и b_2 равны

* Коэффициент при факторе x_j в множественной линейной регрессии.

** Где n — число наблюдений, m — число параметров в уравнении регрессии без учета свободного члена.

$$se(b_1) = \frac{385,63}{6,37} \sqrt{\frac{1-0,803}{(1-0,093)(30-2-1)}} \approx 5,43;$$

$$se(b_2) = \frac{385,63}{108,82} \sqrt{\frac{1-0,803}{(1-0,093)(30-2-1)}} \approx 0,32.$$

Фактические значения t -критерия для коэффициентов регрессии составят

$$t_{b_1} = \frac{30,08}{5,43} \approx 5,54; \quad t_{b_2} = \frac{2,16}{0,32} \approx 6,75.$$

Найдем табличное значение t -критерия. Число степеней свободы $df = 30 - 2 - 1 = 27$. Зададим уровень значимости $\alpha = 0,05$. Тогда $t_{\text{табл}} = 2,0518$. Оба фактических значения t -критерия больше табличного, поэтому каждый из коэффициентов регрессии значим с вероятностью 0,95.

При построении доверительных интервалов коэффициентов множественной регрессии (*второй способ оценки значимости параметров уравнения регрессии*) также используются формула стандартной ошибки параметра и табличное значение t -критерия. Для линейной функции границы доверительного интервала находят по формуле

$$b_j \pm t_{\text{табл}} \cdot se(b_j), \quad (3.86)$$

где $t_{\text{табл}}$ – табличное значение t -критерия. Параметр b_j значим, если в доверительный интервал не попадает нуль.

В нашем примере доверительные интервалы для коэффициентов регрессии составят:

- для коэффициента b_1 при факторе x_1 :
 - нижняя граница: $30,08 - 2,0518 \cdot 5,43 \approx 18,939$;
 - верхняя граница: $30,08 + 2,0518 \cdot 5,43 \approx 41,221$;
- для коэффициента b_2 при факторе x_2 :
 - нижняя граница: $2,16 - 2,0518 \cdot 0,32 \approx 1,503$;
 - верхняя граница: $2,16 + 2,0518 \cdot 0,32 \approx 2,817$.

Нуль не попадает в доверительные интервалы для обоих коэффициентов: $0 \notin [18,939; 41,221]$ и $0 \notin [1,503; 2,817]$, поэтому каждый из них значим.

Доверительные интервалы для коэффициентов регрессии можно проинтерпретировать. Они указывают границы, в которых с заданной долей вероятности находится параметр. Для рассмотренной линейной функции то, что коэффициент b_1 при факторе x_1 находится в границах $[18,939; 41,221]$, означает, что с вероятностью 0,95 при изменении числа занятых на одного человека объем произведенной продукции изменится не меньше, чем на 18,939 тыс. руб., и не больше, чем на 41,221 тыс. руб., при фиксированной стоимости основных фондов. Аналогичный вывод можно сформулировать по доверительному интервалу для коэффициента b_2 .

Если функция регрессии нелинейная, то оценка значимости ее параметров производится для линеаризованной формы. Одна из рассматриваемых выше моделей являлась нелинейной: $\hat{y} = 37,46 \cdot x_1^{0,5} \cdot x_2^{0,4}$

В линеаризованной форме она имела вид:

$$\ln \hat{y} = \ln 37,46 + 0,5 \ln x_1 + 0,4 \ln x_2.$$

Для нахождения ошибки по формуле (3.84) необходимо найти коэффициент детерминации между логарифмами факторов ($r_{\ln x_1 \ln x_2}^2$), а также средние квадратические отклонения (σ) прологарифмированных значений результата и факторов:

$$r_{\ln x_1 \ln x_2}^2 = 0,105; \quad \sigma_{\ln y} = 0,126; \quad \sigma_{\ln x_1} = 0,128; \quad \sigma_{\ln x_2} = 0,189.$$

Коэффициент детерминации, рассчитанный по линеаризованной модели, известен из расчетов п. 3.5 ($R_{\ln y \ln x_1 \ln x_2}^2 = 0,801$). Ошибки коэффициентов регрессии в данном случае будут равны

$$se(b_1) = \frac{0,126}{0,128} \sqrt{\frac{1 - 0,801}{(1 - 0,105)(30 - 2 - 1)}} \approx 0,089;$$

$$se(b_2) = \frac{0,126}{0,189} \sqrt{\frac{1 - 0,801}{(1 - 0,105)(30 - 2 - 1)}} \approx 0,060.$$

Найдем фактические значения критерия Стьюдента:

$$t_{b_1} = \frac{0,5}{0,089} \approx 5,62; \quad t_{b_2} = \frac{0,4}{0,060} \approx 6,67.$$

Они больше табличного значения t -критерия ($t_{табл} = 2,0518$), и, следовательно, коэффициенты нелинейной регрессии являются значимыми. Этот же результат можно получить с помощью построения доверительных интервалов для коэффициентов регрессии: для b_1 : $0 \notin [0,317; 0,683]$; для b_2 : $0 \notin [0,277; 0,523]$.

Значимость дополнительных факторов, включаемых в уравнение множественной регрессии, можно оценить с помощью *частного F-критерия*. В его основе лежит соотношение сокращения остаточной дисперсии при включении дополнительного фактора и остаточной дисперсии в модели, включающей данный фактор, рассчитанных на одну степень свободы:

$$F_{x_j} = \frac{SS_{ост}^{(1)} - SS_{ост}^{(2)}}{1} : \frac{SS_{ост}^{(2)}}{(n-m-1)}, \quad (3.87)$$

где $SS_{ост}^{(1)}$ – остаточная сумма квадратов для модели без фактора x_j ;

$SS_{ост}^{(2)}$ – остаточная сумма квадратов для модели с фактором x_j .

Разделив каждую из дробей на общую сумму квадратов $SS_{общ}$ и произведя простые преобразования, получим

$$F_{x_j} = \frac{R_{yx_1 \dots x_p}^2 - R_{yx_1 \dots x_{j-1} \dots x_p}^2}{1} : \frac{1 - R_{yx_1 \dots x_j \dots x_p}^2}{(n-m-1)} \quad (3.88)$$

или

$$F_{x_j} = \frac{R_{yx_1 \dots j \dots x_p}^2 - R_{yx_1 \dots x_{j-1} x_{j+1} \dots x_p}^2}{1 - R_{yx_1 \dots}^2} \cdot \frac{n-m-1}{1}, \quad (3.89)$$

где $R_{yx_1 \dots}^2$ — коэффициент детерминации для уравнения зависимости показателя y от факторов $x_1 \dots x_p$, включая фактор x_j ;

$R_{yx_1 \dots \dots x_p}^2$ — коэффициент детерминации для уравнения зависимости показателя y от всех факторов $x_1 \dots x_p$, кроме фактора x_j .

Полученное значение F -критерия сравнивают с табличным при $df_1 = 1$ и $df_2 = (n - m - 1)$ степенях свободы. Если фактическое значение F -критерия больше табличного, включение рассматриваемого фактора x_j существенно уменьшает остаточную дисперсию и является статистически значимым. Можно показать, что для линейной множественной регрессии верно соотношение

$$F_{x_j} = t_{b_j}^2, \quad (3.90)$$

т. е. значимость параметра при факторе x_j означает также, что включение данного фактора в модель регрессии статистически оправдано, и наоборот.

Найдем частные F -критерии для модели $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$.

Для фактора x_1 формула расчета частного F -критерия будет иметь вид

$$F_{x_1} = \frac{R_{yx_1 x_2}^2 - r_{yx_2}^2}{1 - R_{yx_1 x_2}^2} \cdot \frac{n-m-1}{1}.$$

Подставим в нее конкретные числа:

$$F_{x_1} = \frac{0,803 - 0,761^2}{1 - 0,803} \cdot \frac{30 - 2 - 1}{1} \approx 30,68.$$

Табличное значение F -критерия при $df_1 = 1$, и $df_2 = 27$ и $\alpha = 0,05$ равно 4,21. Фактическое значение F -критерия выше табличного, следовательно, включение в модель фактора x_1 после фактора x_2 статистически значимо.

Проверим выполнение соотношения (3.90):

$$t_{b_1}^2 = 30,69 \approx F_{x_1}.$$

Незначительное расхождение между значениями объясняется ошибками округления.

Для фактора x_2 имеем

$$F_{x_2} = \frac{0,803 - 0,682^2}{1 - 0,803} \cdot \frac{30 - 2 - 1}{1} = 46,31.$$

Табличное значение F -критерия то же, что и для фактора x_1 : $F_{\text{табл.}} = 4,21$. Оно меньше фактического значения. Это означает, что включение фактора x_1 после x_2 также статистически значимо.

Кроме оценки значимости включения отдельных факторов F -критерий Фишера используется для оценки значимости всей модели в целом. В этом случае его называют также *общим F-критерием*. Под незначимостью модели понимается одновременное равенство всех коэффициентов при факторах нулю:

$$b_1 = b_2 = \dots = b_p = 0.$$

Отметим, что если модель незначима, то незначимы и показатели корреляции, рассчитанные по ней. Действительно, если

$$b_1 = b_2 = \dots = b_p = 0,$$

то

$$\hat{y} = \bar{y} = a$$

и график функции регрессии параллелен оси абсцисс. Тогда

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 0.$$

Фактическое значение F -критерия рассчитывается по формуле

$$F = \frac{SS_{\text{факт}}}{SS_{\text{ост}}} \cdot \frac{n-m-1}{m}. \quad (3.91)$$

На основе этой формулы можно получить формулу расчета F -критерия через коэффициент множественной детерминации:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \quad (3.92)$$

Фактическое значение F -критерия сравнивают с табличным, найденным для $df_1 = m$ и $df_2 = (n - m - 1)$ степеней свободы. Уравнение регрессии значимо, если фактическое значение F -критерия больше табличного.

Для регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ фактическое значение F -критерия Фишера составит

$$F = \frac{0,803}{1-0,803} \cdot \frac{30-2-1}{2} \approx 55,03.$$

Табличное значение F -критерия при $df_1 = 2$ и $df_2 = 30 - 2 - 1 = 27$ степенях свободы и уровне значимости $\alpha = 0,05$ равно 3,35. Так как фактическое значение F -критерия больше табличного, уравнение регрессии следует признать статистически значимым с вероятностью 0,95.

В основе проверки значимости уравнения регрессии и его параметров по общему и частному F -критерию лежит анализ соотношения дисперсий. Поэтому расчет этих критериев может быть оформлен в виде таблицы дисперсионного анализа.

Расчет общего F -критерия целесообразно оформить в виде таблицы дисперсионного анализа (табл. 3.4).

Таблица 3.4

Анализ статистической значимости модели множественной регрессии

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	<i>F</i> -критерий	Фишера
				фактическое значение	табличное значение для $\alpha = 0,05$
Общая	$n-1$	$SS_{\text{общ}} = \sum(y - \bar{y})^2$	—	—	—
Регрессия	m	$SS_{\text{факт}} = \sum(\hat{y} - \bar{y})^2$	$MS_{\text{факт}}$	$F_{\text{факт}} = MS_{\text{факт}} / MS_{\text{ост}}$	$F_{\text{табл}} (df_1 = m; df_2 = n - m - 1)$
Остаточная	$n-m-1$	$SS_{\text{ост}} = \sum(y - \hat{y})^2$	$MS_{\text{ост}}$	—	—

Для модели регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ дисперсионный анализ представлен в табл. 3.5. Приведенное в таблице фактическое значение *F*-критерия Фишера незначительно отличается от найденного нами ранее из-за ошибок округления коэффициента множественной детерминации R^2 .

Таблица 3.5

Анализ статистической значимости модели множественной регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	<i>F</i> -критерий	Фишера
				фактическое значение	табличное значение для $\alpha = 0,05$
Общая	29	4 461 255	—	—	—
Регрессия	52	3 581 435	1 790 718	54,95	3,35
Остаточная	27	879 819,7	32 585,91	—	—

Для построения таблицы дисперсионного анализа значимости фактора x_j , включаемого в модель, введем дополнительные обозначения:

$$\hat{y}^{(1)} = f(x_1, \dots, x_{j-1}, x_j, \dots, x_p) \quad (3.93)$$

для множественной регрессии без фактора x_j и

$$\hat{y}^{(2)} = f(x_1, \dots, x_j, \dots, x_p) \quad (3.94)$$

для множественной регрессии с включенным фактором x_j .

Соответственно расчеты, относящиеся к первому уравнению регрессии, обозначим верхним индексом ⁽¹⁾, ко второму — верхним индексом ⁽²⁾ (табл. 3.6).

Таблица 3.6

Анализ статистической значимости включения фактора x_j после факторов $x_1 \dots x_{j-1} x_{j+1} \dots x_p$ в модель множественной регрессии

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	<i>F</i> -критерий	Фишера
				фактическое значение	табличное значение для $\alpha = 0,05$
Общая	$n - 1$	$SS_{\text{общ}} = \sum (y - \bar{y})^2$	—	—	—
Регрессия со всеми факторами	m	$SS_{\text{факт}}^{(2)} = \sum (\hat{y}^{(2)} - \bar{y})^2$	$MS_{\text{факт}}^{(2)}$	$F_{\text{факт}}^{(2)}$	$F_{\text{табл}}^{(2)}$ ($df_1 = m$; $df_2 = n - m - 1$)
Регрессия без фактора x_j	$m - 1$	$SS_{\text{факт}}^{(1)} = \sum (\hat{y}^{(1)} - \bar{y})^2$	$MS_{\text{факт}}^{(1)}$	$F_{\text{факт}}^{(12)}$	$F_{\text{табл}}^{(12)}$ ($df_1 = m - 1$; $df_2 = n - m - 1$)
Обусловленная включением в модель фактора x_j после факторов $x_1 x_{j-1} x_{j+1} \dots x$	1	$SS_{\text{факт}}^{(2)} - SS_{\text{факт}}^{(1)}$	$MS_{\text{факт}}^{(x_j)}$	$F_{x_j}^{(2)}$	$F_{\text{табл}}^{(x_j)}$ ($df_1 = 1$; $df_2 = n - m - 1$)
Остаточная для регрессии со всеми факторами	$n - m - 1$	$SS_{\text{ост}}^{(2)} = \sum (y - \hat{y}^{(2)})^2$	$MS_{\text{ост}}^{(2)}$	—	—

Первая, вторая и последняя строки табл. 3.6 соответствуют строкам табл. 3.4. Дополнительные строки предназначены для анализа значимости фактора x_j , включаемого в модель. В частности, в третьей строке (регрессия без фактора x_j) приводятся данные о факторной сумме квадратов и дисперсии на одну степень свободы, рассчитанные для первой модели (со всеми факторами, кроме фактора x_j). В предпоследнем столбце этой строки дано значение *F*-критерия, относящееся к гипотетической ситуации, когда остаточные дисперсии на одну степень свободы, рассчитанные для первой и второй моделей, совпадают. *F*-критерий вычисляется по формуле

$$F_{\text{факт}}^{(12)} = \frac{MS_{\text{факт}}^{(1)}}{MS_{\text{ост}}^{(2)}}. \quad (3.95)$$

В четвертой, предпоследней, строке содержится расчет частного *F*-критерия для фактора x_j по формуле (3.87).

Отметим, что третья и четвертая строки табл. 3.6 будут разными для разных факторов. Поэтому если в модели всего p факторов, необходимо составить p таблиц дисперсионного анализа с частным F -критерием.

В частности, для модели $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ необходимо составить две таблицы с частным F -критерием (табл. 3.7 и 3.8). Фактические значения F -критерия Фишера в них несколько отличаются от рассчитанных по формулам (3.89) и (3.92) из-за ошибок округления коэффициентов детерминации.

Сравнивая фактические и табличные значения F -критерия, получаем, что уравнение регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$ статистически значимо и значимо также включение в него каждого из факторов.

Таблица 3.7

**Анализ статистической значимости включения фактора x_1
после фактора x_2 в модель множественной регрессии**

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	<i>F</i> -критерий Фишера фактическое значение	табличное значение для $\alpha = 0,05$
Общая	29	4 461 255	—	—	—
Регрессия со всеми факторами	2	3 581 435	1 790 718	54,95	3,35
Регрессия без фактора x_1	1	2 583 011	258 303 011	79,27	4,21
Обусловленная включением в модель фактора x_1 после фактора x_2	1	998 424	998 424	30,64	4,21
Остаточная для регрессии со всеми факторами	27	879 819,7	32 585,91	—	—

Таблица 3.8

**Анализ статистической значимости включения фактора x_2
после фактора x_1 в модель множественной регрессии**

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	<i>F</i> -критерий Фишера фактическое значение	табличное значение для $\alpha = 0,05$
Общая	29	4 461 255	—	—	—
Регрессия со всеми факторами	2	3 581 435	1 790 718	54,95	3,35

Окончание табл.3.8

Источники вариации	Число степеней свободы (df)	Сумма квадратов (SS)	Дисперсия на одну степень свободы ($MS = SS/df$)	F -критерий фактическое значение	Фишера табличное значение для $\alpha = 0,05$
Регрессия без фактора x_2	1	2 078 561	2 078 561	63,77	4,21
Обусловленная включением в модель фактора x_2 после фактора x_1	1	1 502 874	1 502 874	46,12	4,21
Остаточная для регрессии со всеми факторами	27	879 819,7	32 585,91	-	-

Исследование значимости отдельных параметров модели и значимости всей модели завершается выводами о возможности дальнейшего использования этой модели для анализа и прогноза экономической ситуации. Модель считается пригодной для дальнейшего использования, если значимы все параметры при факторах и значима модель в целом.

Если некоторые параметры при факторах незначимы, то необходимо изменить спецификацию модели и повторно проверить ее на значимость.

3.7. Прогнозирование по модели множественной регрессии

Прогнозирование по модели множественной регрессии проводится аналогично прогнозированию по уравнению парной регрессии. В частности, строят точечный и интервальный прогнозы.

Точечный прогноз – это расчетное значение результата, полученное подстановкой в уравнение множественной регрессии прогнозных значений факторов. Если заданы прогнозные значения факторов $x_1^{np}, x_2^{np}, \dots, x_p^{np}$, то прогнозное значение результата (точечный прогноз) будет равно

$$\hat{y}_{np} = f(x_1^{np}, x_2^{np}, \dots, x_p^{np}). \quad (3.96)$$

Интервальный прогноз – это минимальное и максимальное значения результата, в промежуток между которыми с заданной долей вероятности попадет фактическое значение результата при заданных прогнозных значениях факторов.

Интервальный прогноз для линейной функции вычисляется по формуле

$$\hat{y}_{np} \pm t_{\text{табл}} \cdot se(y_{np}), \quad (3.97)$$

где $t_{\text{табл}}$ — табличное значение t -критерия Стьюдента при $df = n - m - 1$ степенях свободы;
 $se(y_{np})$ — средняя ошибка прогнозного значения результата, вычисляемая по формуле

$$se(y_{np}) = \sqrt{\frac{\sum e_i^2}{n-m-1}} (1 + X_{np}^T (X^T X)^{-1} X_{np}), \quad (3.98)$$

где X — матрица исходных значений факторов вида 3.30;
 X_{np} — матрица-столбец прогнозных значений факторов вида

$$X_{np} = \begin{pmatrix} 1 \\ x_1^{np} \\ x_2^{np} \\ \vdots \\ x_p^{np} \end{pmatrix}$$

Формула (3.98) частично представлена в матричной форме из-за чрезмерной громоздкости обычной записи.

Найдем прогнозные значения для функции регрессии $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$.

Пусть прогнозные значения факторов составят: число занятых (x_1) — 70 человек; стоимость основных фондов (x_2) — 800 тыс. руб. Найдем точечный и интервальный прогноз объема выпуска продукции.

Точечный прогноз будет равен

$$\hat{y}_{np} = 474,09 + 30,08 \cdot 70 + 2,16 \cdot 800 = 4307,69.$$

Ошибка прогноза по формуле (3.98) составит 211,33. Табличное значение t -критерия при числе степеней свободы $df = 27$ и уровне значимости $\alpha = 0,05$ составляет 2,0518. Следовательно, границы прогноза:

- нижняя: $4307,69 - 2,0518 \cdot 211,33 = 3874,1$;
- верхняя: $4307,69 + 2,0518 \cdot 211,33 = 4741,3$.

Таким образом, можно утверждать, что если на предприятии будут заняты 70 человек и стоимость основных фондов составит 800 тыс. руб., то в среднем объем выпускa продукции составит 4307,7 тыс. руб.; с вероятностью 0,95 он будет находиться в границах от 3874,1 тыс. руб. до 4741,3 тыс. руб.

Прогнозирование по нелинейным моделям множественной регрессии также можно осуществлять по формулам (3.96), (3.97), (3.98), предварительно линеаризовав указанные модели. Например, для степенной функции вместо значений результата и факторов в исходных единицах измерения следует взять их логарифмы.

3.8. Анализ случайных остатков в модели регрессии

Использование МНК для нахождения параметров регрессии предполагает выполнение предпосылок, перечисленных в п. 3.3. На практике проверка их выполнения осуществляется уже после того, как были найдены параметры уравнения регрессии. Статистическому анализу подвергаются случайные остатки (e): если справедливо утверждение, что результат (y) является случайной величиной, а факторы (x_1, \dots, x_p) — неслучайными величинами, то закон распределения случайных остатков должен совпадать с законом распределения результата. Так как предполагается, что случайная величина Y_i ^{*} имеет математическое ожидание, равное выровненному значению результата для заданных значений факторов, то случайный остаток E_i имеет математическое ожидание, равное нулю. Стандартные отклонения (σ_i) для i -го значения результата и i -го значения остатка одинаковы.

Предпосылки МНК выполняются, если случайные остатки E_i удовлетворяют следующим условиям:

1. Математическое ожидание случайного остатка равно нулю:

$$ME_i = 0. \quad (3.99)$$

2. Дисперсия случайных остатков одинакова для различных i :

$$\sigma_{E_i}^2 = \sigma_{E_j}^2 = \sigma^2 \quad (3.100)$$

3. Случайные остатки не зависят друг от друга:

$$r_{E_i E_j} = 0, \quad i \neq j. \quad (3.101)$$

4. Случайные остатки не зависят от значений факторов, входящих в модель регрессии:

$$r_{E_i X_j} = 0. \quad (3.102)$$

5. Остатки распределены по нормальному закону распределения.

Первые четыре условия известны как *условия Гаусса—Маркова*.

В реальных экономических расчетах, как правило, каждая случайная величина E_i имеет только одну реализацию. Поэтому выполнение предпосылок МНК проверяют по всей совокупности остатков.

Математическое ожидание каждого случайного остатка E_i должно быть равно нулю. Тогда математическое ожидание их суммы тоже равно нулю. Выполнение этого требования обеспечивается самой методикой расчетов по МНК. Из первого уравнения системы 3.24 следует, что

$$\sum (y - \hat{y}) = 0,$$

т. е.

$$\sum e = 0.$$

*Напомним, что для каждого конкретного набора значений факторов рассматривается своя случайная величина результата Y_i и соответственно остатка E_i .

Постоянство дисперсии случайных остатков называют *гомоскедастичностью остатков*. Напротив, если эта дисперсия не постоянна, то такое явление называют *гетероскедастичностью остатков*.

Проверка выполнения требования гомоскедастичности остатков может быть произведена визуально на основе графика остатков или с помощью специальных критериев.

Для проведения визуального анализа дисперсии остатков необходимо построить график зависимости остатков от выровненного значения результата. В случае гомоскедастичности «облако» остатков находится в области, параллельной оси абсцисс. Все прочие случаи соответствуют гетероскедастичности остатков.

К тестам, позволяющим выявить наличие гетероскедастичности остатков, относят тесты Гольдфельда—Квандта, Парка, Глейзера, Уайта, ранговой корреляции Спирмена и т. д.

Тест Гольдфельда—Квандта применяется, если случайные остатки предполагаются нормально распределенными случайными величинами.

Процедура проверки следующая:

1. Все наблюдения упорядочивают по мере возрастания какого-либо фактора, который, как предполагается, оказывает влияние на возрастание дисперсии остатков.

2. Упорядоченную совокупность делят на три группы, причем первая и последняя должны быть равного объема с числом единиц большим, чем число параметров модели регрессии. Пусть в первую и третью группы отобрано по k единиц.

3. По первой и третьей группе находят параметры уравнений регрессии и остатки по ним.

4. Используя данные об остатках моделей первой и третьей группы, рассчитывают фактическое значение F -критерия Фишера по формуле

$$F = \frac{\sum_{i=1}^k e_i^2}{\sum_{i=n-k+1}^{n-k} e_i^2} = \frac{SS_{\text{ост}}^{(3)}}{SS_{\text{ост}}^{(1)}}. \quad (3.103)$$

В числителе данной формулы просуммированы квадраты остатков регрессии, построенной по третьей группе ($SS_{\text{ост}}^{(3)}$), в знаменателе — по первой ($SS_{\text{ост}}^{(1)}$).

Если предполагается обратная зависимость дисперсии от значения фактора, то числитель и знаменатель в формуле (3.103) меняют местами.

5. Сравнивают фактическое значение F -критерия с табличным, найденным для $df_1 = df_2 = k - m - 1$ степеней свободы. Если F -фактическое больше табличного, то гипотеза об отсутствии гетероскедастичности отклоняется.

Тесты Парка, Глейзера и Уайта предполагают, что дисперсия случайных остатков представляет собой определенную функцию зависимости от некоего фактора (или факторов).

Для *теста Парка* это зависимость вида:

$$\ln e_i^2 = a + b \ln x_{ij} + v_i, \quad (3.104)$$

где x_{ij} — i -е значение j -го фактора;

v_i — случайный остаток.

По *тесту Глейзера* находят параметры целой серии уравнений, задаваемых функцией:

$$|e_i| = a + b x_{ij}^k + v_i, \quad (3.105)$$

где k — какое-либо число: например, $k = -1; -0,5; 0,5; 1$ и т. п.

Тест Уайта заключается в построении квадратичной функции, включающей все факторы, а также их попарные произведения. Для случая с двумя факторами эта функция будет иметь вид*

$$e_i^2 = a + b_{11}x_{1i} + b_{12}x_{1i}^2 + b_{21}x_{2i} + b_{22}x_{2i}^2 + c_{12}x_{1i}x_{2i} + v_i. \quad (3.106)$$

Остатки считаются гетероскедастичными, если параметр b в функциях по тесту Парка (3.104) или тесту Глейзера (3.105) значим (для теста Глейзера — хотя бы при одном значении k). При проверке по тесту Уайта гетероскедастичность случайных остатков имеется, если вся функция (3.106) значима по *F*-критерию Фишера.

Тест ранговой корреляции Спирмена, так же как и ранее рассмотренные тесты, основывается на предположении о зависимости (прямой или обратной) величины дисперсии случайных остатков от значений какого-либо фактора. Для проведения проверки по этому тесту значения остатков, взятые по модулю, и значения этого фактора ранжируют (например, по возрастанию), а затем находят коэффициент корреляции рангов Спирмена:

$$r_{po} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (3.107)$$

где d — разность между рангами i -го остатка и i -го значения фактора.

Полученное значение коэффициента корреляции проверяют на значимость, рассчитывая фактическое значение *t*-критерия Стьюдента (3.108) и сравнивая его с табличным при числе степеней свободы $df = n - 2$:

$$t_{r_{Cn}} = \frac{r_{po} \sqrt{n-2}}{\sqrt{1 - r_{Cn}^2}}. \quad (3.108)$$

*Слагаемое, содержащее произведение двух факторов, можно опустить. Если факторов три и более, то при достаточном объеме исходных данных в модель вводятся попарные произведения факторов.



Рис. 3.1. Зависимость остатков от фактора x_1

Рис. 3.2. Зависимость остатков от фактора x_2

Если фактическое значение критерия больше табличного, то гипотеза о гомоскедастичности остатков отклоняется.

Причинами гетероскедастичности случайных остатков могут быть неверная функциональная форма уравнения регрессии (неверная спецификация модели), неоднородность исследуемой совокупности. Соответственно способами устранения гетероскедастичности являются построение модели иной функциональной формы и (или) разбиение совокупности на однородные группы. Если по каким-то причинам это сделать невозможно или нежелательно, то для нахождения параметров уравнения регрессии можно воспользоваться обобщенным методом наименьших квадратов.

Проверим на гетероскедастичность наилучшую по величине коэффициента множественной детерминации модель нашего примера $\hat{y} = 474,09 + 30,08x_1 + 2,16x_2$. Для применения всех перечисленных выше тестов необходимо высказать предположение о возможной зависимости значений остатков от какого-либо фактора. Рассмотрим графики зависимости остатков от факторов (рис. 3.1, 3.2).

Визуальный анализ не позволяет сделать точного вывода о наличии гетероскедастичности, поэтому рассмотрим отдельно каждый из факторов.

Используем для анализа критерий Гольдфельда—Квандта. Объем совокупности — 30 наблюдений. Упорядочим ее сначала по фактору x_1 и разделим на три равные части по 10 единиц. В первой и третьей частях найдем параметры уравнений множественной регрессии вида $\hat{y} = a + b_1x_1 + b_2x_2$ и рассчитаем случайные остатки по каждому из них. Получим следующие результаты:

Первая группа (минимальные значения x_1): $y = -666,7 + 44,2x_1 + 3,0x_2$ $SS_{ост}^{(1)} = 284\,046$.

Третья группа (максимальные значения x_1): $y = 522,1 + 36,5x_1 + 1,4x_2$ $SS_{ост}^{(3)} = 98\,824$.

Если предполагать, что дисперсия остатков и фактор x_1 находятся в прямой зависимости, то $F = \frac{98\,824}{284\,046} = 0,35$; в обратной —

$F = \frac{284\,046}{98\,824} = 2,87$. Табличное значение F -критерия равно 3,79 при

$df_1 = df_2 = 7$ степенях свободы и уровне значимости 0,05. Следовательно, дисперсия остатков не зависит от величины фактора x_1 .

Проведем такую же проверку по фактору x_2 . Получим следующие результаты:

Первая группа

(минимальные значения x_2): $y^{\wedge} = -320,67 + 48,2 \cdot x_1 + 3,0 \cdot x_2$ $SS_{oem}^{(1)} = 297\,728$.

Третья группа

(максимальные значения x_2): $y^{\wedge} = 1929,9 + 8,7 \cdot x_1 + 1,4 \cdot x_2$ $SS_{oem}^{(3)} = 30\,087$.

Фактическое значение F -критерия для прямой зависимости дисперсии остатков от фактора x_2 равно 0,1; для обратной — 9,9. Следовательно, можно утверждать, что дисперсия остатков находится в обратной зависимости от фактора x_2 .

Проверим гипотезу о гетероскедастичности остатков с помощью других тестов, опираясь на выводы по тесту Гольдфельда—Квандта.

Тест Парка:

$$\ln e^2 = 11,08 - 0,34 \ln x_2 + v; \quad t_b = -0,16.$$

Тест Глейзера:

$$k = 1 \quad |e| = 188,6 - 0,1x_2 + v; \quad t_b = -0,49;$$

$$k = -1 \quad |e| = 103,0 - 17780 \frac{1}{x_2} + v; \quad t_b = 0,3;$$

$$k = 0,5 \quad |e| = 232,0 - 4,03\sqrt{x_2} + v; \quad t_b = -0,45.$$

Тест Уайта:

$$e^2 = 69792,33 - 7593,6x_1 + 66,4x_1^2 + 652,9x_2 - 0,59x_2^2 + v; \quad F = 0,55.$$

Тест ранговой корреляции Спирмена:

$$r_{n} = -0,099 \quad t_{r_{n}} = -0,053.$$

Параметры при факторе x_2 в тестах Парка и Глейзера незначимы ($t_{\text{табл}} = 2,05$). Незначимы также уравнение регрессии в teste Уайта ($F_{\text{табл}} = 4,2$) и коэффициент ранговой корреляции Спирмена. Следовательно, на основании этих тестов можно утверждать, что остатки в выбранной нами модели гомоскедастичны. Несоответствие этого вывода с выводом по тесту Гольдфельда—Квандта можно объяснить тем, что последний предназначен для анализа больших массивов данных, тогда как в рассматриваемом примере наблюдений недостаточно.

Независимость случайных остатков друг от друга, как правило, проверяется в тех исследованиях, где очевидна последовательность, в которой можно упорядочить остатки, например при исследовании динамических рядов. Взаимозависимость случайных остатков называется *автокорреляцией остатков*. Для ее выявления можно рассчитать

парный линейный коэффициент корреляции или использовать другие критерии, которые будут рассмотрены в следующих главах.

Распределение остатков по нормальному закону распределения обеспечивает корректность выводов относительно значимости параметров уравнения регрессии и всего уравнения в целом по *t*- и *F*-критериям. Было, однако, показано, что при нарушении этой предпосылки оценки параметров уравнения регрессии обладают достаточно хорошими свойствами.

3.9. Обобщенный метод наименьших квадратов

Обобщенный метод наименьших квадратов (ОМНК) применяется в тех случаях, когда нарушены предпосылки МНК, касающиеся характера случайных остатков, а именно:

- постоянство дисперсии случайных остатков;
- некоррелированность остатков между собой.

Применение обычного МНК к модели, в которой нарушены эти предпосылки, ведет к тому, что найденные параметры уравнения регрессии не будут эффективными оценками генеральных параметров. Кроме того, их дисперсии также будут рассчитаны со смещением, что приведет к ложным выводам при оценке качества модели (оценке статистической значимости параметров и всей модели в целом) и при проведении прогнозирования по ней.

Для объяснения обобщенного метода наименьших квадратов на гляднее использовать матричную форму записи. Напомним, что формула для расчета вектора-столбца неизвестных параметров с помощью обычного МНК в матричной форме имеет вид (см. п. 3.3)

$$\boldsymbol{B} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \quad (3.109)$$

Сущность обобщенного метода наименьших квадратов состоит в том, чтобы устранить нарушения предпосылок МНК, «корректируя» расчеты параметров уравнения регрессии с учетом значений ковариационной матрицы остатков. Такая «корректировка» может быть проведена с использованием формулы

$$\boldsymbol{B} = (\boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{Y}, \quad (3.110)$$

где $\boldsymbol{\Omega}$ — ковариационная матрица остатков (3.112). Ее элемент Ω_{ij} равен ковариации остатков для *i*-го и *j*-го значений результата:

$$\Omega_{ij} = \sum (e_i - \bar{e}_i)(e_j - \bar{e}_j), \quad i, j \in [1; n] \quad (3.111)$$

В общем виде ковариационная матрица остатков имеет вид

$$\boldsymbol{\Omega} = \begin{pmatrix} \sum (e_1 - \bar{e}_1)(e_1 - \bar{e}_1) & \sum (e_1 - \bar{e}_1)(e_2 - \bar{e}_2) & \sum (e_1 - \bar{e}_1)(e_n - \bar{e}_n) \\ \sum (e_2 - \bar{e}_2)(e_1 - \bar{e}_1) & \sum (e_2 - \bar{e}_2)(e_2 - \bar{e}_2) & \dots \sum (e_2 - \bar{e}_2)(e_n - \bar{e}_n) \\ \sum (e_n - \bar{e}_n)(e_1 - \bar{e}_1) & \sum (e_n - \bar{e}_n)(e_2 - \bar{e}_2) & \sum (e_n - \bar{e}_n)(e_n - \bar{e}_n) \end{pmatrix} \quad (3.112)$$

Суммы на главной диагонали матрицы Ω представляют ковариацию остатков самих с собой, т. е. их дисперсию остатков для i -го наблюдения.

Если выполняется предпосылка о равенстве математического ожидания случайного остатка E_i нулю, то ковариационная матрица остатков Ω может быть записана как

$$\Omega = \begin{pmatrix} \sigma_{E_1}^2 & \sum e_1 e_2 & \sum e_1 e_n \\ \sum e_2 e_1 & \sigma_{E_2}^2 & \sum e_2 e_n \\ \sum e_n e_1 & \sum e_n e_2 & \sigma_{E_n}^2 \end{pmatrix}. \quad (3.113)$$

Если предпосылки МНК о гомоскедастичности и некоррелированности остатков выполняются, то матрица Ω превращается в скалярную и обобщенный метод наименьших квадратов дает тот же результат, что и обычный МНК.

Основная проблема, возникающая при использовании ОМНК, заключается в том, что фактические значения элементов матрицы Ω неизвестны. Поэтому для применения этого метода используют их *оценки*, полученные на основе исследования имеющихся в распоряжении данных. В этом случае говорят о *доступном ОМНК*.

Для оценки элементов ковариационной матрицы остатков Ω выдвигают разные предпосылки об их характере и структуре. Рассмотрим различные подходы к оценке ковариационной матрицы остатков.

Как правило, гетероскедастичность остатков присуща данным пространственных выборок, а их корреляция — временным рядам. Поэтому в зависимости от характера исходных данных часто предполагают или гетероскедастичность остатков при условии отсутствия автокорреляции, или их автокорреляцию при условии гомоскедастичности.

Если остатки только гетероскедастичны, автокорреляционная матрица остатков имеет вид

$$\Omega = \begin{pmatrix} \sigma_{E_1}^2 & 0 & 0 \\ 0 & \sigma_{E_2}^2 & 0 \\ 0 & 0 & \sigma_{E_n}^2 \end{pmatrix}. \quad (3.114)$$

Таким образом, следует дать оценку не всем элементам матрицы Ω , а только ее n элементам, стоящим на главной диагонали.

Обобщенный метод наименьших квадратов можно с учетом формулы (3.114) записать в виде системы нормальных уравнений:

$$\left\{ \begin{array}{l} \sum_i \frac{y_i}{\sigma_i^2} = a \sum_i \frac{1}{\sigma_i^2} + b_1 \sum_i \frac{x_{1i}}{\sigma_i^2} + \dots + b_p \sum_i \frac{x_{pi}}{\sigma_i^2} \\ \sum_i \frac{y_i x_{1i}}{\sigma_i^2} = a \sum_i \frac{x_{1i}}{\sigma_i^2} + b_1 \sum_i \frac{x_{1i}^2}{\sigma_i^2} + \dots + b_p \sum_i \frac{x_{pi} x_{1i}}{\sigma_i^2} \\ \vdots \\ \sum_i \frac{y_i x_{pi}}{\sigma_i^2} = a \sum_i \frac{x_{pi}}{\sigma_i^2} + b_1 \sum_i \frac{x_{1i} x_{pi}}{\sigma_i^2} + \dots + b_p \sum_i \frac{x_{pi}^2}{\sigma_i^2} \end{array} \right. \quad (3.115)$$

В данной системе переменные и их попарные произведения как бы взвешиваются путем домножения на $\frac{1}{\sigma_i^2}$. Поэтому ОМНК для случая гетероскедастичных остатков называют еще *методом взвешенных наименьших квадратов*.

Дисперсия случайного остатка $\sigma_{E_i}^2$ (при условии $ME_i = 0$) равна

$$\sum (e_i - \bar{e}_i)^2 = \sum e_i^2.$$

Поэтому в качестве оценок дисперсий случайных остатков могут быть использованы квадраты остатков, полученных при применении обычного МНК.

Процедура применения ОМНК в данном случае предполагает следующие шаги:

- к исходным данным применяется обычный МНК и вычисляются остатки e_i ;
- делаются предположения относительно функциональной зависимости дисперсии остатков от каких-либо факторов:

$$\hat{e}^2 = f(x_1, x_2, \dots); \quad (3.116)$$

- с помощью МНК находят параметры модели (3.116), используя в качестве фактических значений зависимой переменной случайные остатки e_i , найденные на первом шаге;
- по модели (3.116) рассчитывают выровненные значения остатков \hat{e}_i^2 . Эти значения рассматривают как оценки неизвестных дисперсий случайных остатков $\sigma_{e_i}^2$, то есть диагональных элементов матрицы (3.114);
- используя ОМНК (3.110), находят новые значения параметров уравнения регрессии.

Применение ОМНК для случая гетероскедастичности остатков (3.114) еще более упрощается, если предполагается зависимость дисперсии случайных остатков от квадратов значения какого-то одного фактора x_j :

$$\sigma_{E_i}^2 = \sigma^2 x_{ji}^2, \quad (3.117)$$

где j — номер фактора, i — номер наблюдения.

В этом случае ковариационная матрица остатков будет иметь вид

$$\Omega = \sigma^2 \begin{pmatrix} x_{j1}^2 & 0 & 0 \\ 0 & x_{j2}^2 & 0 \\ 0 & 0 & x_{jn}^2 \end{pmatrix}. \quad (3.118)$$

Диагональные элементы матрицы Ω уже известны и равны квадратам фактических значений фактора x_j . Постоянный множитель σ^2 при расчетах по формуле (3.110) сокращается, поэтому матрицу Ω можно заменить матрицей Ω_1 без этого множителя:

$$\Omega_1 = \begin{pmatrix} x_{j1}^2 & 0 & 0 \\ 0 & x_{j2}^2 & 0 \\ 0 & 0 & x_{jn}^2 \end{pmatrix}. \quad (3.119)$$

Использование значений какого-либо фактора в качестве основы для оценки ковариационной матрицы остатков позволяет экономически проинтерпретировать результаты, полученные по ОМНК. Пусть, например, имеется уравнение множественной линейной регрессии с двумя факторами:

$$y = a + b_1 x_1 + b_2 x_2 + e. \quad (3.120)$$

Предположим также, что установлены гетероскедастичность случайных остатков и зависимость их дисперсий от квадрата фактора x_2 . Тогда применение ОМНК предполагает, что каждый элемент уравнения регрессии (3.120) будет разделен на x_2 :

$$\frac{y}{x_2} = a \frac{1}{x_2} + b_1 \frac{x_1}{x_2} + b_2 + \frac{e}{x_2}. \quad (3.121)$$

Нетрудно убедиться, что применение к уравнению (3.121) обычного МНК даст систему (3.115) с заменой σ_i^2 на x_2^2 . Полученные значения параметров a , b_1 , b_2 в уравнении (3.121) следует правильно интерпретировать. В частности, параметр a является свободным членом и не интерпретируется, а параметры b_1 и b_2 являются коэффициентами условно-чистой регрессии при факторах x_1 и x_2 .

Рассмотрим, например, совокупность страховых организаций. Построим множественную линейную регрессию. В качестве результата возьмем общую страховую сумму (млрд руб.), а в качестве факторов — количество договоров страхования (тыс.) и уставный капитал компаний (млн руб.). Исходные данные, а также результаты последующих вычислений приведены в табл. 3.9 (цифры условные).

Таблица 3.9

Анализ остатков для модели зависимости общей страховой суммы (y , млрд руб.) от количества договоров страхования (x , тыс.) и уставного капитала (z , млн руб.)

п/п	59 x	z	y	\hat{y}	e	59 $/x = x_1$	$z/x = z_1$	59 $y/x = y_1$	\hat{y}	e_1
1	20	390	51,3	52,93	-1,626	0,050	19,500	2,565	2,646	-0,081
2	22	380	57,2	54,43	2,767	0,045	17,273	2,600	2,482	0,118
3	23	380	53,9	55,60	-1,702	0,043	16,522	2,343	2,427	-0,083
4	25	410	63,2	60,43	2,767	0,040	16,400	2,528	2,417	0,111
5	27	420	61,1	63,60	-2,501	0,037	15,556	2,263	2,355	-0,092
6	28	420	67,8	64,77	3,030	0,036	15,000	2,421	2,314	0,107
7	30	550	73,5	77,91	-4,407	0,033	18,333	2,450	2,559	-0,109
8	32	520	80,0	77,75	2,247	0,031	16,250	2,500	2,406	0,094
9	34	410	68,1	70,95	-2,854	0,029	12,059	2,003	2,097	-0,094
10	35	450	79,0	75,45	3,555	0,029	12,857	2,257	2,156	0,101
11	37	470	75,6	79,44	-3,845	0,027	12,703	2,043	2,144	-0,101
12	44	460	91,4	86,80	4,603	0,023	10,455	2,077	1,979	0,099
13	45	480	85,1	89,63	-4,527	0,022	10,667	1,891	1,994	-0,103
14	46	520	98,2	94,12	4,081	0,022	11,304	2,135	2,041	0,094
15	49	510	91,6	96,80	-5,195	0,020	10,408	1,869	1,975	-0,106
16	52	450	101,1	95,32	5,782	0,019	8,654	1,944	1,846	0,098
17	57	460	96,9	101,99	-5,094	0,018	8,070	1,700	1,803	-0,103
18	59	450	110,2	103,50	6,699	0,017	7,627	1,868	1,770	0,098
19	64	500	107,4	113,50	-6,100	0,016	7,813	1,678	1,784	-0,106
20	66	550	126,3	119,99	6,309	0,015	8,333	1,914	1,822	0,092
21	68	590	118,1	125,65	-7,552	0,015	8,676	1,737	1,847	-0,111
22	70	800	149,0	145,43	3,565	0,014	11,429	2,129	2,050	0,079

После применения МНК к исходным данным получим следующее уравнение регрессии:

$$y = -2,85 + 1,17x + 0,083z + e; \quad R^2 = 0,968; \quad F = 285,16.$$

Это уравнение статистически значимо с вероятностью 0,95 ($F_{\text{табл}} = 3,52$), связь между показателями очень тесная. Значимы также параметры при факторах ($t_{\text{факт}} = 13,26$ и $5,33$ при $t_{\text{табл}} = 2,09$, $a = 0,05$). Таким образом, эти характеристики указывают на высокое качество модели. Однако выводы относительно значимости параметров будут надежными, и по модели можно проводить дальней-

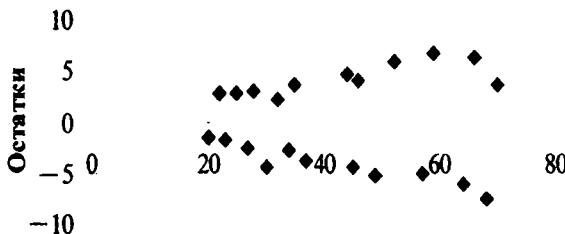


Рис. 3.3. Зависимость случайных остатков от фактора x

ший анализ и прогноз только в том случае, если в результате анализа случайных остатков не будет установлено нарушение предпосылок МНК. Так как данные пространственные, то наиболее возможным нарушением может быть нарушение гомоскедастичности.

Проведем проверку остатков на гетероскедастичность по тесту Глейзера, задав $k = 1$ (в скобках указано фактическое значение t -критерия для коэффициента при факторе):

$$|e| = 0,49 + 0,086x + v; \quad |e| = 0,93 + 0,0066z + v. \quad (6,91) \quad (1,77)$$

Так как коэффициент при факторе x значим ($t_{\text{табл}} = 2,08$), то можно утверждать, что имеет место гетероскедастичность остатков, причем их дисперсия зависит от фактора x . Этот вывод подтверждает график зависимости случайных остатков от фактора x (рис. 3.3).

Так как модуль остатков линейно зависит в соответствии с тестом Глейзера от значения фактора x , можно предположить, что их дисперсия пропорциональна квадрату этого фактора:

$$\sigma_{E_i}^2 = \sigma^2 x^2.$$

Тогда для устранения гетероскедастичности необходимо найти параметры уравнения

$$\frac{y}{x} = a \frac{1}{x} + b_1 + b_2 \frac{z}{x} + \frac{e}{x}.$$

Введем следующие обозначения:

$$\frac{y}{x} = y_1; \quad \frac{1}{x} = x_1; \quad \frac{z}{x} = z_1; \quad \frac{e}{x} = e_1.$$

Уравнение регрессии будет иметь вид

$$y_1 = ax_1 + b_1 + b_2 z_1 + e_1.$$

После применения МНК к этому уравнению получим

$$y_1 = 0,06x_1 + 1,21 + 0,074z_1 + e_1.$$

Проверка остатков этого уравнения по тесту Глейзера дает следующий результат:

$$|e| = 0,1 - 0,000036x + v.$$

Фактическое значение t -критерия для коэффициента при x равно $-0,26$, которое по модулю меньше табличного ($t = 2,08$), поэтому можно говорить об отсутствии гетероскедастичности случайных остатков.

Таким образом, зависимость общей страховой суммы от количества договоров страхования и величины уставного фонда может быть выражена следующим уравнением регрессии:

$$y = 0,06 + 1,21x + 0,074z + e.$$

Параметры этого уравнения регрессии интерпретируют так же, как параметры уравнения, полученного с помощью МНК. То есть при изменении количества договоров на одну тысячу общая страховая сумма в среднем изменится на 1,21 млрд руб. при неизменном уставном капитале, а при изменении уставного капитала на 1 млн руб. общая страховая сумма в среднем изменится на 0,074 млрд руб. при фиксированном количестве договоров страхования.

Если случайные остатки автокоррелированы и предполагается отсутствие гетероскедастичности, то ковариационная матрица остатков примет вид

$$\Omega = \begin{pmatrix} 1 & \sum e_1 e_2 & \sum e_1 e_n \\ \sum e_2 e_1 & 1 & \sum e_2 e_n \\ \sum e_n e_1 & \sum e_n e_2 & 1 \end{pmatrix}. \quad (3.122)$$

В этой матрице неизвестны уже $n(n-1)/2$ элементов, поэтому, как правило, выдвигают предположения относительно взаимосвязи этих элементов между собой, сокращая тем самым количество неизвестных. Например, предполагают, что случайные остатки связаны автокорреляционной зависимостью первого порядка:

$$e_i = \rho_1 e_{i-1} + \delta_i, \quad (3.123)$$

где δ_i — случайные остатки, удовлетворяющие предпосылкам МНК,

ρ_1 — неизвестный параметр.

Можно показать, что в этом случае ковариационная матрица остатков (3.122) примет вид

$$\Omega = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_1^2 & \rho_1^{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_1^{n-2} \\ \rho_1^2 & \rho_1 & 1 & \rho_1^{n-3} \\ \rho_1^{n-1} & \rho_1^{n-2} & \rho_1^{n-3} & 1 \end{pmatrix}, \quad (3.124)$$

где σ^2 — постоянная дисперсия остатков.

Как и в случае автокорреляции остатков, постоянный множитель σ^2 можно опустить, используя для расчетов только матрицу

$$\Omega_1 = \begin{pmatrix} 1 & \rho_1 & \rho_1^{2} & \rho_1^{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_1^{n-2} \\ \rho_1^{2} & \rho_1 & 1 & \rho_1^{n-3} \\ \rho_1^{n-1} & \rho_1^{n-2} & \rho_1^{n-3} & 1 \end{pmatrix}. \quad (3.125)$$

Параметр ρ_1 , как правило, неизвестен. Его оценку производят, исследуя взаимосвязь остатков, полученных после применения к модели регрессии обычного МНК.

Контрольные вопросы

1. Какой вид связей между показателями отражает уравнение регрессии? Дайте краткую характеристику его элементов.
2. Какие требования предъявляются к объему наблюдений, необходимому для построения уравнения регрессии?
3. Какие требования предъявляются к факторам, включаемым в уравнение регрессии?
4. Что такое мультиколлинеарность факторов и как ее выявить?
5. Назовите основные алгоритмы построения модели множественной регрессии.
6. Что такое модель с усеченными данными?
7. Что такое модель с цензурированными данными?
8. Опишите *tobit*-модель.
9. Чем различаются уравнения множественной регрессии в натуральном и стандартизованном масштабе?
10. Каковы свойства стандартизованных переменных?
11. Почему конкретные значения параметров уравнения регрессии называют их оценками?
12. Что означает, что оценка параметра является несмещенной? эффективной? состоятельной?
13. Какие методы могут применяться для нахождения параметров уравнения регрессии?
14. Какими статистическими свойствами должны обладать исходные данные, чтобы выполнялись предпосылки метода наименьших квадратов?
15. Какой должна быть функция регрессии, чтобы применить к ней МНК?
16. Что означает понятие «линеаризация модели»? Как она может быть произведена?
17. Как записывается решение с помощью МНК в матричной форме?
18. В чем сущность метода максимального правдоподобия?
19. В каком случае оценки параметров уравнения регрессии с помощью МНК и ММП совпадут?

20. Назовите показатели силы связи, которые можно рассчитать по уравнению множественной регрессии. В чем их сходство и в чем различие?
21. Для какой функции регрессии константами являются абсолютные показатели силы связи? относительные показатели силы связи (коэффициенты эластичности)?
22. Как связаны между собой коэффициенты уравнения множественной линейной регрессии в натуральном и стандартизованном масштабе?
23. Назовите показатели тесноты связи для уравнения множественной регрессии. В каких границах они изменяются?
24. Для каких нелинейных функций регрессии показатели тесноты связи, найденные для исходной и линеаризованной формы, не равны между собой? Какой вариант расчета предпочтительнее?
25. Как будет изменяться показатель тесноты связи при росте количества параметров уравнения регрессии и при прочих равных условиях? Какой показатель позволяет учесть такое изменение?
26. Поясните смысл показателей частной детерминации.
27. Что понимают под значимостью параметра?
28. Какой критерий используется для оценки значимости параметров уравнения регрессии?
29. Какими способами можно оценить значимость параметров уравнения регрессии с помощью критерия Стьюдента? В чем их сходство и различие?
30. Для чего используется частный F -критерий?
31. Что понимают под значимостью модели регрессии в целом? Какой метод используется для ее оценки? Опишите методику его применения.
32. Каковы последствия, с точки зрения применения модели регрессии на практике, наличия в ней незначимых параметров (факторов)? незначимости всей модели в целом?
33. Что такое точечный и интервальный прогноз? Как они связаны между собой? Какова процедура нахождения прогнозных значений?
34. Почему точечный прогноз всегда необходимо дополнять интервальным?
35. Каким требованиям должны удовлетворять случайные остатки, чтобы выполнялись предпосылки МНК?
36. Поясните термины «гомоскедастичность» и «гетероскедастичность остатков».
37. Какие тесты применяют для выявления гетероскедастичности остатков?
38. В чем сходство и в чем различие тестов Парка, Глейзера, Уайта?
39. Назовите способы устранения гетероскедастичности остатков.
40. В каких случаях применяют обобщенный метод наименьших квадратов?
41. В чем заключается проблема применения ОМНК и каковы пути ее решения?
42. Объясните назначение и сущность метода взвешенных наименьших квадратов.

ГЛАВА 4. ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

4.1. Особенности включения в модели регрессии неколичественных показателей

При изучении связей между показателями результативный признак, являющийся количественной переменной, может зависеть не только от количественных, но и от неколичественных факторных признаков. Например, на величину надоя с одной фуражной коровы (количественный показатель) оказывает влияние качество кормов (плохое, среднее, хорошее — неколичественный показатель). Очевидно, что для включения неколичественной переменной в уравнение регрессии необходимо каким-то образом поставить в соответствие ее качественным значениям числа. Это можно сделать с помощью так называемых фиктивных переменных (*dummy variables*).

Фиктивные переменные — это переменные бинарного типа, т. е. каждая переменная может принимать всего два значения — единица и нуль:

$$z = \begin{cases} 1 \\ 0 \end{cases}$$

Для учета влияния одной неколичественной переменной вводят несколько фиктивных переменных: их должно быть на единицу меньше, чем значений этой переменной.

За каждым (кроме одного) значением неколичественной переменной закрепляют одну фиктивную переменную. Данная фиктивная переменная равна единице, если неколичественная переменная приняла значение, соответствующее этой фиктивной переменной, и равна нулю в противном случае.

Ситуация, когда неколичественная переменная приняла значение, не имеющее в соответствии фиктивной переменной, описывается равенством всех фиктивных переменных нулю. Значение неколичественной переменной, не имеющее соответствующей фиктивной переменной, может быть выбрано произвольно. Так как этому значению соответствует равенство всех фиктивных переменных нулю, его рассматривают как базу сравнения значений результата при разных значениях неколичественной факторной переменной. За это значение обычно принимают то, с которым будут сравнивать значения результата при других значениях неколичественного фактора.

Например, рассматривается влияние уровня квалификации работника на объем выпускаемой им продукции. Пусть имеются три возможных степени квалификации: высшая, средняя, низкая. Для двух из этих значений необходимо ввести две фиктивные переменные. Так как мы ожидаем, что максимальная отдача — от работника с высшей квалификацией, за базу можно принять минимальную (низкую) квалификацию. Для остальных двух значений неколичественной переменной введем две фиктивные переменные — z_1 и z_2 :

$$z_1 = \begin{cases} 1 & \text{высшая квалификация;} \\ 0 & \text{не высшая квалификация.} \end{cases} \quad z_2 = \begin{cases} 1 & \text{средняя квалификация;} \\ 0 & \text{не средняя квалификация.} \end{cases}$$

Тогда, если $z_1 = z_2 = 0$, то неколичественная переменная «уровень квалификации» принимает значение «низкая квалификация».

4.2. Спецификация моделей регрессии с фиктивными независимыми переменными

В общем случае модель с фиктивными переменными имеет вид

$$y = f(x_1, x_2, \dots, x_p, z_{11}, z_{12}, \dots, z_{21}, z_{22}, \dots, z_{j1}, z_{j2}, \dots, e), \quad (4.1)$$

- где y — переменная-результат;
 x_1, x_2, \dots, x_p — количественные переменные-факторы;
 z_{11}, z_{12}, \dots — фиктивные переменные, соответствующие значениям первой неколичественной переменной-фактора;
 z_{21}, z_{22}, \dots — фиктивные переменные, соответствующие значениям второй неколичественной переменной-фактора;
 z_{j1}, z_{j2}, \dots — фиктивные переменные, соответствующие значениям j -й неколичественной переменной-фактора;
 e — случайный остаток.

Включение в модель фиктивных переменных расширяет круг вопросов, подлежащих решению на этапе спецификации модели.

Учет влияния неколичественной переменной, принимающей три и более значений, означает необходимость ввода в уравнение регрессии двух и более фиктивных переменных. Таким образом, значительно увеличивается число параметров, что должно быть подкреплено соответствующим объемом исходных данных (не менее 7 на один параметр, не считая свободного члена).

При включении в уравнение регрессии фиктивных переменных возникает также вопрос о характере влияния количественных факторов на результат при различных значениях неколичественного фактора. Ниже будут рассмотрены различные варианты моделей регрессии с фиктивной переменной.

4.3. Модели регрессии с фиктивными переменными сдвига

Рассмотрим в качестве формы уравнения регрессии линейную функцию. Для простоты возьмем в качестве факторов одну количественную переменную (x_1) и одну фиктивную переменную (z_{11}):

$$y = a + b_1 x_1 + c_{11} z_{11} + \epsilon. \quad (4.2)$$

Из этого уравнения следует, что при $z_{11} = 1$ результат (y) равен

$$y = (a + c_{11}) + b_1 x_1 + \epsilon, \quad (4.3)$$

а при $z_{11} = 0$ результат (y) равен:

$$y = a + b_1 x_1 + \epsilon. \quad (4.4)$$

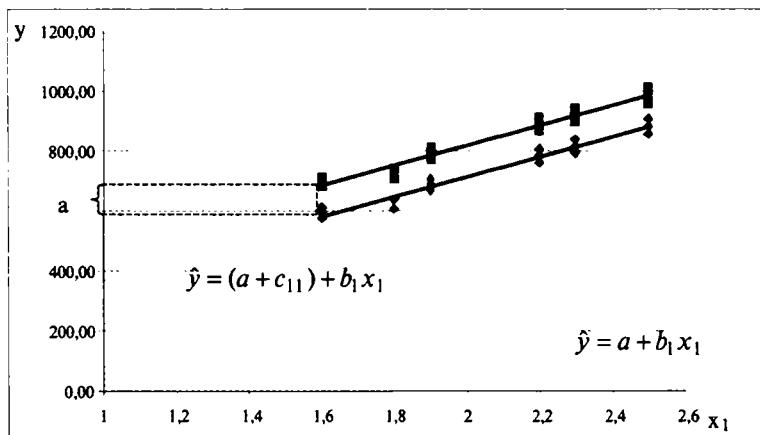


Рис. 4.1. Сила влияния количественного фактора (x_1) на результат (y) не зависит от значения фиктивной переменной (z_{11})

Сравнивая два полученных уравнения (4.3) и (4.4), видим, что они различаются величиной свободного члена. То есть для одного уровня неколичественной переменной уровень результата всегда в среднем будет на c_{11} единиц выше или ниже, чем для другого.

Графически эта ситуация соответствует двум параллельным прямым (рис. 4.1). Отметим, что коэффициент b_1 при количественном факторе остается неизменным. То есть изменение фактора x_1 оказывает одинаковое влияние на результат при разных значениях неколичественной переменной.

Так как изменение значения фиктивной переменной в модели 4.2 приводит к изменению значения результата на некую среднюю величину, не зависящую от значений количественного фактора, такую переменную еще называют *фиктивной переменной сдвига*. Изменение ее значения приводит к переходу от одной параллельной прямой к другой.

Например, рассмотрим в качестве результата стоимость изготовления металлической двери (y , тыс. руб.), в качестве количественного фактора — площадь дверного полотна (x_1 , м²), в качестве неколичественного фактора (z_{11}) — страну — изготовителя замка. С некоторым допущением разделим все дверные замки на две группы: отечественные (более дешевые) и импортные (более дорогие). То есть фиктивная переменная z_{11} будет определена как

$$z_{11} = \begin{cases} 1, & \text{если замок импортный;} \\ 0, & \text{если замок отечественный.} \end{cases}$$

Будем считать, что остальные факторы, оказывающие влияние на стоимость двери (дверные глазки, декоративная отделка, дополнительные пластины усиления и т. д.), являются одинаковыми для всех двер-

рей, а стоимость двери при прочих равных условиях колеблется в зависимости от фирмы-изготовителя.

Исходные данные приведены в табл. 4.1.

Таблица 4.1

Стоимость металлической двери в зависимости от площади дверного полотна и типа замка

Площадь дверного полотна $y, \text{м}^2$	Тип замка z_{11}	Стоимость двери, x_1 тыс. руб.	Площадь дверного полотна $y, \text{м}^2$	Тип замка z_{11}	Стоимость двери, x_1 тыс. руб.
1,6	0	16,656	1,6	1	19,672
1,6	0	17,126	1,6	1	20,131
1,6	0	17,600	1,6	1	20,593
1,8	0	18,463	1,8	1	21,465
1,8	0	17,524	1,8	1	20,569
1,8	0	18,449	1,8	1	21,508
1,9	0	19,352	1,9	1	22,387
1,9	0	19,920	1,9	1	22,915
1,9	0	20,456	1,9	1	23,495
2,2	0	22,051	2,2	1	25,076
2,2	0	22,683	2,2	1	25,703
2,2	0	23,337	2,2	1	26,351
2,3	0	22,952	2,3	1	25,979
2,3	0	23,614	2,3	1	26,672
2,3	0	24,291	2,3	1	27,319
2,5	0	24,752	2,5	1	27,766
2,5	0	25,485	2,5	1	28,515
2,5	0	26,193	2,5	1	29,236

После применения МНК к уравнению (4.2) получим

$$\hat{y} = 1,380 + 9,649x_1 + 3,025z_{11} \quad (4.5)$$

$$(t_{\text{факт}}) (2,05) (30,03) (15,20)$$

Табличное значение t -критерия Стьюдента составляет 2,04 при числе степеней свободы $df = n - m - 1 = 36 - 2 - 1 = 33$ и уровне значимости $\alpha = 0,05$. Следовательно, все параметры уравнения регрессии значимы. То есть оба фактора оказывают значимое воздействие на результат. Это подтверждает также коэффициент множественной детерминации — он равен 0,97.

Проинтерпретируем параметры полученного уравнения регрессии.

Коэффициент b_1 при факторе x_1 равен 9,649, следовательно, с изменением площади дверного полотна на один квадратный метр ее

стоимость в среднем возрастает на 9,649 тыс. руб. при условии, что страна-производитель замка не меняется.

Коэффициент c_{11} при фиктивной переменной равен 3,025, что означает, что дверь с импортным замком в среднем на 3,025 тыс. руб. дороже, чем дверь с отечественным замком, при одной и той же площади дверного полотна.

Подставляя значение фиктивной переменной $z_{11} = 0$, получим уравнение зависимости стоимости двери от ее площади при условии, что замок отечественный, $z_{11} = 1$ — то же уравнение для двери с импортным замком:

$$z_{11} = 0: \quad \hat{y} = 1,380 + 9,649x_1; \quad (4.6)$$

$$z_{11} = 1: \quad \hat{y} = 4,405 + 9,649x_1. \quad (4.7)$$

Отметим, что если находить параметры этих уравнений регрессии отдельно, применяя к исходным данным МНК, получим аналогичный, но не совсем тот же результат.

Так, после применения МНК к исходным данным для дверей с отечественными замками уравнение регрессии имеет вид

$$\hat{y} = 1,400 + 9,640x_1. \quad (4.8)$$

Применение МНК к данным для дверей с импортными замками приводит к следующему результату:

$$\hat{y} = 4,386 + 9,659x_1. \quad (4.9)$$

Различие параметров моделей (4.6), (4.7) и (4.8), (4.9) объясняется тем, что на стоимость дверей оказывают воздействие случайные факторы (колебания курса валют, разные расценки у фирм-производителей). Поэтому минимизация квадратов случайных отклонений в целом по совокупности и по отдельным ее частям приводит к разным результатам.

4.4. Модели регрессии с фиктивными переменными наклона

Рассмотрим другую ситуацию: коэффициент регрессии при количественном факторе зависит от значения фиктивной переменной. То есть можно записать:

$$\hat{y} = a + b_{11}x_1, \quad \text{если } z = 0; \quad (4.10)$$

$$\hat{y} = a + b_{12}x_1, \quad \text{если } z = 1; \quad (4.11)$$

$$b_{11} \neq b_{12}. \quad (4.12)$$

В таком случае говорят, что имеют место структурные изменения в исследуемой зависимости. Для их учета в уравнении регрессии фиктивную переменную вводят как сомножитель при количественной переменной:

$$\hat{y} = a + b_1x_1 + d_{111}x_1z_{11}* \quad (4.13)$$

*Так как параметр d объединяет две переменные — x_1 и z_{11} , он имеет тройной индекс — d_{111} .

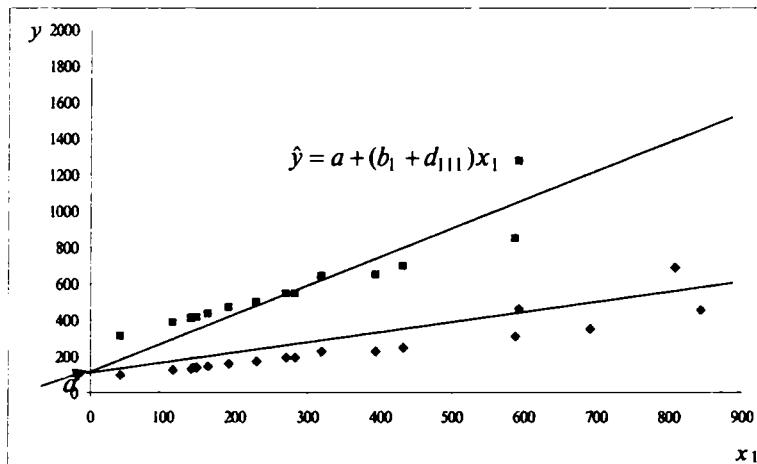


Рис. 4.2. Сила влияния количественного фактора (x_1) на результат (y) зависит от значения фиктивной переменной (z_{11})

Действительно, если рассмотреть это уравнение для $z_{11} = 1$ и для $z_{11} = 0$, получим соответственно

$$\begin{aligned} z_{11} = 0 & \quad \hat{y} = a + b_1 x_1 \\ z_{11} = 1 & \quad \hat{y} = a + (b_1 + d_{111}) x_1; \end{aligned}$$

Следовательно, коэффициент b_{12} из модели (4.11) будет равен $(b_{11} + d_{111})$.

Графически модель (4.13) можно представить в виде двух прямых с разным углом наклона, отражающих зависимость результата от количественного фактора при разных значениях фиктивной переменной (рис.4.2.). Так как речь идет о фиктивной переменной, включение которой позволяет изменить угол наклона прямой, такую переменную называют *фиктивной переменной наклона*.

Соответственно параметр b_1 интерпретируется как сила влияния количественного фактора при одном значении неколичественной переменной (для которой $z_{11} = 0$), а параметр d_{111} — как среднее изменение силы влияния количественного фактора при переходе от одного значения неколичественной переменной к другому (при переходе от $z_{11} = 0$ к $z_{11} = 1$).

Рассмотрим в качестве примера зависимость стоимость проезда в поездах дальнего следования (y , руб.) от расстояния (x_1 , км) и типа вагона (плацкартный или купейный). Очевидно, что стоимость проезда в вагонах разного типа различается, но эта разница зависит от расстояния. Так как в плацкартном вагоне места всегда дешевле, примем этот тип вагона за базу ($z_{11} = 0$). Для купейного вагона соответственно $z_{11} = 1$. Исходные данные приведены в табл. 4.2.

Таблица 4.2

**Стоимость проезда в поезде дальнего следования
из Санкт-Петербурга в зависимости от расстояния и типа вагона**

Станция назначения	Расстояние x_1 , км	Тип вагона z_{11}	Стоимость проезда y , руб.	Станция назначения	Расстояние x_1 , км	Тип вагона z_{11}	Стоимость проезда y , руб.
Мга	42	0	98,1	Мга	42	1	313,4
Волхов	114	0	124,9	Волхов	114	1	384,2
Будогощь	140	0	133,4	Будогощь	140	1	404,4
Приозерск	141	0	137,4	Приозерск	141	1	414,6
Луга	147	0	137,4	Луга	147	1	414,6
М. Вишера	162	0	145,5	М. Вишера	162	1	435,5
Тихвин	192	0	157,6	Тихвин	192	1	465,8
Пикалево	230	0	169,7	Пикалево	230	1	495,3
Подпорожье	272	0	189,8	Подпорожье	272	1	547,3
Псков	284	0	189,8	Псков	284	1	547,3
Бологое	319	0	227,9	Бологое	319	1	643,3
Петрозаводск	394	0	229,2	Петрозаводск	394	1	648,4
Осташков	431	0	249,3	Осташков	431	1	697,4
Вел. Луки	588	0	308,9	Вел. Луки	588	1	850,5
Вологда	592	0	463,4	Вологда	592	1	1275,2
Ярославль	691	0	349,6	Ярославль	691	1	950,7
Тула	808	0	687	Тула	808	1	1834,8
Владимир	843	0	458	Владимир	843	1	1223,5

Применим МНК к модели (4.13):

$$\begin{cases} \sum y = na + b_1 \sum x_1 + d_{111} \sum x_1 z_{11} \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + d_{111} \sum x_1^2 z_{11} \\ \sum yx_1 z_{11} = a \sum x_1 z_{11} + b_1 \sum x_1^2 z_{11} + d_{111} \sum x_1^2 z_{11}^2 \end{cases}$$

Подставим в систему исходные данные:

$$\begin{cases} 17\ 003,1 = 36a + 12\ 780b_1 + 6390d_{111} \\ 8\ 182\ 500 = 12780a + 6\ 652\ 916b_1 + 3\ 326\ 458d_{111} \\ 5\ 998\ 322 = 6390a + 3\ 326\ 458b_1 + 3\ 326\ 458d_{111} \end{cases}$$

После решения системы уравнение регрессии (4.13) примет вид

$$\hat{y} = 112,21 + 0,44x_1 + 1,15x_1 z_{11} \quad (4.14)$$

(т.факт.) (2,23) (4,31) (11,40)

Табличное значение t -критерия равно 2,04 ($\alpha = 0,05$, $df = 33$), следовательно, все найденные параметры уравнения регрессии значимы. Это подтверждает нашу гипотезу о том, что прирост стоимости проезда на один километр различен для плацкартных и купейных вагонов. Теснота связи между результатом и факторами высокая, $R^2 = 0,89$.

Параметр b_1 , равный 0,44, показывает, что с изменением расстояния на один километр стоимость проезда в плацкартном вагоне в среднем изменяется в ту же сторону на 0,44 руб. Параметр d_{111} равен 1,15, т. е. дополнительно за каждый километр пути за проезд в купейном вагоне нужно заплатить на 1,15 руб. больше, чем в плацкартном. Можно сказать иначе. Сумма параметров b_1 и d_{111} равна 1,59, следовательно, с изменением расстояния на один километр стоимость проезда в купейном вагоне меняется в ту же сторону в среднем на 1,59 руб.

4.5. Общий вид модели регрессии с фиктивными переменными

Модели, рассмотренные выше, можно считать частными случаями общей модели. Очевидно, что это будет модель, в которой при разных значениях фиктивной переменной будут разные параметры. Например, для простейшего случая с одним количественным фактором (x_1) и одной фиктивной переменной (z_{11}) необходимо объединить в одну модель два уравнения регрессии:

$$\begin{array}{ll} \text{при } z_{11} = 0 & \hat{y} = a_1 + b_{11}x_1; \\ \text{при } z_{11} = 1 & \hat{y} = a_2 + b_{12}x_1. \end{array} \quad (4.15)$$

(4.16)

Используем для этого модели (4.2) и (4.13):

$$\hat{y} = a + b_1 x_1 + d_{111} x_1 z_{11} + c_{11} z_{11}. \quad (4.17)$$

Действительно,

$$\begin{array}{ll} \text{при } z_{11} = 0 & \hat{y} = a + b_1 x_1; \\ \text{при } z_{11} = 1 & \hat{y} = (a + c_{11}) + (b_1 + d_{111})x_1. \end{array}$$

Параметры уравнений (4.15) и (4.16) можно выразить через параметры уравнения (4.17) следующим образом:

$$\begin{aligned} a_1 &= a; b_{11} = b_1; \\ a_2 &= (a + c_{11}); b_{12} = (b_1 + d_{111}). \end{aligned}$$

Графически модель (4.17) может быть представлена в виде двух прямых, отражающих зависимость результата от количественного фактора при разных значениях фиктивной переменной. Параметры этих прямых различны, поэтому они имеют разный угол наклона (разный коэффициент при количественной переменной) и пересекают ось ординат в разных точках (разный свободный член, рис. 4.3).

Рассмотрим в качестве примера зависимость стоимости проезда (y , руб.) от расстояния (x_1 , км) и типа поезда ($z_{11} = 0$, если проезд осуществляется на электричке; $z_{11} = 1$ — на поезде дальнего следования). Можно ожидать, что характер ценообразования на два этих типа поез-

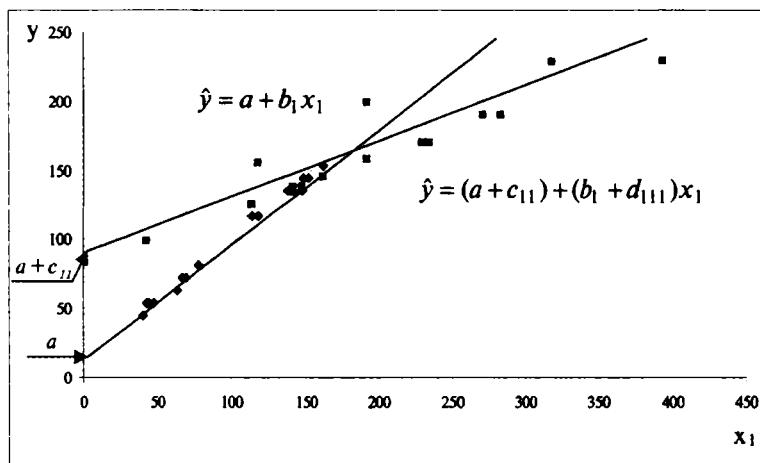


Рис. 4.3. Влияние неколичественного фактора полностью изменяет параметры регрессии результата (y) на количественный фактор (x_1)

дов полностью различен, т. е. для описания зависимости следует применить модель (4.17). Исходные данные приведены в табл. 4.3.

Таблица 4.3

Стоимость проезда в электричках и поездах дальнего следования из Санкт-Петербурга в зависимости от расстояния и типа вагона

Станция назначения	Расстояние x_1 , км	Тип вагона z_{11}	Стоимость проезда y , руб.	Станция назначения	Расстояние x_1 , км	Тип вагона z_{11}	Стоимость проезда y , руб.
Платформа 39 км	39	0	45	Платформа 152 км	152	0	144
Мга	42	0	54	М. Вишера	162	0	153
Платформа 44 км	44	0	54	Мга	42	1	98,1
Платформа 47 км	47	0	54	Волхов	114	1	124,9
Платформа 63 км	63	0	63	Чудово	118	1	155,2
Платформа 67 км	67	0	72	Приозерск	141	1	137,4
Платформа 69 км	69	0	72	Луга	147	1	137,4
Платформа 78 км	78	0	81	М. Вишера	162	1	145,5

Станция назначения	Расстояние x_1 , на z_{11} км	Тип вагона z_{11}	Стоимость проезда y , руб.	Станция назначения	Расстояние x_1 , км	Тип вагона z_{11}	Стоимость проезда y , руб.
Волхов	114	0	117	Новгород	192	1	198,5
Чудово	118	0	117	Тихвин	192	1	157,6
Платформа 138 км	138	0	135	Пикалево	230	1	169,7
Будогощь	140	0	135	Лодейное поле	235	1	169,7
Приозерск	141	0	135	Подпорожье	272	1	189,8
Луга	147	0	135	Псков	284	1	189,8
Платформа 148 км	148	0	144	Бологое	319	1	227,9

После применения МНК получим следующее уравнение регрессии:

$$\hat{y} = 13,92 + 0,86x_1 + 72,10z_{11} - 0,46x_1z_{11}.$$

$$(t_{\text{факт.}}) \quad (2,32) \quad (15,76) \quad (7,54) \quad (-6,99)$$

$$R^2 = 0,96, \quad n = 30$$

Табличное значение t -критерия равно 2,06 при $df = 30 - 3 - 1 = 26$ степенях свободы и уровне значимости $\alpha = 0,05$, следовательно, все параметры регрессии значимы и их можно проинтерпретировать.

Коэффициент 0,86 при факторе x_1 означает, что если ехать на электричке ($z_{11} = 0$), то с изменением расстояния на один километр стоимость проезда изменяется в среднем на 0,86 руб. Если воспользоваться поездом дальнего следования ($z_{11} = 1$), то средняя стоимость одного километра пути будет ниже, чем в электричке ($d_{111} = -0,46$) на 0,46 руб. Иначе говоря, средняя стоимость одного километра пути на поезде дальнего следования составит $(0,86 - 0,46) = 0,40$ руб.

Параметр c_{11} интерпретируется иначе, чем в модели с фиктивной переменной сдвига (4.2). Так как прямые (4.15) и (4.16) непараллельны, параметр c_{11} может рассматриваться только как разница между значениями результата для $z_{11} = 1$ и $z_{11} = 0$ при условии, что количественный фактор x_1 равен нулю.

Для нашего примера можно отметить: так как $c_{11} \neq 0$ и $d_{111} \neq 0$, на коротких расстояниях передвижение на поезде дальнего следования обходится дороже, чем на электричке. С увеличением расстояния более выгодно становится пользоваться поездом дальнего следования (стоимость проезда за километр в нем ниже). Точка пересечения прямых (4.15) и (4.16) имеет координаты, которые можно проинтерпретировать: это расстояние (x_1), для которого стоимость проезда (y) на электричке и поезде дальнего следования равны

$$13,92 + 0,86x_1 = 86,02 + 0,40x_1$$

$$72,10 = 0,46x_1$$

$$x_1 \approx 156,74 \text{ км.}$$

$$y = 13,92 + 0,86 \cdot 156,76 \approx 148,72 \text{ или}$$

$$y = 86,02 + 0,40 \cdot 156,76 \approx 148,72.$$

Стоимость проезда на расстояние 156,74 км одинакова как для электричек, так и для поездов дальнего следования и составляет примерно 156 руб. 74 коп. При расстоянии, меньшем 156,72 км, выгоднее пользоваться электричкой, большем указанной величины — поездом дальнего следования.

При достаточном числе наблюдений (не менее 7 на каждый параметр регрессии, кроме свободного члена) модель (4.17) может быть использована на первом шаге при изучении влияния фиктивной переменной на результат. Незначимость отдельных параметров этой модели свидетельствует о необходимости перехода к моделям (4.2), (4.13) или к модели без фиктивной переменной.

Модели (4.2), (4.13), (4.17) были построены для простейшего случая с двумя факторами: одной фиктивной переменной и одной количественной переменной. Запишем каждую из этих моделей в общей форме:

1. Модель с фиктивными переменными сдвига:

$$\hat{y} = a + b_1 x_1 + \dots + b_p x_p + c_{11} z_{11} + c_{12} z_{12} + \dots + \\ + c_{21} z_{21} + c_{22} z_{22} + \dots + c_{j1} z_{j1} + c_{j2} z_{j2} + \dots \quad (4.18)$$

где a — свободный член;

b_1, \dots, b_p — коэффициенты при количественных факторах x_1, \dots, x_p ;

c_{11}, c_{12}, \dots — коэффициенты при фиктивных переменных z_{11}, z_{12}, \dots , созданных для первого неколичественного фактора;

c_{21}, c_{22}, \dots — коэффициенты при фиктивных переменных z_{21}, z_{22}, \dots , созданных для второго неколичественного фактора;

c_{j1}, c_{j2}, \dots — коэффициенты при фиктивных переменных z_{j1}, z_{j2}, \dots , созданных для j -го неколичественного фактора.

2. Модель с фиктивными переменными наклона:

$$\hat{y} = a + b_1 x_1 + \dots + b_p x_p + \\ + d_{111} x_1 z_{11} + d_{112} x_1 z_{12} + \dots + d_{1j1} x_1 z_{j1} + d_{1j2} x_1 z_{j2} + \dots + \\ + d_{211} x_2 z_{11} + d_{212} x_2 z_{12} + \dots + d_{2j1} x_2 z_{j1} + d_{2j2} x_2 z_{j2} + \dots + \quad (4.19) \\ + \dots + \\ + d_{pj1} x_p z_{11} + d_{pj2} x_p z_{12} + \dots + d_{pj1} x_p z_{j1} + d_{pj2} x_p z_{j2} + \dots$$

Для удобства чтения уравнение (4.19) записано в несколько строк. В первой приведено обычное уравнение множественной регрессии с количественными факторами $x_1 \dots x_p$. В другие строки начиная со вто-

рой включены (с параметрами d) попарные произведения количественных факторов на фиктивные переменные.

3. Общая форма модели с фиктивными переменными объединяет модели (4.18) и (4.19):

$$\begin{aligned}\hat{y} = & a + b_1 x_1 + \dots + b_p x_p + \\ & + c_{11} z_{11} + c_{12} z_{12} + \dots + c_{21} z_{21} + c_{22} z_{22} + \dots + c_{j1} z_{j1} + c_{j2} z_{j2} + \dots + \\ & + d_{111} x_1 z_{11} + d_{112} x_1 z_{12} + \dots + d_{1j1} x_1 z_{j1} + d_{1j2} x_1 z_{j2} + \dots + \\ & + d_{211} x_2 z_{11} + d_{212} x_2 z_{12} + \dots + d_{2j1} x_2 z_{j1} + d_{2j2} x_2 z_{j2} + \dots + \\ & + \dots + \\ & + d_{p11} x_p z_{11} + d_{p12} x_p z_{12} + \dots + d_{pj1} x_p z_{j1} + d_{pj2} x_p z_{j2} + \dots\end{aligned}. \quad (4.20)$$

Частным случаем модели с фиктивными переменными является модель, не содержащая количественных факторов. Рассмотрим этот вид моделей на примере модели с одной фиктивной переменной (z_{11}). Очевидно, что здесь не может быть фиктивной переменной наклона, так как отсутствует количественный фактор. Поэтому модель может быть единственного вида — с фиктивной переменной сдвига:

$$\hat{y} = a + c_{11} z_{11}. \quad (4.21)$$

Так как фиктивная переменная принимает всего два значения, может быть рассчитано всего два выровненных значения результата: при $z_{11} = 0$ и при $z_{11} = 1$. Это будут соответственно величины a и $(a + c_{11})$. Эти величины можно рассматривать как средние значения результата по группам, образованным двумя значениями фиктивной переменной.

Существенным недостатком моделей с фиктивными переменными является значительное число параметров, требующее соответствующего увеличения числа наблюдений. Например, если неколичественный фактор имеет три значения, то для линейной модели с фиктивными переменными сдвига требуются два дополнительных слагаемых. Если же рассматриваются модели с несколькими количественными факторами и фиктивными персмнными наклона, то число параметров еще больше возрастает. Как правило, объем наблюдений, находящихся в распоряжении исследователя, весьма незначителен и не может обеспечить статистическую значимость столь «длинной» модели.

Одним из выходов из этого положения является объединение нескольких значений неколичественной переменной одним, если эти значения не различаются по силе влияния на результат. Например, местоположение квартиры определяется ее адресом: городом, улицей, номером дома и номером квартиры. Это индивидуальная характеристика рассматриваемого объекта. Она может быть отражена в неколичественной переменной «местоположение», имеющей столько значений, сколько адресов. Такие индивидуальные характеристики можно укрупнить, объединив квартиры по принадлежности к номеру дома, еще крупнее — к номеру квартала, к району города, или, наконец, раз-

делить все квартиры на те, которые расположены в центре города и те, которые находятся вне его.

Модели с фиктивными переменными могут применяться для анализа временных рядов, в частности для изучения периодических колебаний. Если рассматриваются сезонные колебания продаж (например, меховой одежды), можно построить модель зависимости объема продаж от квартала, месяца и т. п. Чем шире будет интервал времени, рассматриваемый в качестве фактора, тем короче будет модель, что повышает ее статистическую надежность. Однако слишком широкие интервалы ведут к значительному усреднению характеристик влияния неколичественной переменной, что понижает аналитическую ценность модели.

4.6. Исследование структурных изменений с помощью теста Чоу

Как было показано выше, при разных значениях фиктивной переменной получаются разные уравнения регрессии. Целесообразность применения двух уравнений регрессии вместо одного можно оценить, не прибегая к вводу фиктивных переменных. Для этого используется тест Чоу.

Пусть имеется n наблюдений, позволяющих охарактеризовать зависимость результата от одного или нескольких количественных факторов. У исследователя есть основания предполагать, что совокупность неоднородна с точки зрения числовых характеристик воздействия факторов на результат. Предполагается также, что однородность может быть достигнута, если совокупность разбить по определенному критерию на две части.

Для проверки этих предположений находят параметры трех уравнений регрессии. Первое уравнение строится для всей совокупности наблюдений, второе и третье — для соответствующих выделенных подмножеств совокупности наблюдений. Для каждого из этих уравнений находят остаточную сумму квадратов $SS_{\text{ост}}$:

$$SS_{\text{ост}} = \sum (y - \hat{y})^2 = \sum \epsilon^2.$$

Обозначим остаточную сумму квадратов для общего уравнения регрессии через SS_0 , для уравнений регрессий по подмножествам наблюдений — через SS_1 и SS_2 .

Тогда равенство выполняется, если параметры всех трех функций совпадают.

В противном случае $SS_0 = SS_1 + SS_2$. Чем больше разница между двумя частями этого неравенства, тем больше различия между двумя подмножествами с точки зрения параметров уравнений регрессии. Существенность различий проверяют с помощью F -критерия.

Фактическое значение F -критерия находят по формуле

$$F = \frac{(SS_0 - (SS_1 + SS_2)) \cdot (n - m_1 - m_2 - 2)}{(SS_1 + SS_2) \cdot (m_1 + m_2 + 1)},$$

где m_1 и m_2 — количество параметров (без свободного члена) в уравнениях, построенных по подмножествам, n — число наблюдений по всей совокупности.

Табличное значение F -критерия находят для степеней свободы $df_1 = m_1 + m_2 + 1$ и $df_2 = n - m_1 - m_2 - 2$. Если фактическое значение окажется больше табличного, то имеют место структурные сдвиги и целесообразно строить уравнение регрессии с соответствующей фиктивной переменной.

Применение теста Чоу рассмотрим на примере из п. 4.5, в котором исследовалась зависимость стоимости проезда от расстояния и типа поезда.

Нам необходимо найти параметры уравнения

$$\hat{y} = a + bx_1$$

для массивов данных:

- для всех данных ($n = 30$);
- для данных о стоимости проезда в электричках ($n = 17$);
- для данных о стоимости проезда в поездах дальнего следования ($n = 13$).

Применяя МНК к данным табл. 4.3, получим следующие уравнения регрессии и значения сумм квадратов остатков:

- по всем типам поездов:

$$\hat{y} = 44,32 + 0,60x_1; \quad SS_0 = 8285,58;$$

- для проезда в электричках:

$$\hat{y} = 13,92 + 0,86x_1; \quad SS_1 = 124,80;$$

- для проезда в поездах дальнего следования:

$$\hat{y} = 86,03 + 0,40x_1; \quad SS_2 = 2459,26.$$

Фактическое значение F -критерия равно

$$F = \frac{8285,58 - (124,80 + 2459,26)}{(124,80 + 2459,26)} \cdot \frac{30 - 1 - 1 - 2}{1+1+1} \approx 19,12.$$

Табличное значение F -критерия равно 2,98 (при $\alpha = 0,05$ и $df_1 = 1+1+1 = 3$ и $df_2 = 30 - 1 - 1 - 2 = 26$ степенях свободы).

Так как фактическое значение F -критерия больше табличного, следует признать существенность различия характеристик зависимости стоимости проезда от расстояния для разных типов поездов. Следовательно, для каждого типа поезда следует строить свое уравнение регрессии или объединить их в одно, используя фиктивную переменную. Это и было сделано в п. 4.5.

Тест Чоу можно использовать также, если исследуется зависимость уровня ряда от времени, т. е. строится уравнение тренда вида $\hat{y} = f(t)$. В этом случае деление совокупности наблюдений на две части производится относительно определенного момента времени, в который, по мнению исследователя, произошли какие-либо структурные изменения.

Контрольные вопросы

1. Можно ли учесть в уравнении регрессии неколичественные факторы? Каким образом?
2. Дайте определение фиктивной переменной.
3. Сколько фиктивных переменных нужно ввести, если имеются два неколичественных фактора, причем один из них имеет три возможных значения, а другой — два?
4. Как интерпретируется коэффициент регрессии при фиктивной переменной сдвига?
5. Как интерпретируется коэффициент регрессии при фиктивной переменной наклона?
6. Каков общий вид модели регрессии с одной количественной и одной фиктивной переменной?
7. В какой модели (моделях) с фиктивными переменными учитывается различие влияния количественных факторов на результат при разных значениях неколичественного фактора?
8. Назовите достоинства и недостатки моделей с фиктивными переменными.
9. Пусть имеется уравнение регрессии с одним количественным и одним неколичественным фактором, выраженным тремя фиктивными переменными. Сколько возможных значений у неколичественного фактора? Как на основе заданного уравнения регрессии найти уравнения парной регрессии, содержащие только количественный фактор? Сколько будет таких уравнений и почему?
10. Какова область применения теста Чоу?
11. Какие показатели сравниваются между собой по тесту Чоу? Какой статистический критерий в этом случае используется?
12. Опишите методику применения теста Чоу.

ГЛАВА 5. МОДЕЛИ ДИСКРЕТНОГО ВЫБОРА

На практике часто встречаются задачи, в которых результативная переменная принимает дискретные значения: например, нужно оценить риск бедности или вероятность того или иного претендента быть избранным в президенты. Использование линейных регрессионных моделей в данном случае некорректно. В данной главе рассматриваются модели, которые могут быть использованы для моделирования случаев, где зависимая переменная принимает дискретные значения.

5.1. Модели бинарного выбора. Использование линейной регрессии

Предположим, исследователю нужно понять, какие факторы влияют на бедность*. Естественно, что бедность объясняется многими факторами, такими как наличие работы, пол, возраст, уровень образования, место проживания, состояние здоровья и др. Пусть будет k объясняющих переменных. Таким образом, набор объясняющих переменных можно представить вектором $\mathbf{X} = (x_1, \dots, x_k)'$. Пусть имеются данные по выборке n домохозяйств $i = 1, \dots, n$. Тогда бедность можно представить следующим образом:

$$y_i = 1, \text{ если } i\text{-й индивид (домохозяйство) признан бедным}; \\ y_i = 0, \text{ если } i\text{-й индивид (домохозяйство) не является бедным}.$$

Попробуем использовать регрессионную модель для того, чтобы объяснить y_i через векторы $(x_{i0}, x_{i1}, \dots, x_{ik})'$, где $x_{i0} \equiv 1$ — свободный член. Линейная модель выглядит следующим образом:

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i = x_i'\beta + \epsilon_i, \quad (5.1)$$

где $x_i = (x_{i0}, x_{i1}, \dots, x_{ik})'$. В рамках классической линейной регрессионной модели (*classical linear regression model*) так как y_i принимает значение 1 и 0, имеем

$$E(\epsilon_i | x_i) = (1 - x_i'\beta)P(y_i = 1 | x_i) + (-x_i'\beta)P(y_i = 0 | x_i) = \\ = (1 - x_i'\beta)P(y_i = 1 | x_i) - x_i'\beta(1 - P(y_i = 1 | x_i)) = 0. \quad (5.2)$$

Решая уравнение (5.2) относительно $P(y_i = 1 | x)$, получим

$$P(y_i = 1 | x_i) = x_i'\beta, \\ P(y_i = 0 | x_i) = 1 - x_i'\beta. \quad (5.3)$$

Таким образом, линейная модель подразумевает то, что $x_i'\beta$ является вероятностью, следовательно, принимает значения в интервале $[0, 1]$. На практике это условие чаще нарушается, чем выполняется, что является серьезным недостатком модели. Еще одним недостатком модели является то, что ошибки ϵ_i не распределены по нормальному за-

* В данном случае имеется в виду бедность по абсолютному критерию, когда бедными считаются индивидуумы или домохозяйства, собственных ресурсов которых недостаточно для удовлетворения базовых потребностей.

кону и гетероскедастичны. Дисперсия ошибки может быть представлена следующим образом:

$$\begin{aligned} V(\varepsilon_i | x_i) &= E(\varepsilon_i^2 | x_i) = (1 - x_i \beta)^2 P(y_i = 1 | x_i) + (-x_i \beta)^2 [1 - P(y_i = 1 | x_i)] \\ &= (1 - x_i \beta)^2 x_i \beta + (x_i \beta)^2 [1 - P(y_i = 1 | x_i)] = x_i \beta (1 - x_i \beta). \end{aligned} \quad (5.4)$$

Это означает, что дисперсия ошибки непостоянна и зависит от объясняющих переменных. Наблюдения, для которых значение $P(y_i = 1 | x_i)$ близко к 0 или 1, будут иметь относительно небольшую дисперсию, тогда как наблюдения, для которых значение $P(y_i = 1 | x_i)$ близко к $\frac{1}{2}$, будут иметь большую дисперсию. Заметим, что дисперсия ошибки также зависит от параметров модели β .

Введение в модели бинарного выбора

Чтобы избежать проблем, связанных с линейной моделью, был создан класс моделей *бинарного выбора* (так называемые *одномерные дихотомические модели**). Данные модели описывают в основном прямую вероятность того, что $y_i = 1$, хотя они часто выводятся из предполагаемой гипотезы о существовании латентной переменной (смотри ниже). В общем, мы имеем:

$$P(y_i = 1 | x_i) = G(x_i, \beta) \quad (5.5)$$

для некоторой функции $G(\cdot)$. Это уравнение показывает значение вероятности того, что $y_i = 1$ в зависимости от вектора x_i , содержащего индивидуальные характеристики. Ясно, что функция $G(\cdot)$ должна принимать значения в интервале $[0, 1]$.

Одним из наиболее легких путей достижения такого результата является небольшое преобразование уравнения (5.1), называемого *моделью линейной вероятности*, со следующей функцией распределения:

$$\begin{aligned} F(w) &= 0, w < 0; \\ F(w) &= w, 0 \leq w \leq 1; \\ F(w) &= 1, w > 1. \end{aligned} \quad (5.6)$$

Если вычисленные значения уравнения регрессии получаются больше 1 или меньше 0, они приравниваются к 1 или 0 соответственно. Это не очень хорошо, так как мы можем предсказать событие с вероятностью 1 или 0, тогда как всегда есть вероятность того, что событие может не произойти или произойти, соответственно. В то время как процедура оценивания может давать несмещенные оценки, предсказанные значения всегда смещены. В силу перечисленных свойств данная модель на практике используется крайне редко.

*Следует отметить отличие данного сорта моделей от многомерных (*multivariate*) построений, которые описывают совокупность нескольких одновременных бинарных выборов.

Один из способов избежать этой проблемы состоит в использовании в качестве функции $G(\cdot)$ в уравнении (5.5) одной из функций распределения, так как они наиболее соответствуют самой идее вероятности и получаемые значения лежат в интервале $[0, 1]$. Часто используют стандартную функцию нормального распределения:

$$F(w) = \Phi(w) = \int_{-\infty}^w \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt, \quad (5.7)$$

что отражается в названии модели — *probit-модель (probit model)*, и стандартную логистическую функцию распределения, задаваемую уравнением

$$F(w) = L(w) = \frac{e^w}{1+e^w}, \quad (5.8)$$

что привело к образованию *logit-модели (logit model)*.

Эти две последние модели получили широкое применение на практике. Стандартное нормальное и стандартное логистическое распределения имеют нулевое математическое ожидание, тогда как дисперсии равны 1 и $\pi^2/3$ соответственно. Эти два распределения очень похожи, и соответствующие модели выдают похожие результаты.

Величина коэффициентов в бинарных моделях имеет иную интерпретацию, нежели величина коэффициентов в обычных линейных регрессиях. Один из возможных путей состоит в переходе к предельным эффектам коэффициентов, которые можно проинтерпретировать также, как и обычные коэффициенты. Предельные эффекты позволяют также сравнивать и параметры модели между собой. Предельные эффекты вычисляются как частные производные вероятности того, что $y_i = 1$, по непрерывной независимой переменной x_{ik} , например. Для указанных трех моделей получим

$$\begin{aligned} \frac{\partial x_i \beta}{\partial x_{ik}} &= \beta_k; \text{ (или } 0); \\ \frac{\partial \Phi(x_i \beta)}{\partial x_{ik}} &= \Phi(x_i \beta) \beta_k; \\ \frac{\partial L(x_i \beta)}{\partial x_{ik}} &= \frac{e^{x_i \beta}}{(1+e^{x_i \beta})^2} \beta_k = \Lambda(x_i \beta) [1 - \Lambda(x_i \beta)] \beta, \end{aligned}$$

где $\Phi(\cdot)$ — плотность стандартного нормального распределения; $\Lambda(\cdot)$ — стандартная функция логистического распределения. В двух последних случаях эффект изменения x_{ik} зависит от величины x_i . Во всех случаях знак перед коэффициентом сохраняется в предельном эффекте. Для дискретных независимых переменных, например для фиктивной переменной, предельный эффект вычисляется иным образом: определяется разность вероятностей наступления события $y_i = 1$ для двух случаев, когда фиктивная переменная принимает значение 1 и 0 при фиксированном значении остальных переменных.

Предположение о существовании латентной переменной

Модель бинарного выбора может быть выведена исходя из предположения о существовании одной ненаблюдаемой *латентной переменной* y_i^* :

$$y_i^* = x'_i \beta + \varepsilon_i. \quad (5.9)$$

Если остатки ε_i являются независимыми и одинаково распределенными по стандартному закону нормального распределения с единичной дисперсией, то получим *probit*-модель. Стандартное логистическое распределение остатков с дисперсией $\frac{\pi^2}{3}$ приводит к *logit*-модели.

На практике мы наблюдаем следующие величины:

$$\begin{aligned} y_i &= 1, \text{ если } y_i^* > 0, \\ y_i &= 0, \text{ если } y_i^* \leq 0. \end{aligned}$$

В данной конструкции следует обратить внимание на две вещи. Во-первых, на предположение об известной дисперсии ε_i . Рассмотрим регрессию $y_i^* = x'_i \beta + \sigma \varepsilon_i$. Преобразуем данное уравнение: $\frac{y_i^*}{\sigma} = x'_i \left(\frac{\beta}{\sigma} \right) + \varepsilon_i$. Полученная модель по сути та же самая, что и модель (5.9) с теми же данными. Наблюдаемые значения зависимой переменной те же: 0 и 1. Это означает, что мы ничего не знаем о дисперсии σ в данных, следовательно, она не может быть оценена.

Во-вторых, на предположение о нуле как пороге отнесения к той или иной категории. Пусть a — ненулевой порог, а α — неизвестный свободный член и в данном случае x и β не включают в себя константу. Тогда

$$P(y_i = 1) = P(y_i^* > a) = P(\alpha + x'_i \beta + \varepsilon_i > a) = P[(\alpha - a) + x'_i \beta + \varepsilon_i > 0]. \quad (5.10)$$

Так как параметр α неизвестен, разность $(\alpha - a)$ тоже остается неизвестной.

С двумя нормализациями вероятность того, что $y_i = 1$, есть вероятность того, что латентная переменная y_i^* больше 0 может быть выражена как

$$P(y_i = 1) = P(y_i^* > 0) = P(x'_i \beta + \varepsilon_i > 0) = P(-\varepsilon_i < x'_i \beta) = F(x'_i \beta), \quad (5.11)$$

где F — функция распределения $-\varepsilon_i$, или в случае симметричного распределения — функция распределения ε_i .

Оценивание параметров моделей бинарного выбора

Для оценивания параметров β модели (5.5) обычно используют метод максимального правдоподобия. В предположении о том, что наблюдения y_i независимы, можно записать общее уравнение правдоподобия:

$$L(\beta) = \prod_{i=1}^N P(y_i = 1 | x_i; \beta)^{y_i} P(y_i = 0 | x_i; \beta)^{1-y_i} \quad (5.12)$$

где β включено для того, чтобы подчеркнуть, что данная функция правдоподобия — это функция от β . Обычно удобно работать с логарифмом функции правдоподобия. Подставив $P(y_i = 1 | x_i; \beta) = F(x_i; \beta)$, получим

$$\log L(\beta) = \sum_{i=1}^N y_i \log F(x_i; \beta) + \sum_{i=1}^N (1 - y_i) \log(1 - F(x_i; \beta)). \quad (5.13)$$

Дифференцируя равенство (5.13) по β , получим:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^N \left[\frac{y_i - F(x_i; \beta)}{F(x_i; \beta)(1 - F(x_i; \beta))} f(x_i; \beta) \right] x_i = 0, \quad (5.14)$$

где $f = F'$ — плотность распределения.

Для *logit*-модели уравнение (5.14) упрощается:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^N \left[y_i - \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right] x_i = 0. \quad (5.15)$$

Отсюда мы можем найти вероятность того, что $y_i = 1$:

$$\hat{p}_i = \frac{\exp(x_i \hat{\beta})}{1 + \exp(x_i \hat{\beta})}. \quad (5.16)$$

Уравнение правдоподобия (5.14) является системой нелинейных (относительно β) уравнений и решается обычно итерационными методами. Можно показать*, что для *probit*- и *logit*-моделей данная функция является вогнутой по β , следовательно, решение уравнения (5.14) дает оценку максимального правдоподобия параметров β **.

Качество модели

Для моделей дискретного выбора не существует унифицированного показателя качества «подгонки» модели, такого, как R^2 для моделей линейных регрессий. Однако существует несколько показателей, которые наиболее распространены на практике.

Часто процесс проверки качества «подгонки» модели базируется на сравнении аналогичной модели, содержащей только константу в качестве объясняющих переменных. Обозначим через $\log L_f$ значение функции правдоподобия исходной модели, через $\log L_c$ обозначим значение функции правдоподобия той же модели с нулевыми параметрами, но с константой. Естественно, что $\log L_f \geq \log L_c$. Чем больше их разность, тем лучше должна быть модель. На этой идее основаны показатели качества модели *Epseudo R²* и *McFadden R²*:

$$Epseudo R^2 = 1 - \frac{1}{1 + 2(\log L_f - \log L_c) / n}; \quad (5.17)$$

$$McFadden R^2 = 1 - \frac{\log L_f}{\log L_c}, \quad (5.18)$$

* См., например: Greene William H. Econometric Analysis. Ch. 21, 2004.

** Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс. 6-е изд., М., 2004.

где n — количество наблюдений. Логарифм правдоподобия является суммой вероятностей, следовательно, справедливо соотношение $\log L_c \leq \log L_f < 0$, что приводит к тому, что значения этих двух показателей ограничены нулем и единицей. Чем больше значение этих показателей, тем лучше модель. Следует отметить, что данные показатели редко достигают значений, превышающих 0,5.

Альтернативным методом оценки качества модели является вычисление процента угаданных значений. Для этого обычно составляют таблицу попаданий и промахов (табл. 5.1). При этом важную роль играет пороговое значение, определяющее, имеет ли наблюдение предсказанный положительный (при превышении порогового значения) или отрицательный (в противном случае) исход. Использование порогового значения 0,5 не всегда является самым лучшим. Например, в случае сравнительной несбалансированности выборки (значительное преобладание нулей над единицами или наоборот) указанное правило может не позволить предсказать значение, равное единице (или нулю). Поэтому в таких случаях лучше использовать пороговое значение, отражающее долю положительных исходов.

Таблица 5.1

Перекрестная таблица фактических и предсказанных исходов

		Фактические		(\hat{y}_i)	Всего	
		Предсказанные				
y_i	0	n_{00}	n_{01}	n_0		
	1	n_{10}	n_{11}	n_1		
	Всего	n_0	n_1	n		

На основе табл. 5.1 можно вычислить процент правильно угаданных исходов как

$$w = \frac{n_{00} + n_{11}}{n} \cdot 100\%.$$

Интуитивно понятно, что чем ближе значение w к 100%, тем лучше предсказательная сила модели. На практике значение w больше 70% является нормой и модель считается обладающей хорошей предсказательной силой. При значениях меньше 50% предсказательную силу модели считают слабой.

Проверка гипотез

Для данных моделей проверка гипотез о наличии ограничений на коэффициенты может проводиться с помощью обычных тестов для линейных регрессий: *тест Вальда (Wald test)*, *тест множителей Лагранжа (LM test)* или *отношения правдоподобия (LR test)*. В качестве аналога *F*-теста в линейной регрессии о совместной незначимости всех коэффициентов в бинарных моделях используют *LR*-тест:

$$LR = -2(\log \hat{L}_c - \log \hat{L}_f) \sim \chi^2(k),$$

где k — количество ограничений (в данном случае это количество независимых переменных в модели).

Пример. Выбор карьеры молодыми людьми в 16 лет в Великобритании

В Великобритании выбор карьеры совершается по достижении 16 лет. В этом возрасте все подростки сдают национальные экзамены. Через несколько месяцев после сдачи экзаменов они решают, продолжать обучение в школе или нет. Если они решают оставить школу, то они могут пойти работать на полный рабочий день или совмещать работу с учебой.

Исследуем факторы, влияющие на данное решение, используя данные обследования UK National Child and Development Survey. Эти данные относятся к индивидуумам, рожденным в Великобритании в марте 1958 г. ($n = 3423$). Описание переменных:

SCHOOL — решение, принятное в 16 лет: 1 — оставаться в школе; 0 — в противном случае;

ABLE7 — общие способности индивидуума, измеренные в 7 лет;

LOGINC — логарифм семейного дохода (в 16 лет);

CTRATIO — количество учеников на одного учителя в школе (индикатор качества школы);

OLDSIB — количество старших братьев и сестер (в 16 лет);

YNGSIB — количество младших братьев и сестер (в 16 лет);

ETOT — средняя успеваемость в 16 лет (до принятия решения);

FEMALE — 1 — для девочек, 0 — для мальчиков.

Чтобы узнать, какие характеристики влияют на решение индивидуума оставаться в школе или не оставаться, были оценены бинарные *probit*- и *logit*-модели, где зависимая переменная SCHOOL принимает значение 1, если подросток решает продолжить обучение в школе, 0 — в противном случае. Результаты представлены в табл. 5.2.

Таблица 5.2
Оценки моделей бинарного выбора

Переменная	<i>Logit</i>		<i>Probit</i>	
	оценка	станд. откл.	оценка	станд. откл.
Константа	-3,795	(0,684)	-2,061	(0,384)
<i>ABLE7</i>	0,041	(0,003)	0,023	(0,002)
<i>LOGINC</i>	0,676	(0,117)	0,397	(0,066)
<i>CTRATIO</i>	-0,243	(0,024)	-0,142	(0,014)
<i>OLDSIB</i>	-0,294	(0,074)	-0,178	(0,043)
<i>YNGSIB</i>	-0,063	(0,038)	-0,040	(0,022)
<i>ETOT</i>	0,222	(0,017)	0,124	(0,009)
<i>FEMALE</i>	-0,059	(0,086)	-0,033	(0,050)
$\log L_f$	-1645,635		-1651,797	

Окончание табл. 5.2

Переменная	<i>Logit</i>		<i>Probit</i>	
	оценка	станд. откл.	оценка	станд. откл.
$\log L_c$	-2127,798		-2127,798	
<i>Epseudo R</i> ²	0,2198		0,2176	
<i>McFadden R</i> ²	0,2266		0,2237	

Переменные YNGSIB и FEMALE оказались статистически незначимыми на 5%-ном уровне значимости. Полученные знаки при коэффициентах согласуются со здравым смыслом. Например, положительный знак при коэффициенте для переменной *LOGINC* (логарифм семейного дохода) говорит о положительной зависимости между уровнем дохода и решением продолжить обучение в школе.

Модель выбирают исходя из максимума логарифмического правдоподобия ($\log L_f$). Для данного случая *logit*-модель подходит лучше. Ее будем анализировать.

Значение коэффициентов *logit*-модели непосредственно интерпретировать нельзя, так как их величины искажены функцией распределения. Для этого нужно вычислить предельные эффекты.

Вычислим, например, эффект от одного дополнительного младшего родственника. Будем считать, что переменная YNGSIB непрерывна (хотя это не совсем так). Тогда один из способов рассчитать предельный эффект:

$$\frac{\partial L(y|x;\beta)}{\partial x_{ik}} = \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2} \beta_k = \Lambda(x_i\beta)[1 - \Lambda(x_i\beta)]\beta = 0,237(-0,06) = 0,014,$$

т. е. с каждым последующим родственником вероятность пойти в школу уменьшается на 1,4%.

Для примера можно оценить вероятность остаться в школе для мальчика при средних значениях остальных переменных:

$$\begin{aligned} \Pr(SCHOOL = 1 | FEMALE = 0) &= \Lambda[-3,795 + 0,04 \cdot \overline{ABLE7} + \\ &+ 0,68 \cdot \overline{LOGINC} - 0,24 \cdot \overline{CTRATIO} - 0,29 \cdot \overline{OLDSIB} - \\ &- 0,06 \cdot \overline{YNGSIB} + 0,22 \cdot \overline{ETOT}] = \Lambda(-1,167768) = 0,237 \end{aligned}$$

Отсюда видно, что при средних значениях остальных переменных мальчик, скорее всего, учебу не продолжит (даже с учетом порогового значения, равного доле положительных исходов $0,31 = \frac{1072}{3423}$).

Рассмотрим предсказательную силу данной модели. Для этого составим перекрестную таблицу фактических и предсказанных исходов при пороговом значении, равном 0,31.

Таблица 5.3

Характеристики предсказательной силы модели

Фактические	Прогнозные (\hat{y}_i)		Всего
	0	1	
y_i	0	1598	753
	1	235	837
Всего	1833	1590	3423

Процент правильно угаданных исходов равен $w = \frac{1598 + 837}{3423} \times 100\% = 70,95\%$, что указывает на неплохую предсказательную силу модели. В идеале этот показатель, как отмечалось, стремится к 100%.

Ошибки спецификации модели

Модели бинарного выбора строятся, как было указано ранее, при некоторых предположениях: 1) остатки гомоскедастичны и 2) распределены по стандартному закону распределения в *probit*-модели и по стандартному логистическому закону распределения в случае *logit*-модели.

Что происходит при нарушении данных предположений? Например, в случае линейной регрессии нарушение предположения о гомоскедастичности остатков приводит к потере эффективности, но не к смещению и потере состоятельности оценок. В моделях бинарного выбора нарушение данного предположения приводит к потере эффективности (т. е. смещению вычисленных статистик), асимптотическому смещению оценок и потере состоятельности оценок*.

К смещению и несостоятельности оценок приводит и пропуск существенных переменных в модели бинарного выбора**. Напомним, что одним из признаков пропуска важных переменных является нарушение условия о нормальности остатков (в случае *probit*-модели). К сожалению, проверка того, являются ли остатки распределенными по логистическому закону распределения, является нетривиальной задачей, и на практике этим не занимаются.

5.2. Модели множественного выбора

Модели множественного выбора (*multinomial, multi-response models*) используются в тех случаях, когда имеется более чем две альтернативы. Такие модели получили распространение для описания вероятности выбора каждой из возможных альтернатив как функции от независимых переменных. Важной задачей является описание этих вероятно-

* Подробнее об этом см.: Greene W. H. Econometric Analysis; Verbeek M. A Guide to Modern Econometrics; Maddala G. S. Limited-dependent and Qualitative Variables in Econometrics.

** Подробнее о проблеме спецификации см. там же.

стей с использованием ограниченного числа неизвестных параметров логически состоятельным способом с учетом того, что вероятности изменяются в пределах от нуля до единицы и сумма вероятностей наступления всех событий должна быть равна единице.

Следует отметить отличие данного типа моделей от многомерных (*multivariate*) построений, которые описывают совокупность нескольких одновременных бинарных выборов.

Модели множественного выбора можно разделить на *модели с упорядоченными альтернативами* (*ordered response models*) и *модели с неупорядоченными альтернативами* (*unordered response models*). Модели с упорядоченными альтернативами применяются тогда, когда логически оправдано упорядочение альтернатив, т. е. подразумевается существование ненаблюданной латентной переменной, которая управляет выбором между альтернативами. Следовательно, результат будет чувствителен к способу упорядочения альтернатив. Модели же с неупорядоченными альтернативами нечувствительны к тому способу, которым будут перенумерованы альтернативы. Часто они основываются на предположении, что каждая альтернатива имеет случайный уровень полезности и индивидуум выбирает альтернативу, приносящую наибольшую полезность.

5.2.1. Модели множественного выбора с упорядоченными альтернативами

Рассмотрим выбор между M альтернативами, пронумерованными от 1 до M . Если существует логическое упорядочивание этих альтернатив, то может использоваться дискретная модель с упорядоченными альтернативами. Эта модель также основывается на предположении о существовании одной ненаблюданной латентной переменной y^* :

$$y^* i = x_i' \beta + \varepsilon_i; \\ y_i = j, \text{ если } \gamma_{j-1} < y^* i \leq \gamma_j \quad (5.19)$$

для неизвестных γ_j , таких, что $\gamma_0 = -\infty$, $\gamma_1 = 0$ и $\gamma_M = \infty$. Следовательно, вероятность выбора альтернативы под номером j есть вероятность того, что латентная переменная y^* находится между двумя границами — γ_{j-1} и γ_j . В предположении о том, что остатки ε_i являются независимыми и одинаково распределенными по стандартному нормальному закону, получаем *упорядоченную probit-модель* (*ordered probit model*). Логистическое распределение остатков дает *упорядоченную logit-модель* (*ordered logit model*). В случае $M = 2$ получаем модели бинарного выбора.

Рассмотрим в качестве примера случай трех альтернатив. Предположим, что существует логическое упорядочение альтернатив. Тогда разумно ввести индексную функцию y_i^* , связанную с наблюдаемой функцией y_i таким образом, что большие ее значения y_i^* соотносятся в среднем с большими значениями y_i . Будем считать, что функция y_i^* ли-

нейна по независимым переменным x_i . Тогда имеем следующую модель с упорядоченными альтернативами:

$$\begin{aligned} y_i^* &= x'_i \beta + \varepsilon_i \\ \text{если } y_i^* <= 0 &\Rightarrow y_i = 1, \\ \text{если } 0 < y_i^* <= \gamma &\Rightarrow y_i = 2, \\ \text{если } y_i^* > \gamma &\Rightarrow y_i = 3. \end{aligned} \quad (5.20)$$

В данном случае одна из границ приравнена к нулю, что фиксирует положение, при этом нужно также нормировать и шкалу y_i^* . Один из очевидных путей состоит в том, чтобы зафиксировать дисперсию остатков ε_i . Для *упорядоченной probit-модели* это означает, что остатки ε_i одинаково распределены по нормальному стандартному закону. Вероятности, лежащие в основе модели, могут быть получены как

$$\begin{aligned} P\{y_i = 1|x_i\} &= P\{y_i^* \leq 0|x_i\} = \Phi(-x'_i \beta), \\ P\{y_i = 3|x_i\} &= P\{y_i^* > \gamma|x_i\} = 1 - \Phi(\gamma - x'_i \beta) \end{aligned}$$

и

$$P\{y_i = 2|x_i\} = \Phi(\gamma - x'_i \beta) - \Phi(-x'_i \beta),$$

где пороговое значение γ является неизвестным параметром, который вычисляется совместно с вектором коэффициентов β . Параметры оценивают методом максимального правдоподобия. Интерпретация коэффициентов β производится с учетом модели с латентной переменной или в терминах предельных эффектов для соответствующих вероятностей, как в случае бинарных моделей. Предположим, что один из коэффициентов β (k -й, например) имеет положительный знак. Это означает, что латентная переменная y_i^* возрастает с увеличением x_{ik} . Следовательно, имеется вероятность того, что $y_i = 3$ будет расти, в то время как вероятность $y_i = 1$ будет падать. Поведение средних категорий остается неизвестным: соответствующая вероятность может как расти, так и падать.

Для случая M вариантов выбора вероятности получаем аналогичным способом:

$$\begin{aligned} P\{y_i = 0|x_i\} &= P\{y_i^* \leq 0|x_i\} = \Phi(-x'_i \beta) \\ P\{y_i = 1|x_i\} &= \Phi(\gamma_1 - x'_i \beta) - \Phi(-x'_i \beta) \\ P\{y_i = 2|x_i\} &= \Phi(\gamma_2 - x'_i \beta) - \Phi(\gamma_1 - x'_i \beta) \\ P\{y_i = M|x_i\} &= P\{y_i^* > \gamma_{M-1}|x_i\} = 1 - \Phi(\gamma_{M-1} - x'_i \beta) \end{aligned} \quad (5.21)$$

Для того чтобы все вероятности были положительными, необходимо выполнение условия $0 < \gamma_1 < \gamma_2 < \dots < \gamma_{M-1}$.

На рис. 5.1 представлены графически приведенные выше вероятности (случай пяти вариантов выбора).

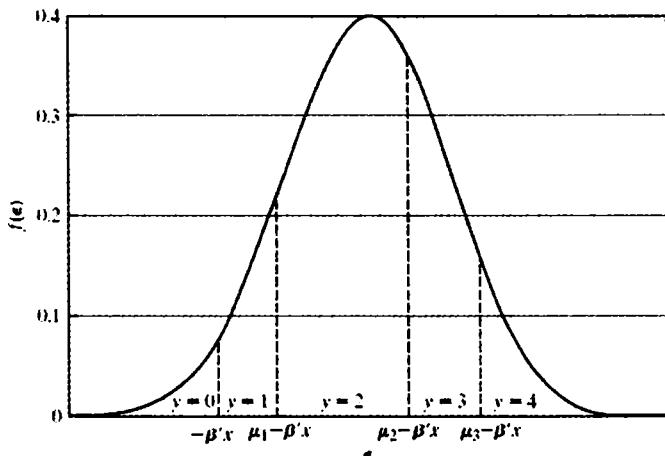


Рис. 5.1. Представление вероятностей в упорядоченной *probit*-модели.
(Источник: Green, 2004, Fig. 21.4)

Замечание о стандартизации

Для иллюстрации ситуации, когда необходимо провести нормализацию, воспользуемся моделью

$$\begin{aligned} y^*_i &= \beta_1 + x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim NID(0, \sigma^2); \\ y_i &= 1, \text{ если } y^*_i \leq \gamma_1, \\ &= 2, \text{ если } y^*_i < \gamma_1 \leq \gamma_2, \\ &= 3, \text{ если } y^*_i > \gamma_2, \end{aligned}$$

где константа не представлена в векторе x_i . Так как мы можем наблюдать, примет ли переменная y_i значения 1, 2 или 3, единственное, что вычисляют наши данные, — это вероятности этих трех событий при данных x_i . Эти значения вероятностей дает нам функция максимального правдоподобия. Покажем это на примере вероятности того, что $y_i = 1$ (при данном векторе x_i), которая имеет вид:

$$P\{y_i = 1|x_i\} = P\{\beta_1 + x_i^\top \beta + \varepsilon_i \leq \gamma_1|x_i\} = \Phi\left(\frac{\gamma_1 - \beta_1}{\sigma} - x_i^\top \beta \left(\frac{\beta}{\sigma}\right)\right).$$

Отсюда видно, что изменение параметров β , β_1 , σ и γ не приведет к получению различных значений вероятностей, пока отношения β/σ и $(\gamma_1 - \beta_1)/\sigma$ будут оставаться постоянными. Таким образом, возникает проблема идентификации: разные комбинации значений параметров ведут к одному и тому же значению функции правдоподобия, т.е. не существует единственного максимума. Чтобы обойти эту проблему, налагаются ограничения стандартизации (часто удобной является нормализация $\sigma = 1$ и $\beta_1 = 0$, или $\gamma_1 = 0$).

Оценивание осуществляется при помощи метода максимального правдоподобия, где перечисленные выше вероятности включены в функцию правдоподобия.

$$\log L(\beta, \gamma) = \sum_{i:y_i=0} \log(\Pr(y_i = 0|x_i, \beta, \gamma)) + \sum_{i:y_i=1} \log(\Pr(y_i = 1|x_i, \beta, \gamma)) + \dots + \sum_{i:y_i=M} \log(\Pr(y_i = M|x_i, \beta, \gamma)).$$

Еще раз отметим необходимость оценивания пороговых значений совместно с оцениванием параметров модели.

Интерпретация коэффициентов

Интерпретация коэффициентов проходит в терминах модели с подразумеваемой латентной переменной или в терминах влияния на соответствующие вероятности, как в моделях бинарного выбора. При этом для тестирования значимости индивидуальных коэффициентов применимы стандартные нормальные тесты.

Марно Вербик* отмечает, что интерпретация коэффициентов данной модели является не столь очевидной, как это может показаться. Грин** говорит о том, что «многие авторы попросту указывают коэффициенты и t -отношения, иногда с некоторыми комментариями по поводу значимости эффектов, но достаточно редко делают предположения о том, каковы по силе и направлению будут данные эффекты».

Предположим, что в рассматриваемой модели k -й коэффициент β_k является положительным. Это означает, что латентная переменная y^* возрастает, если значение x_{ik} растет. Соответственны вероятности того, что $y_i = 3$, будет возрастать, и одновременно вероятность того, что $y_i = 1$, будет падать. Воздействие на промежуточные категории неоднозначно: имеется вероятность того, что $y_i = 2$ может как возрастать, так и падать.

Пределные эффекты

Для сравнения влияния независимых переменных x_i на y используются предельные эффекты.

Пределные эффекты вычисляются по следующим формулам:

$$\begin{aligned} \frac{\partial \Pr\{y_i = 0|x_i\}}{\partial x_i} &= -\phi(x'_i \beta) \beta \\ \frac{\partial \Pr\{y_i = 1|x_i\}}{\partial x_i} &= [\phi(-x'_i \beta) - \phi(\mu - x'_i \beta)] \beta. \\ \frac{\partial \Pr\{y_i = 2|x_i\}}{\partial x_i} &= \phi(\mu - x'_i \beta) \beta \end{aligned} \quad (5.22)$$

* См.: Verbeek M. A Guide to Modern Econometrics. 2nd ed. John Wiley and Sons, Ltd, 2004. P. 204.

** Green (2003). P. 741.

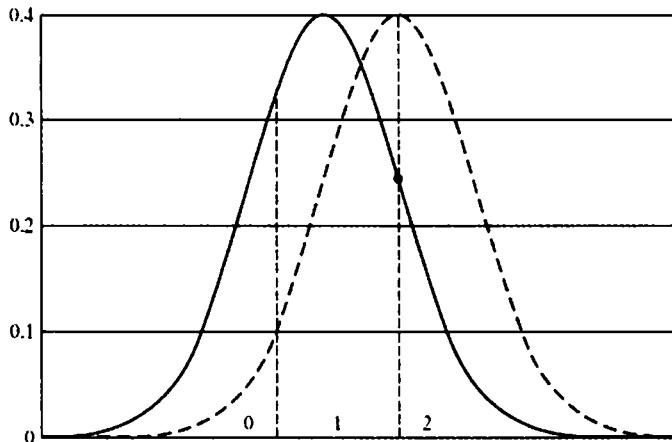


Рис. 5.2. Эффект изменения одного из объясняющих факторов на предсказанные вероятности.
(Источник: Greene, 2003, fig. 21.5)

На рис. 5.2 показана ситуация изменения значения одного из факторов, влияющих на выбор.

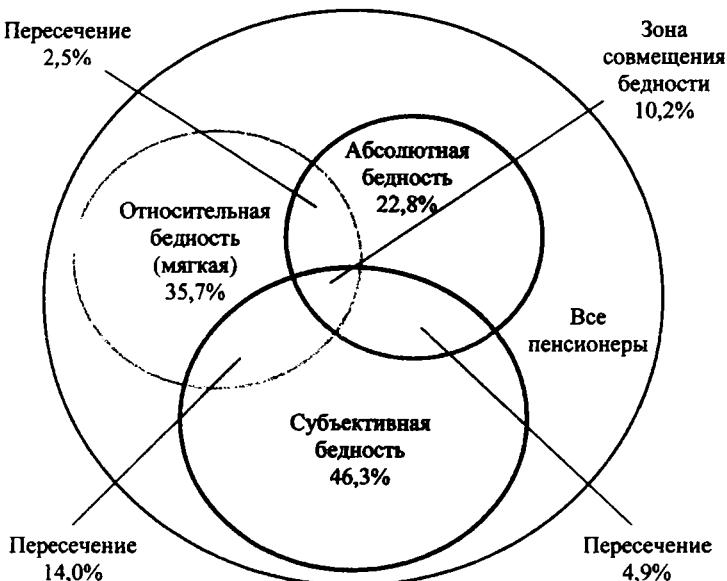
Видно, что увеличение значения одного из объясняющих факторов при прочих равных (β и μ) эквивалентно смещению распределения вправо (пунктирная линия). То есть предполагая, что коэффициент β положителен для данного фактора x , вероятность $\Pr\{y = 1\}$ должна снизиться. Другими словами, предельный эффект для указанной вероятности имеет знак, противоположный знаку коэффициента β . Используя ту же логику рассуждений, изменение в вероятности выбора $\Pr\{y = 3\}$ (или в $\Pr\{y = M\}$ для общего случая) должно иметь тот же знак, что и коэффициент.

Изменение вероятностей промежуточных значений остается неопределенным до проведения специальных расчетов.

Предельные эффекты по всем параметрам в сумме дают нуль (так как сумма вероятностей по всем альтернативам должна быть равна единице).

Данный подход не может быть использован для оценивания эффекта влияния искусственной (*dummy*) переменной. С этой целью следует сравнивать вероятности для двух различных значений искусственной переменной при прочих независимых переменных, равных их выборочным средним.

Пример. Изучение бедности на примере пенсионеров Ленинградской области. Данная работа выполнена по результатам исследования домохозяйств Ленинградской области в рамках программы по борьбе с бедностью SPRILLO.



На практике для изучения бедности используют три основных подхода: абсолютный, относительный и субъективный*. Каждый из них обладает своими преимуществами и недостатками. Поэтому предпочтительнее комбинированный подход, основанный на построении многокритериальной линии бедности, объединяющей абсолютную, относительную и субъективную линии бедности. Он позволяет, с одной стороны, использовать достоинства каждой из концепций бедности, а с другой — нивелировать их недостатки.

Основанием для построения комбинированного критерия бедности является гипотеза о том, что все рассматриваемые критерии бедности равноправны между собой, т. е. ни один критерий не является предпочтительным при оценке бедности. Таким образом, можно всех пенсионеров в выборке поделить на 4 группы (рис. 5.3):

- не бедные ни по одному критерию;
- бедные только по одному критерию;
- бедные только по двум критериям;
- бедные по трем критериям (крайне бедные).

Исходя из данной гипотезы области пересечений только двух критерев из трех можно рассматривать как общую совокупность. Также

* Подробнее об этом см.: Овчарова Л. Н., Прокофьева Л. М., Токсанбаева М. С. и др. Бедность: альтернативные подходы к определению и измерению. М.: Московский центр Карнеги, 1998.

можно поступить и с областями, состоящими только из одного критерия. Таким образом, получаем описанные выше группы.

Важным является то, что эти группы можно ранжировать исходя из того соображения, что быть бедным по одному критерию предпочтительнее, чем быть бедным по двум, тем более по трем критериям. Тогда появляется возможность оценить вероятность попадания пенсионера в каждую из этих областей, применяя модели множественного выбора с упорядоченными альтернативами.

После анализа данных в модель вошли следующие переменные:

BEDN_L – истинная бедность:

- 1 – небедные;
- 2 – бедные по одной концепции;
- 3 – бедные по двум концепциям;
- 4 – бедные по трем концепциям (крайне бедные);

DEMONS – дискретная переменная, отражающая тип семьи, в которой проживает пенсионер. Она представлена следующими индикаторными переменными:

DEM04 – одинокий пенсионер;

DEM05 – супружеская пара в пенсионном возрасте;

DEM06 – пенсионеры, проживающие в других семьях без детей;

DEM09 – пенсионеры, проживающие в семьях с детьми;

DOL_LEK – доля расходов на лекарства в доходе пенсионера;

DOL_NERAB – доля неработающих членов в домохозяйстве;

DOL_PIT – доля расходов на питание в доходе пенсионера;

EDU – образование, которое было закодировано следующим образом:

- 1 – начальное образование и ниже;
- 2 – неполное среднее;
- 3 – среднее полное;
- 4 – ПТУ;
- 5 – среднее профессиональное (специальное);
- 6 – неполное высшее;
- 7 – полное высшее образование и выше;

JIL_USL – плохое и очень плохое состояние жилья: 1 – плохое и очень, 0 – в противном случае;

LPH – роль ЛПХ в домохозяйстве: 2 – основной источник питания и доходов, 1 – дополнительный источник, 0 – не влияет;

MESTO – тип населенного пункта, который представлен следующими индикаторами:

MES01 – город с населением 50 000 человек и более;

MES02 – город с населением от 20 000 до 50 000 человек;

MES03 – город с населением менее 20 000 человек;

MES04 – поселок городского типа;

MES05 – деревня;

RISK_VOSPR – риск восприятия угрозы алкоголизма;

SEX – пол: 1 – мужчина, 0 – женщина;

Получили следующую модель:

Таблица 5.3

Probit-модель для всех пенсионеров старше трудоспособного возраста

Переменная	Коэффициент	Ст. ошибка	z-статистика	Вероятность
DOL_NERAB	1,539561	0,120556	12,77055	0,0000
DOL_PIT	-1,346449	0,273231	-4,927868	0,0000
DOL_LEK	0,586553	0,243702	2,406845	0,0161
EDU	-0,153900	0,014957	-10,28965	0,0000
JIL_USL	0,541523	0,072313	7,488639	0,0000
RISK_VOSPR	0,520777	0,118060	4,411119	0,0000
DEM05	-0,290579	0,067007	-4,336557	0,0000
MES02	0,200321	0,075272	2,661284	0,0078
MES04	-0,211358	0,099346	-2,127504	0,0334
MES05	0,198497	0,068821	2,884246	0,0039
SEX	0,177724	0,067381	2,637587	0,0083
LPH	0,095798	0,044818	2,137481	0,0326
Пороговые значения				
LIMIT_1:C(13)	0,644602	0,184882	3,486565	0,0005
LIMIT_2:C(14)	1,624163	0,187614	8,656921	0,0000
LIMIT_3:C(15)	2,553219	0,191599	13,32581	0,0000
log L_f			-1902,979	

Из табл. 5.3 видно, что все коэффициенты статистически значимы на 5%-ном уровне значимости.

Значимость модели проверялась по LR - статистике, которая подчиняется распределению по χ^2 .

H_0 : введение объясняющих переменных (кроме константы) не является значимым,

$$\begin{aligned} H_1: & \text{ не } H_0, \\ LR = -2(\ln L_0 - \ln L_f) &= 498,16, \\ \chi^2_{0.05, 12} &= 21,03. \end{aligned}$$

Следовательно, можно отвергнуть гипотезу H_0 , т. е. признать, что модель статистически значима.

Probit-модель с упорядоченными альтернативами оценивает *очки* (*scores*) S как линейную функцию от независимых переменных.

Предсказываемые вероятности будут зависеть от значения S и нормально распределенного остатка. В соответствии с оцененными пороговыми значениями:

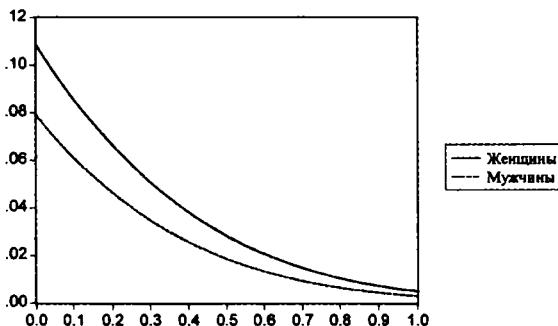


Рис. 5.3. Предсказанная вероятность риска крайней бедности для мужчин и женщин

$$P(BEDN_L = 1) = P(S + u \leq LIMIT_2)$$

$$P(BEDN_L = 2) = P(LIMIT_2 < S + u \leq LIMIT_3)$$

$$P(BEDN_L = 3) = P(LIMIT_3 < S + u \leq LIMIT_4)$$

$$P(BEDN_L = 4) = P(S + u > LIMIT_4)$$

Приведем график предсказанной вероятности крайней бедности для мужчин и женщин (рис. 5.3).

Из рис. 5.3 видно, что у женщин риск попадания в число крайне бедных больше, чем у мужчин. Это можно объяснить тем, что у мужчин средняя продолжительность жизни меньше, чем у женщин, которые после смерти своих мужей остаются вдовами. Из модели видно, что для одиноких пенсионеров риск попадания в число крайне бедных больше, чем для супружеских пар в пенсионном возрасте (коэффициент предпеременной DEM05 (супружеские пары в пенсионном возрасте) отрицательный).

Рассмотрим подробнее влияние переменной DOL_PIT (доля затрат на питание). Коэффициент перед данной переменной получился отрицательным, что не соответствует классическому правилу о том, что у бедных доля расходов на питание в среднем выше, чем у небедных. Чтобы понять, как ведет себя эта переменная, перейдем к предсказанию вероятности принадлежности к каждой группе бедности в зависимости от доли расходов на питание.

На рис. 5.4 показаны вероятности принадлежности к некоторой группе бедности для мужчин (для женщин картина аналогичная).

Во-первых, можно заметить, что с увеличением доли расходов на питание вероятность оказаться бедным по одному критерию увеличивается. Увеличение вероятности попадания в группу крайней бедности при уменьшении доли затрат на питание можно объяснить тем, что крайне бедным приходится жить за счет личного подсобного хозяйства, т. е. они почти ничего не тратят на еду. Это подтверждают эмпирические данные: для 16,7% крайне бедных личное подсобное хозяйство

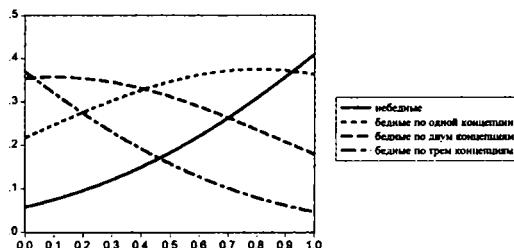


Рис. 5.4. Вероятности принадлежности
к некоторой группе бедности для мужчин

является основным источником питания и доходов при среднем значении 7,1% для всей выборки пенсионеров, т. е. значимость этого фактора повышается более чем в 2 раза.

Предельные эффекты

Интерпретация коэффициентов в моделях с дискретной зависимой переменной не столь очевидна, как в моделях линейных регрессий. Но можно перейти к предельным эффектам коэффициентов, которые имеют интерпретацию, близкую к интерпретации коэффициентов в линейных регрессиях, т. е. показывают, на сколько изменится вероятность попасть в ту или иную категорию при увеличении независимой переменной на единицу (табл. 3.20).

Таблица 5.4

Предельные эффекты для *probit*-модели
для пенсионеров старше трудоспособного возраста

Переменная	Небедные	Предельные эффекты		
		Бедные по		
		одной концепции	двум кон- цепциям	трем кон- цепциям
Доля неработающих	-0,5712	0,0570	0,3171	0,1971
Доля расходов на питание	0,4995	-0,0498	-0,2774	-0,1723
Доля расходов на лекарства	-0,2176	0,0217	0,1208	0,0751
Уровень образования	0,0571	-0,0057	-0,0317	-0,0197
Состояние жилища	-0,2159	0,0240	0,1176	0,0743
Риск восприятия алкоголизма	-0,2082	0,0233	0,1133	0,0717
Домохозяйства без детей	0,928	-0,0068	-0,0539	-0,0322
Город с населением от 20 до 50 тыс. человек	-0,0893	0,0114	0,0473	0,0306

Окончание табл. 5.4

Переменная	Небедные	Предельные эффекты		
		Бедные по		
		одной концепции	двум концепциям	трем концепциям
ПГТ	0,0634	-0,0038	-0,0375	-0,0221
Деревня	-0,0886	0,0113	0,0469	0,0304
Пол	-0,0809	0,0106	0,0426	0,0277
ЛПХ	-0,0355	0,0035	0,0197	0,0123

Например, для пенсионера, проживающего в домашнем хозяйстве, где хотя бы один член имеет склонность к алкоголизму, его вероятность попасть в группу крайне бедных больше на 0,07; в группу бедных по двум концепциям увеличивается на 0,11; в группу бедных по одной концепции увеличивается на 0,02; в группу небедных уменьшается на 0,11 по сравнению с аналогичным пенсионером, проживающим в других домохозяйствах.

На основе данной модели можно определить группу риска, представленную крайне бедными пенсионерами, что позволит проводить более точную адресную политику по борьбе с бедностью.

5.2.2. Модели множественного выбора с неупорядоченными альтернативами

В некоторых случаях не существует естественного упорядочивания между альтернативами, например при моделировании способа передвижения (автобус, поезд, машина, велосипед, пешком). Общей стартовой точкой для данных моделей является предположение о существовании случайной полезности, которая влияет на выбор альтернатив. Предполагается, что случайные полезности являются линейными функциями от наблюдаемых характеристик (индивидуальной или иной природы) и имеют аддитивно-разделяемую структуру. Предполагается, что индивидуум выбирает ту альтернативу, которая дает ему максимальную полезность.

Для формализации предположим, что выбор осуществляется между M альтернативами ($j = 1, 2, \dots, M$), проиндексированными в произвольном порядке. Также предположим, что уровень полезности, которая достигается индивидуумом i для каждой из альтернатив, задается при помощи U_{ij} , где $j = 1, 2, \dots, M$. Полезность каждой альтернативы является линейной функцией от наблюдаемых характеристик (индивидуальной или иной природы). Тогда альтернатива j будет выбрана индивидуумом i только в том случае, когда будет давать ему максимум полезности, т.е. если $U_{ij} = \max\{U_{i1}, \dots, U_{iM}\}$. Естественно, что данные уровни полезности являются ненаблюдаемыми и необходимо сделать

несколько дополнительных предположений для того, чтобы сделать эту конструкцию пригодной для работы с ней.

Предположим, что полезности ведут себя следующим образом: $U_{ij} = \mu_{ij} + \varepsilon_{ij}$, где μ_{ij} является неслучайной функцией наблюдаемых и небольшим количеством неизвестных параметров, а ε_{ij} является ненаблюдаемым остаточным членом. Тогда

$$\begin{aligned} P\{y_i = j\} &= P\{U_{ij} = \max(U_{i1}, \dots, U_{iM})\} = \\ &= P\left\{\mu_{ij} + \varepsilon_{ij} > \max_{k=1, \dots, J, k \neq j} (\mu_{ik} + \varepsilon_{ik})\right\}. \end{aligned} \quad (5.23)$$

Для удобства предположим, что все ε_{ij} взаимно независимы, а также распределены по закону распределения Вейбулла (Weibull). Тогда функция распределения каждого из остатков ε_{ij} имеет вид $F(t) = \exp\{-e^{-t}\}$, причем $F(t)$ не содержит неизвестных параметров. В этом случае можно показать, что

$$P\{y_i = j\} = \frac{\exp\{\mu_{ij}\}}{\exp\{\mu_{i1}\} + \exp\{\mu_{i2}\} + \dots + \exp\{\mu_{iM}\}}. \quad (5.24)$$

Данная структура автоматически предполагает, что $0 \leq P\{y_i = j\} \leq 1$, а также то, что $\sum_{j=1}^M P\{y_i = j\} = 1$.

Обычно один из уровней полезности принимают равным нулю, например $\mu_{i1} = 0$, и полагают, что μ_{ij} является линейной функцией от наблюдаемых переменных ($\mu_{ij} = x'_{ij} \beta$). Тогда получим

$$P\{y_i = j\} = \frac{\exp\{x'_{ij} b\}}{1 + \exp\{x'_{i2} b\} + \dots + \exp\{x'_{iM} b\}}, \quad j = 1, 2, \dots, M. \quad (5.25)$$

Данное выражение представляет собой *logit*-модель с множественными альтернативами (*Multinomial Logit Model (MNL)* или *Independent Logit Model*). Здесь вероятность того, что индивидуум выберет альтернативу j , зависит от объясняющих переменных и от коэффициентов β .

Несмотря на привлекательность аналитического выражения для MNL-модели, она имеет один большой недостаток, который вытекает из предположения, что все ошибки ε_{ij} независимы между собой. То есть подразумевается, что уровни полезности (зависящие от наблюдаемых характеристик) любых двух альтернатив независимы. Данное утверждение не совсем корректно, если две или более альтернативы похожи. Типичным примером будет разбивка категории «путешествовать на автобусе» на две: «путешествовать на синем автобусе» и «путешествовать на красном автобусе».

Другими словами, отношение вероятностей двух альтернатив не зависит от природы других альтернатив. Предположим, что первая альтернатива описывает путешествие на машине, а вторая — путешествие на (синем) автобусе. Тогда соотношение шансов (*odds ratio*)

$P\left\{y_i = 2\right\} / P\left\{y_i = 1\right\} = \exp\{x'_{i2}\beta\}$, независимо от того, является ли третьей альтернативой красный автобус или поезд. Данное свойство привлекательно с вычислительной точки зрения, но в то же время вызывает проблемы при «поведенческой» интерпретации модели.

Если предполагается, что ϵ_{ij} являются независимыми и распределенными по стандартному нормальному закону, то получим *probit-модель с множественными альтернативами* (*multinomial probit model*). Тогда вероятности должны содержать $(M - 1)$ -мерный интеграл, который весьма труден для вычислений.

В некоторых случаях выбор между M альтернативами можно разделить на два или несколько последовательных выбора. В данном случае можно думать о проблеме выбора, состоящей из нескольких последовательных этапов, хотя необходимо помнить, что модель возникает как модификация стохастической спецификации первоначальной *MNL (conditional)* модели, а не как «новая» модель описания поведения. Такие модели называются *моделями с группировкой* (*Nested Models*) и выступают как одна из возможных альтернатив для *MNL*-модели. Например, задачу о выборе транспортного средства между автомобилем, поездом и автобусом можно разделить на задачу выбора между общественным и личным транспортом, а общественный транспорт, в свою очередь, можно разделить на выбор между поездом и автобусом.

Контрольные вопросы

1. Какие модели называют моделями дискретного выбора?
2. Обоснуйте некорректность применения классической линейной модели в случае, если результирующая переменная принимает дискретные значения.
3. Охарактеризуйте модели бинарного выбора.
4. В чем сходство и в чем различие *logit-* и *probit-*моделей?
5. Какова интерпретация коэффициентов в моделях бинарного выбора?
6. Объясните модель бинарного выбора, используя предположение о латентной переменной.
7. Какой метод применяют для оценки параметров модели бинарного выбора?
8. Какие показатели используют для оценки качества модели дискретного выбора? Какая идея лежит в их основе?
9. Какие тесты используют при проверке гипотез о наличии ограничений на коэффициенты и о совместной незначимости коэффициентов в моделях бинарного выбора?
10. Какие модели называют моделями множественного выбора?
11. Назовите модели, являющиеся моделями с упорядоченными альтернативами.
12. Как интерпретируются коэффициенты в моделях с упорядоченными альтернативами?
13. Охарактеризуйте модели множественного выбора с неупорядоченными альтернативами.

ГЛАВА 6. МОДЕЛИРОВАНИЕ ИЗОЛИРОВАННОГО ДИНАМИЧЕСКОГО РЯДА

6.1. Компоненты динамического ряда

В качестве источников информации в эконометрике широко используют динамические (временные) ряды. Модели по рядам динамики могут строиться на основе:

- изолированного динамического ряда, т. е. изучается один динамический ряд: например, по данным о численности занятых за ряд лет строится модель динамики численности занятых;
- системы взаимосвязанных рядов динамики, т. е. один из рядов рассматривается как моделируемый объект, а другие — как его факторы: например, строится модель прибыли в зависимости от объема реализации, численности работающих, фондооруженности труда и т. п.

При построении моделей по временным рядам необходимо учитывать компоненты (составные части) динамического ряда.

Уровни динамического ряда в конкретный период времени t принимают те или иные значения в результате действия разных факторов. Одни из них являются основными, формирующими величину уровня (y_t) на данном этапе исторического развития, а другие — случайными, несущественными с точки зрения содержания, его материальной природы. Фактическую величину уровня динамического ряда (y_t) можно представить как функцию трех компонент:

- тенденции ряда*, обусловленной влиянием общих факторов, определяющих основное направление развития явления за длительный период времени — тренд ряда;
- периодических колебаний*, вызванных особенностями существования явления в одни периоды по сравнению с другими;
- случайных колебаний*, связанных с действием разного рода второстепенных факторов — случайная компонента.

Символически это можно представить как

$$y_t = f(T, -, \xi),$$

где y_t — фактический уровень динамического ряда в период времени t ;

T — тренд ряда;

P — периодические колебания (циклические, сезонные);

ξ — случайная составляющая.

Рассматриваемые компоненты динамического ряда необязательно присущи каждому временному ряду. Могут быть ряды динамики, в которых отсутствует как тенденция, так и периодические колебания. В этом случае уровни ряда $y_t = f(\xi)$. Они являются функцией случайной компоненты: колеблются вокруг среднего уровня, что характерно для так называемого *стационарного ряда*. На графике такой ряд представляет собой ломаную линию, параллельную оси времени (рис. 6.1).

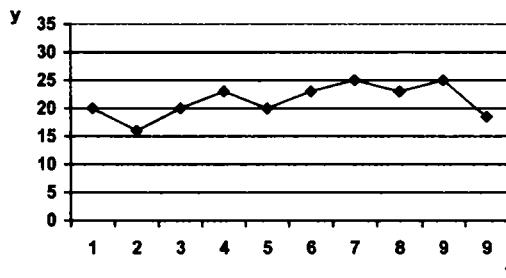


Рис. 6.1. Ряд без тенденции и периодических колебаний

Модель уровня такого динамического ряда имеет вид

$$y_t = \bar{y} + \xi,$$

где y_t — уровни динамического ряда;

\bar{y} — средний за период уровень ряда;

ξ — случайная составляющая, определяемая как $\xi = y_t - \bar{y}$.

Такие ряды в экономике сравнительно редки. Чаще имеют место ряды с тенденцией. В основном ряды без тенденции наблюдаются при изучении динамики показателей из относительных и средних величин. Например, доля единого социального налога в процентах к фонду оплаты труда на предприятиях России представляет собой во многих случаях подобный ряд, так как при наличии регрессивной шкалы налогообложения (ставки налога снижаются с ростом оплаты труда) доходы подавляющей части работников облагаются по единой ставке (попадают в первую группу шкалы налогообложения).

Большинство динамических рядов в экономике характеризуется тенденцией и случайными колебаниями (рис. 6.2).

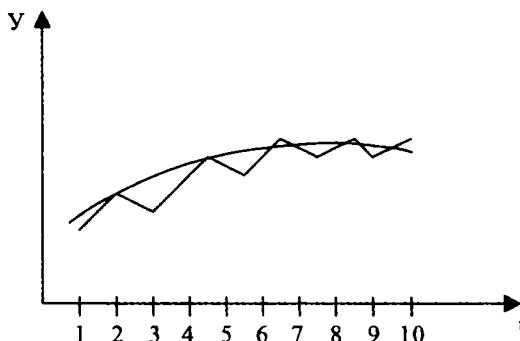


Рис. 6.2. Ряд с тенденцией (плавная линия) и случайными колебаниями (ломаная линия)

Модель уровня такого ряда имеет вид

$$y_t = f(T) + \xi,$$

где $f(T)$ – математическая функция, характеризующая закономерность развития явления во времени, т. е. описывающая тенденцию развития явления – тренд ряда;
 ξ – случайные колебания.

Если обозначить теоретическое значение уровня ряда, соответствующее определенной математической функции тренда, как \hat{y}_t , то случайные колебания ξ составят величину

$$\xi_t = y_t - \hat{y}_t.$$

Например, за 1995–2003 гг. динамический ряд потребления овощей и продовольственных бахчевых культур на душу населения России составил (в кг):

1995 г.	1996 г.	1997 г.	1998 г.	1999 г.	2000 г.	2001 г.	2002 г.	2003 г.
76	75	79	78	83	86	89	91	94

Наблюдающаяся тенденция к росту может быть описана уравнением тренда вида

$$\hat{y}_t = 71,805e^{0.0294t},$$

где t принимает значения 1, 2, ..., 9.

Согласно этой тенденции теоретическое значение уровня ряда в 2003 г. (\hat{y}_t) составило 93,6 кг (в уравнение подставлено значение $t = 9$). Так как фактическое значение в этот год (y_t) составило 94 кг, то величина случайной составляющей (ξ) окажется равной 0,4 кг (94 – 93,6).

При изучении динамики явления за продолжительный период времени уровни ряда могут обнаруживать регулярные колебания, повторяющиеся через равные промежутки времени спады или подъемы. Такие колебания принято называть *периодическими* (рис. 6.3 и 6.4).

Если период колебаний насчитывает несколько лет, то такие периодические колебания считают *циклическими*. Например, солнечная активность проявляется с периодом 10–11 лет. В сфере предпринима-

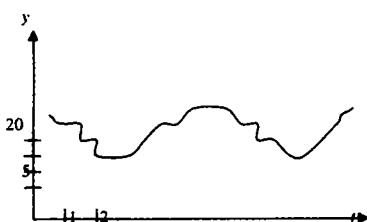


Рис. 6.3. Ряд с периодическими и случайными колебаниями
 $y_t = f(C, \xi)$



Рис. 6.4. Ряд с тенденцией, периодическими и случайными колебаниями $y_t = f(C, \xi)$

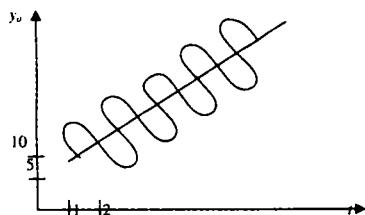


Рис. 6.5. Аддитивная модель

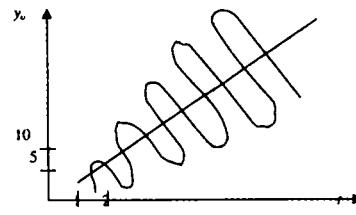


Рис. 6.6. Мультипликативная модель

тельства могут иметь место экономические циклы, включающие в себя рост, спад, свертывание и затем оживление экономической деятельности. Длина цикла зависит от вида деятельности и охватывает нередко 3–12 лет. Так, в ряде стран цена и производство свинины подвергаются регулярным колебаниям, цикл которых длится около 3 лет*.

Регулярные колебания в течение года называются *сезонными* (обозначаются S): например, колебания спроса на одежду с изменением сезона года (весна, лето, осень, зима), колебания цен на сельскохозяйственную продукцию и т. п. Цикл колебаний равен году. Наличие сезонных колебаний означает, что на протяжении ряда лет в одни и те же кварталы (месяцы) года наблюдается рост или снижение уровня ряда. Так, рождественские праздники обуславливают рост товарооборота в декабре и январе.

В отличие от периодических случайные колебания не носят регулярный характер и связаны с действием разного рода случайных причин.

Рассматриваемые компоненты динамического ряда позволяют представить уровень динамического ряда в виде *аддитивной* или *мультипликативной модели*:

$$y_t = T + S + \xi \text{ — аддитивная модель;}$$

$$y_t = T \times S \times \xi \text{ — мультипликативная модель.}$$

Выбор модели зависит от характера периодических колебаний. Если амплитуда, например, сезонных колебаний остается во времени постоянной, то применяется аддитивная модель. Если же амплитуда колебаний изменяется во времени, то рассматривается мультипликативная модель (рис. 6.5 и 6.6).

На рис. 6.5 показаны тенденция к увеличению (прямая линия) и периодические колебания — отклонения от тенденции, равные по всей длине динамического ряда, т. е. с одинаковой амплитудой волны.

На рис. 6.6 также присутствуют тенденция уровней ряда и периодические колебания, амплитуда которых возрастает во времени.

В аддитивной модели компоненты ряда выражены в тех же единицах измерения, что и рассматриваемый в динамике признак. Так, если y_t измеряется в тысячах тонн, то и составные части ряда тоже

*См.: Ланге О. Введение в эконометрику: пер. с польск. М. Прогресс, 1964. С. 145.

выражены в тысячах тонн. Пусть $y_t = 45$ тыс. т, а согласно тенденции $\hat{y}_t = 40$ тыс. т. Если имеют место периодические колебания, то разница $y_t - \hat{y}_t = 5$ тыс. т характеризует периодическую и случайную составляющие. Предположим, что периодическая компонента составила 12 тыс. т, т. е. тренд вместе с периодической составляющей равен 52 тыс. т. Однако с учетом случайных колебаний фактическое значение $y_t = 45$ тыс. т. Следовательно, случайная компонента составила величину $\xi = 45 - 52 = -7$, или, иначе, $\xi = 5 - 12 = -7$ тыс. т.

При мультипликативной модели периодическая и случайная составляющие выражены в относительных величинах. Так, при сезонных колебаниях S — это индекс сезонности. Пусть $y_t = 38$ тыс. т, а согласно тенденции $\hat{y}_t = 31$ тыс. т. Индекс сезонности для периода t составил, например, 119,4%. Тогда тренд с учетом сезонности окажется равным 37 тыс. т ($31 \cdot 1,194$), а случайная компонента по абсолютной величине составит 1 тыс. т (38 тыс. т — 37 тыс. т). Ее можно представить в виде относительной величины: $\xi_t = \frac{38}{37} = 1,027$. В этом случае

мультипликативное разложение уровня динамического ряда окажется следующим:

$$38 = 31 \cdot 1,194 \cdot 1,027, \text{ т. е.}$$

$$y_t = \hat{y}_t \cdot P \cdot \xi_t$$

Если случайную компоненту представить абсолютной величиной, то получим модель смешанного типа:

$$y_t = \hat{y}_t + P + \xi_t, \text{ т. е.}$$

$$38 = 31 + 1,194 + 1.$$

Рассмотренные компоненты динамического ряда учитываются как при построении модели изолированного временного ряда, так и при построении регрессионных моделей на основе системы взаимосвязанных рядов динамики, что и будет изложено далее.

6.2. Автокорреляция уровней динамического ряда и характеристика его структуры

При наличии тенденции в ряду динамики уровни ряда характеризуются автокорреляцией, т. е. каждый последующий уровень ряда зависит от предыдущего. Например, цена на товар сегодня, как правило, зависит от цены вчерашнего дня. Корреляционная связь между последовательными значениями уровней динамического ряда называется *автокорреляцией уровней ряда*.

Для измерения автокорреляции уровней динамического ряда используется коэффициент автокорреляции уровней:

$$r_{y_t, y_{t-\tau}} = \frac{\overline{(y_t - \bar{y}_t)(y_{t-\tau} - \bar{y}_{t-\tau})}}{\sigma_{y_t} \cdot \sigma_{y_{t-\tau}}},$$

где y_t — фактические уровни динамического ряда;

- y_{t-1} – уровни того же динамического ряда, но сдвинутые на τ шагов во времени;
- τ – величина лага (сдвига во времени), принимающая значения 1, 2, 3 и т. д. и определяющая порядок коэффициента автокорреляции.

При $\tau = 1$ рассчитывается коэффициент автокорреляции *первого порядка*, т. е. измеряется корреляция текущих значений уровней динамического ряда (y_t) с предшествующими уровнями (y_{t-1}).

При $\tau = 2$ изучается зависимость текущих уровней ряда (y_t) с уровнями этого же ряда, сдвинутыми на 2 временных шага (y_{t-2}), т. е. рассчитывается коэффициент автокорреляции второго порядка, а при $\tau = 3$ – соответственно третьего порядка, при $\tau = k$ – коэффициент автокорреляции *k-го порядка*. Чем длиннее динамический ряд, тем выше может быть порядок коэффициента автокорреляции уровней.

Коэффициент автокорреляции уровней ряда практически рассчитывается по формуле линейного коэффициента корреляции. Поэтому его величина изменяется в пределах от -1 до $+1$. Чем ближе его величина к ± 1 , тем сильнее зависимость текущих уровней динамического ряда от предыдущих.

Если ряд характеризуется четко выраженной тенденцией, то для него коэффициент автокорреляции первого порядка приближается к $+1$. Так, для рассмотренного ранее ряда динамики потребления овощей на душу населения коэффициент автокорреляции уровней первого порядка составил 0,948, демонстрируя тесную связь последующих уровней ряда от предыдущих. Методика его расчета представлена в табл. 6.1.

Таблица 6.1

**Расчет коэффициента автокорреляции первого порядка
для ряда динамики потребления овощей и продовольственных
бахчевых культур на душу населения России**

Годы	t	y_t	y_{t-1}	$y_t y_{t-1}$	y_t^2	y_{t-1}^2
1995	1	76	–	–	–	–
1996	2	75	76	5700	5625	5776
1997	3	79	75	5925	6241	5625
1998	4	78	79	6162	6084	6241
1999	5	83	78	6474	6889	6084
2000	6	86	83	7138	7396	6889
2001	7	89	86	7654	7921	7396
2002	8	91	89	8099	8281	7921
2003	9	94	91	8554	8836	8281
Σ	–	675*	657	55 706	57 273	54 213

*Определено без уровня первой строки.

Поскольку в примере рассчитывается коэффициент автокорреляции первого порядка, т. е. когда $\tau = 1$, формула его расчета приобретает вид:

$$r_{y_t y_{t-1}} = \frac{(\bar{y}_t y_{t-1} - \bar{y}_t \bar{y}_{t-1})}{\sigma_{y_t} \cdot \sigma_{y_{t-1}}},$$

где y_t — уровни ряда в момент времени t ;

y_{t-1} — те же уровни ряда, но сдвинутые на год, т. е. уровни ряда в момент времени $t-1$ (предыдущий год).

Поскольку оба ряда (y_t и y_{t-1}) для расчета коэффициента автокорреляции должны быть одинаковой длины, то первая строка табл. 6.1 в расчетах не участвует. В итоговой строке табл. 6.1 даны необходимые суммы для подсчета отдельных элементов формулы коэффициента автокорреляции уровней:

$$\bar{y}_t = \sum_{t=2}^n y_t \quad (n-1) = 675 - 8 = 84,375,$$

$$\bar{y}_{t-1} = \sum_{t=2}^n y_{t-1} \quad (n-1) = 657 - 8 = 82,125,$$

$$\sigma_{y_t} = \sqrt{\sum_{t=2}^n y_t^2 / (n-1) - (\bar{y}_t)^2} = \sqrt{\frac{57273}{8} - 84,375^2} = 6,32332,$$

$$\sigma_{y_{t-1}} = \sqrt{\sum_{t=2}^n y_{t-1}^2 / (n-1) - (\bar{y}_{t-1})^2} = \sqrt{\frac{54273}{8} - 82,125^2} = 5,66651.$$

Соответственно коэффициент автокорреляции уровней составит

$$r_{y_t y_{t-1}} = \frac{6963,25 - 84,375 \cdot 82,125}{6,323 \cdot 5,667} = 0,948.$$

Методика расчета коэффициентов автокорреляции более высоких порядков та же, но при этом число коррелируемых пар уменьшается. В нашем примере их 8 (с $t = 2$ по $t = 9$). Если же увеличим лаг до 2 лет, т. е. $\tau = 2$, то останется 7 коррелируемых пар (с $t = 3$ по $t = 9$), при $\tau = 3$ будет 6 коррелируемых пар (с $t = 4$ по $t = 9$). Ввиду уменьшения числа наблюдений при расчете коэффициента автокорреляции уровней увеличение величины лага небеспредельно: принято считать, что максимальная величина лага должна быть не более чем $n/4$ (n — длина динамического ряда). Для нашего примера при $n = 9$ максимальная величина лага составит 2 года ($\tau = 2$).

Для расчета коэффициента автокорреляции второго порядка составим табл. 6.2.

Таблица 6.2

**Расчет коэффициента автокорреляции уровней второго порядка
(для ряда динамики, представленного в табл. 6.1)**

t	y_t	y_{t-1}	$y_t y_{t-1}$	y_t^2	y_{t-1}^2
1	76	—	—	—	—
2	75	—	—	—	—
3	79	76	6004	6241	5776
4	78	75	5850	6084	5625
5	83	79	6557	6889	6241
6	86	78	6708	7396	6084
7	89	83	7387	7921	6889
8	91	86	7826	8281	7396
9	94	89	8366	8836	7921
Σ	600*	566	48 698	51 648	45 932

*Подсчитано без первых двух строк.

Так как теперь в расчете участвует 7 коррелируемых пар (y_t и y_{t-2}), первые две строки табл. 6.2 не принимаются во внимание. Коэффициенты автокорреляции разных порядков принято обозначать как r_1 , r_2 , r_3 , ..., r_k , где 1, 2, ..., k указывают на номер порядка коэффициента автокорреляции. Формула расчета коэффициента автокорреляции второго порядка следующая:

$$r_2 = (\bar{y}_t y_{t-2} - \bar{y}_t \cdot \bar{y}_{t-2}) / \sigma_{y_t} \cdot \sigma_{y_{t-2}},$$

где $\bar{y}_t = \sum_{t=3}^n y_t / (n-2) = 600 / 7 = 85,714$,

$$\bar{y}_{t-2} = \sum_{t=3}^n y_{t-2} / (n-2) = 566 / 7 = 80,857,$$

$$\bar{y}_t y_{t-2} = \sum_{t=3}^n y_t y_{t-2} / (n-2) = 48698 / 7 = 6956,857,$$

$$\sigma_{y_t} = \sqrt{\sum_{t=3}^n y_t^2 / (n-2) - (\bar{y}_t)^2} = \sqrt{\frac{51648}{7} - 85,714^2} = 5,603,$$

$$\sigma_{y_{t-2}} = \sqrt{\sum_{t=3}^n y_{t-2}^2 / (n-2) - (\bar{y}_{t-2})^2} = \sqrt{\frac{45932}{7} - 80,857^2} = 4,885.$$

Соответственно имеем

$$r_2 = \frac{6956,857 - 85,714 \cdot 80,857}{5,603 \cdot 4,885} = 0,960.$$

В рассмотренном примере уровни динамического ряда имеют тенденцию к возрастанию и коэффициенты автокорреляции приближаются к +1. Аналогичная картина будет наблюдаться и при тенденции к уменьшению уровней динамического ряда. Например, лесовосстановление в России за 1995–2002 гг. характеризуется тенденцией к снижению. Уровни ряда (в тыс. га) составили:

1995 г.	1996 г.	1997 г.	1998 г.	1999 г.	2000 г.	2001 г.	2002 г.
1454	1110	1092	1019	964	973	960	887

Коэффициенты автокорреляции первого и второго порядка оказались $r_1 = 0,812$ и $r_2 = 0,885$, что подтверждает наличие тенденции в ряду динамики. При этом r_1 и $r_2 > 0$, хотя ряд и имеет тенденцию к снижению. Чем тенденция по ряду динамики более четкая, тем ближе r_1 и r_2 к +1.

Для стационарного динамического ряда с небольшими колебаниями уровней r_1 достаточно близок к 0 и может принимать небольшое отрицательное значение. Так, предположим, что уровни ряда приняли следующие значения (последовательно во времени):

3; 1; 2; 1; 2; 1; 3; 3; 2; 3; 1; 2; 1; 1; 3; 3; 2; 2; 1; 3; 3; 2; 2; 3; 1; 2; 2; 1; 3; 1.

Коэффициент автокорреляции первого порядка составил $-0,209$, а второго порядка $-0,056$.

Серию коэффициентов автокорреляции уровней ряда с последовательным увеличением величины лага принято называть *автокорреляционной функцией (АКФ)*.

Для стационарного временного ряда с увеличением величины лага взаимосвязь y_t и $y_{t-\tau}$ ослабевает и АКФ характеризуется монотонным убыванием, что графически должно представлять затухающую кривую (рис. 6.7).

В нашем примере АКФ для стационарного ряда составила $r_1 = -0,209$; $r_2 = 0,056$; $r_3 = -0,114$; $r_4 = -0,356$; $r_5 = 0,057$; $r_6 = -0,074$; $r_7 = -0,003$. Однако при ограниченной длине динамического ряда поведение АКФ в виде рис. 6.7 не всегда соблюдается.

АКФ дает представление о внутренней структуре динамического ряда. С помощью АКФ можно определить наличие или отсутствие в ряду динамики периодических колебаний и соответственно величину периода колебаний: она равна той величине лага (τ), при которой коэффициент автокорреляции уровней наибольший.

Предположим, что объем продажи товара за 18 месяцев характеризуют следующим образом (рис. 6.8).

График показывает наличие тенденции, а также периодических колебаний. Это подтверждает и АКФ:

$$\begin{aligned} r_1 &= 0,863; \quad r_2 = 0,829; \quad r_3 = 1; \\ r_4 &= 0,812; \quad r_4 = 0,755; \quad r_6 = 1. \end{aligned}$$

Достаточно высокое значение коэффициента автокорреляции первого порядка (равное 0,863) означает наличие тенденции в ряду динамики. Вместе с тем максимальное значение коэффициента автокорре-

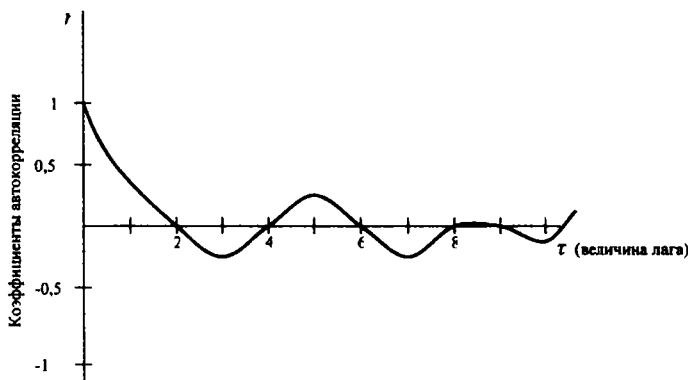


Рис. 6.7. Коррелограмма АКФ, т. е. график АКФ

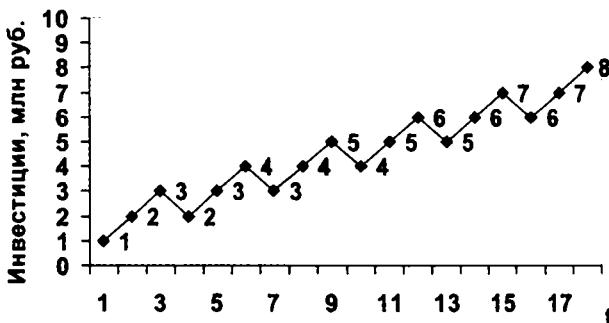


Рис. 6.8. Динамика объема продаж

ляции наблюдается при лаге 3 и кратном ему лаге 6, т. е. для ряда характерна регулярная колеблемость уровней через 3 месяца: подъем 3-го месяца сменяется спадом в следующий месяц. Иными словами, волнобразное изменение объема продаж повторяется через 3 месяца, что и демонстрирует АКФ. Для динамического ряда с монотонной тенденцией к возрастанию (или уменьшению) уровней АКФ имеет значения, близкие к +1, которые медленно снижаются с увеличением величины лага. Например, за 60 кварталов динамика объема продаж характеризовалась уравнением тренда:

$$\hat{y}_t = 101,659 - 0,9254t,$$

где y — объем продаж в тыс. руб., $t = 1, 2, \dots, 60$.

Коэффициент детерминации для него составил 0,973, характеризуя хорошее качество описания тенденции ряда: отклонения фактических уровней ряда от теоретических, обусловленных тенденцией, составляют всего 2,7%. АКФ для данного ряда оказалась: $r_1 = 0,991; r_2 = 0,984; r_3 = 0,980; r_4 = 0,979; r_5 = 0,973; r_6 = 0,968; r_7 = 0,963; r_8 = 0,965; r_9 = 0,963; r_{10} = 0,962; r_{11} = 0,959; r_{12} = 0,957; r_{13} = 0,952; r_{14} = 0,955; r_{15} = 0,943$.

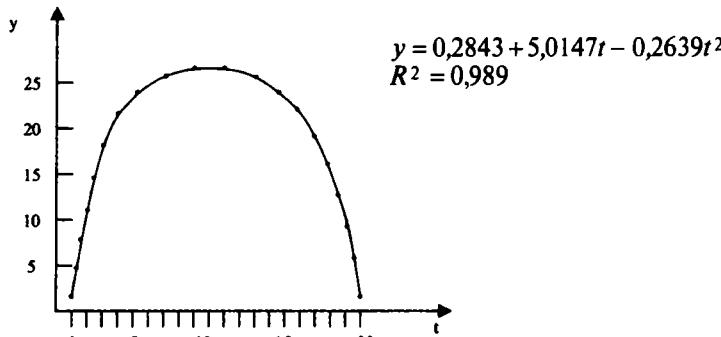


Рис. 6.9. Тренд в виде параболы второй степени

Если ряд характеризуется сменой тенденций, то АКФ примет значения, стремительно уменьшающиеся с увеличением величины лага и сопровождаемые иногда сменой знака коэффициента автокорреляции. Так, например, динамический ряд описывается параболой второго порядка (рис. 6.9).

АКФ оказывается следующей:

Лаг	r	Коррелограмма		
		-1	0	+1
1	0,887			xxxxxxxxx
2	0,535			xxxxx
3	0,192			xx
4	-0,104		x	
5	-0,350			xxxx

Иными словами, знание АКФ может помочь при подборе модели рассматриваемого динамического ряда.

6.3. Модели тенденции развития

6.3.1. Общая характеристика моделей тенденции

Закономерность изменения уровней динамического ряда во времени может быть представлена в виде модели тенденции. При ее построении уровни динамического ряда рассматриваются как функция времени (t) и случайной компоненты (ξ). Тогда модель уровня динамического ряда можно выразить как

$$y_t = \bar{y} + (\hat{y}_t - \bar{y}) + (y_t - \hat{y}_t), \quad (6.1)$$

где y_t — фактический уровень динамического ряда в период времени t ;
 \bar{y} — средний уровень динамического ряда за весь период времени;

\hat{y}_t — теоретический уровень динамического ряда, обусловленный тенденцией развития, т. е. трендом ряда.

В этой модели $(\hat{y}_t - \bar{y})$ характеризует эффект тенденции, а $(y_t - \hat{y}_t)$ — случайную составляющую ξ . Ввиду того, что $\bar{y} + (\hat{y}_t - \bar{y}) = \hat{y}_t$, данную модель уровня временного ряда можно представить как

$$y_t = \hat{y}_t + \xi_t, \quad (6.2)$$

где \hat{y}_t — модель тенденции, когда уровни ряда рассматриваются как функция времени t : $\hat{y}_t = f(t)$.

Совершенно ясно, что практическая значимость модели тенденции будет тем выше, чем меньше будут остаточные колебания (случайная составляющая $\xi_t = y_t - \hat{y}_t$).

Построение модели тенденции (уравнения тренда) включает в себя следующие этапы работы:

- выбор математической функции, описывающей тенденцию;
- оценка параметров модели;
- проверка адекватности выбранной функции и оценка точности модели;
- расчет точечного и интервального прогнозов.

В настоящее время компьютерные программы анализа временных рядов содержат достаточно широкий набор математических функций для построения уравнения тренда. Все многообразие их можно свести в три группы:

- функции с монотонным характером возрастания (убывания) и отсутствием пределов роста (снижения);
- кривые с насыщением, т. е. устанавливается нижняя или верхняя граница изменения уровней ряда;
- S-образные кривые, т. е. кривые с насыщением, имеющие точку перегиба.

В первую группу функций входят полиномы k -й степени:

$$\hat{y}_t = a + b_1 t + b_2 t^2 + \dots + b_k t^k. \quad (6.3)$$

При $k = 1$ получаем линейный тренд: $\hat{y}_t = a + b_1 t$, который часто записывают как $\hat{y}_t = a + bt$.

По содержанию линейный тренд означает, что уровни динамического ряда изменяются с одинаковой скоростью, т. е. с равным абсолютным приростом (параметр b). В этом можно убедиться, подставляя в уравнение линейного тренда порядковые значения $t(1, 2, 3, \dots, k)$: теоретические уровни ряда \hat{y} , будут изменяться на величину параметра b , т. е. в арифметической прогрессии.

Например, уравнение тренда для индексов потребительских цен за 12 месяцев года составило: $\hat{y}_t = 99,9 + 1,9t$, где $t = 1, 2, \dots, 12$. Из уравнения видно, что ежемесячно цены возрастили в среднем на 1,9 процентных пункта.

При $k = 2$ получаем параболу второй степени:

$$\hat{y}_t = a + b_1 t + b_2 t^2. \quad (6.4)$$

Данная функция рекомендуется для моделирования тенденции, если временной ряд характеризуется постоянным абсолютным ускорением, т. е. постоянными являются вторые разности (приросты абсолютных приростов). В этом случае скорость ряда изменяется линейно:

t	$\hat{y}_t = a + b_1 t + b_2 t^2$	Скорость $\Delta = y_t - y_{t-1}$	Ускорение $\Delta^{(1)} = \Delta y_t - \Delta y_{t-1}$
0	a	—	—
1	$a + b_1 + b_2$	$b_1 + b_2$	—
2	$a + 2b_1 + 4b_2$	$b_1 + 3b_2$	$2b_2$
3	$a + 3b_1 + 9b_2$	$b_1 + 5b_2$	$2b_2$
4	$a + 4b_1 + 16b_2$	$b_1 + 7b_2$	$2b_2$
5	$a + 5b_1 + 25b_2$	$b_1 + 9b_2$	$2b_2$

Как видим, параметр a означает начальный уровень ряда динамики при $t = 0$. Параметр b_1 представляет собой средний абсолютный прирост за весь период времени, если t обозначено так, что $\Sigma t = 0$ (при обозначении t как ряд натуральных чисел, что наиболее распространено при компьютерной обработке, параметр b_1 такой интерпретации не имеет). Параметр b_2 характеризует половину абсолютного ускорения динамического ряда.

Например, динамика численности детей в возрасте 7 лет характеризуется по району за последние 15 лет уравнением тренда

$$\hat{y}_t = 323,7 + 10,8t - 1,6t^2,$$

где y — тыс. чел., $t = 1, 2, \dots, 15$.

Следовательно, ежегодно, начиная с $t = 4$ численность детей сокращается в среднем с ускорением 3,2 тыс. чел.

При $k = 3$ имеем параболу третьей степени:

$$y_t = a + b_1 t + b_2 t^2 + b_3 t^3. \quad (6.5)$$

Этот вид тренда предполагает, что по временному ряду стабильны третии разности (Δ'''), т. е. приrostы вторых приростов ($\Delta''' = \Delta''_t - \Delta''_{t-1}$), а абсолютные ускорения имеют линейную тенденцию:

t	$\hat{y}_t = a + b_1 t + b_2 t^2 + b_3 t^3$	$\Delta_t = y_t - y_{t-1}$	$\Delta''_t = \Delta_t - \Delta_{t-1}$	$\Delta''' = \Delta''_t - \Delta''_{t-1}$
0	a	—	—	—
1	$a + b_1 + b_2 + b_3$	$b_1 + b_2 + b_3$	—	—
2	$a + 2b_1 + 4b_2 + 8b_3$	$b_1 + 3b_2 + 7b_3$	$2b_2 + 6b_3$	—
3	$a + 3b_1 + 9b_2 + 27b_3$	$b_1 + 5b_2 + 19b_3$	$2b_2 + 12b_3$	$6b_3$
4	$a + 4b_1 + 16b_2 + 64b_3$	$b_1 + 7b_2 + 37b_3$	$2b_2 + 18b_3$	$6b_3$
5	$a + 5b_1 + 25b_2 + 125b_3$	$b_1 + 9b_2 + 61b_3$	$2b_2 + 24b_3$	$6b_3$

Полиномы высоких степеней требуют достаточно длинных динамических рядов, чтобы параметры тренда были статистически надежными: на каждый параметр при t должно приходиться не менее 6–7 временных единиц. Следовательно, парабола уже третьей степени должна содержать не менее 20 лет (если уровни ряда представлены по годам), что предполагает достаточно стабильную экономику.

Чаще отдают предпочтение функциям с меньшим числом параметров. Среди них широкое применение находит показательная функция

$$\hat{y}_t = ab^t \quad (6.6)$$

или равносильная ей экспонента

$$\hat{y}_t = e^{a+bt} \text{ (либо } y = ae^{bt}), \quad (6.7)$$

которые характеризуются стабильным коэффициентом (темпом) роста:

t	1	2	3	4	6
$y = ab^t$	ab	ab^2	ab^3	ab^4	ab^5
Коэффициент роста	—	b	b	b	b

Например, за ряд лет динамика прибыли характеризуется уравнением вида $\hat{y}_t = 13,5 \cdot 1,5^t$, где $t = 1, 2, \dots, n$. Следовательно, ежегодно прибыль возрастает в среднем на 50% (коэффициент роста — 1,5). Данный тренд в виде экспоненты примет выражение $y = e^{2,603 + 0,405t}$, где $e^{2,603} = 13,5$ и $e^{0,405} = 1,5$. Рост по экспоненте означает геометрическую прогрессию уровней динамического ряда, что в экономике возможно сравнительно небольшой период времени (ограничены ресурсы, меняются условия рынка).

Если стабильными оказываются коэффициенты опережения темпов роста, то динамический ряд может быть описан логарифмической параболой

$$\hat{y}_t = ab^t c^{t^2} \quad (6.8)$$

Свое название данная функция получила ввиду того, что, прологарифмировав ее, получим параболу второй степени:

$$\lg y = \lg a + \lg b + t^2 \lg c.$$

Для этой функции темпы роста изменяются в одно и то же число раз (c^2):

t	$\hat{y}_t = ab^t c^{t^2}$	Коэффициент роста $-k_i$	Коэффициент опережения k_i / k_{i-1}
1	abc	—	—
2	$ab^2 c^4$	bc^3	—
3	$ab^3 c^9$	bc^5	c^2
4	$ab^4 c^{16}$	bc^7	c^2
5	$ab^5 c^{25}$	bc^9	c^2

Например, дебиторская задолженность за ряд лет характеризуется уравнением $\hat{y}_t = 1,47 \cdot 1,30^t \cdot 1,05^{e^2}$. Следовательно, имеет место ускоренное увеличение дебиторской задолженности с коэффициентом опережения темпов роста 1,05², или 1,1025, т. е. темпы роста ежегодно возрастили в среднем в 1,1025 раза.

При моделировании тенденции используются и другие функции, приводимые к линейному виду. Так, при замедленном росте уровней ряда может использоваться полулогарифмическая кривая

$$\hat{y}_t = a + b \ln t. \quad (6.9)$$

В 1990-е гг. XX в. по этой функции развивалось в стране потребление картофеля.

Предполагая разную меру пропорциональности изменений уровней во времени, может быть использована степенная функция

$$\hat{y}_t = at^b. \quad (6.10)$$

При $b > 0$ она характеризует непрерывный рост уровней с падающими темпами роста, а при $b < 0$ — их ускоренное снижение. Величина t^b означает базисный коэффициент роста:

t	$y = at^b$	Базисный коэффициент роста
1	a	1
2	$a2^b$	2^b
3	$a3^b$	3^b
4	$a4^b$	4^b
5	$a5^b$	5^b

Поэтому степенная функция практически сообщает о величине среднего коэффициента роста:

$$\bar{K} = \sqrt[b]{t^b}. \quad (6.11)$$

Например, обеспеченность городского населения Республики Коми жильем (в м² общей площади на человека) за 1990–1999 гг. характеризовалась уравнением вида: $\hat{y}_t = 15,876t^{0,08}$, где $t = 1, 2, \dots, 10$. Следовательно, за весь период обеспеченность населения жильем выросла в 1,202 раза (10^{0,08}), т. е. ежегодно она возрастала в среднем на 2,07%

$$(\bar{K} = \sqrt[9]{1,202} = 1,027).$$

К кривым с насыщением можно отнести гиперболы вида:

$$y = a + \frac{b}{t}; \quad (6.12)$$

$$y = a + \frac{b}{c+t}. \quad (6.13)$$

Равносторонняя гипербола ($y = a + \frac{b}{t}$) при $b > 0$ означает, что уровни ряда снижаются во времени и асимптотически приближаются к параметру a .

Например, индексы потребительских цен (декабрь текущего года к декабрю предыдущего года) за 1998–2003 гг. по России изменились по гиперболе вида:

$$\hat{y}_t = 95,557 + \frac{87,126}{t} \quad r^2 = 0,9897.$$

Уравнение характеризует падающую тенденцию ИПЦ, при которой ИПЦ не может быть меньше 95,6%. Тренд описывает 99% вариации ИПЦ, и лишь 1% ее связан с действием случайных факторов.

Если $b < 0$, то уравнение тренда $y = a + \frac{b}{t}$ характеризует тенденцию к росту уровней ряда с асимптотической границей, равной параметру a . Так, численность мужчин старше трудоспособного возраста в Санкт-Петербурге за 1979–1995 гг. характеризовалась повышающейся тенденцией: $\hat{y}_t = 296,92 - \frac{89,90}{t}$, из которой следует, что численность мужчин этой возрастной категории за этот период не превышала 296,9 тыс. чел. Этот максимум выдерживался и для 1996 и 1997 гг., а в 1998 г. он превысил эту величину, составив 303,1 тыс. чел.

Гипербола всегда $y = a + \frac{b}{c+t}$ при $b > 0$ и $c > 0$ характеризует падающую тенденцию с нижней асимптотой, равной параметру a . При $b < 0$ данная кривая означает рост уровней ряда, который происходит до определенного предела, описываемого параметром a . Рассматриваемая гипербола предполагает более плавное замедление изменения уровней, чем равносторонняя гипербола $y = a + \frac{b}{t}$.

Среди гипербол нередко используется так называемая обратная функция $y = \frac{1}{a+bt}$. Свое название она получила в связи с тем, что при сведении ее к линейному виду используются обратные значения y , т. е. $\frac{1}{y} = a + bt$. Следует отметить, что если $\frac{1}{y}$ имеет экономический смысл, то параметры данной функции интерпретируются аналогично линейному тренду. Например, предположим, что динамика трудоемкости продукции (y) характеризуется уравнением $y = \frac{1}{1+2t}$. Оно означает снижение трудоемкости и рост производительности труда на 2 единицы. В иных случаях параметры обратной функции экономически не интерпретируются.

При $b > 0$ ряд характеризуется понижающейся тенденцией, а при $b < 0$ — повышающейся.

Среди кривых с насыщением может использоваться модифицированная экспонента

$$y = c \pm ab^t, \quad (6.14)$$

где c — асимптота (верхняя для функции $y = c - ab^t$ и нижняя для функции $y = c + ab^t$). Так, при изучении тенденции роста уровня механизации труда целесообразно учитывать ограничение роста (показатель уровня механизации труда не может быть больше 100%). Если изучается динамика детской смертности, то можно установить нижнюю асимптоту, т. е. минимальное значение детской смертности, исходя из достигнутых условий жизни.

Модифицированная экспонента характеризуется постоянным отношением последовательных во времени приростов. Величина этого отношения равна параметру b .

Так, модифицированная экспонента роста уровня механизации труда $\hat{y}_t = 100 - 12,7 \cdot 0,895^t$ означает, что ежегодно скорость ряда снижается в 0,895 раза, или на 10,5%. Верхняя граница уровня механизации труда — 100%.

Величина 100% характеризует уровень ручного труда. Поэтому в уравнении (номер уравнения) интерпретируется и параметр a : $a = 12,7\%$ означает начальный уровень ручного труда. Соответственно 87,3% составит начальный уровень механизированного труда.

Модифицированная экспонента служит базовой кривой для других кривых с насыщением, а именно S-образных кривых: логистической кривой и кривой Гомперца. Тенденция развития явления в S-образных кривых охватывает три этапа: вначале довольно медленный рост, который затем убыстряется, далее сменяется уменьшением роста и приближением уровня ряда к предельному значению, т. е. к уровню насыщения.

Если в модифицированной экспоненте вместо y ввести обратную величину $\frac{1}{y}$, то получим логистическую кривую вида

$$y = \frac{1}{c + ab^t}, \quad (6.15)$$

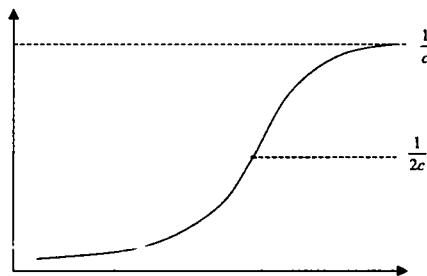


Рис. 6.10. Логистическая кривая Перла—Рида

которую называют кривой Перла—Рида. В ней верхняя асимптота составит величину $\frac{1}{c}$ (рис. 6.10).

Точка перегиба у этой кривой равна $t_p = \frac{1}{\ln b} \ln \frac{c}{a}$. Значение y в точке перегиба равно $\frac{1}{2c}$. При практических расчетах исследователь может

не иметь в полном виде S -образную кривую. Тогда точка перегиба находится за пределами наблюдаемых величин уровней ряда. В этом случае верхняя асимптота является теоретическим максимумом и ориентироваться на него в дальнейшем прогнозе достаточно проблематично.

Однако чаще сегодня применяется логистическая кривая вида

$$y = \frac{c}{1 + be^{-at}}, \quad (6.16)$$

где c — верхняя асимптота;

b и a — параметры функции;

e — основание натурального логарифма.

Механизм развития производства новых товаров описывается иногда этой кривой.

Г. Тинтнер* применил данную функцию для описания тенденции роста численности населения Швеции за 100 лет по десятилетним интервалам с 1850 по 1950 г.:

$$y = \frac{10\,328\,806}{1 + 2,1176e^{-0,14t}}.$$

Согласно этой кривой верхняя асимптота роста численности населения Швеции составила 10 328 806 человек (справочно: в 2005 г. население Швеции — 9,0 млн человек).

Максимальное значение показателя c соответствует на графике отрезку кривой, параллельному оси абсцисс. Минимальное значение функции 0 при t стремится к минус бесконечности, что обычно отсутствует при использовании модели тенденции в прогнозных расчетах.

К классу S -образных кривых относится также кривая Гомперца:

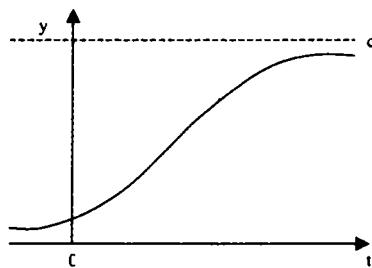
$$y = ca^{bt} \quad (6.17)$$

Она нашла применение в страховых расчетах и экстраполяции численности населения.

Верхняя асимптота соответствует значению параметра c , а нижняя равна 0, если $\ln a < 0$ (рис. 6.11, 6.12).

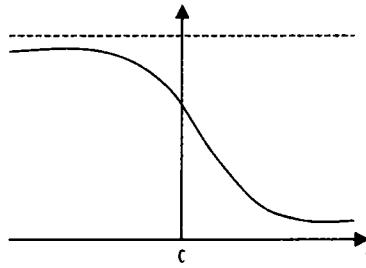
Если $\ln a > 0$, то кривая имеет нижнюю асимптоту, равную величине параметра c (рис. 6.13, 6.14):

*См.: Г. Тинтнер. Введение в эконометрию: пер. с нем. М.: Статистика, 1965. С. 291.



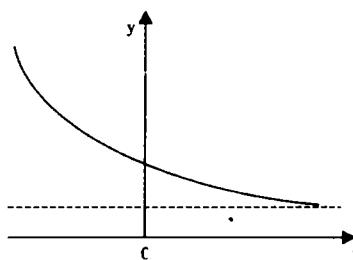
$$\ln a < 0; b < 1$$

Рис. 6.11. Кривая Гомперца



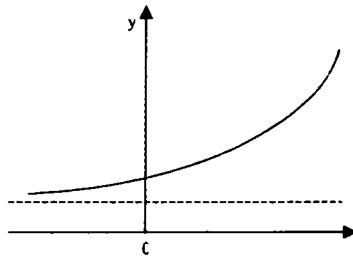
$$\ln a < 0; b > 1$$

Рис. 6.12. Кривая Гомперца



$$\ln a > 0; b < 1$$

Рис. 6.13. Кривая Гомперца



$$\ln a > 0; b > 1$$

Рис. 6.14. Кривая Гомперца

Кривая Гомперца основана на модифицированной экспоненте. Прологарифмировав уравнение кривой Гомперца, получим после замены переменных уравнение модифицированной экспоненты:

$$\lg y = \lg c + b' \lg a; \lg y = Y;$$

$$\lg c = c' \text{ и } \lg a = a' \rightarrow Y = c' + a'b'.$$

Параметр c' будет характеризовать уровень насыщения. Точкой перегиба данной кривой будет

$$t_p = \frac{1}{\ln b} \ln\left(-\frac{1}{\ln a}\right)$$

со значением функции $y_{t_p} = \frac{c}{e}$, где e — основание натурального логарифма.

Например, затраты на строительство автомобильных дорог описаны в работе Льюис* в виде кривой Гомперца: $\hat{y}_t = 4644,5 \times 0,0961435^{0,93176^t}$. Уравнение тренда показывает предельное значение

*См.: Льюис К. Д. Методы прогнозирования экономических показателей: пер. с англ. М.: Финансы и статистика, 1986. С. 111–112.

затрат 4644,5 ден. ед. Точка перегиба составляет 12 лет, ей соответствуют затраты 1708,6 ден. ед. Далее прирост затрат постепенно падает.

6.3.2. Оценка параметров уравнения тренда

При использовании полиномов разных степеней оценка параметров уравнения тренда производится методом наименьших квадратов (МНК) точно так же, как оценки параметров уравнения регрессии на основе пространственных данных. В качестве зависимой переменной рассматриваются уровни динамического ряда, а в качестве независимой переменной — фактор времени t , который обычно выражается рядом натуральных чисел: 1, 2, ..., n .

Оценка параметров нелинейных функций проводится МНК после линеаризации, т. е. приведения их к линейному виду. Рассмотрим применение МНК для некоторых нелинейных функций, которые не излагались подробно в разделе регрессии.

Для оценки параметров показательной кривой $y = ab^t$ или экспоненты $y = e^{a+bt}$ (либо $y = ae^{bt}$) путем логарифмирования функции приводятся к линейному виду: $\ln y = \ln a + t \ln b$ или экспоненты: $\ln y = a + bt$. Далее строится система нормальных уравнений:

$$\begin{cases} \sum \ln y = n \ln a + \ln b \sum t \\ \sum t \ln y = \ln a \sum t + \ln b \sum t^2 \end{cases} \quad (6.18)$$

Пример: Уровень безработицы в Псковской области за 1998–2003 г. характеризуется данными (в %):

1998 г.	1999 г.	2000 г.	2001 г.	2002 г.	2003 г.
15,9	13,3	12,5	10,3	8,2	8,1

Исходя из графика была выбрана показательная кривая $y = ab^t$. Для построения системы нормальных уравнений были рассчитаны вспомогательные величины:

$$\begin{aligned} \sum \ln y &= 14,40795; \sum_{t=1}^6 t = 21, \\ \sum t \ln y &= 47,91946; \sum_{t=1}^6 t^2 = 91. \end{aligned}$$

Система нормальных уравнений составила:

$$\begin{cases} 14,40795 = 6 \ln a + 21 \ln b \\ 47,91946 = 21 \ln a + 91 \ln b \end{cases}$$

Решая ее, получим:

$$\begin{aligned} \ln a &= 2,903001, \quad a = e^{2,903} = 18,229, \\ \ln b &= -0,14334, \quad b = e^{-0,143} = 0,866. \end{aligned}$$

Соответственно имеем экспоненту $\hat{y}_t = 18,229e^{-0,143t}$ или показательную кривую $\hat{y}_t = 18,229 \cdot 0,866^t$.

За рассматриваемый период времени уровень безработицы снижался: в среднем темп снижения составил 86,6%. Экспонента достаточно хорошо описывает исходный ряд: коэффициент детерминации оказался 0,966, т. е. данный тренд описывает 96,6% колеблемости уровня ряда, и лишь 3,4% ее связаны со случайными факторами.

Некоторую специфику имеет оценка параметров кривых с насыщением: модифицированной экспоненты, логистической кривой, кривой Гомперца. По этим функциям должна быть сначала определена асимптота. Если она может быть задана исследователем на основе анализа временного ряда, то другие параметры могут быть оценены по МНК.

Пример. Уровень механизации труда (в %) характеризуется динамическим рядом (табл. 6.3)

Таблица 6.3

Расчет параметров модифицированной экспоненты

$$y = c - ab^t$$

Годы	y	$Y = c - y$	$\ln Y$	t	$t \ln Y$	t^2	\hat{y}_t
1998	82	18	2,890	1	2,890	1	82,4
1999	85	15	2,708	2	5,416	4	85,6
2000	89	11	2,399	3	7,197	9	88,2
2001	91	9	2,197	4	8,788	16	90,3
2002	92	8	2,079	5	10,395	25	92,0
2003	93	7	1,946	6	11,676	36	93,5
2004	94	6	1,792	7	12,544	49	94,6
2005	96	4	1,386	8	11,088	64	95,6
Σ	722	—	17,397	36	69,994	204	722,2

Так как уровень механизации труда не может превышать 100%, то имеется объективно заданная верхняя асимптота $c = 100$. Для оценки параметров a и b приведем рассматриваемую функцию к линейному виду:

$$c - y = ab^t;$$

обозначим $(c-y)$ через Y и прологарифмируем:

$$\ln Y = \ln a + t \ln b.$$

Далее применим МНК и получим систему нормальных уравнений:

$$\begin{cases} \sum \ln Y = n \ln a + \ln b \sum t \\ \sum t \ln Y = \ln a \sum t + \ln b \sum t^2. \end{cases}$$

Для нашего примера исходя из данных итоговой строки табл. 6.3 имеем

$$\begin{cases} 17,397 = 8 \ln a + 36 \ln b \\ 69,994 = 36 \ln a + 204 \ln b. \end{cases}$$

Решив ее, получим: $\ln a = 3,06311$; $\ln b = -0,19744$. Соответственно потенцируя, получим: $a = 21,394$, $b = 0,8208$, т. е. уравнение $Y = 21,394 \cdot 0,8208^t$.

Переходя от Y к исходным уровням ряда, уравнение модифицированной экспоненты составит

$$\hat{y}_t = 100 - 21,394 \cdot 0,8208^t,$$

где параметр $b = 0,8208$ показывает средний коэффициент снижения уровня использования ручного труда за 1998–2005 гг. Расчетные значения y_t , т. е. \hat{y}_t , могут быть найдены путем подстановки в уравнение $\hat{y}_t = 100 - 21,394 \cdot 0,8208^t$ соответствующих значений t либо на основе уравнения $\ln Y = 3,06311 - 0,19744 t$ при компьютерной обработке определяется $\ln Y_t$ далее $100 - e^{\ln Y_t}$. Так, при $t = 8$ $\ln Y = 1,48363$ и $100 - e^{1,48363} = 100 - 4,40892 = 95,59108 = 95,6$ (см. табл. 6.3, последняя графа). Ввиду некоторой смещенности оценок (так как МНК применяется к логарифмам) $\sum y_t \neq \sum \hat{y}_t$, хотя в примере эти величины достаточно близки друг другу.

Если асимптота c не задана, то оценка параметров модифицированной экспоненты усложняется. В этих случаях могут использоваться разные методы оценивания: метод трех сумм, метод трех точек*, с помощью регрессии** — метод Брианта***.

Для логистической кривой Перла — Рида ($y = \frac{1}{c + ab^t}$) аналогично

параметры a и b могут быть найдены МНК, если асимптота c задана. Тогда данная функция преобразовывается в линейную из логарифмов: $\frac{1}{y} = c + ab^t$; $\frac{1}{y} - c = ab^t$; обозначим $\frac{1}{y} - c$ через Y и прологарифмируем,

$\ln Y = \ln a + t \ln b$. Далее параметры a и b определяются МНК, как и в примере по табл. 6.3.

Для логистической кривой вида $y = \frac{c}{1 + be^{-at}}$ параметры a и b могут быть оценены МНК, если асимптота c задана, ибо в этом случае функция линеаризуема: $1 + be^{-at} = \frac{c}{y}$; $\frac{c}{y} - 1 = be^{-at}$; обозначим через Y величину $(\frac{c}{y} - 1)$ и прологарифмируем: $Y = be^{-at}$ и $\ln Y = \ln b - at$. Далее применяя МНК, оцениваем параметры a и b .

При практических расчетах значение верхней асимптоты логистической кривой может быть определено исходя из существа развития

*См.: Четыркин Е. М. Статистические методы прогнозирования. М. Статистика, 1975. С. 114–122.

**См. там же. С. 125–130.

***См.: Льюис К. Д. Методы прогнозирования экономических показателей. М. Финансы и статистика, 1986. С. 107–109.

явления, различного рода ограничений для его роста (нормативы потребления, законодательные акты), а также графически.

Если верхняя асимптота не задана, то для оценки параметров могут использоваться разные методы: Фишера, Юла, Родса, Нейра и др. Сравнительная оценка и обзор этих методов изложены в работе Е. М. Четыркина*.

Покажем на примере расчет параметров логистической кривой по методу Фишера.

Пример. Производство продукции характеризуется следующими данными (табл. 6.4).

Таблица 6.4

Расчет параметров логистической кривой

t	y_t	$z_t = \frac{1}{2} \ln \frac{y_{t+1}}{y_{t-1}}$	$Y = \frac{403}{y_t} - 1$	$\ln Y$	\hat{y}_t
1	12	—	32,583	3,484	12,2
2	28	0,788	13,393	2,595	26,2
3	58	0,661	5,948	1,783	54,3
4	105	0,572	2,838	1,043	104,2
5	182	0,453	1,214	0,194	176,8
6	260	0,282	0,55	-0,598	256,4
7	320	0,163	0,259	-1,349	321,0
8	360	0,086	0,119	-2,125	361,7
9	380	—	0,061	-2,805	383,5
Σ	1705	—	—	2,222	1696,3

Метод Фишера основан на определении производной для логистической кривой $y_t = \frac{c}{1 + be^{-at}}$. Дифференцируя данную функцию по t , получим

$$\frac{dy_t}{dt} = ay_t(1 - \frac{y_t}{c}) \rightarrow \frac{dy_t}{dt} \cdot \frac{1}{y_t} = a(1 - \frac{y_t}{c}) = a - \frac{a}{c}y_t.$$

Обозначим темп прироста логистической кривой $\frac{dy_t}{dt} \cdot \frac{1}{y_t}$ через z_t :

$z_t = a - \frac{a}{c}y_t$, т. е. для z_t имеем линейную функцию с параметрами a и $\frac{a}{c}$.

Чтобы ее решить, необходимо оценить z_t . Предполагая, что интервалы между уровнями в ряду динамики равны, Фишер предложил приближенно оценивать z_t в виде $z_t = \frac{1}{2} \ln \frac{y_{t+1}}{y_{t-1}}$, где $t = 2, 3, \dots, n - 1$. Для наше-

*См.: Четыркин Е. М. Статистические методы прогнозирования. М.: Статистика, 1975. С. 126–133.

го примера значения z_t представлены в графе 3 табл. 6.4. Далее применяем МНК к уравнению $z_t = a - \frac{a}{c} y_t$, т. е. строим регрессию z_t от y_t , броя данные от $t = 2$ до $t = 8$. Уравнение регрессии составит $z_t = 0,806 - 0,002 y_t$. Исходя из него находим параметры a и c для логистической кривой. Параметр $a = 0,806$, т. е. свободный член уравнения $z_t = 0,806 - 0,002 y_t$. Данное уравнение статистически значимо: F -критерий Фишера равен 689,6, $R^2 = 0,996$. Соответственно для него значимы и параметры: t критерий для параметра a 47,2 и $-26,2$ для параметра $\frac{-a}{c}$.

Так как $-\frac{a}{c} = -0,002$, то $\frac{a}{c} = 0,002$ и $c = \frac{a}{0,002} = \frac{0,806}{0,002} = 403$, т. е. верхняя асимптота производства продукции составляет 403 ед.

После того как найдены параметры a и c , находим параметр b . Для этого нашу функцию $y_t = \frac{c}{1+be^{-at}}$ представим как $\frac{c}{y_t} - 1 = be^{-at}$. Обозначим через Y выражение в левой части равенства, т. е. $Y = \frac{c}{y_t} - 1$. Тогда имеем уравнение $Y = be^{-at}$. Прологарифмируем его: $\ln Y = \ln b - at$.

В этом уравнении свободным членом является $\ln b$. Его можно определить из первого уравнения системы нормальных уравнений, а именно: $\ln b = \bar{\ln Y} + \bar{at}$. Для нашего примера имеем: $\ln b = \frac{1}{9}(2,222 + 0,806 \cdot 5) = 4,2769$. Соответственно $b = e^{4,2769} = 72,016$. Таким образом, логистическая кривая запишется как:

$$y_t = \frac{403}{1 + 72,016e^{-0,806t}}.$$

Теоретические значения данной функции представлены в графе 6 табл. 6.4 (найдены путем подстановки соответствующих значений t). Они достаточно близко подходят к исходным данным: коэффициент корреляции между ними $-0,999$, $\sum y_t \neq \sum \hat{y}_t$, ввиду того что в расчетах использовались логарифмы.

Параметры кривой Гомперца $y_t = ca^{bt}$ также могут быть оценены МНК, если асимптота c задана, ибо в этом случае данная функция сводима к линейному виду: $\frac{y_t}{c} = a^{bt}$; прологарифмировав, получим

$$\lg\left(\frac{y_t}{c}\right) = bt \lg a; \text{ вторично прологарифмируем } \lg\left(\lg\frac{y_t}{c}\right) = \lg b + \lg(\lg a).$$

Обозначив $\lg\left(\lg\frac{y_t}{c}\right)$ как y^* , $\lg b$ — как B и $\lg(\lg a)$ через A , кривая Гомперца принимает линейный вид: $y^* = A + Bt$, для оценки параметров которой применим МНК.

6.3.3. Оценка адекватности модели тенденции

Модель тенденции считается адекватной реальному процессу, если теоретические (найденные по уравнению тренда) уровни ряда достаточно близко подходят к фактическим их значениям, т. е. y_t и \hat{y}_t мало отличаются друг от друга. Для оценки адекватности модели проводится анализ остатков ($e_t = y_t - \hat{y}_t$).

Модели тенденции можно сравнивать по величине остаточной суммы квадратов:

$$S_2 = \sum (y_t - \hat{y}_t)^2. \quad (6.19)$$

Чем меньше эта величина, тем в большей мере уравнение тренда подходит для описания тенденции временного ряда.

Обратимся к рассмотренному на с. 142 примеру динамики безработицы. Предположим, что было рассчитано не только уравнение экспоненты, но и уравнение линейного тренда $\hat{y}_t = 17,033 - 1,614t$.

Для линейного тренда остаточная сумма квадратов составила 1,802, а для экспоненты — 1,215. Следовательно, экспонента лучше описывает тенденцию ряда.

Другим показателем при выборе функции тренда является коэффициент детерминации R^2 . Чем выше R^2 , тем соответственно выше вероятность того, что данная модель тенденции описывает исходные данные. В нашем примере R^2 для экспоненты составил 0,966, а для линейного тренда — 0,962, подтверждая еще раз, что экспонента в большей мере подходит для описания тенденции.

Величина $1 - R^2$ отражает влияние случайной составляющей, т. е. показывает, какая доля вариации уровней динамического ряда не связана с тенденцией. Так, в нашем примере для экспоненты $1 - R^2 = 0,034$. Это означает, что лишь 3,4% вариации уровней динамического ряда не связано с рассматриваемой тенденцией.

Однако рассмотренные критерии адекватности модели тенденции (S^2 и R^2) могут привести к неправильным выводам, если не учитывать статистическую значимость параметров уравнения тренда. Общеизвестно, что можно по ряду из n точек построить полином степени $(n-1)$ и он пройдет через все точки ряда. В этом случае $S^2 = 0$, а $R^2 = 1$. Однако в таком случае уравнение $y_t = a + b_1 t + b_2 t^2 + \dots + b_{n-1} t^{n-1}$ описывает исходные уровни ряда, но не является моделью тенденции, ибо отражает не только тенденцию, но и влияние случайной компоненты.

Если для нашего примера построить параболу третьей степени $y_t = a + b_1 t + b_2 t^2 + b_3 t^3$, то получим следующее уравнение:

$$\hat{y}_t = 16,967 - 0,912t - 0,404t^2 + 0,051t^3,$$

t -критерий: 6,6 -0,3 -0,4 0,6

Как видим, все параметры при факторе времени не значимы. Между тем $R^2 = 0,979$, а $S^2 = 1,002$. Аналогичную картину дает и парабола второй степени:

$$\hat{y}_t = 18,25 - 2,527t + 0,130t^2,$$

t -критерий: 16,3 -3,5 1,3, $R^2 = 0,975$.

Параметр при t^2 статистически незначим. Поэтому, хотя для этой функции R^2 выше, а $S^2 = 1,002$ ниже, чем для экспоненты, параболу второй степени нельзя считать лучшей формой тренда.

Уравнение тренда хорошо описывает тенденцию, если отсутствует автокорреляция в остатках ($y_t - \hat{y}_t$), т. е. остатки текущего периода не коррелируют с остатками предыдущего периода.

Измерить автокорреляцию в остатках можно с помощью *коэффициента автокорреляции остатков*:

$$r_{a_e} = \frac{\overline{e_t e_{t-1}} - \overline{e_t} \cdot \overline{e_{t-1}}}{\sigma_{e_t} \cdot \sigma_{e_{t-1}}}, \quad (6.20)$$

где $e_t = y_t - \hat{y}_t$, т. е. остатки текущего периода;

e_{t-1} — остатки предыдущего периода.

Иными словами, автокорреляция в остатках оценивается так же, как и автокорреляция уровней ряда, с тем лишь отличием, что в расчетах используются остаточные величины e_t , а не уровни динамического ряда y_t .

Пример. Динамика численности детей в возрасте от 8 до 13 лет в N -ской области за последние 15 лет характеризуется параболой второй степени:

$$\hat{y}_t = 110,5 + 10,8t - 0,8t^2,$$

где y_t — численность детей, тыс. чел.;

$t = 1, 2, 3, \dots, 15$;

$R^2 = 0,967$; все параметры уравнения тренда статистически значимы.

Остаточные величины и расчет остатков представлены в табл. 6.5.

Таблица 6.5

Расчет коэффициента автокорреляции остатков

t	e_t	e_{t-1}	$e_t e_{t-1}$	e_t^2	e_{t-1}^2
1	6	—	—	—	—
2	0,2	6	1,2	0,04	36
3	-2,6	0,2	-0,52	6,76	0,04
4	-4	-2,6	10,4	16	6,76
5	-5	-4	20	25	16
6	-3,1	-5	15,5	9,61	25
7	0,7	-3,1	2,17	0,49	9,61
8	2,8	0,7	1,96	7,84	0,49
9	2,5	2,8	7	6,25	7,84
10	2,9	2,5	7,25	8,41	6,25
11	3,3	2,9	9,57	10,89	8,41

t	e_t	e_{t-1}	$e_t e_{t-1}$	e_t^2	e_{t-1}^2
12	1,6	3,3	5,28	2,56	10,89
13	-1,0	1,6	-1,6	1	2,56
14	-4	-1	4	16	1
15	-0,3	-4	1,2	0,09	16
Σ	-6*	0,3	79,07	110,94	146,85

*Без первой строки.

Используя итоговую строку табл. 6.5, рассчитаем отдельные составляющие формулы коэффициента автокорреляции остатков:

$$\bar{e}_t = \frac{\sum_{t=2}^{15} e_t}{n-1} = \frac{-6}{14} = -0,42857,$$

$$\bar{e}_{t-1} = \frac{\sum_{t=2}^{15} e_{t-1}}{n-1} = \frac{0,3}{14} = 0,021429,$$

$$\bar{e}_t e_{t-1} = \frac{\sum_{t=2}^{15} e_t e_{t-1}}{n-1} = \frac{79,07}{14} = 5,647857,$$

$$\sigma_{e_t} = \sqrt{\frac{\sum_{t=2}^{15} e_t^2}{n-1} - (\bar{e}_t)^2} = \sqrt{\frac{110,94}{14} - (-0,42857)^2} = 2,782196,$$

$$\sigma_{e_{t-1}} = \sqrt{\frac{\sum_{t=2}^{15} e_{t-1}^2}{n-1} - (\bar{e}_{t-1})^2} = \sqrt{\frac{146,85}{14} - 0,021429^2} = 3,238646.$$

Соответственно коэффициент автокорреляции остатков окажется равным

$$r_{a_e} = \frac{5,647857 - (-0,42857)(0,021429)}{2,782196 \cdot 3,238646} = 0,627824.$$

Его величина не столь мала, чтобы утверждать об отсутствии автокорреляции остатков. Очевидно, данное уравнение тренда не является наилучшим, ибо нарушена предпосылка МНК об отсутствии автокорреляции остатков.

Чтобы вывод об отсутствии (наличии) автокорреляции в остатках не был субъективным, принято рассчитывать *критерий Дарбина—Уотсона*:

$$D - W = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}. \quad (6.21)$$

Критерий Дарбина — Уотсона и коэффициент автокорреляции остатков связаны между собой соотношением:

$$D - W \cong 2(1 - r_{ae}). \quad (6.22)$$

Убедиться в этом можно, если преобразуем формулу коэффициента автокорреляции остатков и раскроем числитель формулы критерия Дарбина — Уотсона.

Параметры уравнения тренда обычно оцениваются МНК и в соответствии с предпосылками МНК $\sum_{t=1}^n e_t = 0$ и $\bar{e}_t = \frac{\sum_{t=1}^n e_t}{n} = 0$, что имеет место и в нашем примере, если учтем в расчетах первую строку табл. 6.5.

Поскольку ряд e_{t-1} сдвинут по отношению к ряду e_t на один временной интервал, то вариация в этих рядах мало отличается друг от друга и можно предположить, что $\sigma_{e_t} \approx \sigma_{e_{t-1}}$. Тогда коэффициент автокорреляции остатков можно представить в следующем виде:

$$r_{ae} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}. \quad (6.23)$$

Данная формула является приближенной, ибо если e_1 и e_n сильно различаются между собой, то $\sigma_{e_t} \neq \sigma_{e_{t-1}}$.

Для нашего примера (по данным табл. 6.5) имеем:

$$\sum_2^{15} e_t e_{t-1} = 79,07; \sum_1^{15} e_t^2 = 110,94 + 6^2 = 146,94.$$

Соответственно коэффициент автокорреляции остатков окажется равным

$$r_{ae} = \frac{79,07}{146,94} = 0,538111,$$

что несколько ниже найденного ранее значения (0,627824), но также демонстрирует заметную связь соседних величин остатков, т. е. их автокорреляцию.

Далее преобразуем формулу критерия Дарбина — Уотсона

$$\begin{aligned}
 D - W &= \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n e_t^2 - 2 \sum_{t=2}^n e_t e_{t-1} + \sum_{t=2}^n e_{t-1}^2}{\sum_{t=1}^n e_t^2} = \\
 &= \frac{\sum_{t=1}^n e_t^2 - e_1^2 + \sum_{t=1}^n e_{t-1}^2 - e_n^2 - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2},
 \end{aligned}$$

разделим каждое слагаемое числителя на $\sum_{t=1}^n e_t^2$ и предположим, что

$$\sum e_t^2 = \sum e_{t-1}^2:$$

$$1 - \frac{e_1^2}{\sum_{t=1}^n e_t^2} + 1 - \frac{e_n^2}{\sum_{t=1}^n e_t^2} - 2 \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} = 2(1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}) - \frac{e_1^2 + e_n^2}{\sum_{t=1}^n e_t^2}.$$

При большем числе наблюдений $e_1^2 + e_n^2$ значительно меньше $\sum_{t=1}^n e_t^2$.

Поэтому приближенно можно считать, что $D - W \approx 2(1 - r_{a_e})$. Из этого соотношения видно, что при полной положительной автокорреляции остатков ($r_{a_e} = 1$) критерий $D - W = 0$, а при полной отрицательной автокорреляции ($r_{a_e} = -1$) критерий $D - W = 4$. Если же автокорреляция в остатках отсутствует, т. е. $r_{a_e} = 0$, то $D - W = 2$. Иными словами, критерий Дарбина—Уотсона изменяется в пределах: $0 \leq D - W \leq 4$.

Дарбин и Уотсон разработали пороговые значения показателя $D - W$, позволяющие принять или отвергнуть гипотезу об отсутствии автокорреляции в остатках.

При заданном числе наблюдений n (длина динамического ряда) и m параметров при t в уравнении тренда (или m объясняющих переменных в уравнении регрессии) установлены при 5%-ном уровне значимости верхняя $D - W_u$ (u — upper) и нижняя (l — low) границы критерия.

Фактическое значение критерия сравнивается с табличными. Если $D - W < 2$, то возможны следующие варианты:

- 1) при $D - W < l$ нижней границы $D - W_l$ нулевая гипотеза об отсутствии автокорреляции отвергается и делается вывод о наличии положительной автокорреляции в остатках;
- 2) при $D - W > D - W_u$ (верхней границы) нулевая гипотеза об отсутствии автокорреляции не отвергается, т. е. делается вывод об

отсутствии корреляционной связи последующих остатков с предыдущими;

- 3) при $D - W_f \leq D - W \leq D - W_u$ нельзя ни отвергнуть, ни принять нулевую гипотезу об отсутствии автокорреляции в остатках, т. е. значение $D - W$ попало в область неопределенности и необходимы дальнейшие исследования, например по большему числу наблюдений.

Если фактическое значение $D - W > 2$, что означает отрицательную автокорреляцию, то с пороговыми табличными значениями сравнивается величина $4 - D - W$.

При этом возможны следующие варианты:

- 1) при $4 - D - W < D - W_f$ отвергается нулевая гипотеза об отсутствии автокорреляции и делается вывод о наличии отрицательной автокорреляции в остатках;
- 2) при $4 - D - W > D - W_u$ нулевая гипотеза об отсутствии автокорреляции в остатках принимается;
- 3) при $D - W_f \leq 4 - D - W_u$ нельзя сделать определенного вывода о наличии или отсутствии автокорреляции в остатках по имеющимся данным.

По величине критерия Дарбина—Уотсона можно определить величину коэффициента автокорреляции остатков исходя из соотношения $D - W \approx 2(1 - r_{a_1})$.

Отсюда $\frac{D - W}{2} \approx 1 - r_{a_1}$ и соответственно

$r_{a_1} \approx 1 - \frac{D - W}{2}$. Поэтому если $D - W > 2$, то $r_{a_1} < 0$, а при $D - W < 2$ имеем $r_{a_1} > 0$.

Таким образом, если фактическое значение критерия Дарбина—Уотсона не слишком отличается от 2, то можно сделать вывод об отсутствии автокорреляции в остатках. Использование критерия Дарбина—Уотсона можно показать графически (рис. 6.15).

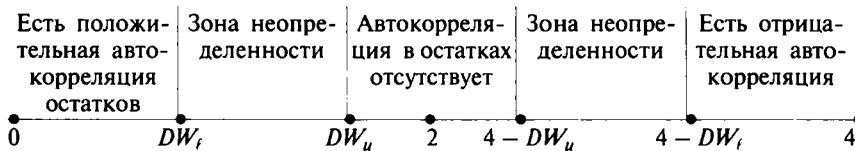


Рис. 6.15. Схема применения критерия Дарбина—Уотсона

Пример 6.1. Найдем критерий Дарбина—Уотсона применительно к данным табл. 6.5, используя формулу

$$D - W = \frac{\sum_{t=2}^{15} (e_t - e_{t-1})^2}{\sum_{t=1}^{15} e_t^2}.$$

$$\sum_{t=2}^{15} (e_t - e_{t-1})^2 = (0,2 - 6)^2 + (-2,6 - 0,2)^2 + \dots + (-4 - (-1))^2 + (-0,3 - (-4))^2 = 99,65;$$

$\sum_{t=1}^{15} e_t^2 = 146,94$, как было уже рассчитано ранее. Отсюда имеем

$$D - W = \frac{99,65}{146,94} = 0,678168, \text{ что } < 2, \text{ т. е. } r_{a_1} > 0, \text{ что и было показано}$$

ранее.

Учитывая, что $n = 15$, а $m = 2$ (тренд был найден в виде параболы второй степени), нижняя граница $D - W_\ell = 0,95$, а верхняя $D - W_u = 1,54$, т. е. фактическое значение критерия $D - W$ меньше его нижней границы $D - W_\ell$, что подтверждает вывод о наличии в остатках автокорреляции и означает плохой подбор модели тенденции к исходным данным, хотя $R^2 = 0,967$ и параметры уравнения по t -критерию значимы.

Пример 6.2. Динамика просроченной задолженности за 12 месяцев характеризовалась на предприятии уравнением тренда $\hat{y}_t = 410,1 - 6,5t$ при $t = 1, 2, \dots, 12$; $R^2 = 0,921$. Критерий Дарбина—Уотсона составил 2,226. Требуется оценить автокорреляцию в остатках.

Так как фактическое значение 2, то речь идет об оценке отрицательной автокорреляции в остатках. Приближенно

$$a_1 = \frac{1 - 2,226}{2} = -0,113, \text{ что означает слабую автокорреляцию в остатках.}$$

Убедимся в этом через сравнение фактического значения $4 - D - W$ с табличными значениями при 5%-ном уровне значимости $4 - D - W = 4 - 2,226 = 1,774$. При $n = 12$ и $m = 1$ $D - W_\ell = 0,97$ и $D - W_u = 1,33$. Так как $4 - D - W_u$ (верхней границы), то автокорреляция в остатках отсутствует и линейный тренд достаточно хорошо описывает исходный динамический ряд. Ответ на данный вопрос можно дать и графически. С этой целью необходимо нанести на числовую ось в интервале $[0; 4]$ табличные значения критерия $D - W$, а именно: $D - W_\ell = 0,97$; $D - W_u = 1,33$; $4 - D - W_u = 2,67$; $4 - D - W_\ell = 3,03$ и посмотреть, в какой интервал попадает фактическое значение критерия $D - W = 2,226$ (рис. 6.16).

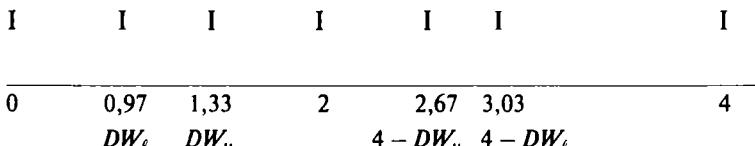


Рис. 6.16.

Фактическое значение $D - W = 2,226$ попадает в промежуток от $D - W_u$ до $4 - D - W_u$. Следовательно, можно принять нулевую гипотезу об отсутствии автокорреляции в остатках.

Критерий Дарбина—Уотсона используется и для оценки автокорреляции по регрессионным моделям, построенным по временным рядам. Однако его нельзя использовать, если в модели среди объясняющих переменных содержатся лаговые значения результативного признака. Например, инвестиции в основной капитал в текущем периоде рассматриваются как функция не только разных экономических переменных (прибыли, процентной ставки), но и достигнутого в предыдущий период размера инвестиций.

Для комплексной оценки адекватности модели тенденции можно пользоваться и другими характеристиками, которые обычно рассматриваются при оценке качества регрессионных моделей: например, средней ошибкой аппроксимации, показателями асимметрии и эксцесса для остаточных величин.

6.4. Моделирование периодических колебаний

6.4.1. Ряд Фурье

Одним из методов моделирования временного ряда с периодическими колебаниями является **ряд Фурье**. Его построение зависит от наличия или отсутствия тенденции в ряду динамики. При отсутствии тенденции, т. е. при стационарном динамическом ряде, методика построения ряда Фурье применяется непосредственно к уровням динамического ряда. Если же в ряде динамики наблюдается тенденция, то ряд Фурье применяется к отклонениям от тенденции. Соответственно эти различия учитываются и при прогнозировании:

- по стационарному временному ряду прогноздается по ряду Фурье (п. 6.4.1.1);
- по ряду с тенденцией производится суммарный прогноз, т. е. сначала строится прогноз исходя из тенденции развития уровней ряда и далее к нему прибавляется прогноз по ряду Фурье отклонений от тренда (п. 6.4.1.2).

6.4.1.1. Ряд Фурье по стационарному ряду

Стационарный ряд с периодическими колебаниями представлен на рис. 6.17.

Уровни ряда варьируют вокруг среднего значения (\bar{y}), а их колебания (волны) повторяются. Интервал времени, необходимый для того, чтобы динамический ряд начал повторяться, называется *периодом* и обозначен на графике P . Его величина (расстояние между пиками или впадинами) составляет на графике 10 месяцев (12 — 2). Если ряд имеет период P , то он, как правило, имеет также период $2P, 3P$ и т. п. В общем случае для стационарного периодического временного ряда справедливо равенство $y_t = y_t + cp$, где $c = 1, 2,$

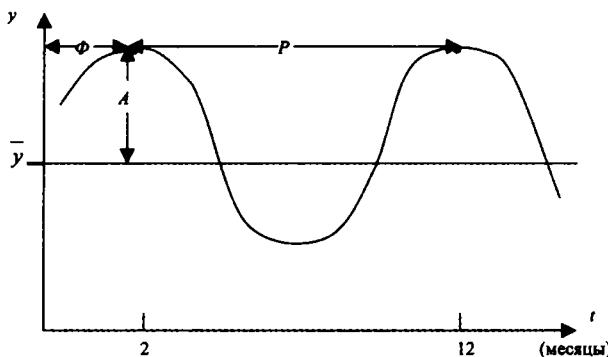


Рис. 6.17. Стационарный ряд с периодическими колебаниями

Величина, обратная периоду, называется частотой *динамического ряда* (f): $f = 1/P$. Частота указывает на число повторений цикла в единицу времени: $f = \frac{1}{10}$ в месяц (по графику).

Отклонение от среднего уровня до пика (или впадины) называется *амплитудой временного ряда* (на графике A).

Расстояние между началом отсчета времени ($t = 0$) и ближайшим пиковым значением называется *фазой* (Φ).

Стационарный периодический временной ряд можно задать четырьмя параметрами: периодом (P) или частотой (f), амплитудой (A), фазой (Φ) и средним значением (\bar{y}), что может быть представлено в виде

$$y_t = \bar{y} + A \cos W(t - \Phi),$$

где W — угловая частота, измеряемая в радианах в единицу времени и равная $W = 2\pi f$; $0 \leq W \leq 2\pi$; Φ — фаза.

Пусть, например, имеем ряд (рис. 6.18):

Для этого ряда $\bar{y} = 20$; $P = 12$ мес. $f = 1/12$; $A = y_{\max} - \bar{y} = 30 - 20 = 10$; $\Phi = 6$ мес. Учитывая, что $y_t = \bar{y} + A \cos W(t - \Phi)$, где $W = 2\pi f = 2\pi / 12 = \pi/6$, получим:

$$\text{для } t = 6 \text{ мес } y_t = 20 + 10 \cos \frac{\pi}{6} (6 - 6) = 20 + 10 \cos 0 = 20 + 10 \cdot 1 = 30;$$

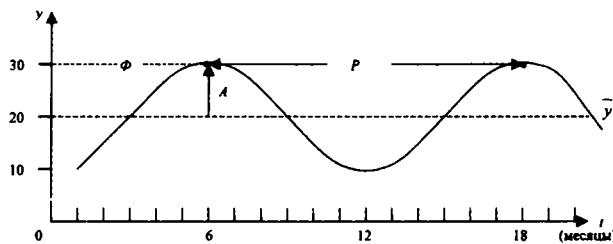


Рис. 6.18. Периодический стационарный ряд

для $t = 12$ мес. $y_t = 20 + 10 \cos \frac{\pi}{6} (12 - 6) = 20 + 10 \cos \pi = 20 + 10(-1) = 10$;

для $t = 6$ мес. $y_t = 20 + 10 \cos \frac{\pi}{6} (12 - 6) = 20 + 10 \cos \pi = 20 + 10(-1) = 30$.

Рассмотренное выражение $y_t = \bar{y} + A \cos W(t - \Phi)$ называется гармоническим представлением ряда и часто записывается через синусы и косинусы без упоминания о фазе:

$$y_t = \bar{y} + a \cos Wt + b \sin Wt, \quad (6.24)$$

где $a = A \cos \Phi$ и $b = A \sin \Phi$.

Ввиду того что $\cos^2 x + \sin^2 x = 1$, $a^2 + b^2 = A$, т. е. существует взаимосвязь между амплитудой колебаний и параметрами гармоники. Параметры гармоники также связаны с фазой ряда:

$$\operatorname{tg} \Phi = \frac{\sin \Phi}{\cos \Phi} = \frac{b}{a} \text{ или } \operatorname{arctg} \frac{b}{a} = \Phi.$$

Теоретически стационарный временной ряд с периодическими колебаниями может быть представлен как сумма среднего значения и ряда синусоид и косинусоид, что и называется рядом Фурье:

$$y_t = \bar{y} + \sum_{i=1}^{\infty} a_i \cos W_i t + \sum_{i=1}^{\infty} b_i \sin W_i t. \quad (6.25)$$

Анализируемые ряды динамики обычно имеют конечную длину N . Поэтому ряд Фурье приобретает вид

$$y_t = \bar{y} + \sum_{i=1}^n a_i \cos W_i t + \sum_{i=1}^n b_i \sin W_i t, \quad (6.26)$$

где $n = N/2$ (N — длина временного ряда).

Заменив \bar{y} параметром a_0 , ряд Фурье принимает вид

$$y_t = a_0 + \sum_{i=1}^n a_i \cos W_i t + \sum_{i=1}^n b_i \sin W_i t. \quad (6.27)$$

Оценка параметров данного уравнения обычно дается МНК. Покажем его применение для случая одной гармоники:

$$y_t = a_0 + a_1 \cos t + b_1 \sin t, \quad (6.28)$$

где t принимает значения от 0 с последующим увеличением на $\frac{2\pi}{N}$.

Система нормальных уравнений примет вид

$$\begin{cases} Na_0 + a_1 \sum \cos t + b_1 \sum \sin t = \sum y_t \\ a_0 \sum \cos t + a_1 \sum \cos^2 t + b_1 \sum \sin t \cos t = \sum y_t \cos t \\ a_0 \sum \sin t + a_1 \sum \cos t \sin t + b_1 \sum \sin^2 t = \sum y_t \sin t \end{cases} \quad (6.29)$$

В этой системе $\sum \cos t = \sum \sin t = 0$. Поэтому из первого уравнения системы получаем, что $a_o = \frac{\sum y_t}{N} = \bar{y}$. Так как $\sum \sin t \cos t = 0$ (табл. 6.6), то из второго уравнения системы получим оценку параметра a_1 , а из третьего – параметра b_1 :

$$a_1 = \frac{\sum y_t \cos t}{\sum \cos^2 t}; \quad (6.30)$$

$$b_1 = \frac{\sum y_t \sin t}{\sum \sin^2 t}. \quad (6.31)$$

Ввиду того что $\cos^2 t = \frac{1 + \cos 2t}{2}$,

$$\sum \cos^2 t = \frac{1}{2} \sum (1 + \cos 2t) = \frac{1}{2} (N + \sum \cos 2t) = \frac{1}{2} (N + 0) = \frac{N}{2}.$$

$$\text{Аналогично } \sum \sin^2 t = \frac{N}{2}, \text{ ибо } \sum \sin^2 t = \sum (1 - \cos^2 t) = N - \frac{N}{2} = \frac{N}{2}.$$

Следовательно, параметры гармонии определяются как

$$a_1 = \frac{2 \sum y_t \cos t}{N} \text{ и } b_1 = \frac{2 \sum y_t \sin t}{N}.$$

Ряд Фурье с двумя гармониками имеет вид

$$y_t = a_o + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t. \quad (6.32)$$

При этом параметры a_o , a_1 и b_1 соответствуют тем значениям, которые были найдены при рассмотрении одной гармоники. Параметры a_2 и b_2 найдем аналогично:

$$a_2 = \frac{2}{N} \sum y_t \cos 2t \text{ и } b_2 = \frac{2}{N} \sum y_t \sin 2t.$$

В общем виде для i гармоник параметры ряда Фурье определяются по формулам

$$a_i = \frac{2}{N} \sum y_t \cos W_i t \text{ и } b_i = \frac{2}{N} \sum y_t \sin W_i t.$$

Чаще всего описание временного ряда не превышает четырех гармоник.

Пример. Производство товара К по месяцам характеризуется данными (единиц):

Номер месяца	1	2	3	4	5	6	7	8	9	10
y_t	22	24	23	14	6	5	6	8	15	17
месяца	ii	12	13	14	15	16	17	18	19	20
y_t	24	25	24	18	8	5	9	14	19	23

Графическое представление этого временного ряда дано на рис. 6.19.

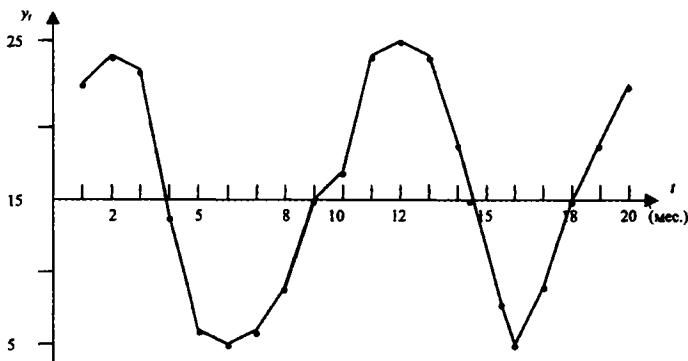


Рис. 6.19. Периодический ряд динамики производства товара К

Перед нами стационарный динамический ряд, для которого $\bar{y} = 15,45$, $\sigma_y^2 = 52,1475$.

Расчеты для определения параметров ряда Фурье представлены в табл. 6.6.

Таблица 6.6

Расчет параметров по ряду Фурье

№ п/п	y_t	t	$\cos t$	$\sin t$	$\cos 2t$	$\sin 2t$	$\cos 3t$	$\sin 3t$	$\cos 4t$	$\sin 4t$
1	22	0	1	0	1	0	1	0	1	0
2	24	$0,1\pi$	0,951	0,309	0,809	0,588	0,588	0,809	0,309	0,951
3	23	$0,2\pi$	0,809	0,588	0,309	0,951	-0,309	0,951	-0,809	0,588
4	14	$0,3\pi$	0,588	0,809	-0,309	0,951	-0,951	0,309	-0,809	-0,588
5	6	$0,4\pi$	0,309	0,951	-0,809	0,588	-0,809	-0,588	0,309	-0,951
6	5	$0,5\pi$	0	1	-1	0	0	-1	1	0
7	6	$0,6\pi$	-0,309	0,951	-0,809	-0,588	0,809	-0,588	0,309	0,951
8	8	$0,7\pi$	-0,588	0,809	-0,309	-0,951	0,951	0,309	-0,809	0,588
9	15	$0,8\pi$	-0,809	0,588	0,309	-0,951	0,309	0,951	-0,809	-0,588
10	17	$0,9\pi$	-0,951	0,309	0,809	-0,588	-0,588	0,809	0,309	-0,951
11	24	π	-1	0	1	0	-1	0	1	0
12	25	$1,1\pi$	-0,951	-0,309	0,809	0,588	-0,588	-0,809	0,309	0,951
13	24	$1,2\pi$	-0,809	-0,588	0,309	0,951	0,309	-0,951	-0,809	0,588
14	18	$1,3\pi$	-0,588	-0,809	-0,309	0,951	0,951	-0,309	-0,809	-0,588
15	8	$1,4\pi$	-0,309	-0,951	-0,809	0,588	0,809	0,588	0,309	-0,951
16	5	$1,5\pi$	0	-1	-1	0	0	1	1	0
17	9	$1,6\pi$	0,309	-0,951	-0,809	-0,588	-0,809	0,588	0,309	0,951
18	14	$1,7\pi$	0,588	-0,809	-0,309	-0,951	-0,951	-0,309	-0,809	0,588
19	19	$1,8\pi$	0,809	-0,588	0,309	-0,951	-0,309	-0,951	-0,809	-0,588
20	23	$1,9\pi$	0,951	-0,309	0,809	-0,588	0,588	-0,809	0,309	-0,951
Σ	309	-	0	0	0	0	0	0	0	0

Отсчет t ведем с 0, прибавляя каждый раз величину $\frac{2\pi}{N}$, т. е. в нашем случае $0,1\pi$ (графа t). Таблица содержит значения $\cos t, \sin t, \cos 2t, \sin 2t, \cos 3t, \sin 3t, \cos 4t, \sin 4t$ для расчета параметров уравнения с четырьмя гармониками:

$$y_t = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t + \\ + a_3 \cos 3t + b_3 \sin 3t + a_4 \cos 4t + b_4 \sin 4t.$$

Чтобы воспользоваться ранее приведенными формулами a_i и b_i , были найдены по данным табл. 6.6 следующие значения:

$$\begin{array}{ll} \Sigma y \cos t = 6,667; & \Sigma y \sin t = -17,948; \\ \Sigma y \cos 2t = 92,883; & \Sigma y \sin 2t = 26,577; \\ \Sigma y \cos 3t = -2,698; & \Sigma y \sin 3t = -10,568; \\ \Sigma y \cos 4t = -16,753; & \Sigma y \sin 4t = 11,274. \end{array}$$

Ввиду того что в нашем примере $\frac{2}{N} = 0,1$, a_i и b_i составят

$$\begin{array}{ll} a_1 = 0,6667; & b_1 = -1,7948; \\ a_2 = 9,2883; & b_2 = -2,6577; \\ a_3 = -0,2698; & b_3 = -1,0568; \\ a_4 = -1,6753; & b_4 = 1,1274. \end{array}$$

Соответственно ряд Фурье представит собой следующее выражение:

$$y_t = \bar{y} + \sum_{i=1}^4 C_{it},$$

где C_{it} — гармоники вида

$$C_{it} = a_i \cos W_i t + b_i \sin W_i t.$$

Для нашего примера соответствующие гармоники даны в табл. 6.7.

Таблица 6.7

Четыре периодические составляющие динамического ряда производства продукции К

Номер гармоники	Гармоническая функция	R^2
1	$0,6667 \cos t - 1,7948 \sin t$	0,0351
2	$9,2883 \cos 2t - 2,6577 \sin 2t$	0,930
3	$-0,2698 \cos 3t - 1,0568 \sin 3t$	0,942
4	$-1,6753 \cos 4t + 1,1274 \sin 4t$	0,976

Ряд Фурье с одной гармоникой, тогда

$$\hat{y}_t = 15,45 + 0,6667 \cos t - 1,7948 \sin t,$$

а с двумя: $\hat{y}_t = 15,45 + 0,6667 \cos t - 1,7948 \sin t + 9,2883 \cos 2t - 2,6577 \sin 2t$.

Аналогично записывается модель с тремя и четырьмя гармониками.

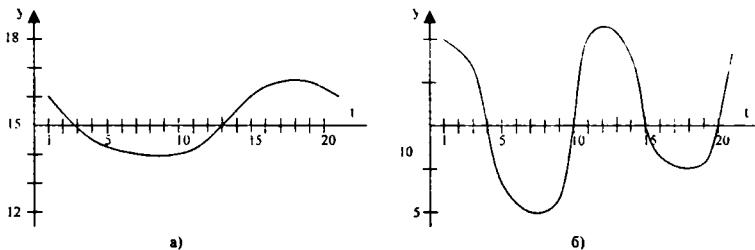


Рис. 6.20: а) ряд с одной гармоникой; б) ряд с двумя гармониками

Как видно из последней графы табл. 6.7, модель с двумя гармониками хорошо описывает исходный динамический ряд. Хотя R^2 при увеличении числа гармоник и возрастает, но параметры модели для третьей гармоники по t -критерию Стьюдента оказываются статистически незначимыми. Поэтому при выборе модели лучше предпочесть модель с двумя гармониками.

Как видно из графика (рис. 6.19), для рассматриваемого временного ряда амплитуда колебаний (A) приближается к 10, что и имеет место для уравнения с двумя гармониками:

$$A_2 = \sqrt{a_2^2 + b_2^2} = \sqrt{9,2883^2 + 2,6577^2} = 9,66,$$

т. е. интервал, через который ряд начинает повторяться, равен 10 месяцам.

Для уравнения с одной гармоникой период повторения составит 20 месяцев и, естественно, выравненный динамический ряд (ряд \hat{y}_t) плохо аппроксимирует исходные данные (рис. 6.20).

Для прогноза, используя ряд Фурье с двумя гармониками, в уравнение подставляется следующее по порядку значение t : в примере на 21-й месяц $t = 2\pi$. Учитывая что $\cos 2\pi = \cos 4\pi = 1$, а $\sin 2\pi = \sin 4\pi = 0$, прогноз окажется следующим:

$$\begin{aligned} y_p &= 15,45 + 0,6667 \cos 2\pi - 1,7948 \sin 2\pi + \\ &+ 9,2883 \cos 4\pi - 2,6577 \sin 4\pi = 25,4 \text{ ед.} \end{aligned}$$

6.4.1.2. Ряд Фурье по ряду с тенденцией

В экономике чаще встречаются динамические ряды с тенденцией. В этом случае при наличии периодических колебаний ряд Фурье может быть использован, если привести ряд к стационарному виду. Для этой цели можно найти линейный тренд ($\hat{y}_t = a + bt$) и применить ряд Фурье к остаточным величинам ($e_t = y_t - \hat{y}_t$). Возможен иной путь: ряд Фурье строится по первым разностям, что равносильно учету линейного тренда. Иными словами, по ряду динамики определяются цепные абсолютные приrostы: $\Delta_t = y_t - y_{t-1}$, которые далее используются как информационная база для построения ряда Фурье.

Пример. Динамика средних цен на товар по месяцам года характеризуется графиком (рис. 6.21).

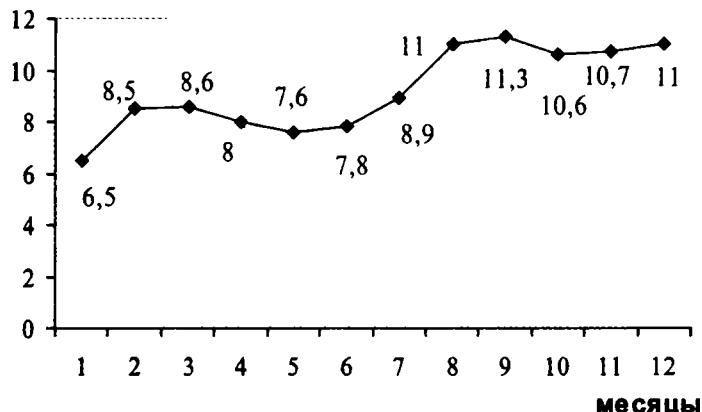


Рис. 6.21. Ряд с тенденцией и периодической составляющей

Как видно из графика, ряд имеет тенденцию. Уравнение линейного тренда составило $\hat{y}_t = 6,683 + 0,388t$, где $t = 1, 2, \dots, 12$. Оно описывает 73,7% вариации средних цен ($R^2 = 0,737$) и статистически значимо, ибо $F = 28$ при табличном значении 4,84 (для 5%-ного уровня значимости). Подставив в данное уравнение соответствующие значения t , получим расчетные величины средних цен (\hat{y}_t) и остатки ($e_t = y_t - \hat{y}_t$). График остатков представлен на рис. 6.22.

Остатки (e_t) представляют собой стационарный ряд и хорошо описываются рядом Фурье с двумя гармониками:

$$e_t = 0,123 \cos t - 0,296 \sin t - 0,137 \cos 2t + 1,005 \sin 2t; R^2 = 0,880.$$

В данном уравнении свободный член (a_0) отсутствует, ибо $\sum e_t = 0$.

Модель рассматриваемого динамического ряда представит собой систему уравнений:

$$\begin{cases} y_t = a + bt + e_t, \\ e_t = a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t. \end{cases}$$

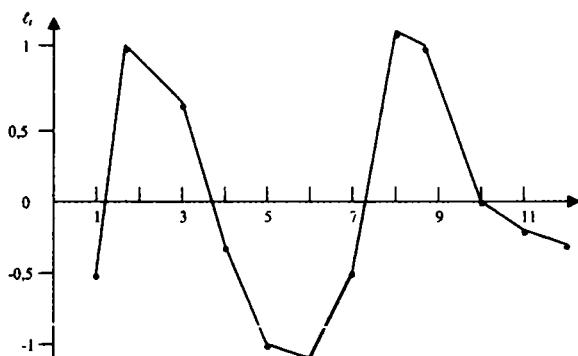


Рис. 6.22. График остатков от линейного тренда

В нашем примере модель ряда имеет вид:

$$\begin{cases} y_t = 6,683 + 0,388t + e_t \\ e_t = 0,123 \cos t - 0,296 \sin t - 0,137 \cos 2t + 1,005 \sin 2t. \end{cases}$$

Подставив в данную систему значения t (для тренда — 1, 2, ..., 12; для остатков — $0, \frac{\pi}{6}, \frac{2\pi}{6}, \dots, \frac{11\pi}{6}$), найдем теоретические значения уровней ряда динамики, которые тесно коррелируют с исходными данными ($R = 0,9855$).

Аналогично поступим и для прогноза на следующий месяц года, построим суммарный прогноз: прогноз по тренду плюс прогноз по ряду Фурье для остаточных величин.

В рассматриваемом примере прогноз на 14-й месяц составит:

- a) по тренду: $6,683 + 0,388 \cdot 14 = 12,122$;
- б) по остаткам: $0,123 \cos \frac{13\pi}{6} - 0,296 \sin \frac{13\pi}{6} \cdot 0,137 \cos \frac{13\pi}{3} + 1,005 \sin \frac{13\pi}{3} = 0,123 \cdot 0,866 - 0,296 \cdot 0,5 - 0,137 \cdot 0,5 + 1,005 \cdot 0,866 = 0,760$;
- в) итого: 12,88.

Ряд Фурье может использоваться также для отображения и прогнозирования динамики с сезонными колебаниями. При этом амплитуда колебаний не должна превышать 4 квартала или 12 месяцев, ибо сезонные — это внутригодичные колебания. Вместе с тем сезонные колебания могут изучаться и с помощью иных моделей, позволяющих не только учесть сезонность, но и измерить ее количественно, что имеет, несомненно, практическое значение.

6.4.2. Аддитивная модель сезонности

Аддитивная модель предполагает агрегирование отдельных компонент уровней динамического ряда. В зависимости от того, есть или нет тенденция в ряду динамики, она может иметь следующий вид:

$$y_t = \bar{y} + S + \xi \quad \text{— при отсутствии тенденции;}$$

$$y_t = \hat{y}_t + S + \xi \quad \text{— при наличии тенденции;}$$

где y_t — уровень динамического ряда в период времени t ;

\bar{y} — средний уровень динамического ряда;

\hat{y}_t — теоретический уровень ряда согласно тенденции;

S — сезонная составляющая, измеренная в тех же единицах, что и уровень ряда;

ξ — случайная компонента, измеренная в тех же единицах, что и уровень ряда.

6.4.2.1. Аддитивная модель при отсутствии тенденции

При отсутствии тенденции в ряду динамики общая колеблемость уровней ряда раскладывается на две составляющие: влияние сезонности S и влияние случайности ξ . Тогда имеем равенство

$$(y_t - \bar{y}) = (\bar{y}_{S_j} - \bar{y}) + (y_t - \bar{y}_{S_j}), \quad (6.33)$$

где \bar{y}_{S_j} — средний уровень соответствующего периода внутри года (месяца, квартала) за ряд лет.

В данном равенстве величина $(\bar{y}_{S_j} - \bar{y})$ отражает влияние сезонности, а величина $(y_t - \bar{y}_{S_j})$ характеризует влияние случайной компоненты. Средние значения \bar{y} и \bar{y}_{S_j} определяются по средней арифметической простой: $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$; $\bar{y}_{S_j} = \frac{1}{K} \sum_{t=1}^k y_t$,

где n — число уровней в ряду динамики (квартале-годы, месяце-годы);

K — число лет в ряду динамики.

Пример. Реализация детских велосипедов по магазину характеризуется данными (табл. 6.8):

Таблица 6.8

Кварталы	Годы			Итого	\bar{y}_{S_j}	S_j
	1	2	3			
I	25	30	26	81	27	-62,25
II	125	120	133	378	126	36,75
III	180	162	180	522	174	84,75
IV	30	30	30	90	30	-59,25
Итого	360	342	369	1071	89,25	0

Итоговые данные по годам не обнаруживают четкой тенденции, и можно применить рассмотренную ранее модель. Общий среднеквартальный уровень (\bar{y}) составил $89,25$ ($\frac{1071}{12}$), и если бы не было влияния сезонности, то каждый квартал реализация составляла бы 89 единиц.

Однако под воздействием сезонности она была в I и IV кварталах существенно ниже, а во II и III кварталах — выше среднего уровня. Измерение сезонности при аддитивной модели означает расчет абсолютных показателей сезонности:

$$S_j = \bar{y}_{S_j} - \bar{y} \text{ (см. последнюю графу табл. 6.8).}$$

Сумма абсолютных показателей сезонности за год равна нулю ($\sum S_j = 0$). Уровень ряда соответствующего квартала (месяца) можно представить с помощью линейной модели с фиктивными переменными:

$$y_t = a + b_1 z_1 + b_2 z_2 + b_3 z_3 + \xi_t,$$

где z_1, z_2, z_3 – фиктивные переменные для I, II и III кварталов, принимающие значение 1 для рассматриваемого квартала и 0 – для остальных.

Так, $z_1 = 1$ только для I квартала, $z_2 = 1$ – для II квартала и $z_3 = 1$ – для III квартала. Применяя к матрице исходных данных (y_t, z_1, z_2, z_3) МНК, получим оценку параметров a, b_1, b_2, b_3 . В данной модели сравнение ведется с четвертым кварталом, для которого $z=0$.

Параметры модели интерпретируются следующим образом. Параметр $a = \bar{y}_S$ для IV квартала, т. е. \bar{y}_{S_4} ; параметры $b_j = \bar{y}_{S_j} - \bar{y}_{S_4}$, т. е. показывают, на сколько средний уровень j -го квартала ниже или выше среднего уровня за IV квартал. Такая интерпретация параметров обусловлена спецификой фиктивных переменных и применением к модели МНК.

Использование МНК в нашем случае приводит к системе нормальных уравнений:

$$\begin{cases} \sum y_t = na + b_1 \sum z_1 + b_2 \sum z_2 + b_3 \sum z_3 \\ \sum y_t z_1 = a \sum z_1 + b_1 \sum z_1^2 + b_2 \sum z_1 z_2 + b_3 \sum z_1 z_3 \\ \sum y_t z_2 = a \sum z_2 + b_1 \sum z_1 z_2 + b_2 \sum z_2^2 + b_3 \sum z_2 z_3 \\ \sum y_t z_3 = a \sum z_3 + b_1 \sum z_1 z_3 + b_2 \sum z_2 z_3 + b_3 \sum z_3^2 \end{cases} \quad (6.34)$$

В этой системе $n = 12$; $\sum z_1 = \sum z_2 = \sum z_3 = 3$ (число исследуемых лет); $\sum z_1 z_2 = \sum z_1 z_3 = \sum z_2 z_3 = 0$; $\sum y_t z_1 = \sum y_{j=1}$ (итог за I квартал по строке 1 табл. 6.8); $\sum y_t z_2 = \sum y_{j=2}$ (итог по строке 2 табл. 6.8); $\sum y_t z_3 = \sum y_{j=3}$. Таким образом, искомая система уравнений составит

$$\begin{cases} \sum y_t = 12a + 3b_1 + 3b_2 + 3b_3 = 1071 \\ \sum y_t z_1 = 3a + 3b_1 = 81 \\ \sum y_t z_2 = 3a + 3b_2 = 378 \\ \sum y_t z_3 = 3a + 3b_3 = 522 \end{cases}$$

Вычтем из первого уравнения три последующих: $\sum y_{j=4} = 3a$ и получим $a = \frac{\sum y_{j=4}}{3} = \bar{y}_{S_{j=4}}$, т. е. параметр a отражает средний уровень за IV квартал.

Разделив уравнения 2, 3 и 4 на 3 и подставив значение параметра a , найдем оценки параметров b_1, b_2 и b_3 , а именно:

$$\frac{\sum y_{j=1}}{3} - \bar{y}_{S_{j=4}} = \bar{y}_{S_{j=1}} - \bar{y}_{S_{j=4}} = b_1;$$

$$\frac{\sum y_{j=2}}{3} - \bar{y}_{S_{j=4}} = \bar{y}_{S_{j=2}} - \bar{y}_{S_{j=4}} = b_2;$$

$$\frac{\sum y_{j=3}}{3} - \bar{y}_{S_{j=4}} = \bar{y}_{S_{j=3}} - \bar{y}_{S_{j=4}} = b_3.$$

Используя данные табл. 6.8, получим:

$$b_1 = 27 - 30 = -3;$$

$$b_2 = 126 - 30 = 96;$$

$$b_3 = 174 - 30 = 144.$$

Таким образом, модель уровней ряда для табл. 6.8 составит:

$$y_t = 30 - 3z_1 + 96z_2 + 144z_3 + \xi_t.$$

Ее можно получить, не обращаясь к данным табл. 6.8, а используя матрицу исходных данных вида

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \dots \\ y_n \end{pmatrix}; Z = \begin{pmatrix} z_1 & z_2 & z_3 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

На основании модели $\hat{y}_t = 30 - 3z_1 + 96z_2 + 144z_3$ могут быть найдены средние значения для каждого квартала \bar{y}_{S_j} , а именно: $\bar{y}_{S_j} = b_j + a$, т. е.

$$\bar{y}_{S_I} = b_1 + a = -3 + 30 = 27;$$

$$\bar{y}_{S_{II}} = b_2 + a = 96 + 30 = 126;$$

$$\bar{y}_{S_{III}} = b_3 + a = 144 + 30 = 174 \text{ (см. табл. 6.8).}$$

Иными словами, имея модель уровней динамического ряда, одновременно имеем значения средних уровней для каждого квартала. Это позволяет исходя из модели оценить сезонные колебания по их абсолютной величине: $S_j = \bar{y}_{S_j} - \bar{y}$. Учитывая, что $\bar{y} = 89,25$, получим те же значения S_j , что и по табл. 6.8.

Используя модель с фиктивными переменными при решении на компьютере, мы получаем также оценку качества модели через R^2 и F-критерий Фишера. В рассматриваемом примере $R^2 = 0,993$; $F = 403$, т. е. модель описывает 99,3% вариации исходных данных, а на долю случайной колеблемости приходится 0,7%.

Рассматриваемая модель может быть использована в прогнозировании путем подстановки в нее соответствующего значения z_j . Поскольку наш ряд динамики не имеет тенденции, то прогнозные значе-

ния составляют те же величины, что и средние уровни для каждого квартала, т. е. \bar{y}_{S_j} .

6.4.2.2. Аддитивная модель при наличии тенденции

При наличии тенденции в ряду динамики общая колеблемость уровней ряда раскладывается на три составляющие:

$$(y_t - \bar{y}) = \begin{array}{l} \text{общая} \\ \text{вариация} \end{array} + \begin{array}{l} \text{влияние} \\ \text{тенденции} \end{array} + \begin{array}{l} \text{влияние} \\ \text{сезонности} \end{array} + \begin{array}{l} \text{влияние} \\ \text{случайности} \end{array} \quad (6.35)$$

где y_S — тренд с учетом сезонности, т. е. уровень ряда, обусловленный одновременно влиянием тенденции и сезонности.

Графически влияние этих составляющих представлено на рис. 6.23.

Чем больше угол наклона линии тренда (\hat{y}_t) к среднему значению ряда (\bar{y}), тем больше влияние тенденции (рис. 6.23, а). Чем больше плавная кривая (y_S) отклоняется от линии тренда (\hat{y}_t), тем значительнее влияние сезонности (рис. 6.23, б). Чем ближе фактические уровни ряда (y_t) подходят к плавной линии точек y_S , тем меньше влияние случайности.

Существуют разные подходы расчета отдельных составляющих рассматриваемой аддитивной модели, которые зависят от того, как найдены выровненные данные (\hat{y}_t), отражающие тенденцию, а именно:

- а) путем исключения сезонности из данных;
- б) включая сезонность, т. е. выравнивая непосредственно исходные уровни динамического ряда.

Чаще предпочтение отдается первому подходу, при котором вначале производится выравнивание динамического ряда методом скользящих средних для выделения сезонных колебаний (S), а далее, исключив их, определяется тренд без сезонных колебаний ($T = \hat{y}_t$).

Построение модели включает в себя следующие расчеты:

1. Нахождение сглаженных уровней динамического ряда методом скользящих средних (\hat{y}_t).
2. Оценка сезонной компоненты (S_j) и ее корректировка (\hat{S}_j).
3. Элиминирование сезонной компоненты из исходных данных временного ряда ($y_t - \hat{S}_j$).

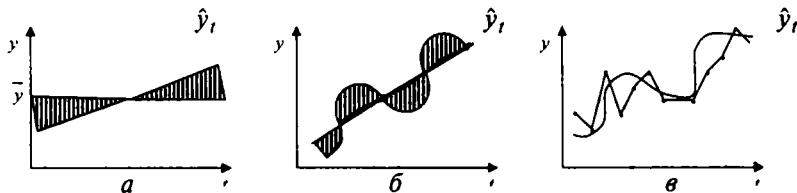


Рис. 6.23. Разложение колеблемости уровней динамического ряда на составляющие

4. Построение уравнения линейного тренда по уровням ряда с элиминированием сезонности.
5. Расчет выровненных значений трендовой составляющей (\hat{y}_t).
6. Расчет теоретических уровней ряда с учетом сезонности ($\hat{y}_t + \hat{S}_j$).
7. Расчет случайной компоненты (ξ), позволяющий оценить далее качество построенной модели.

Построение аддитивной модели для временного ряда с тенденцией рассмотрим на примере (табл. 6.9).

Пример. Число официально зарегистрированных безработных в районе (y_t — тыс. чел.), а также расчет сглаженных уровней методом скользящих средних (\tilde{y}) и сезонной компоненты (S_j) представлены в табл. 6.9.

Таблица 6.9

Расчет абсолютных показателей сезонности для аддитивной модели

Кварталы, j	2003 г.			2004 г.			2005 г.		
	y_t	\tilde{y}_t	S_j	y_t	\tilde{y}_t	S_j	y_t	\tilde{y}_t	S_j
I	10	—	—	9	8,125	0,875	7	6,625	0,375
II	8	—	—	7	7,875	-0,875	5	6,25	-1,25
III	6	8,625	-2,625	5	7,5	-2,5	4	—	—
IV	11	8,375	2,625	10	7	3	8	—	—

Ввиду того что сезонность характеризует внутригодичные колебания при сглаживании уровней ряда (y_t) методом скользящих средних, период скольжения должен быть равен году, чтобы можно было погасить влияние сезонности.

В нашем примере сглаживание ряда должно быть произведено с помощью четырехчленной скользящей средней. Однако выравнивание ряда скользящими средними с четным периодом скольжения (4 или 12 — по месячным данным) не позволяет отнести сглаженный уровень (\tilde{y}_t) к конкретному периоду времени (кварталу, месяцу). Так, первая скользящая средняя составит $(10 + 8 + 6 + 11) / 4 = 8,75$. Она относится к периоду между II и III кварталами 2003 г. Вторая скользящая средняя, относящаяся к середине между III и IV кварталами, составит $(8 + 6 + 11 + 9) / 4 = 8,5$.

Чтобы иметь оценку сглаженного уровня для конкретного квартала (месяца), проводится операция центрирования, т. е. находится средняя величина из двух смежных скользящих средних. Так, используя среднюю арифметическую простую из двух ранее найденных скользящих средних, получим центрированную скользящую среднюю, которую можно отнести к III кварталу 2003 г.: $(8,75 + 8,5) / 2 = 8,625$.

Рассмотренная процедура центрирования может быть упрощена за счет расчета центрированной скользящей средней (\tilde{y}_t) по формуле

$$\tilde{y}_t = \frac{1/2 y_1 + y_2 + y_3 + y_4 + 1/2 y_5}{4}.$$

Для нашего примера центрированная скользящая средняя для III квартала 2003 г. окажется равной

$$\tilde{y}_3 = \frac{1/2 \cdot 10 + 8 + 6 + 11 + 1/2 \cdot 9}{4} = 8,625,$$

т. е. столько же, сколько было показано ранее. Соответственно следующая центрированная скользящая средняя составит

$$\tilde{y}_4 = \frac{1/2 \cdot 8 + 6 + 11 + 9 + 1/2 \cdot 7}{4} = 8,375.$$

Аналогично находим и другие центрированные скользящие средние (табл. 6.9). При этом слаженный ряд сокращается на 4 уровня, что отражено в табл. 6.9 прочерком для I и II кварталов 2003 г. и III и IV кварталов 2005 г.

Слаженные уровни (\tilde{y}_t) характеризуют движение числа безработных, в котором погашено влияние сезонности. Измерить сезонность в виде абсолютной величины можно как $S_j = y_t - \tilde{y}_t$ (табл. 6.9).

Анализируя абсолютные показатели сезонности (S_j), видим, что под воздействием сезонного фактора численность безработных снижается во II и III кварталах и возрастает в IV и I кварталах.

Поскольку анализируются данные за ряд лет (как правило, не менее трех), то для каждого периода года (квартала, месяца) получается несколько показателей сезонности: в нашем примере — по два для каждого квартала. Поэтому, чтобы иметь оценку сезонной составляющей для одноименных периодов, рассчитываются средние показатели сезонности как средняя арифметическая простая:

$$\bar{S}_j = \frac{1}{n} \sum S_j,$$

где j — номер периода.

Сезонные колебания взаимопогашаются в течение года. Поэтому $\sum \bar{S}_j$ должна быть равна нулю. При практических расчетах возможна незначительная погрешность, когда $\sum \bar{S}_j \neq 0$. В этом случае проводится корректировка сезонной компоненты:

$$\hat{S}_j = \bar{S}_j \pm \Delta_{\text{поправки}}, \quad (6.36)$$

где \hat{S}_j — скорректированная величина сезонной компоненты;

$$\Delta_{\text{поправки}} = \frac{1}{4} (\sum \bar{S}_j - 0).$$

Если $\Delta_{\text{поправки}} > 0$, то на эту величину необходимо уменьшить значения \bar{S}_j . Если же $\Delta_{\text{поправки}} < 0$, то на эту величину увеличиваются значения \bar{S}_j .

Для нашего примера корректировка сезонной компоненты представлена в табл. 6.10.

Таблица 6.10
Корректировка сезонной компоненты для аддитивной модели

Кварталы	I	II	III	IV	Сумма
\bar{S}_j	0,625	-1,0625	-2,5625	2,8125	-0,1875
$\hat{S}_j = \bar{S}_j + 0,0469$	0,6719	-1,0156	-2,5156	2,8594	0

Найденная сезонная компонента (\hat{S}_j) используется далее в анализе для:

- исключения сезонности из данных, чтобы получить более ясную картину тенденции;
- включения сезонности в модель для прогноза.

Исключение сезонности из данных временного ряда по аддитивной модели производится как

$$y_t - \hat{S}_j = Y_t,$$

где Y_t — уровни динамического ряда без сезонности, т. е. они отражают тенденцию со случайной составляющей (табл. 6.11).

Таблица 6.11
Разложение уровней ряда по аддитивной модели

Годы	Кварталы	y_t	\hat{S}_j	Y_t	\hat{y}_t	$y_S = \hat{y}_t + \hat{S}_j$	$\xi_t = y_t - y_S$
2003	I	10	0,67	9,33	9,41	10,08	-0,08
	II	8	-1,02	9,02	9,06	8,04	-0,04
	III	6	-2,51	8,51	8,71	6,20	-0,20
	IV	11	2,86	8,14	8,37	11,23	-0,23
2004	I	9	0,67	8,33	8,02	8,69	0,31
	II	7	-1,02	8,02	7,67	6,65	0,35
	III	5	-2,51	7,51	7,33	4,82	0,18
	IV	10	2,86	7,14	6,98	9,84	0,16
2005	I	7	0,67	6,33	6,63	7,30	-0,30
	II	5	-1,02	6,02	6,29	5,27	-0,27
	III	4	-2,51	6,51	5,94	3,43	0,57
	IV	8	-2,86	5,14	5,59	8,45	-0,45
	Σ	90	0	90	90	90	0

Проведя аналитическое выравнивание уровней ряда без сезонности (Y_t) по линейному тренду, получим $\hat{Y}_t = 9,755 - 0,347t + \xi_t$, где $t=1,2,\dots,12$. Данное уравнение отражает собственно влияние тенденций, ибо оно найдено по уровням ряда при устранении воздействия се-

зонного фактора. Подставив в него соответствующие значения t , получим теоретические (\hat{y}_t) значения уровней ряда согласно тенденции (табл. 6.11).

Полученное уравнение тренда $Y_t = 9,755 - 0,347t + \xi_t$, позволяет иметь прогнозное значение (например, на I квартал 2006 г.) уровня ряда согласно тенденции, т. е. при элиминировании сезонности. Для нашего примера $Y_{1\text{кв.}2006г.} = 9,755 - 0,347 \cdot 13 = 5,244$ тыс. чел.

Далее для прогноза численности безработных I квартала 2006 г. необходимо включить в прогноз сезонную компоненту ($\hat{S}_{1\text{квартала}}$), т. е. находим тренд с учетом сезонности (y_S):

$$y_{S_{1\text{кв.}2006г.}} = \hat{y}_{t=13} + \hat{S}_{1\text{кв.}}$$

По данным примера суммарный прогноз численности безработных на I квартал 2006 г. окажется равным

$$y_S = 5,244 + 0,672 = 5,916 \text{ тыс. чел.}$$

Прибавив сезонную компоненту (\hat{S}_j) к уровням ряда согласно тенденции (\hat{y}_t), получим по модели тренд с учетом сезонной волны (y_S):

$$y_S = \hat{y}_t + \hat{S}_j \text{ (табл. 6.11).}$$

Зная тренд с учетом сезонности, можно определить величину случайной составляющей для каждого уровня динамического ряда как $\xi_t = y_t - y_S$ (последняя графа табл. 6.11). Случайная компонента ξ_t представляет собой абсолютную величину ошибки модели для каждого уровня временного ряда. Обобщенную характеристику роли случайной компоненты для оценки качества аддитивной модели можно получить, рассчитав по модели коэффициент детерминации R^2 и на его основе — величину $1 - R^2$.

Для рассматриваемого примера $\sum \xi_t^2 = 1,0678$, а $\sum (y_t - \bar{y})^2 = 55$. Соответственно $R^2 = 1 - \frac{1,0678}{55} = 0,981$, т. е. аддитивная модель в нашем примере объясняет 98,1% вариации исходных уровней динамического ряда. На долю случайной составляющей приходится 1,9% колеблемости уровней.

6.4.2.3. Применение фиктивных переменных к аддитивной модели с тенденцией

Аддитивная модель уровней динамического ряда при наличии тенденции и сезонности может быть построена как модель регрессии с включением в нее фактора времени (t) и фиктивных переменных (z). Так, при квартальном разрезе информации модель примет вид

$$y_t = a + bt + c_1 z_1 + c_2 z_2 + c_3 z_3 + \xi_t. \quad (6.37)$$

Фактор времени t в этой модели позволит учесть влияние тенденции. Сезонный фактор представлен фиктивными переменными, число которых должно быть на единицу меньше числа рассматриваемых периодов года. При квартальных данных за ряд лет $z = 3$, а при помесяч-

ных $z = 11$. Каждая из фиктивных переменных принимает значения 1 — для рассматриваемого квартала и 0 — для остальных. Предположим, что z_1 , z_2 и z_3 соответствуют учету сезонного фактора в I, II и III кварталах соответственно, т. е.

$$z_1 = \begin{cases} 1 - \text{I квартала}, \\ 0 - \text{для остальных}; \end{cases}$$

$$z_2 = \begin{cases} 1 - \text{II квартала}, \\ 0 - \text{для остальных}; \end{cases}$$

$$z_3 = \begin{cases} 1 - \text{III квартала}, \\ 0 - \text{для остальных}; \end{cases}$$

Так как фиктивные переменные z принимают только значения 1 и 0, то практически мы имеем модель тенденции для каждого квартала:

- для I квартала: $y_t = (a + c_1) + bt + \xi_t$;
- для II квартала: $y_t = (a + c_2) + bt + \xi_t$;
- для III квартала: $y_t = (a + c_3) + bt + \xi_t$;
- для IV квартала: $y_t = a + bt + \xi_t$.

Иными словами, параметры при фиктивных переменных c_1, c_2, c_3 отражают изменение уровня ряда соответствующего квартала под воздействием сезонности по сравнению с IV кварталом.

Рассмотрим применение данной модели к нашему примеру с численностью безработных. Матрица исходных данных составит

$$y_t = \begin{pmatrix} 10 \\ 8 \\ 6 \\ 11 \\ 9 \\ \dots \\ 8 \end{pmatrix}, \quad t = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \dots \\ 12 \end{pmatrix}, \quad z_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ - \\ \dots \\ 0 \end{pmatrix}, \quad z_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \quad z_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}.$$

Применяя традиционный МНК, получим следующее уравнение:

$$\hat{y}_t = 12,417 - 0,344t - 2,031z_1 - 3,688z_2 - 5,010z_3$$

$$t \quad 38,5 \quad -11 \quad -6,7 \quad -12,5 \quad -17,3$$

$$R^2 = 0,984 \quad f = 108,25$$

Статистически значимо как само уравнение (по F -критерию Фишера), так и отдельные его параметры (по t -критерию Стьюдента). Уравнение описывает 98,4% вариации исходных уровней. Доля случайной компоненты в общей дисперсии y_t составляет 5,6%.

Параметр $b = -0,344$ указывает на тенденцию снижения уровней ряда при элиминировании сезонности. Его величина по содержанию и численно практически совпадает с величиной параметра b в уравнении

ний тренда по данным с устранением сезонности, найденным ранее по табл. 6.11: $Y_t = 9,755 - 0,347t + \xi_t$. Иными словами, ежеквартально независимо от сезона уровни ряда снижаются в среднем на 0,34 тыс. чел.

Параметры c_1, c_2, c_3 показывают, что в I, II и III кварталах уровни ряда независимо от влияния тенденции были в среднем ниже, чем в IV квартале, на соответствующие величины. Так, если в нашу модель подставим значения t , а фактор сезонности уберем, принимая $z_1 = z_2 = z_3 = 0$, т. е. на уровне IV квартала, то получим условные значения уровней ряда без сезонности. Например, для I квартала они составят: 12,073 в 2003 г.; 10,698 в 2004 г.; 9,323 в 2005 г. Фактические значения уровней ряда соответственно были (y_t): 10; 9 и 7. Разность этих величин покажет влияние сезонности I квартала по отношению к IV вместе со случайной составляющей. Усреднив их, найдем среднюю оценку отклонения сезонности I квартала по сравнению с IV: $(-2,073 - 1,698 - 2,323)/3 = -2,031$, что соответствует величине параметра c_1 . Аналогично можно убедиться в правильности выводов по модели и для параметров c_2, c_3 .

Параметр $a = 12,417$ характеризует уровень IV квартала 2002 г. вместе с сезонной компонентой. Прогноз по данной модели на I квартал 2006 г. составит ту же величину, что и по предыдущей модели, а именно 5,914 тыс. чел. (при учете в параметрах большего количества знаков после запятой 5,916 тыс. чел.). Для рассматриваемой модели несколько выше коэффициент детерминации — 0,985 (в предыдущей модели — 0,981). Это говорит о ее предпочтении для целей прогноза.

В аналитическом плане аддитивная модель с тенденцией на первый взгляд более информативна. Она дает исследователю конкретные значения сезонной составляющей. В модели же с фиктивными переменными параметры c_1, c_2, c_3 фиксируют не уровни сезонности соответствующего квартала, а их отличие от воздействия сезонности в IV квартале.

Вместе с тем, зная параметры c_1, c_2 и c_3 , можно определить показатели сезонности. Исходя из содержания параметров при фиктивных переменных имеем:

$$c_1 = S_1 - S_4,$$

$$c_2 = S_2 - S_4,$$

$$c_3 = S_3 - S_4,$$

где S_1, S_2, S_3 и S_4 — показатели сезонности соответствующих кварталов.

Тогда $c_1 + c_2 + c_3 = S_1 + S_2 + S_3 - 3S_4$.

Но так как $\sum_1^4 S = 0$, можно записать, что $c_1 + c_2 + c_3 = 0 - S_4 - 3S_4$

или $c_1 + c_2 + c_3 = 0 - 4S_4$. Отсюда получим значение сезонной компоненты для IV квартала:

$$S_4 = -\frac{1}{4}(c_1 + c_2 + c_3).$$

Для нашего примера, используя значения параметров модели, получим $S_4 = -\frac{1}{4}(-2,031 - 3,688 - 5,010) = 2,682$. Далее определяем сезонные компоненты I, II и III кварталов:

$$S_1 = c_1 + S_4;$$

$$S_2 = c_2 + S_4;$$

$$S_3 = c_3 + S_4.$$

По данным примера имеем

$$S_1 = -2,031 + 2,682 = 0,651;$$

$$S_2 = -3,688 + 2,682 = 1,006;$$

$$S_3 = -5,010 + 2,682 = -2,328;$$

$$S_4 = 2,682; \quad \Sigma S = -0,001 \text{ (при более точном счете получим 0).}$$

Численно сезонная компонента в модели с фиктивными переменными очень близка к тому, что было найдено для аддитивной модели с тенденцией (табл. 6.10).

Таким образом, исследуя сезонность по квартальным данным, видим, что модель с фиктивными переменными по своему содержанию адекватна аддитивной модели с тенденцией, рассмотренной в п. 6.4.2.2.

Вместе с тем в практических исследованиях сезонность изучают по месячным данным, ибо сезонность может проявлять себя и внутри квартала. В этом случае применение модели с фиктивными переменными (а их будет 11) потребует информации не менее чем за 7–8 лет, чтобы на каждый параметр модели приходилось достаточное число степеней свободы и можно было получить надежные оценки параметров. При ограниченной по числу лет информации изучение сезонности по месячным данным целесообразно вести по аддитивной модели.

6.4.3. Мультипликативная модель сезонности

В мультипликативной модели уровень динамического ряда рассматривается как произведение его компонент:

$$y_t = \hat{y}_t \cdot K_S \cdot E_t, \quad (6.38)$$

где y_t – фактические уровни динамического ряда;

\hat{y}_t – теоретические значения уровней динамического ряда согласно тенденции;

K_S – коэффициент сезонности;

E_t – коэффициент случайной компоненты.

В данной модели $\hat{y}_t \cdot K_S$ представляет собой тренд с учетом сезонной волны (y_s), т. е. уровень ряда, обусловленный влиянием как тенденции, так и сезонности: $y_s = \hat{y}_t \cdot K_S$. Используя величину y_s , мультипликативную модель можно представить как

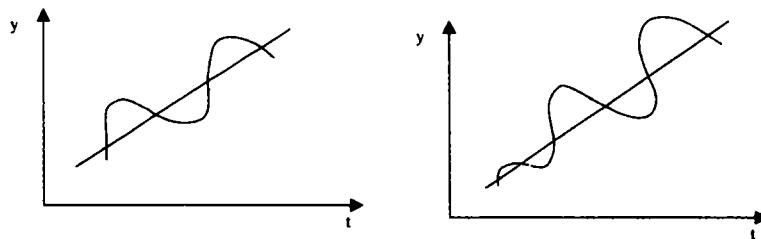


Рис. 6.24. Сезонность: сравнение аддитивной и мультипликативной модели с линейной тенденцией:

$$y_t = \hat{y}_t \cdot \frac{y_S}{\hat{y}_t} \cdot \frac{y_t}{y_S}, \quad (6.39)$$

$$\frac{y_S}{\hat{y}_t} = K_S \text{ and } \frac{y_t}{y_S} = E_t.$$

Как видим, отличие мультиплекативной модели от аддитивной состоит в том, что в мультиплекативной модели сезонная и случайная составляющие определены в виде относительных величин (коэффициентов), а в аддитивной модели — в виде абсолютных величин (тысяч рублей, тонн, человек и т. п.). И хотя сезонная компонента в анализе может быть определена и как абсолютная величина (S), и как относительная (K_S), это не означает, что по одному и тому же динамическому ряду обе модели одинаково возможны. Эти модели в практических расчетах дадут близкие результаты, если амплитуда колебаний уровней ряда слабо изменяется во времени.

Ввиду того что в мультиплексной модели сезонность выражена в процентах, при наличии тенденции в ряду динамики амплитуда сезонных колебаний меняющаяся. Так, если коэффициент сезонности примет значение для I квартала 1,2, или 120%, то при повышающейся тенденции в ряду динамики прирост в 20% будет для I квартала каждого года представлять увеличивающуюся сезонную волну (рис. 6.24).

Методика построения мультипликативной модели содержит в целом те же шаги расчетов, что и по аддитивной модели:

1. Нахождение слаженных уровней динамического ряда методом скользящих средних (\tilde{y}_t).
 2. Оценка сезонной составляющей в виде коэффициентов сезонности (K_{S_j}) и их корректировка (\hat{K}_{S_j}).
 3. Элиминирование сезонной компоненты из исходных данных временного ряда ($\frac{y_t}{\hat{K}_{S_j}}$).
 4. Построение уравнения линейного тренда по уровням ряда без сезонности.

5. Расчет выровненных значений трендовой составляющей (\hat{y}_t).
6. Расчет теоретических уровней ряда с учетом сезонности $y_S(\hat{y}_t \cdot \hat{K}_{S_j})$.
7. Расчет случайной компоненты $E_t(\frac{y_t}{y_S})$. Для аналитических целей случайная компонента может быть найдена и по абсолютной величине:

$$\xi_t = y_t - y_S.$$

Пример. Прибыль компании за последние три года характеризуется данными (тыс. ден. ед.):

Таблица 6.12

Кварталы	1-й год	2-й год	3-й год
1	6	10	17
2	7	14	22
3	9	18	25
4	15	25	35
Σ	37	67	99

Таблица демонстрирует наличие тенденции (уровни каждого года выше предыдущего), сезонности с возрастающей амплитудой колебаний (9, 15 и 18) по годам. Поэтому целесообразно исследовать мультипликативную модель. Результаты расчетов представлены в табл. 6.13.

Таблица 6.13

Мультипликативная модель сезонности

	y_t	\tilde{y}_t	K_{S_j}	\hat{K}_{S_j}	Y_t	\hat{y}_t	y_S	E_t
1	2	3	4	5	6	7	8	9
1	6	—	—	0,782	7,67	6,09	4,76	1,26
2	7	—	—	0,922	7,59	7,99	7,37	0,95
3	9	9,75	0,923	0,976	9,22	9,90	9,66	0,93
4	15	11,125	1,348	1,320	11,36	11,81	15,59	0,96
5	10	13,125	0,762	0,782	12,79	13,72	10,73	0,93
6	14	15,5	0,903	0,922	15,18	15,62	14,41	0,97
7	18	17,625	1,021	0,976	18,44	17,53	17,11	1,05
8	25	19,5	1,282	1,320	18,94	19,43	25,66	0,97
9	17	21,375	0,795	0,782	21,73	21,35	16,69	1,02
10	22	23,5	0,936	0,922	23,86	23,25	21,44	1,03
11	25	—	—	0,976	25,61	25,16	24,56	1,02
12	35	—	—	1,320	26,52	27,07	35,73	0,98

В графе 3 приведены центрированные скользящие средние, найденные аналогично тому, как было показано по аддитивной модели на примере табл. 6.9. Они характеризуют четкую тенденцию уровней ряда к возрастанию независимо от наличия сезонности. Измерим сезонность, исчислив коэффициент сезонности $K_{S_j} = \frac{y_t}{\bar{y}_t}$ (см. графу 4 табл. 6.13).

В графе 4 приведено 8 коэффициентов сезонности (по два для каждого квартала). Чтобы иметь оценку сезонности конкретного квартала независимо от года, рассчитаем средние коэффициенты сезонности (по средней арифметической простой):

$$\bar{K}_{S_j} = \frac{1}{2} \sum K_{S_j},$$

где j – номер квартала.

В результате получим следующие средние коэффициенты сезонности:

$$\text{I квартал} = (0,762+0,795)/2 = 0,779;$$

$$\text{II квартал} = (0,903+0,936)/2 = 0,919;$$

$$\text{III квартал} = (0,923+1,021)/2 = 0,972;$$

$$\text{IV квартал} = (1,348+1,282)/2 = 1,315.$$

Средняя величина коэффициентов сезонности должна быть равна 1, а за год $\sum K_{S_j} = 4$. Поскольку в примере $\sum K_{S_j} = 3,985$, то требуется корректировка коэффициентов сезонности. Поправочный коэффициент составит величину 1,003764. Соответственно умножив на эту величину средние коэффициенты сезонности, получим скорректированные коэффициенты \hat{K}_{S_j} (см. графу 5 табл. 6.13) для каждого квартала. Их сумма по таблице равна 12, что соответствует данным за три года.

Чтобы устраниТЬ влияние сезонности, разделим фактические уровни ряда (y_t) на скорректированные коэффициенты сезонности:

$$Y_t = \frac{y_t}{\hat{K}_{S_j}} \quad (\text{см. графу 6 табл. 6.13}). \quad \text{Эти данные будут отражать влияние}$$

тенденции и случайности. Проведем их выравнивание, найдя линейный тренд: $\hat{y}_t = 4,179 + 1,907t$, и оценим по кварталам трендовую составляющую (результаты представлены в графике 7).

Далее найдем тренд с учетом сезонной волны как $y_S = \hat{y}_t \cdot \hat{K}_{S_j}$ (см. график 8).

Рассчитаем ошибки по мультипликативной модели как $E_t = \frac{y_t}{y_S}$

(см. график 9). Чем они ближе к 1, тем лучше модель описывает исходный временной ряд. Величина $(1-E_t)$ показывает, какую долю составляет случайная компонента в теоретическом значении уровня временного ряда. Как видно из таблицы, в большинстве случаев влияние слу-

чайной компоненты не превышает 5% (лишь в первой строке оно достаточно весомо — 26%). В целом можно считать, что рассмотренная мультиплекативная модель хорошо описывает исходные данные. Это подтверждает и расчет среднего коэффициента случайной составляющей: $\bar{E} = \frac{1}{n} \sum E_i = \frac{12,07}{12} = 1,0058$. Незначительное отклонение его величины от 1 свидетельствует о хорошем качестве модели. Этот же вывод получим и найдя коэффициент корреляции фактических уровней ряда (y_t) и теоретических (y_s): $r_{y,y_s} = 0,9968$.

Прогноз по мультиплекативной модели на I квартал четвертого года можно получить следующим образом:

$$y_p = (\hat{Y} = a + bt_p) \cdot \hat{K}_{S_1}.$$

Подставляя значения параметров и величину коэффициента сезонности для I квартала, получим

$$y_p = (4,1799 + 1,9074t_p) \cdot 0,782 = (4,1799 + 1,9074 \cdot 13) \cdot 0,782 = 22,7.$$

Контрольные вопросы

1. Как используются показатели динамики для обоснования типа модели тренда?
2. Дайте определение автокорреляции уровней и поясните, как она используется при моделировании динамического ряда.
3. Что такое автокорреляционная функция и как рассчитывается выборочная оценка коэффициента автокорреляции?
4. Как интерпретируются параметры модели в виде показательной кривой?
5. Как оцениваются параметры S-образных кривых?
6. Как оценить соответствие модели характеру тренда?
7. Что такое автокорреляция в остатках и как она измеряется?
8. В чем причины появления автокорреляции в остатках?
9. С какой целью используется критерий Дарбина—Уотсона?
10. В каких пределах принимает значения критерий Дарбина—Уотсона?
11. Опишите алгоритм построения аддитивной модели временного ряда.
12. Дайте определение сезонной декомпозиции и корректировки временных рядов.
13. В чем состоит отличие аддитивных моделей при наличии и отсутствии тренда в ряду динамики?
14. В чем отличие подходов к оцениванию сезонности в аддитивной и мультиплекативной моделях?
15. Как используются фиктивные переменные при моделировании сезонных колебаний?
16. Как применяется ряд Фурье для оценки периодических колебаний?
17. Как оцениваются параметры модели в виде ряда Фурье?
18. Как используется ряд Фурье при наличии тенденции в ряду динамики?

ГЛАВА 7. МОДЕЛИ РЕГРЕССИИ ПО ВРЕМЕННЫМ РЯДАМ

7.1. Специфика изучения взаимосвязей по рядам динамики

При построении модели регрессии по рядам динамики необходимо помнить, что высокая корреляция между уровнями временных рядов может иметь место и при отсутствии реальной связи между явлениями. Иными словами, может иметь место *ложная корреляция* — установление связи там, где ее на самом деле нет. Поэтому при построении регрессионных моделей по рядам динамики требуется их предварительная специальная обработка.

Наличие ложной корреляции связано с тенденцией каждого из рядов динамики, с автокорреляцией их уровней. Если ряды динамики характеризуются наличием тренда, то при построении модели регрессии надо исключить тренд. В противном случае корреляция уровней рядов динамики будет преувеличена (коэффициент корреляции будет близок к +1 при одинаковой тенденции в рядах и к -1 при противоположной тенденции). Предположим, что строится регрессия личных сбережений граждан от доходов населения по данным за ряд лет. Коэффициент детерминации при этом составил 0,95. Может показаться, что получен хороший результат и уравнение регрессии пригодно для прогноза. Однако, анализируя остатки ($y - \hat{y}_x$), мы обнаружим наличие в них автокорреляции. Следовательно, наше уравнение регрессии содержит систематическую погрешность, так как не учитывает влияние тенденции. Высокое значение коэффициента детерминации указывает лишь на то, что обоим рядам свойственна тенденция к повышению уровней.

Если ряды динамики характеризуются не только тенденцией, но и периодическими колебаниями, то при построении модели регрессии следует учесть обе компоненты динамических рядов. Это означает, что из первоначальных данных должна быть исключена как тенденция, так и периодическая составляющая. Модель регрессии в этом случае строится либо по остаточным величинам, либо с включением в нее обоих компонент динамического ряда наряду с экономическими переменными.

Кроме того, изучая параллельные временные ряды, можно столкнуться с таким явлением, как временной лаг, т. е. запаздывание уровней одного ряда относительно другого. Например, спрос на товары длительного пользования может зависеть от доходов предыдущих лет. Поэтому при изучении связи по рядам динамики сначала рассчитывается взаимная корреляционная функция, представляющая собой множество коэффициентов корреляции между уровнями рядов (y_t) и ($x_{t-\tau}$), сдвинутыми относительно друг друга на τ моментов времени. Величина лага определяется по наибольшему коэффициенту корреляции. Если временной лаг существует, то он должен быть учтен в модели регрессии.

Определенные трудности при построении модели регрессии по временным рядам возникают в связи с проблемой мультиколлинеарности факторов, когда за счет тенденции объясняющие переменные оказываются тесно связанными между собой. Выходом из создавшегося положения может явиться построение модели регрессии по отклонениям от тренда.

Однако можно строить регрессию и по уровням рядов динамики, если удается при этом устраниТЬ автокорреляцию в остатках, применив, например, обобщенный метод наименьших квадратов.

7.2. Учет тенденции при построении модели регрессии

Методы учета тенденции при построении модели регрессии по временным рядам делятся на две группы:

- методы исключения тенденции из уровней динамического ряда и построение модели по остаточным величинам;
- включение в модель регрессии фактора времени.

7.2.1. Методы исключения тенденции

Теоретически возможны два подхода для исключения тенденции из уровней временного ряда:

- метод последовательных разностей;
- метод отклонений от тренда.

Наиболее точным из них является *метод отклонений от тренда*, ибо тенденция учитывается в виде уравнения тренда, описывающего закономерность изменения уровней ряда во времени. *Метод последовательных разностей* учитывает тенденцию, представленную полиномом соответствующей степени. Так, если тенденция линейная, то регрессия строится по первым разностям, т. е. абсолютным приростам; если же тенденция характеризуется параболой второй степени, то для модели регрессии используются вторые разности, т. е. абсолютные ускорения.

Поскольку тренд может быть описан любой математической функцией, а не только полиномом k -го порядка, теоретически более оправданым является учет тенденции в модели регрессии методом отклонений от тренда. Вместе с тем построение модели регрессии по последовательным разностям имеет право на существование и как наиболее простой способ учета тенденции будет рассмотрен далее.

7.2.1.1. Метод последовательных разностей

Если в ряде динамики имеется четко выраженная линейная тенденция, то ее можно устранить, перейдя от исходных уровней ряда (y_t) к цепным абсолютным приростам (Δ_t), т. е. первым разностям. Объясняется это тем, что линейный тренд характеризуется постоянным абсолютным приростом. Его величина в уравнении $\hat{y}_t = a + bt$ соответствует параметру b . Но так как $y_t = \hat{y}_t + \xi_t$, то первые разности в линейном тренде будут варьировать за счет случайной составляющей вокруг

своей константы — параметра b . Тенденция в уровнях временного ряда будет устранена.

Так, цепной абсолютный прирост можно представить как

$$\begin{aligned}\Delta_t &= y_t - y_{t-1}, \text{ и, если } y_t = a + bt + \xi_t, \text{ то} \\ \Delta_t &= (a + bt + \xi_t) - (a + b(t-1) + \xi_{t-1}) = b + (\xi_t - \xi_{t-1}).\end{aligned}\quad (7.1)$$

Если ряд динамики характеризуется тенденцией в виде параболы второй степени, то для ее устранения можно заменить исходные уровни ряда на вторые разности (Δ''), т. е. на величину абсолютных ускорений. Как известно, парабола второй степени характеризуется постоянным абсолютным ускорением (вторыми разностями), а первые разности имеют линейную тенденцию. Поэтому для динамического ряда с тенденцией в виде параболы $\hat{y}_t = a + bt + ct^2$ вторые разности (Δ'') будут колебаться вокруг величины $2c$ за счет случайной ошибки (ξ_t), ибо всегда $y_t = \hat{y}_t + \xi_t$. Соответственно тенденция из исходных данных временного ряда будет устранена.

Так, абсолютное ускорение представим как

$$\begin{aligned}\Delta''_t &= \Delta_t - \Delta_{t-1}, \text{ и если } y_t = a + bt + ct^2 + \xi_t, \text{ то} \\ \Delta_t &= y_t - y_{t-1} = (a + bt + ct^2 + \xi_t) - (a + b(t-1) + \\ &\quad + c(t-1)^2 + \xi_{t-1}) = (b - c) + 2ct + (\xi_t - \xi_{t-1}),\end{aligned}\quad (7.2)$$

т. е. первые разности (Δ_t) являются линейной функцией от времени t . Вторые разности (Δ''_t) окажутся равными

$$\begin{aligned}\Delta''_t &= (b - c) + 2ct + (\xi_t - \xi_{t-1}) - [(b - c) + 2c(t-1) + \\ &\quad + (\xi_{t-1} - \xi_{t-2})] = 2c + (\xi_t - 2\xi_{t-1} + \xi_{t-2}).\end{aligned}\quad (7.3)$$

Они не зависят от фактора времени t и могут быть использованы для построения регрессии по временным рядам.

Аналогично можно показать, что если тенденция характеризуется полиномом третьей степени, то для модели регрессии следует использовать третьи разности, чтобы исключить тенденцию из уровней временного ряда. Однако модели регрессии по вторым, третьим разностям мало информированы с точки зрения их интерпретации и последующего использования в прогнозировании. Поэтому ограничимся рассмотрением регрессии по первым разностям.

Исследуем два динамических ряда с линейными тенденциями. Модель линейной регрессии примет вид:

$$\Delta y_t = a + b\Delta x_t + u_t,\quad (7.4)$$

где $\Delta y_t, \Delta x_t$ — первые разности;

u_t — случайная ошибка.

Параметр b в модели характеризует среднее изменение скорости ряда y_t с изменением абсолютного прироста ряда x_t , на единицу.

Следует заметить, что если модель будет характеризоваться высоким показателем R^2 и отсутствием автокорреляции в остатках, то для прогнозирования конкретных значений y_t можно перейти к уравнению вида

$$y_p = y_n + a + b(x_p - x_n),$$

где y_p — прогнозное значение уровня ряда;
 y_n — конечный уровень динамического ряда;
 x_p и x_n — то же по ряду x .

В данном уравнении величина $x_p - x_n = \Delta x_p$ оценивает прогнозное значение скорости ряда x , а $y_p - y_n = \Delta y_p$ — прогнозное значение скорости ряда y .

Прогнозное значение фактора x_p может быть дано либо по модели $x_t = f(z_t)$, где z_t — объясняющая переменная x_i , либо по тренду $\hat{x}_p = a + bt$. От того, насколько хорошо спрогнозировано значение фактора x_p , зависит качество прогноза y_p .

Пример. Инвестиции в основной капитал (y_t — млн руб.) и прибыль предыдущего года (x_t — млн руб.) характеризуются по предприятию за 9 лет следующими данными:

Годы	1	2	3	4	5	6	7	8	9
x_t	2,4	2,9	3,2	3,6	4	4,5	4,8	5,1	5,5
y_t	0,7	1	1,2	1,4	1,7	2,1	2,3	2,5	2,9

Если к этим данным применить МНК, то получим уравнение регрессии:

$$\hat{y}_t = -1,047 + 0,701x_t;$$

$R^2 = 0,9948$; $D - W = 1,15$. Значение R^2 , близкое к 1, обусловлено наличием линейной тенденции в рассматриваемых временных рядах. Модель регрессии по критерию Дарбина—Уотсона не позволяет отклонить гипотезу об отсутствии автокорреляции в остатках: табличные значения его при 5%-ном уровне значимости составили: 0,82 — нижнее и 1,32 — верхнее.

Чтобы устранить из данных тенденцию, найдем первые разности (табл. 7.1).

Таблица 7.1

**Первые разности временных рядов инвестиций
в основной капитал (Δy_t) и прибыли (Δx_t)**

Годы	1	2	3	4	5	6	7	8	9
Δx		0,5	0,3	0,4	0,4	0,5	0,3	0,3	0,4
Δy		0,3	0,2	0,2	0,3	0,4	0,2	0,2	0,4

Используя МНК, получим уравнение регрессии:

$$\Delta y_t = -0,0231 + 0,7692 \Delta x_t + u_t; R^2 = 0,5245; D - W = 1,4005.$$

Коэффициент регрессии $b = 0,7692$ показывает, что увеличение абсолютного прироста прибыли на 1 млн руб. приводит в среднем к увеличению абсолютного прироста инвестиций в основной капитал на 769,2 тыс. руб.

В данном уравнении регрессии отсутствует автокорреляция в остатках: коэффициент автокорреляции остатков составил 0,02 и критерий Дарбина—Уотсона превышает верхнее табличное значение 1,33.

Прогноз на 10-й год выполним по уравнению

$$y_p = y_n + a + b(x_p - x_n),$$

где $y_n = 2,9$; $x_n = 5,5$;

x_p найдем по уравнению линейного тренда:

$$\hat{x}_t = 2,075 + 0,385t \text{ при } t = 10, \quad x_p = 5,925.$$

Соответственно $y_p = 2,9 - 0,0231 + 0,7692(5,925 - 5,5) = 3,2$ млн руб.

7.2.1.2. Метод отклонений от тренда

Как уже указывалось, метод отклонений от тренда является более точным методом исключения тенденции из данных временного ряда. Это связано не только с тем, что тенденция выражается в виде уравнения тренда любой математической функции. Рассматриваемые для модели регрессии ряды динамики могут иметь разные тенденции: например, ряд x , описывается гиперболой, а ряд y , — параболой. В этом случае метод отклонений от тренда позволяет исключить из каждого временного ряда соответствующую ему тенденцию.

Алгоритм построения регрессии при применении метода отклонений следующий:

1. Для каждого временного ряда определяются уравнение тренда и теоретические значения \hat{y}_t , \hat{x}_t .
2. По каждому из рядов находятся остаточные величин

$$dy = y_t - \hat{y}_t; \quad dx = x_t - \hat{x}_t.$$

3. Строится модель регрессии $dy = f(dx)$.

В линейной регрессии $dy = a + bdx + u$, параметр b показывает, как в среднем изменяется величина случайных отклонений по ряду y , с изменением случайных колебаний ряда x , на единицу. Если при этом оба ряда характеризуются линейной тенденцией, то параметр $a = 0$, ибо $\sum dy = \sum dx = 0$. Тогда модель линейной регрессии примет вид $dy = bdx + u$, и параметр b будет выступать коэффициентом пропорциональности. Его величина будет показывать, во сколько раз случайные отклонения по ряду y , в среднем выше (ниже) случайных отклонений по ряду x .

Для прогноза конкретных значений у можно перейти к уравнению, связывающему между собой уровни временных рядов. С этой целью в модель регрессии $dy = a + bdx$ подставим значения dy и dx , раскрыв их содержание, т. е. $dy = y_t - \hat{y}_t$, и $dx = x_t - \hat{x}_t$.

Тогда имеем, например, для линейной регрессии $dy = a + bdx$, т. е. $(y_t - \hat{y}_t) = a + b(x_t - \hat{x}_t)$ или $y_t = \hat{y}_t + a + b(x_t - \hat{x}_t)$.

Данную модель можно использовать для прогноза:

$$y_p = \hat{y}_{t=p} + a + b(x_{p=p} - \hat{x}_{t=p}),$$

где y_p — прогнозное значение y ;

$\hat{y}_{t=p}$ — прогноз y по тренду при $t = p$;

$x_{p=p}$ — прогнозное значение x ;

$\hat{x}_{t=p}$ — прогноз x исходя из уравнения тренда при $t = p$.

Результат прогноза зависит от качества прогноза фактора x и от качества трендовых моделей, используемых в прогнозировании.

Пример. По данным за 10 месяцев рассматривается зависимость прибыли предприятия (y_t — тыс. руб.) от затрат на мероприятие по охране труда (x_t — тыс. руб.) (табл. 7.2).

Таблица 7.2

Расчет остаточных величин для построения модели регрессии

$$dy = a + bdx + u_t$$

Период t	y_t	\hat{y}_t	dy	x_t	\hat{x}_t	$<dx$
1	250	259,3	-9,3	9	10,1	-1,1
2	305	294	11,0	15	12,9	2,1
3	314	316,4	-2,4	13	14,8	-1,8
4	338	333,4	4,6	18	16,4	1,6
5	354	347,1	6,9	19	17,7	1,3
6	363	358,8	4,2	20	18,8	1,2
7	375	369	6,0	22	19,9	2,1
8	376	378	-2,0	19	20,8	-1,8
9	376	386,2	-10,2	20	21,7	-1,7
10	385	393,6	-8,6	21	22,5	-1,5

Каждый из рядов имеет повышающуюся тенденцию, которая достаточно хорошо описывается степенной функцией. Для ряда прибыли уравнение тренда составило $\hat{y}_t = 259,28t^{0.1813}$, $R^2 = 0,9699$, а для ряда затрат на мероприятие по охране труда $\hat{x}_t = 10,111t^{0.3469}$, $R^2 = 0,851$. Автокорреляция в остатках отсутствует: $r_{a_1} = 0,076$ для ряда y_t и $r_{a_1} = -0,067$ для ряда x_t .

Если коррелировать исходные уровни рядов динамики, то R^2 составит 0,925, а уравнение регрессии окажется $\hat{y}_t = 165,42 + 10,12x$. Однако ввиду наличия в каждом из рядов четкой тенденции можно предположить, что результаты регрессионно-корреляционного анализа завышены. Поэтому применим метод устранения тенденции, найдя отклонения от тренда: $dy = y_t - \hat{y}_t$ и $dx = x_t - \hat{x}_t$ (см. табл. 7.2).

Применяя к рядам dy и dx МНК, получим уравнение линейной регрессии $dy = -0,1913 + 3,8349dx$, $R^2 = 0,7865$ (при компьютерной обработке в расчетах использованы dy и dx с точностью до 0,00001, если использовать данные табл. 7.2, то $dy = -0,135 + 3,8593dx$, $R^2 = 0,7862$). Полученное уравнение регрессии показывает, что при устранении из исходных уровней временных рядов тенденции связь между остаточными величинами имеет место. Поэтому данную модель можно использовать в прогнозировании. На это указывает и отсутствие в модели автокорреляции остатков (коэффициент автокорреляции — 0,325, а критерий Дарбина—Уотсона — 1,76; сравнивая с табличными значениями величину $4 - DW$ при $\alpha = 0,05$ и числе степеней свободы 10,

увидим, что фактическое значение $D - W$ превышает верхнюю границу 1,32).

Для прогноза на 11-й месяц воспользуемся уравнением вида:

$$y_p = \hat{y}_{t=p} + a + b(x_p - \hat{x}_{t=p}),$$

где $\hat{y}_{t=p} = 259,28t^{0,1813} = 259,28 \cdot 110,1813 = 400,47$;

$$\hat{x}_{t=p} = 10,111t^{0,3469} = 10,111 \cdot 110,3469 = 23,23;$$

$x_p = 22$ (принято представлять как $x_n = 21$ и плюс $\Delta n = 1$).

Тогда y_p составит

$$y_p = 400,47 + (-0,1913) + 3,8349(22 - 23,23) = 395,57 \text{ тыс. руб.}$$

7.2.2. Включение в модель регрессии фактора времени

Модель регрессии по временным рядам может быть построена по исходным данным с *включением* в нее как отдельной независимой переменной *фактора времени* t , т. е. для двух связанных рядов динамики строится модель вида

$$y = a + bx + ct + \xi, \text{ где } t = 1, 2, 3, \dots, n.$$

Включая в регрессию фактор времени t , устранием тенденцию из уровней временных рядов. Это объясняется спецификой множественной регрессии: коэффициенты регрессии показывают изолированное влияние на результат соответствующего фактора при неизменном уровне других факторов. В рассматриваемом двухфакторном уравнении регрессии коэффициент регрессии b характеризует «чистое» воздействие переменной x на y в условиях неизменной тенденции, т. е. при устранении тенденции.

Математически доказано, что если временные ряды характеризуются линейной тенденцией, то включение в модель фактора времени t равносильно построению модели регрессии по отклонениям от трендов с последующим переходом от нее к исходным уровням временного ряда зависимой переменной y .

Уравнение регрессии $\hat{y} = a + bx + ct$ может быть построено двумя путями:

- применяя метод наименьших квадратов, получаем оценки параметров a , b и c (именно так строится данная модель при компьютерной обработке);
- последовательно включаем в модель линейную тенденцию ряда y и линейную регрессию остаточных величин $dy = bdx + \xi$, где dy , dx — остаточные величины от линейных тенденций.

Рассмотрим второй подход построения линейной модели регрессии с включением фактора времени t . Алгоритм построения модели следующий:

1. Строится линейное уравнение тренда для ряда y :

$$\hat{y}_t = a^* + c^*t.$$

2. Строится линейное уравнение тренда для ряда x :

$$\hat{x}_t = A + Bt.$$

3. Находятся остаточные величины dy и dx :

$$dy = y_t - (a^* + c^* t);$$

$$dx = x_t - (A + Bt).$$

4. Строится регрессия по отклонениям от трендов:

$dy = bdx + \xi$ (свободный член в уравнении отсутствует, ибо при линейных трендах $\sum dy = \sum dx = 0$),

5. Определяется модель для y_t :

$$y_t = \hat{y}_t + dy$$

или

$$y_t = a^* + c^* t + b(x_t - A - Bt) + \xi,$$

откуда имеем

$$y_t = (a^* - bA) + (c^* - bB)t + bx + \xi.$$

Данное уравнение соответствует уравнению регрессии

$$y_t = a + bx + ct + \xi,$$

где

$$a = a^* - bA; \quad c = c^* - bB.$$

Рассмотренный подход к построению модели регрессии позволяет понять, что уравнение регрессии с включением фактора времени t : $y_t = a + bx + ct + \xi$ учитывает линейные тенденции для временных рядов y_t и x_t . Кроме того, строя регрессию по отклонениям от линейных трендов, мы получаем остатки ξ , те же, что и в регрессии с включением линейного фактора времени t . Поэтому при наличии в рядах линейных тенденций целесообразно строить модель регрессии по исходным уровням рядов, включая в нее фактор времени t . В этом случае модель регрессии по отклонениям от трендов неинформативна. Тем более надо учесть, что регрессия по отклонениям от линейных трендов является составной частью регрессии с включением фактора времени t (шаг 4 при втором подходе).

В регрессии $\hat{y}_t = a + bx + ct$ параметр b характеризует, на сколько единиц изменяется в среднем y при изменении x на одну единицу в условиях неизменной тенденции; параметр c показывает средний абсолютный прирост y в условиях неизменного уровня объясняющей переменной x .

Пример. По предприятию имеются данные за 12 месяцев года об уровне рентабельности (y , %) и объеме продаж продукции A (x , тыс. ед.):

Месяц t	1	2	3	4	5	6	7	8	9	10	11	12
y	5	6	6	7	8	10	11	11	13	12	13	15
x	1,1	2,1	2,2	3,1	2,2	3,3	4,3	4	5,1	4,2	5	6

Построим модель регрессии с включением в нее фактора времени t :

$$y_t = a + bx + ct + \xi,$$

где $t = 1, 2, \dots, 12$.

Для оценки параметров применим МНК.

Результаты оказались следующими:

$$R = 0,9879;$$

$$R^2 = 0,9759;$$

$$\hat{y}_t = 3,073 + 0,846x + 0,565t; \quad F = 182,3;$$

t -критерий 6,0 2,33 3,82.

В целом уравнение регрессии значимо: $F_{\text{факт.}} = 182,3$ превышает табличное значение при $\alpha = 0,05$ и числе степеней свободы 2 и 9 ($F_{\text{табл.}} = 4,26$). Статистически значимыми являются и параметры уравнения регрессии, ибо фактические значения t -критерия Стьюдента превышают табличное значение $t_{\text{табл.}} = 2,26$ при $\alpha = 0,05$ и $df = 9$.

Эти же результаты получим и применяя метод последовательного включения в модель регрессии линейной тенденции ряда y_t и линейной регрессии отклонений от трендов.

Так, в рассматриваемом примере уравнение линейного тренда для ряда y_t составило

$$\hat{y}_t = 3,954 + 0,892t,$$

где $t = 1, 2, \dots, 12$.

Уравнение хорошо описывает тенденцию: $R^2 = 0,961$, коэффициент автокорреляции в остатках 0,0083. Получено линейное уравнение тренда для ряда x_t :

$$\hat{x}_t = 1,041 + 0,386t, \quad R^2 = 0,897.$$

В целом оно тоже адекватно описывает тенденцию.

Для каждого ряда были рассчитаны отклонения от трендов:

$$dy = y_t - \hat{y}_t, \quad \text{и} \quad dx = x_t - \hat{x}_t.$$

Получено уравнение регрессии по отклонениям трендов:

$$dy = 0,846dx + \xi; \quad R^2 = 0,378.$$

Были получены оценки параметров регрессии с включением фактора времени:

$$a = 3,954 - 0,846 \cdot 1,041 = 3,073;$$

$b = 0,846$ (совпадает с величиной коэффициента регрессии)

по уравнению $dy = bdx + \xi$ и

по уравнению $y_t = a + bx + ct + \xi$;

$$c = 0,892 - 0,846 \cdot 0,386 = 0,565.$$

Данные оценки совпадают с тем, что было получено по МНК.

Остаточные величины (ξ_t) по модели для отклонений от трендов совпадают с остатками для регрессии с включением фактора времени. Автокорреляция в остатках небольшая: $r_{a_1} = 0,282$.

В рассматриваемой модели параметр b показывает, что рост объема продаж на 1 тыс. ед. в условиях неизменной тенденции способствует росту уровня рентабельности на 0,846 процентных пункта. Параметр c характеризует среднемесячный прирост рентабельности независимо от изменения объема продаж, т. е. обусловленный влиянием других факторов, не учитываемых в регрессии.

В примере рассмотрены два динамических ряда. Принцип введения в модель фактора времени сохраняется и при изучении трех и более связанных рядов динамики. Так, если строится регрессия $y = f(x_1, x_2, x_3)$, то включение в нее фактора времени t приводит чаще всего к модели вида $\hat{y}_t = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + ct$. В ней параметры b_1, b_2 и b_3 показывают изолированное воздействие каждой объясняющей переменной на результат y , а параметр c — средний абсолютный прирост y в условиях неизменности значений переменных x_1, x_2 и x_3 .

Время в качестве независимой переменной часто вводится в виде линейного члена, даже если другие переменные подвергаются логарифмированию или иному преобразованию. Например, производственная функция с включением фактора времени часто записывается как

$$P = aK^{b_1} L^{b_2} e^{ct},$$

где P — объем продукции;

K — основной капитал;

L — занятость;

e — основание натурального логарифма;

t — фактор времени, взятый как ряд натуральных чисел $1, 2, \dots, n$.

Линеаризуем данную зависимость, прологарифмировав производственную функцию:

$$\ln P = \ln a + b_1 \ln K + b_2 \ln L + ct.$$

Здесь фактор времени t введен в модель линейно. Рассматриваемая производственная функция нелинейна относительно оцениваемых параметров. В ней параметры b_1 и b_2 являются коэффициентами эластичности, показывая, на сколько процентов повышается объем продукции при увеличении соответствующего фактора (K и L) на 1% в условиях неизменной тенденции.

Параметр c обычно интерпретируется как автономный рост объема продукции в условиях неизменности факторов производства K и L . Так, если $c = 0,0175$, то $e^{0,0175} = 1,01765$ и, следовательно, ежегодно (если t — годы) объем продукции возрастает в среднем при неизменных уровнях затрат капитала и труда в 1,01765 раза, или на 1,76%.

Если тенденция в рядах динамики характеризуется полиномом второй и более высокой степени, то в модель регрессии вводятся t и t^2 , а иногда t в более высокой степени. В этом случае рассматривается регрессия вида $\hat{y} = a + bx + ct + dt^2$ (при двух временных рядах) или

$$\hat{y} = a + \sum_{j=1}^p b_j x_j + ct + dt^2 \quad (\text{при } p \text{ временных рядах}).$$

Вводя в модель регрессии фактор времени в виде t, t^2, \dots, t^k , предполагают, что коэффициенты при переменных остаются неизменными и характеризуют силу связи результата y с соответствующей объясняющей переменной x .

Если предполагается, что в регрессии коэффициенты при независимой переменной подвержены изменению во времени, то в модель можно ввести преобразованные переменные $t x$ (где t означает время). Оценка параметров дается МНК.

Модель регрессии с включением фактора времени в нее как независимой переменной не всегда эффективна ввиду возможной мультиколлинеарности факторов. Если временные ряды, используемые в регрессии, характеризуются четкой тенденцией ($R^2 > 0,9$), то корреляция t и x_j может превышать корреляцию x_j с y и параметры регрессии при объясняющих переменных x оказываются ненадежными и экономически неинтерпретируемыми.

Время может быть учтено в регрессии и через использование лаговых переменных, т. е. запаздывающих переменных, сдвинутых на определенный интервал времени. Например, спрос на недвижимость в значительной мере определяется доходом не текущего, а предыдущих периодов. Вопросы, связанные с построением моделей регрессии с лаговыми переменными, рассматриваются в гл. 8.

7.3. Обобщенный метод наименьших квадратов при построении модели регрессии по временным рядам

Даже учитя тенденцию во временных рядах, модель регрессии может содержать автокорреляцию в остатках. Методы устранения автокорреляции в остатках могут быть разные. Они зависят от причин автокорреляции. Автокорреляция в остатках может быть следствием неправильной спецификации модели: не учтена важная объясняющая переменная, неправильно выбрана форма связи. В этом случае можно попытаться изменить математическую функцию регрессии (например, линейную на степенную), уточнить набор объясняющих переменных. Однако если эти попытки не увенчались успехом и автокорреляция в остатках имеет место, то для ее устраниния можно применить обобщенный метод наименьших квадратов (ОМНК).

ОМНК можно использовать как для парной, так и для множественной регрессии. Для простоты и уяснения сути проблемы рассмотрим регрессию двух временных рядов:

$$y_t = a + bx_t + \xi_t. \quad (7.5)$$

Для периода времени $t-1$ справедливо равенство

$$y_{t-1} = a + bx_{t-1} + \xi_{t-1}. \quad (7.6)$$

Если имеет место автокорреляция в остатках, т. е. последующие по времени остатки зависят от предыдущих, то регрессия остатков может быть представлена как

$$\xi_t = c + d\xi_{t-1} + V_t, \quad (7.7)$$

где V_t — случайная ошибка для линейной регрессии остатков.

Но так как $\xi_t = y_t - \hat{y}_t$, где \hat{y}_t — теоретические уровни линейной регрессии y от x , то $\sum \xi_t = 0$ и $\bar{\xi}_t = 0$. Полагая, что $\bar{\xi}_{t-1} \equiv \bar{\xi}_t$, имеем $c = \bar{\xi}_t - d\bar{\xi}_{t-1} = 0$. Параметр d определим по формуле

$$d = \frac{\text{cov}(\xi_t, \xi_{t-1})}{\sigma_{\xi_{t-1}}^2}, \quad (7.8)$$

где $\text{cov}(\xi_t, \xi_{t-1}) = \overline{\xi_t \xi_{t-1}} - \bar{\xi}_t \cdot \bar{\xi}_{t-1} = \overline{\xi_t \xi_{t-1}}$.

Тогда $d = \frac{\overline{\xi_t \xi_{t-1}}}{\sigma_{\xi_{t-1}}^2}$. Предполагая, что $\sigma_{\xi_t}^2 \equiv \sigma_{\xi_{t-1}}^2$, можно записать, что

$$d = \frac{\overline{\xi_t \xi_{t-1}}}{\sigma_{\xi_{t-1}}^2} = \frac{\sum \xi_t \xi_{t-1}}{\sum \xi_{t-1}^2}, \quad (7.8)$$

т. е. d — коэффициент автокорреляции остатков первого порядка. Обозначим его через ρ . Тогда регрессия остатков примет вид

$$\xi_t = \rho \xi_{t-1} + V_t, \quad (7.10)$$

где ρ — коэффициент автокорреляции остатков первого порядка; V_t — случайная ошибка, удовлетворяющая всем предпосылкам МНК.

Предполагая, что ρ известен, вычтем из уравнения (7.5) уравнение (7.2), умноженное на ρ :

$$y_t - \rho y_{t-1} = a(1-\rho) + b(x_t - \rho x_{t-1}) + (\xi_t - \rho \xi_{t-1}). \quad (7.11)$$

Обозначим новую зависимую переменную $(y_t - \rho y_{t-1})$ через y_t^* , а объясняющую переменную $(x_t - \rho x_{t-1})$ — через x_t^* . Примем также, что $a(1-\rho) = a^*$. Учитывая, что $\xi_t - \rho \xi_{t-1} = V_t$, получим:

$$y_t^* = a^* + b x_t^* + V_t. \quad (7.12)$$

Так как ошибки V_t удовлетворяют предпосылкам МНК (они не содержат автокорреляцию), то оценки a^* и b будут обладать свойствами несмещенных оценок и могут быть получены обычным МНК.

Уравнение (7.12) возможно только при $t \geq 2$, так как при $t=1$ отсутствует лаговая переменная. Чтобы не уменьшать число степеней свободы, рекомендуется для первого периода времени ($t=1$) использовать поправку Прайса—Уинстена:

$$x_1^* = \sqrt{1-\rho^2} \cdot x_1 \text{ и } y_1^* = \sqrt{1-\rho^2} \cdot y_1. \quad (7.13)$$

Таким образом, ОМНК предполагает, что вместо исходных переменных y_t и x_t используются взвешенные переменные $PY = y^*$ и $PX = x^*$, где P — веса. В матричном виде модель регрессии принимает вид: $PY = PXB + P\xi$. В ней матрица весов P составит

$$P = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & 0 & 0 & 0 \\ -\rho & 1 & 0 & 0 & 0 & 0 \\ 0 & -\rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\rho & 1 & 0 \\ 0 & 0 & 0 & 0 & -\rho & 1 \end{bmatrix}$$

Иными словами, матрица исходных данных трансформируется:

$$y^* = \begin{bmatrix} y_1 \sqrt{1-\rho^2} \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_n - \rho y_{n-1} \end{bmatrix}, \quad x^* = \begin{bmatrix} x_1 \sqrt{1-\rho^2} \\ x_2 - \rho x_1 \\ x_3 - \rho x_2 \\ \vdots \\ x_n - \rho x_{n-1} \end{bmatrix}$$

Для длинных динамических рядов поправка Прайса—Уинстена может не применяться. Тогда матрица весов не содержит первую строку рассмотренной матрицы P и в расчетах используется $n-1$ преобразованных наблюдений y_t^* и x_t^* .

К преобразованным переменным y^* , x^* применяется традиционный МНК и оцениваются параметры a^* и b . Далее из соотношения $a^* = a(1-\rho)$ можно найти параметр a как

$$a = a^* / (1-\rho). \quad (7.14)$$

ОМНК распространяется и на случай множественной регрессии:

$$y_t = a + b_1 x_{1t} + b_2 x_{2t} + \dots + b_p x_{pt} + \xi_t.$$

Если имеет место автокорреляция остатков и $\xi_t = \rho \xi_{t-1} + V_t$, то

$$\begin{aligned} y_t - \rho y_{t-1} &= a(1-\rho) + b_1(x_{1t} - \rho x_{1,t-1}) + b_2(x_{2t} - \rho x_{2,t-1}) + \\ &\dots + b_p(x_{pt} - \rho x_{p,t-1}) + (\xi_t - \rho \xi_{t-1}), \end{aligned} \quad (7.15)$$

где $\xi_t = \rho \xi_{t-1} + V_t$.

Или исходя из прежней символики строим модель вида

$$y_t^* = a^* + b_1 x_{1t}^* + b_2 x_{2t}^* + \dots + b_p x_{pt}^* + V_t. \quad (7.16)$$

Применяя к переменным $y_t^*, x_{1t}^*, x_{2t}^*, \dots, x_{pt}^*$ традиционный МНК, найдем оценки параметров $b_1, b_2, \dots, b_p x_{pt}^*$. Свободный член модели определим как $a = a^* / (1-\rho)$. Далее можно написать искомую модель регрессии $\hat{y}_t = a + b_1 x_{1t} + \dots + b_p x_{pt}$, в которой устранена автокорреляция остатков.

Пример. По данным за 1995–2003 гг. по Тамбовской области рассматривается зависимость потребления растительного масла на душу населения (y , кг) от потребления овощей (x , кг).

Таблица 7.3

Исходные данные и результаты анализа

Годы	y_t	x_t	ξ_t	ξ_{t-1}	y_t^*	x_t^*	\hat{y}
1	2	3	4	5	6	7	8
1995	7,2	100	-0,272	-	5,97	82,916	8,03
1996	7,5	101	-0,108	-0,272	11,525	156,9	8,13
1997	7,9	103	0,021	0,108	12,093	159,46	8,33
1998	8,6	105	0,450	0,021	13,016	162,578	8,53
1999	9,5	121	-0,818	0,450	14,307	179,696	10,10
2000	10,9	121	0,582	-0,818	16,210	188,64	10,10
2001	10,3	120	0,117	0,582	16,393	187,64	10
2002	10,1	119	0,053	0,117	15,858	186,081	9,9
2003	10,7	124	-0,025	0,053	16,346	190,522	10,4

Результаты регрессионного анализа оказались следующими:

$$\hat{y}_t = -6,0791 + 0,1355x, \quad R^2 = 0,9188, \quad F = 79,2,$$

$$t = -3,5 \quad +8,9.$$

Уравнение регрессии статистически значимо. Исследуем автокорреляцию остатков. Значения остатков приведены в графе 4 табл. 7.3. Коэффициент автокорреляции остатков найдем по формуле

$$r_{a_1} = \frac{\sum e_i e_{i-1}}{\sum e_i^2} = \frac{-0,734}{1,314} = -0,55901.$$

Его величина не столь мала, а по критерию Дарбина—Уотсона не может быть отвергнута нулевая гипотеза об отсутствии автокорреляции в остатках (величина 4 — DW попадает в зону неопределенности). Для устранения автокорреляции в остатках применим ОМНК.

Определим преобразованные значения зависимой и объясняющей переменных, взвесив их на коэффициент автокорреляции остатков. В качестве весов ρ будем использовать полученное значение $r_{a_1} = -0,559$. Для 1995 г. преобразованные значения составят

$$y_1^* = \sqrt{1-\rho^2} = 7,2\sqrt{1-(-0,559)^2} = 5,97,$$

$$x_1^* = \sqrt{1-\rho^2} = 100\sqrt{1-(-0,559)^2} = 82,916.$$

Остальные переменные со значениями $t = 2, 3, \dots, 9$ преобразуются по формулам

$$y_t^* = y_t - \rho y_{t-1};$$

$$x_t^* = x_t - \rho x_{t-1}.$$

Так, при $t = 2$ получаем

$$y_2^* = 7,5 - (-0,559)7,2 = 11,525;$$

$$x_2^* = 101 - (-0,559)100 = 156,9.$$

Значения y_t^* и x_t^* приведены в графах 6 и 7 табл. 7.3. К преобразованным переменным y^* и x^* применяем обычный МНК. Результаты регрессии оказываются следующими:

$$y^* = -2,8156 + 0,0984x^*, \quad R^2 = 0,961, \quad F = 172,3.$$

В данном уравнении величина $(-2,8156)$ — параметр a^* . Переидем от него к параметру a :

$$a = a^* / (1 - \rho) = -1,806.$$

В окончательном виде уравнение регрессии составит

$$\hat{y} = -1,806 + 0,0984x \text{ (см. графу 8 табл. 7.3).}$$

Для него $R^2 = 0,845$, что несколько ниже, но ниже и автокорреляция в остатках: $r_{a_1} = 0,239$.

Рассмотренный ОМНК базируется на предположении о том, что коэффициент автокорреляции остатков ρ известен. Однако на практике точное значение ρ не известно и используются его оценки $\hat{\rho}$.

В качестве оценки $\hat{\rho}$ может использоваться значение $r_{e_t e_{t-1}}$, полученное на основе критерия Дарбина—Уотсона. Как было показано ранее, $DW \equiv 2(1 - r_{e_t e_{t-1}})$.

Соответственно имеем

$$r_{e_t e_{t-1}} \equiv 1 - \frac{DW}{2}.$$

Данный метод оценивания дает неплохие результаты при достаточно большом числе наблюдений.

Более точную оценку коэффициентов регрессии при ОМНК дает двухшаговая процедура Дарбина. Суть ее заключается в следующем. Уравнение (7.11) можно записать в виде

$$y_t = a(1 - \rho) + bx_t + \rho y_{t-1} - \rho b x_{t-1} + V_t, \quad (7.17)$$

где $V_t = \xi_t - \rho \xi_{t-1}$.

Тогда имеем модель регрессии, в которой ρ входит в число оцениваемых параметров.

К уравнению (7.17) можно применить обычный МНК, так как остатки V_t не содержат автокорреляции. Первый шаг процедуры Дарбина состоит в применении к модели (7.17) традиционного МНК для определения оценки коэффициента автокорреляции остатков ρ при переменной y_{t-1} .

Далее на втором шаге оценка ρ используется для вычисления преобразованных переменных $y_t^* = y_t - \rho y_{t-1}$ и $x_t^* = x_t - \rho x_{t-1}$. К этим преобразованным переменным применяется обычный МНК, т. е. строится уравнение (7.8), в котором коэффициент при x^* служит оценкой коэффициента регрессии b , а величина $a^* / (1 - \rho)$ оценивает параметр a .

Для нашего примера двухшаговая процедура Дарбина—Уотсона приводит к следующим результатам:

на первом шаге получено уравнение регрессии обычным МНК:

$$\hat{y}_t = -8,479 + 0,087x_t + 0,098x_{t-1} - 0,331y_{t-1};$$

соответственно считаем, что $\rho = -0,331y_{t-1}$;

на втором шаге находим преобразованные переменные x_t^* и y_t^* с учетом поправки Прайса—Уинстена и обычным МНК получаем регрессию:

$$y_t^* = -4,038 + 0,109x_t; \quad R^2 = 0,959; \\ t \quad -3,2 \quad 12,8,$$

параметр a составит $-3,707$, т. е. в окончательном виде уравнение регрессии имеет вид

$$y_t = -3,707 + 0,109x_t + V_t.$$

Подставив в данное уравнение соответствующие значения x_t , получим остатки $(y_t - \hat{y}_t)$, для которых отсутствует автокорреляция: $r_{V_t V_{t-1}} = 0,033$. Для нового уравнения $R^2 = 0,7$, что ниже предыдущих результатов, но статистически значимо ($F = 16,3$ при $0,05 F_{1,7} = 5,59$).

Другим методом, позволяющим оценить ρ , является *итеративная процедура Кохрейна—Оркэтта*. Применимально к модели парной регрессии $y_t = a + bx_t + \xi_t$, она состоит в следующем.

На первом шаге оцениваются параметры регрессии обычным МНК, а также остаточные величины, и по ним на основе модели (7.10) $\xi_t = \rho\xi_{t-1} + V_t$ оценивают МНК параметр ρ . Далее на втором шаге переходят к модели по преобразованным переменным $y_t^* = y_t - \rho y_{t-1}$ и $x_t^* = x_t - \rho x_{t-1}$, т. е. рассматривается модель

$$\hat{y}_t^* = a' + b'x_t^*, \quad (7.18)$$

оценка параметров которой дается МНК. Далее полагая, что $a = a' / (1 - \rho)$ и $b' = b$, вновь возвращаемся к исходному линейному уравнению и находим новые оценки остатков, к которым вновь строим регрессию (7.10):

$$\xi_t = \rho\xi_{t-1} + V_t.$$

Соответственно получаем новую оценку коэффициента автокорреляции остатков ρ . Процесс продолжается до тех пор, пока разность между последующей и предыдущей оценками ρ не будет меньше заданного числа.

Итерационная процедура Кохрейна—Оркэтта представляет собой метод коррекции статистических выводов регрессии динамических рядов при наличии автокорреляции ошибок.

Применим итеративную процедуру Кохрейна—Оркэтта к нашему примеру. Ранее (с. 187) уже было приведено уравнение регрессии и соответственно остатки по нему ξ_t (см. графу 4 табл. 7.3). Применение к модели (7.10) МНК приводит к формуле расчета коэффициента автокорреляции остатков ρ :

$$\rho = \frac{\sum_{t=2}^n \xi_t \xi_{t-1}}{\sum_{t=2}^n \xi_{t-1}^2}, \quad (7.19)$$

что не совпадает с расчетом коэффициента автокорреляции остатков первого порядка по формуле, используемой в критерии Дарбина—Уотсона. Для нашего примера найденное по формуле (7.19) составит

$-0,55927$, а исходя из формулы (7.8): $-0,55901$. Результаты, естественно, близкие, так как $\sum \xi_{t-1}^2$ мало отличаются от $\sum \xi_t^2$.

Полагая $\rho = 0,55927$, мы приедем к значениям y^* и x^* для 1996–2003 гг., которые будут достаточно близки к данным граф 6, 7 табл. 7.3. Уравнение регрессии для них составит $\hat{y}^* = -10,0799 + 0,139123x^*$

Считая, что полученные оценки параметров могут быть использованы для исходной регрессии, т. е. $a = -10,0799/(1+0,55927) = -6,4647$ и $b = 0,139123$, найдем теоретические значения y , т. е. \hat{y} , и остаточные величины ξ_t . Применяя к ним формулу (7.19), найдем новое значение $\rho = -0,59576$. Так как новая оценка ρ не совпадает с предыдущей, то процедуру расчетов можно продолжать.

Практически процедура Кохрейна—Оркэтта равносильна применению ОМНК. Что касается оценки автокорреляции остатков, то поскольку ее истинное значение исследователю неизвестно, можно использовать более простую процедуру оценивания в виде двухшаговой процедуры Дарбина.

Если мы предположим, что $\rho = 1$, т. е. имеет место полная положительная автокорреляция в остатках, то ОМНК будет сведен к методу последовательных разностей.

Обратимся к уравнению (7.11). При $\rho = 1$ оно может быть записано как

$$y_t - y_{t-1} = b(x_t - x_{t-1}) + (\xi_t - \xi_{t-1}), \quad (7.20)$$

где $y_t - y_{t-1} = \Delta y_t$; $x_t - x_{t-1} = \Delta x$, и $\xi_t - \xi_{t-1} = \Delta \xi_t$, т. е. мы имеем уравнение регрессии по первым разностям:

$$\Delta y_t = b\Delta x_t + V_t, \quad (7.21)$$

где $V_t = \Delta \xi_t$.

Из уравнения (7.21) по МНК оценивается коэффициент b .

Таким образом, если ρ стремится к 1, а величина критерия Дарбина—Уотсона стремится к 0, то использование регрессии по первым разностям устраняет автокорреляцию в остатках.

Если предположить, что $\rho = -1$, т. е. имеет место полная отрицательная автокорреляция в остатках, то уравнение (7.11) примет вид

$$y_t + y_{t-1} = 2a + b(x_t + x_{t-1}) + (\xi_t - \xi_{t-1}) \quad (7.22)$$

или $(y_t + y_{t-1}) / 2 = a + b(x_t + x_{t-1}) / 2 + (\xi_t - \xi_{t-1}) / 2$.

В данном уравнении $(y_t + y_{t-1}) / 2$ и $(x_t + x_{t-1}) / 2$ — средние величины за два смежных периода. Используя их в качестве новых преобразованных переменных y_t^* и x_t^* , параметры a и b можно оценить МНК. В этом случае ОМНК приводит к модели регрессии по скользящим средним.

Однако предположение о том, что $\rho = \pm 1$, представляет собой довольно редкое на практике явление. Поэтому в большинстве случаев ОМНК применяют или используя величину ρ исходя из критерия Дар-

бина—Уотсона по двухшаговой процедуре Дарбина, или по процедуре Кохрейна—Оркэтта, или по другой итерационной процедуре.

ОМНК позволяет строить модель регрессии по исходным уровням временных рядов, не используя методы исключения тенденции или включения в модель фактора времени.

7.4. Учет сезонности при построении модели регрессии

При отсутствии тенденции в рядах динамики и при наличии сезонной компоненты в модели регрессии зависимая переменная (y_t) может рассматриваться как функция объясняющих переменных $\langle x_1, x_2, \dots, x_p \rangle$ и сезонного фактора z , который можно ввести в модель в виде фиктивных переменных 0 и 1: например, для квартальных данных в модель вводятся фиктивные переменные z_1, z_2 и z_3 , учитывающие информацию I, II и III кварталов:

$z_1 = 1$ — для первого квартала,

$z_2 = 1$ — для второго квартала,

$z_3 = 1$ — для третьего квартала,

в остальных случаях $z_1 = z_2 = z_3 = 0$.

В общем виде линейная модель регрессии имеет вид:

$$y_t = a + \sum_{j=1}^p b_j x_{jt} + \sum_{i=1}^k c_i z_{ti},$$

где $k = 1, 2, 3$ (для квартальных данных), $k = 1, 2, 3, \dots, 11$ (для месячных данных). Построение подобного рода моделей требует достаточно длинных рядов динамики, чтобы параметры модели были статистически надежны.

Коэффициенты регрессии b_j ($j = 1, 2, \dots, p$) показывают воздействие соответствующего фактора x_j на y при устраниении сезонности. Влияние сезонности оценивается через коэффициенты c_i , которые представляют собой разность среднего уровня y для i -го квартала по сравнению с базовым, т. е. если за базу сравнения взят IV квартал, то $c_1 = \bar{y}_I - \bar{y}_{IV}$ при условии, что факторы x_1, x_2, \dots, x_p неизменны. Соответственно $c_2 = \bar{y}_{II} - \bar{y}_{IV}$, $c_3 = \bar{y}_{III} - \bar{y}_{IV}$. Чтобы оценить влияние сезонности I квартала по сравнению со II, можно использовать разность $c_2 - c_1$, а влияние III квартала по сравнению со II — $(c_3 - c_2)$ и т. п.

Практическое использование рассматриваемой модели затруднено рядом обстоятельств:

- ограничено число объясняющих переменных (x) ввиду сравнительно коротких динамических рядов в исследовании. Так, по данным за 5–7 лет, в поквартальном разрезе в модели может быть учтена одна объясняющая переменная;
- сезонность по y и по x может не совпадать по кварталам (например, пик доходов может приходиться на зиму, а пик расходов — на лето);

- коэффициент регрессии при x может подвергаться сезонности, а в модели он предполагается независимым от сезонной компоненты;
- ряды динамики обнаруживают тенденцию.

Рассмотрим содержание линейной модели $y_t = a + bx_t + c_1z_1 + c_2z_2 + c_3z_3 + \xi$, на следующем примере.

Пример. Объем продаж товара фирмой (y — тыс. ед.) исследуется в зависимости от объема продаж его дочерним предприятием (x — тыс. ед.) по данным за 5 лет.

Таблица 7.4

Годы	Кварталы								Итого	
	I		II		III		IV			
	y	x	y	x	y	x	y	x	y	x
2001	11	9	15	10	6	8	12	9	44	36
2002	11	10	16	9	4	3	13	11	44	33
2003	10	7	14	10	7	8	12	11	43	36
2004	10	12	16	9	8	11	13	12	47	44
2005	11	8	18	16	7	6	12	12	48	42
Итого	53	46	79	54	32	36	62	55	226	191
В среднем	10,6	9,2	15,8	10,8	6,4	7,2	12,4	11,0	45,2	38,2

Таблица показывает, что объем продаж практически стабилен по годам, но сильно варьирует по кварталам.

Взяв за базу для сравнения IV квартал, построим модель зависимости объема продаж фирмой от объема продаж дочерним предприятием с учетом сезонности:

$$\hat{y}_t = a + bx_t + c_1z_1 + c_2z_2 + c_3z_3,$$

где $z_1 = 1$ для I квартала и 0 для остальных;

$z_2 = 1$ для II квартала и 0 для остальных;

$z_3 = 1$ для III квартала и 0 для остальных.

Результаты регрессии оказались следующими:

$$\hat{y}_t = 8,896 + 0,319x_t - 1,227z_1 + 3,464z_2 - 4,789z_3$$

$$t \quad 8,3 \quad 3,5 \quad -2,15 \quad 6,3 \quad -7,4$$

$$R^2 = 0,955.$$

Они означают, что независимо от сезонного фактора с ростом объема продаж дочерним предприятием на 1 тыс. ед. объем продаж головной организации возрастает в среднем на 0,319 тыс. ед. Для отдельных кварталов модель регрессии составит:

для I квартала — $\hat{y}_t = 7,669 + 0,319x_t$;

для II квартала — $\hat{y}_t = 12,360 + 0,319x_t$;

для III квартала — $\hat{y}_t = 4,107 + 0,319x_t$;

для IV квартала — $\hat{y}_t = 8,896 + 0,319x_t$.

Коэффициенты регрессии при z показывают влияние сезонности: изменение объема продаж в головной организации в соответствующем квартале по сравнению с IV, когда $z_1 = z_2 = z_3 = 0$ в условиях неизменности объема продаж в дочернем предприятии. Все коэффициенты регрессии при сезонном факторе z согласовываются в примере с изменением средних уровней результата по отдельным кварталам в сравнении с IV:

По модели	По итоговой строке таблицы
$c_1 = -1,227$	$\bar{y}_I - \bar{y}_{IV} = 10,6 - 12,4 = -1,8$
$c_2 = +3,464$	$\bar{y}_{II} - \bar{y}_{IV} = 15,8 - 12,4 = 3,4$
$c_3 = -4,789$	$\bar{y}_{III} - \bar{y}_{IV} = 6,4 - 12,4 = -6$

Наибольшие изменения в средних уровнях наблюдаются в III и II кварталах (по сравнению с IV). Автокорреляция в остатках по модели практически отсутствует: $r_{a_1} = -0,165$.

Если в рядах динамики наблюдаются тенденция и сезонная компонента одновременно, то модель регрессии может иметь вид (при двух временных рядах)

$$\hat{y}_t = a + bx_t + ct + d_1z_1 + d_2z_2 + d_3z_3.$$

Так, если в нашем примере в модель ввести фактор времени $t = 1, 2, 3, \dots, 20$, то результаты окажутся следующими:

$$\begin{array}{l} \hat{y}_t = 8,874 + 0,291x_t + 0,027t - 1,195z_1 + 3,512z_2 - 4,866z_3 \\ t \quad 8,1 \quad 2,9 \quad 0,7 \quad -2,05 \quad 6,2 \quad -7,3 \\ R^2 = 0,957. \end{array}$$

Сравнивая эту модель с предыдущей, видим, что включение в модель фактора времени t в нашем случае нецелесообразно: t -критерий Стьюдента для параметра при t всего 0,7 при табличном значении 2,15 (для $\alpha = 0,05$), а R^2 выше всего на 0,2 процентных пункта. Этот вывод соответствует исходным данным примера, в котором предусмотрено отсутствие тенденции.

Следует заметить, что при наличии четкой тенденции как в ряду зависимой, так и в рядах объясняющих переменных модель с сезонной составляющей в виде линейно введенных фиктивных переменных не всегда может давать хорошие результаты ввиду мультиколлинеарности факторов, особенно если корреляция объясняющих переменных с фактором времени выше их корреляции с результативным признаком. В этом случае коэффициенты регрессии часто статистически незначимы или имеют неинтерпретируемые знаки.

В качестве демонстрации модели регрессии с учетом тенденции и сезонности рассмотрим взаимосвязь импорта и экспорта России по квартальным данным за 2000–2004 гг. Исходные данные представлены в табл. 7.5.

Таблица 7.5

Исходные данные

Кварталы	Экспорт, млрд долл.					Импорт, млрд долл.				
	2000 г.	2001 г.	2002 г.	2003 г.	2004 г.	2000 г.	2001 г.	2002 г.	2003 г.	2004 г.
I	24,4	25,6	21,9	31,1	37,3	10	11,3	12,3	16	19,6
II	25	26,2	26,3	31,7	43,2	10,4	13,6	14,8	18,2	22,7
III	26,6	25,6	28,9	34,9	48,5	11,1	13,2	15,7	19,6	24,7
IV	29,5	24,6	30,2	38,2	54,5	13,4	15,6	18,1	22,3	29,3

Источник данных: Факты. Оценки, прогнозы // Вопросы статистики. 12. 2005. С. 55–66.

Построим регрессию импорта (y_t) на экспорт (x_t), введя в модель фактор времени t (1, 2, 3, ..., 20) и сезонную компоненту в виде фиктивных переменных z_1 , z_2 и z_3 , принимающих значение 1 соответственно для I, II и III кварталов и 0 — для остальных. Применяя к данным табл. МНК, получим следующие результаты:

$$\hat{y}_t = 4,615 + 0,271x_t - 2,528z_1 - 1,545z_2 - 1,742z_3 + 0,461t$$

$t_{\text{критерий}}$	6,7	9,7	-6,8	-4,3	-5,0	11,6
R^2	= 0,992.					

Модель показывает, что независимо от тенденции и сезонности росту экспорта на 1 млрд долл. соответствует увеличение импорта на 271 млн долл. Ежеквартально под воздействием тенденции импорт возрастает в среднем на 461 млн долл. При абстрагировании от влияния тенденции и размера экспорта сезонность проявляется отчетливо: в IV квартале объем импорта выше, чем в предыдущих кварталах (все коэффициенты регрессии при сезонной компоненте меньше нуля и статистически значимы по t -критерию Стьюдента).

Сезонность можно рассматривать и в более укрупненном плане: например, осенне-зимний период и весенне-летний период. В этом случае фиктивная переменная может принимать значение 1, например, для I и IV кварталов, а для II и III кварталов — значение 0 (или наоборот). Такой подход уменьшает число вводимых в модель фиктивных переменных, что может быть полезным при сравнительно коротких рядах динамики.

Контрольные вопросы

- Что такое ложная корреляция?
- Почему регрессия по временным рядам требует специальных методов построения модели?
- В чем состоит смысл метода первых разностей?
- Опишите алгоритм построения модели регрессии по отклонениям от тренда.
- Как учитывается тенденция при построении в нее фактора времени?

6. В каком случае модель по отклонениям от тренда совпадает с моделью с включением в нее фактора времени?
7. В чем смысл использования обобщенного метода наименьших квадратов (ОМНК) при построении модели регрессии по временным рядам?
8. Какова цель поправки Прайса — Винстена?
9. Что представляет собой метод Кохрейна — Оркэтта?
10. Опишите алгоритм обобщенного метода наименьших квадратов для построения модели регрессии по временным рядам.

ГЛАВА 8. МОДЕЛИ С ЛАГОВЫМИ ПЕРЕМЕННЫМИ

8.1. Общая характеристика

До сих пор мы рассматривали модели по временным рядам, в которых $y_t = f(x_t)$. Между тем в моделях временных рядов зависимая переменная y_t , может быть связана не только со значениями объясняемых переменных x в момент времени t , но и с их значениями в предыдущие моменты времени. Так, например, потребление товаров длительного пользования зачастую зависит не только от доходов текущего, но и предыдущих периодов. Аналогично величина основных производственных фондов зависит от размера инвестиций не только текущего года, но и предыдущих лет. В этом случае строятся модели с лаговыми объясняющими переменными. Например,

$$c_t = a + b_1 y_t + b_2 y_{t-1} + \xi_t,$$

где c_t — потребление в период времени t ;

y_t — доход в период времени t ;

y_{t-1} — доход в предыдущий период $t - 1$.

В данной модели лаговой является переменная y_{t-1} , т. е. доход за предыдущий период времени. Возможна ситуация, когда объясняющая переменная x влияет на результат у не сразу же, а с определенным запаздыванием во времени, превышающим один временной интервал. Так, выпуск специалистов высшей квалификации зависит от приема в вузы четырех-пятилетней давности.

Объясняющие переменные, взятые в модели регрессии с запаздыванием во времени, называются *лаговыми переменными*. Величина интервала запаздывания называется *лагом*. Так, в модели $y_t = a + b_1 x_{t-4} + x_t$ лаговая переменная взята с лагом, равным 4.

Вместе с тем в правой части модели лаговой может быть и зависимая переменная. Например, спрос на товар может зависеть не только от дохода, но и от достигнутого спроса на него в предыдущий период времени. Или ставка банковского кредита может зависеть не только от объема денежной массы в наличии, но и от достигнутого ранее процента банковского кредита. В этом случае строятся модели с лаговой зависимой переменной. Например,

$$C_t = a + b_1 y_t + b_2 C_{t-1} + \xi_t,$$

где C_t — потребление в период времени t ;

y_t — доход в период времени t ;

C_{t-1} — потребление в предыдущий период времени $t - 1$.

Модели регрессии по временным рядам с лаговыми переменными принято называть *динамическими моделями*. Их можно подразделить на три класса:

1) модели с лаговыми объясняющими переменными, или, иначе, *модели с распределенными лагами*:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \dots + b_k x_{t-k} + \xi_t;$$

2) модели с лаговыми зависимыми переменными – *модели авторегрессии*:

$$y_t = a + b x_t + c_1 y_{t-1} + \dots + c_k y_{t-k} + \xi_t;$$

3) модели с лаговыми зависимыми и независимыми переменными, т. е. авторегрессионные модели с распределенными лагами:

$$y_t = a + b_1 y_{t-1} + \dots + b_k y_{t-k} + c_0 x_t + c_1 x_{t-1} + \dots + c_k x_{t-k} + \xi_t.$$

Центральным вопросом при построении моделей с лаговыми переменными является выбор величины лага и числа лаговых переменных. Теоретически трудно определить величину лага. Определенную помощь может оказать взаимная корреляционная функция: рассчитывается множество коэффициентов корреляции между уровнями временных рядов (y_t) и (x_t), сдвинутыми относительно друг друга на последовательно увеличивающиеся интервалы времени. Величина лага определяется по максимальному значению коэффициента корреляции. Например, продажа товара за две декады двумя филиалами фирмы характеризуется данными (тыс. ден. ед.):

Числа месяца	Филиал № 1	Филиал № 2	Числа месяца	Филиал № 1	Филиал № 2
1	5	9	11	9,5	13
2	4,5	10,8	12	8	14
3	4	13,5	13	7,6	15
4	4,1	14,5	14	7,5	20
5	5	16	15	7,6	24
6	7	14,7	16	10	25
7	8	14	17	12,2	26
8	9,7	12	18	15	26,3
9	10	11,9	19	15,6	26,4
10	11	12	20	16	27,1

Примем объем продаж филиалом № 1 за y_t , а филиалом № 2 – за x_t . Если прокоррелировать y_t и x_t , то коэффициент корреляции между ними составит 0,6912. Последовательно сдвигая уровни ряда x_t на один временной интервал, коэффициенты корреляции окажутся:

Величина лага							
1	2	3	4	5	6	7	
0,7738	0,867	0,9445	0,9553	0,8562	0,5977	0,0724	

Следовательно, объем продаж филиалом № 1 в наибольшей мере коррелирует с объемом продаж по филиалу № 2 с интервалом в 4 дня. Уравнение регрессии составит: $\hat{y}_t = -0,4575 + 0,6977x_{t-4}$; $R^2 = 0,913$ и $F = 146,2$, что статистически значимо. Оно позволяет по данным филиала № 2, взятым на 4 дня раньше, предсказывать объем продаж по филиалу № 1. Так, например, при объеме продаж за 2-е число в 10,8

тыс. ден. ед. по филиалу № 2 объем продаж по филиалу № 1 составит 6-го числа 7,1 тыс. ден. ед. Соответственно подставляя в уравнение регрессии информацию об объеме продаж филиалом № 2 за 3-е – 16-е числа получим объем продаж по филиалу № 1 на 7-е – 20-е числа.

Выбор величины лага и количества лагов проводится обычно экспериментально: строятся модели с разным числом лагов и их величиной и изучается значимость коэффициентов регрессии при лаговых переменных; останавливаются на модели, для которой все коэффициенты регрессии при лаговых переменных будут статистически значимыми по *t*-критерию Стьюдента.

Построение моделей с лаговыми переменными имеет свою специфику. Дело не только в выборе величины лага и их числа. Во многих случаях оценка параметров моделей с лаговыми переменными не может быть проведена с помощью традиционного МНК ввиду нарушения ряда его предпосылок и требует специальных методов оценивания. При наличии двух и более лаговых переменных возникает проблема мультиколлинеарности факторов, ибо, как правило, x_t , x_{t-1} , x_{t-2} , x_{t-k} или y_{t-1} , y_{t-2} , ..., y_{t-k} связаны между собой, особенно при наличии тенденции в рядах динамики. Это снижает точность оценок коэффициентов при лаговых переменных и требует видоизменять приемы оценивания.

8.2. Модели с распределенными лагами

8.2.1. Интерпретация параметров модели с распределенными лагами

Модели с распределенными лагами бывают двух типов:

- с конечным числом лагов:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \dots + b_k x_{t-k} + \xi_t;$$

- с бесконечным числом лагов:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \dots + \xi_t.$$

Практическое применение чаще имеют модели с конечным числом лагов, т. е. модели, в которых число лагов экспериментально определено.

Предположим рассматривается модель, в которой $k = 4$, т. е. $\hat{y}_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + b_3 x_{t-3} + b_4 x_{t-4}$. Данная модель означает, что изменение во времени t объясняющей переменной x будет влиять на значения результативного признака y в течение 4 следующих моментов времени.

Коэффициент b_0 называют краткосрочным мультипликатором, так как он характеризует среднее изменение результата y при изменении x_t на 1 единицу своего измерения в фиксированный момент времени t .

В момент времени $t+1$ воздействие объясняющей переменной x на результат y составит $(b_0 + b_1)$ единиц, а в момент времени $t+2$ общее изменение y составит $(b_0 + b_1 + b_2)$ единиц.

Любую сумму коэффициентов $\sum_{j=0}^h b_j$, где ($h < k$) называют промежуточным мультипликатором, а сумму всех коэффициентов регрессии $\sum_{j=0}^k b_j$ – долгосрочным мультипликатором, который характеризует общее изменение y через k интервалов времени под воздействием изменения x в момент t на 1 единицу.

При $k = 4$ долгосрочный мультипликатор составит $b_0 + b_1 + b_2 + b_3 + b_4$. Он характеризует общее среднее изменение y через 4 временных интервала при увеличении x в момент времени t на 1 единицу, а промежуточные мультипликаторы:

- $b_0 + b_1$ – изменение y в момент времени $t + 1$;
- $b_0 + b_1 + b_2$ – изменение y в момент времени $t + 2$;
- $b_0 + b_1 + b_2 + b_3$ – изменение y в момент времени $t + 3$.

Если все коэффициенты регрессии имеют одинаковые знаки, т. е. характеризуются односторонним изменением y в исследуемые k моментов времени, то можно определять относительные коэффициенты модели β_j , т. е. $\beta_j = \frac{b_j}{\sum b_j}$, где $0 < \beta_j < 1$, а $\sum \beta_j = 1$. Иными словами,

β_j характеризует долю общего изменения y в момент времени $t + j$.

Предположим, что регрессия основных производственных фондов (y – млн руб.) в зависимости от размера инвестиций (x – млн руб.) характеризуется уравнением

$$\hat{y}_t = 0,8 + 0,7x_t + 1,0x_{t-1} + 1,5x_{t-2} + 0,6x_{t-3} + 0,2x_{t-4}, \text{ где } t \text{ – года.}$$

Анализ уравнения показывает, что рост инвестиций на 1 млн руб. в текущем периоде приводит к росту основных производственных фондов:

- в том же периоде на 0,7 млн руб. (краткосрочный мультипликатор),
- через 1 год на $0,7+1 = 1,7$ млн руб.;
- через 2 года на $0,7+1+1,5 = 3,2$ млн руб.;
- через 3 года на 3,8 млн руб. (промежуточный, как и предыдущие два, мультипликатор);
- через 4 года на 4 млн руб. (долгосрочный мультипликатор).

Относительные коэффициенты модели составят:

$$\begin{aligned}\beta_0 &= 0,7/4 = 0,175; \\ \beta_1 &= 1/4 = 0,25; \\ \beta_2 &= 1,5/4 = 0,375; \\ \beta_3 &= 0,6/4 = 0,15; \\ \beta_4 &= 0,2/4 = 0,05.\end{aligned}$$

Следовательно, в текущем году реализуется 17,5% воздействия увеличения инвестиций на рост основных производственных фондов,

а через год – еще 25%. Через 2 года – еще 37,5%, через 3 года – еще 15% и через 4 года – еще 5%.

Относительные коэффициенты модели (β_j) можно использовать как весовые коэффициенты для расчета *средней величины лага* по средней арифметической:

$$\bar{j} = \sum_{j=0}^k j\beta_j.$$

Величина \bar{j} показывает средний интервал времени, в течение которого будет происходить изменение зависимой переменной y под воздействием изменения объясняющей переменной x в момент времени t . Чем меньше величина среднего лага, тем быстрее реагирует результат y на изменение x . И наоборот, высокое значение среднего лага показывает, что воздействие объясняющей переменной на результат будет сказываться с течением длительного промежутка времени. В рассматриваемом примере величина среднего лага составит

$$\bar{j} = 0 \cdot 0,175 + 1 \cdot 0,25 + 2 \cdot 0,375 + 3 \cdot 0,15 + 4 \cdot 0,05 = 1,65 \text{ года.}$$

Следовательно, основная часть эффекта увеличения инвестиций проявляется через 1,65 года. Кроме среднего лага, можно рассчитывать *медианный лаг* j_M , т. е. тот период времени, в течение которого с момента времени t будет реализована половина общего эффекта воздействия объясняющей переменной x на результат y . Для медианного лага справедливо равенство

$$\sum_{j=0}^{M_c} \beta_j = 0,5.$$

В нашем примере медианный лаг составляет два года, т. е. увеличение инвестиций в период времени t на 1 млн руб. приводит к росту размера основных производственных фондов через 2 года на величину, составляющую половину долгосрочного мультиплексора, т. е. на 2 млн руб. Наибольший аналитический интерес представляет расчет величины медианного лага для моделей с большим числом лаговых переменных.

8.2.2. Оценка параметров моделей с распределенными лагами

Модель с конечным числом лагов при правильной ее спецификации может быть оценена обычным МНК. В этом случае в уравнении:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \dots + b_k x_{t-k} + \xi_t,$$

переменные $x_t, x_{t-1}, \dots, x_{t-k}$ рассматриваются как объясняющие переменные обычной множественной регрессии.

Вместе с тем применение МНК к моделям с конечным числом лагов может быть реально затруднено ввиду следующих причин:

1) при наличии тенденции переменные $x_t, x_{t-1}, \dots, x_{t-k}$ тесно связаны между собой, что вызывает мультиколлинеарность факторов, которая может привести к неинтерпретируемым знакам у коэффициентов регрессии и к снижению их точности;

2) возможна автокорреляция остатков, так как МНК применяется к временным рядам с тенденцией.

Поэтому нередко для оценки параметров модели с распределенным конечным числом лагов используются специальные методы преобразования, как и для модели с бесконечным числом лагов. Разработаны разные методы оценивания параметров моделей с распределенными лагами, которые учитывают характер распределения коэффициентов регрессии при лаговых объясняющих переменных. Иными словами, методы оценивания параметров модели с распределенными лагами основаны на изучении структуры лага. Так, предполагая полиномиальное распределение лаговых коэффициентов, используется *метод Алмон*, а при гипотезе геометрической прогрессии для лаговых коэффициентов применяется *преобразование Койка*.

8.2.2.1. Полиномиально распределенные лаги Алмон

В 1965 г. Ширли Алмон предложила способ оценки параметров модели с распределенными лагами на основе гипотезы, что лаговые коэффициенты регрессии аппроксимируются полиномом соответствующей степени от величины лага. Это значит, что в модели параметр b_j рассматривается как функция: $b_j = c_0 + c_1 j + c_2 j^2 + \dots + c_m j^m$. При этом априори выдвигается предположение о степени полинома. Как правило, используется многочлен невысокой степени ($m \leq 4$).

Предположим, что b_j имеет распределение в виде параболы второй степени, т. е. $b_j = c_0 + c_1 j + c_2 j^2$. Тогда каждый из коэффициентов можно представить как

$$\begin{aligned} b_0 &= c_0; \\ b_1 &= c_0 + c_1 + c_2; \\ b_2 &= c_0 + 2c_1 + 4c_2; \\ b_3 &= c_0 + 3c_1 + 9c_2; \\ &\vdots \\ b_k &= c_0 + kc_1 + k^2 c_2. \end{aligned}$$

Подставим эти соотношения для b_j в модель с распределенными лагами:

$$\begin{aligned} y_t &= a + c_0 x_t + (c_0 + c_1 + c_2)x_{t-1} + (c_0 + 2c_1 + 4c_2)x_{t-2} + \\ &+ (c_0 + 3c_1 + 9c_2)x_{t-3} + \dots + (c_0 + kc_1 + k^2 c_2)x_{t-k} + \xi_t. \end{aligned}$$

Перегруппируем слагаемые с одинаковыми значениями с:

$$\begin{aligned} y_t &= a + c_0(x_t + x_{t-1} + x_{t-2} + \dots + x_{t-k}) + \\ &+ c_1(x_{t-1} + 2x_{t-2} + 3x_{t-3} + \dots + kx_{t-k}) + \\ &+ c_2(x_{t-1} + 4x_{t-2} + 9x_{t-3} + \dots + k^2 x_{t-k}) + \xi_t. \end{aligned}$$

Будем рассматривать слагаемые в скобках при c_0 , c_1 и c_2 как новые переменные z , т. е. модель с распределенными лагами примет вид:

$$y_t = a + c_0 z_0 + c_1 z_1 + c_2 z_2 + \xi_t,$$

где z_0, z_1 и z_2 определяются как

$$z_0 = \sum_{j=0}^k x_{t-j};$$

$$z_1 = \sum_{j=1}^k jx_{t-j};$$

$$z_2 = \sum_{j=1}^k j^2 x_{t-j}.$$

Оценка параметров при преобразованных переменных z дается традиционным МНК. При этом случайные отклонения ξ_t , удовлетворяют предпосылкам МНК. Далее на основе параметров c_0, c_1 и c_2 переходим к оценке параметров b_j , используя выражения коэффициентов через коэффициенты полинома:

$$b_j = c_0 + c_1 j + c_2 j^2.$$

В общем виде при степени полинома m модель регрессии с распределенными лагами составит

$$y_t = a + c_0(x_t + x_{t-1} + \dots + x_{t-k}) + c_1(x_{t-1} + 2x_{t-2} + 3x_{t-3} + \dots + kx_{t-k}) + \\ + c_2(x_{t-1} + 4x_{t-2} + 9x_{t-3} + \dots + k^2 x_{t-k}) + \dots + \\ + c_m(x_{t-1} + 2^m x_{t-2} + 3^m x_{t-3} + \dots + k^m x_{t-k}) + \xi_t,$$

или иначе: $y_t = a + c_0 z_0 + c_1 z_1 + c_2 z_2 + \dots + c_m z_m + \xi_t$.

Как видим, в данной модели переменные z_1, z_2, \dots, z_m представляют собой линейную комбинацию переменных x_t и k лаговых переменных, веса при которых подчиняются полиномиальному распределению:

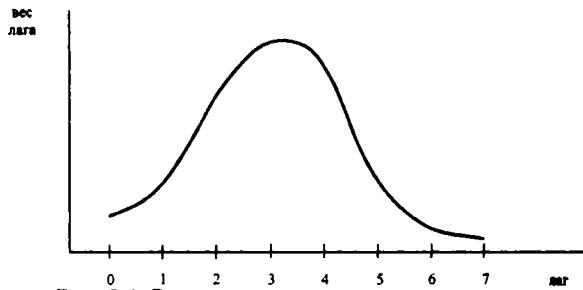


Рис. 8.1. Распределение лаговых переменных

В матричном виде можно записать, что $b = Hc$, где

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 2^m \\ 1 & 3 & 9 & 3^m \\ \dots & & & \\ 1 & k & k^2 & k^m \end{bmatrix} -$$

— матрица весов при лаговых коэффициентах b_j ;

c — вектор коэффициентов при переменных z . Тогда модель в целом принимает вид

$$y = XHc + \xi = Zc + \xi.$$

Стандартная ошибка коэффициентов регрессии при лаговых переменных определится как

$$\mu_{b_j} = S \sqrt{H[z'z]^{-1} H}.$$

Далее через t -критерий Стьюдента оценивается значимость коэффициентов b_j .

Качество модели оценивается через коэффициент детерминации R^2 для уравнения регрессии y_t от преобразованных переменных z , т. е. по модели

$$y = Zc + \xi.$$

Таким образом, применение метода Алмон включает в себя следующие этапы работы:

- 1) определение максимальной величины лага (k);
- 2) определение степени полинома (m), описывающего распределение коэффициентов регрессии (b_j) в зависимости от величины лага;
- 3) расчет преобразованных переменных z_j ;
- 4) расчет параметров линейной регрессии y от преобразованных переменных z , т. е. оценка c_j ;
- 5) переход к исходным параметрам (b_j) модели с распределенными лагами.

Теоретически достаточно сложно определить максимальную величину лага (k). В основном для этой цели используется экспериментальный путь: строится уравнение с большим числом последовательных лагов и, постепенно его уменьшая, изучается значимость коэффициентов регрессии при лаговых объясняющих переменных.

Определение степени полинома m также связано с рядом трудностей. Формально можно изучать графически структуру лага (рис. 8.2).

Если с ростом величины лага j коэффициенты b_j описываются кривыми, представленными на рис. 8.2а–г, то в расчетах могут быть использованы полиномы второй, третьей или четвертой степени. Рис. 8.2д предполагает линейную зависимость от величины лага, а рис. 8.2, е показывает перевернутую V-образную структуру лага (например, при изучении капитальных вложений Де Лью в 1962 г. предложил подобную структуру лага). Однако учитывая, что оценки b_j по МНК часто затруднены, исследователь не располагает подобными графиками. Поэтому степень полинома задается исследователем исходя из соответствующих теоретических соображений и результатов предыдущих исследований.

Пример. По данным за 32 квартала об объеме продукции (y – млн руб.) и инвестициях в основной капитал (x – млн руб.) строится модель с распределенными лагами:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + b_3 x_{t-3} + b_4 x_{t-4} + \xi_t.$$

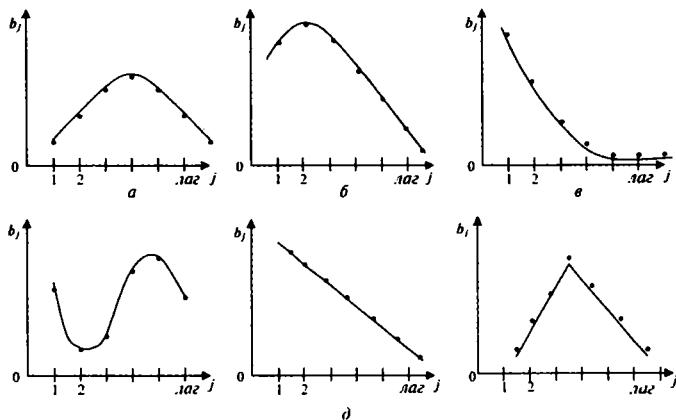


Рис. 8.2. Возможные распределения лаговых коэффициентов регрессии

Таблица 8.1

квартала	y_t	x_t	x_{t-1}	x_{t-2}	x_{t-3}	x_{t-4}	z_0	z_1	z_2
1	5,2	0,87	—	—	—	—	—	—	—
2	5,6	0,9	0,87	—	—	—	—	—	—
3	6,5	1,05	0,9	0,87	—	—	—	—	—
4	6,4	1,04	1,05	0,9	0,87	—	—	—	—
5	6,5	1,05	1,04	1,05	0,9	0,87	4,91	9,32	27,26
6	7,0	1,08	1,05	1,04	1,05	0,9	5,12	9,88	29,06
7	7,4	1,12	1,08	1,05	1,04	1,05	5,34	1,05	31,44
8	7,8	1,16	1,12	1,08	1,05	1,04	5,45	10,59	31,53
9	8,1	1,17	1,16	1,12	1,08	1,05	5,58	10,84	32,16
10	8,0	1,14	1,17	1,16	1,12	1,08	5,67	11,17	33,17
11	8,5	1,17	1,14	1,17	1,16	1,12	5,76	11,44	34,18
12	8,6	1,2	1,17	1,14	1,17	1,16	5,84	11,6	34,82
13	8,8	1,2	1,2	1,17	1,14	1,17	5,88	11,64	34,86
14	8,9	1,24	1,2	1,2	1,17	1,14	5,95	11,67	34,77
15	8,9	1,22	1,24	1,2	1,2	1,17	6,03	11,92	35,56
16	9,3	1,26	1,22	1,24	1,2	1,2	6,12	12,1	36,18
17	9,4	1,23	1,26	1,22	1,24	1,2	6,15	12,22	36,5
18	9,3	1,23	1,23	1,26	1,22	1,24	6,18	12,37	37,09
19	9,6	1,26	1,23	1,23	1,26	1,26	6,2	12,35	37,01

Окончание табл. 5.2

квартала	y_t	x_t	x_{t-1}	x_{t-2}	x_{t-3}	x_{t-4}	z_0	z_1	z_2
20	9,7	1,28	1,26	1,23	1,23	1,26	6,26	12,45	37,41
21	9,7	1,3	1,28	1,26	1,23	1,23	6,3	12,41	37,07
22	9,8	1,32	1,3	1,28	1,26	1,23	6,39	12,56	37,44
23	10,0	1,32	1,32	1,3	1,28	1,26	6,48	12,8	38,2
24	10,2	1,33	1,32	1,32	1,3	1,28	6,55	12,98	38,78
25	10,3	1,33	1,33	1,32	1,32	1,3	6,6	13,13	39,29
26	10,4	1,35	1,33	1,33	1,32	1,32	6,65	13,23	39,65
27	10,5	1,35	1,35	1,33	1,33	1,32	6,68	13,28	39,76
28	10,6	1,36	1,35	1,35	1,33	1,33	6,72	13,36	40
29	10,5	1,32	1,36	1,35	1,35	1,33	6,71	13,43	40,19
30	10,6	1,35	1,32	1,36	1,35	1,35	6,73	13,49	40,51
31	10,7	1,38	1,35	1,32	1,36	1,35	6,76	13,47	40,47
32	11	1,4	1,38	1,35	1,32	1,36	6,81	13,48	40,42

Предполагая квадратичную зависимость b_j от величины лага $b_j = c_0 + c_1 j + c_2 j^2$, имеем следующие соотношения:

$$b_0 = c_0;$$

$$b_1 = c_0 + c_1 + c_2;$$

$$b_2 = c_0 + 2c_1 + 4c_2;$$

$$b_3 = c_0 + 3c_1 + 9c_2;$$

$$b_4 = c_0 + 4c_1 + 16c_2.$$

Соответственно модель с распределенными лагами примет вид

$$y_t = a + c_0 z_0 + c_1 z_1 + c_2 z_2 + \xi_t.$$

Расчет преобразованных переменных z_j представлен в табл. 8.1, где

$$z_0 = x_t + x_{t-1} + x_{t-2} + x_{t-3} + x_{t-4};$$

$$z_1 = x_{t-1} + 2x_{t-2} + 3x_{t-3} + 4x_{t-4};$$

$$z_2 = x_{t-1} + 4x_{t-2} + 9x_{t-3} + 16x_{t-4}.$$

Применяя к данным об y , z_0 , z_1 и z_2 обычный МНК, получим следующее уравнение:

$$y_t = -4,7115 + 3,7713 z_0 - 2,2668 z_1 + 0,5065 z_2 + \xi_t.$$

$$t \quad -15,0 \quad 5,8 \quad -2,6 \quad 2,6$$

Все параметры уравнения статистически значимы ($t_{0,05} = 2,06$ при $df = 24$). $R^2 = 0,9955$ указывает на хорошее качество модели.

Далее найдем коэффициенты регрессии исходной модели, т. е. b_j , используя выражения b_j через коэффициенты c_0 , c_1 и c_2 :

$$b_0 = 3,7713;$$

$$b_1 = 3,7713 + (-2,2668) + 0,5065 = 2,011;$$

$$b_2 = 3,7713 - 2 \cdot 2,2668 + 4 \cdot 0,5065 = 1,2637;$$

$$b_3 = 3,7713 - 3 \cdot 2,2668 + 9 \cdot 0,5065 = 1,5294;$$

$$b_4 = 3,7713 - 4 \cdot 2,2668 + 16 \cdot 0,5065 = 2,8081.$$

Модель регрессии с распределенными лагами составит

$$\hat{y}_t = -4,711 + 3,771x_t + 2,011x_{t-1} + 1,264x_{t-2} + 1,529x_{t-3} + 2,808x_{t-4}$$

$$R^2 = 0,9955.$$

Стандартные ошибки коэффициентов регрессии по модели следующие:

$$m_{b_0} = 0,651; m_{b_1} = 0,296; m_{b_2} = 0,4045; m_{b_3} = 0,300; m_{b_4} = 0,494.$$

Для свободного члена a стандартная ошибка составила 0,313. Соответственно по t -критерию Стьюдента все параметры оказались статистически значимыми: $t_a = -15,0$; $t_{b_0} = 5,8$; $t_{b_1} = 6,8$; $t_{b_2} = 3,1$; $t_{b_3} = 5,1$; $t_{b_4} = 5,7$.

Модель показывает, что рост инвестиций в текущем периоде на 100 тыс. руб. способствует росту объема продукции в том же периоде в среднем на 377 тыс. руб., а через квартал – на 578 тыс. руб. В целом же через год прирост объема продукции за счет роста инвестиций на 100 тыс. руб. ожидается в размере 1,138 млн руб. ($3,771 + 2,011 + 1,264 + 1,529 + 2,808 = 11,383$).

Определив относительные коэффициенты регрессии β_j , увидим, что половина воздействия фактора на результат реализуется с лагом в один квартал:

$$\beta_0 = \frac{3,771}{11,383} = 0,331;$$

$$\beta_1 = \frac{2,011}{11,383} = 0,177;$$

$$\beta_2 = \frac{1,264}{11,383} = 0,111;$$

$$\beta_3 = \frac{1,529}{11,383} = 0,134;$$

$$\beta_4 = \frac{2,808}{11,383} = 0,247.$$

На графике рассматриваемые коэффициенты регрессии представляют собой параболу второй степени:

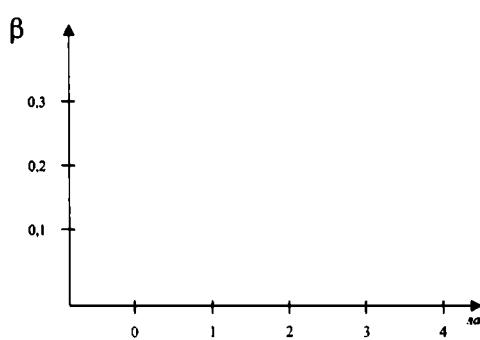


Рис. 8.3. Коэффициент регрессии

Если к исходным данным нашего примера применить традиционный МНК, то результаты окажутся следующими:

$$y_t = -4,782 + 4,390x_t + 1,193x_{t-1} + 1,372x_{t-2} + 2,001x_{t-3} + 2,479x_{t-4}$$

$$R^2 = 0,9957.$$

Хотя коэффициент детерминации здесь даже чуть-чуть выше, но коэффициенты регрессии при лаговых переменных x_{t-1} и x_{t-2} оказываются статистически незначимыми:

$$t_a = -14,7; t_{b_0} = 5,1; t_{b_1} = 1,2; t_{b_2} = 1,7; t_{b_3} = 3,1; t_{b_4} = 4,1.$$

Кроме того, применяя метод Алмон, получаем стандартные ошибки коэффициентов регрессии меньше, чем при традиционном МНК:

$$m_{b_0} = 0,868; m_{b_1} = 0,983; m_{b_2} = 0,811; m_{b_3} = 0,654; m_{b_4} = 0,610.$$

8.2.2.2. Метод Койка

Для модели с бесконечным числом лаговых значений объясняющей переменной

$$y_t = a + b_0x_t + b_1x_{t-1} + b_2x_{t-2} + \dots + \xi_t, \quad (8.1)$$

оценка параметров не представляется возможной без какого-либо допущения относительно поведения коэффициентов при лаговых переменных. Одним из допущений является предположение о том, что после некоторой длины лага (например, k) коэффициенты распределенного лага начнут убывать геометрически с одинаковым темпом λ ($0 < \lambda < 1$). Тогда уравнение (8.1) может быть записано в виде

$$y_t = a + b_0x_t + b_1x_{t-1} + \dots + b_kx_{t-k} + b_k\lambda x_{t-k-1} + b_k\lambda^2 x_{t-k-2} + \dots + U_t. \quad (8.2)$$

В уравнении (8.2) первые k коэффициентов распределенного лага являются свободными (принимают любые значения), а остальные лаговые коэффициенты убывают в геометрической прогрессии.

Если в уравнении (8.2) предположить, что убывание лаговых коэффициентов в геометрической прогрессии происходит сразу же, а не через интервал времени k , то получим следующую модель:

$$y_t = a + b_0x_t + b_0\lambda x_{t-1} + b_0\lambda^2 x_{t-2} + \dots + U_t. \quad (8.3)$$

Коэффициенты данной модели согласовываются с коэффициентами уравнения (8.1), а именно:

$$b_j = b_0\lambda^j; j = 0, 1, 2, \dots \quad (8.4)$$

Это означает, что, оценив три параметра уравнения (8.3), т. е. a , b_0 и λ , можно перейти к модели (8.1): a и b_0 определены по модели (8.3), $b_1 = b_0\lambda$; $b_2 = b_0\lambda^2$; $b_3 = b_0\lambda^3$ и т. д.

Однако наличие в модели (8.3) бесконечного числа лаговых переменных затрудняет практическую ее реализацию, ибо исследователь имеет дело, как правило, с конечным числом лагов. Оценка параметров модели (8.3) возможна, если применить преобразование Койка.

Предполагая, что в модели (8.1) все лаговые коэффициенты имеют одинаковый знак и уменьшаются в геометрической прогрессии, Л. М. Койк предложил для оценки параметров модели (8.3) следующую процедуру:

- построить модель (8.3) для момента времени $(t - 1)$, т. е. получим уравнение

$$y_{t-1} = a + b_0 x_{t-1} + b_0 \lambda x_{t-2} + b_0 \lambda^2 x_{t-3} + \dots + \xi_{t-1}; \quad (8.5.)$$

- умножить уравнение (8.5) на λ , т. е. будем иметь

$$\lambda y_{t-1} = \lambda a + b_0 \lambda x_{t-1} + b_0 \lambda^2 x_{t-2} + b_0 \lambda^3 x_{t-3} + \dots + \lambda \xi_{t-1}; \quad (8.6)$$

- вычесть из уравнения (8.3) уравнение (8.6):

$$y_t - \lambda y_{t-1} = (1 - \lambda)a + b_0 x_t + (\xi_t - \lambda \xi_{t-1}).$$

После преобразования получим

$$y_t = (1 - \lambda)a + b_0 x_t + \lambda y_{t-1} + U_t, \quad (8.7)$$

где $U_t = \xi_t - \lambda \xi_{t-1}$.

Уравнение (8.7) получило название *преобразование Койка*, так как Л. М. Койк впервые (1954 г.) предложил данный подход к оцениванию параметров модели с распределенными лагами.

Практически в модели (8.7) от уравнения с распределенными лагами с бесконечным их числом (8.1) Л. М. Койк перешел к модели авторегрессии, для которой требуется оценить всего три параметра: a , b_0 и λ . Далее из соотношения (8.4) найдем параметры исходной модели (8.1).

Рассмотренный подход нашел широкое применение в исследовании кумулятивного эффекта рекламы на объем продаж, т. е. текущий объем продаж рассматривается в зависимости от расходов на рекламу текущего периода, объема продаж в предыдущий период времени и ошибки U_t^* .

Преобразование Л. М. Койка может быть использовано и при решении модели (8.2), когда несколько первых коэффициентов остаются свободными, а для оставшихся лагов реализуется преобразование Койка. Например, считая, что b_0 и b_1 остаются свободными, а начиная с b_2 все лаговые коэффициенты убывают с одинаковым темпом, можно записать:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \lambda b_2 x_{t-3} + \lambda^2 b_2 x_{t-4} + \dots + U_t.$$

Далее применив преобразование Койка, получим, $y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \lambda b_2 x_{t-3} + \lambda^2 b_2 x_{t-4} + \dots + U_t$, т. е. переходим к модели авторегрессии с распределенными лагами.

Преобразование Койка приводит к существенным упрощениям, ибо вместе с уменьшением числа оцениваемых параметров устраняется и проблема мультиколлинеарности факторов: теперь в модели (8.7) содержатся две независимые переменные x_t и y_{t-1} .

Модель Л. М. Койка позволяет анализировать краткосрочный и долгосрочный мультиплекторы. Краткосрочным мультиплектором является параметр b_0 , а долгосрочным – сумма коэффициентов регрессии, представляющая собой сумму геометрической прогрессии:

*См. подробнее вложение в кн.: Бернд Э. Практика эконометрики: классика и современность: пер. с англ. М.: ЮНИТИ, 2005. С. 457–467.

$$\sum_{j=0}^{\infty} b_j = b_0 + b_0\lambda + b_0\lambda^2 + b_0\lambda^3 + \dots = b_0(1 + \lambda + \lambda^2 + \lambda^3 + \dots) = b_0 \cdot \frac{1}{1-\lambda}.$$

Например, по Великобритании для периода 1924–1938 гг. была построена модель $C_t = 0,18R_t + 0,81C_{t-1}$, где C_t – потребление в период времени t ; R_t – доход в период времени t ; C_{t-1} – потребление в период времени $t-1$ *.

Данное уравнение означает, что краткосрочная склонность к потреблению составляет 0,18, а долгосрочная склонность к потреблению равна $0,18/(1-0,81) = 0,95$. Отсутствие в модели свободного члена не изменяет суть интерпретации краткосрочного и долгосрочного мультипликаторов, хотя, естественно, сказывается на величине параметров модели. Краткосрочный мультипликатор 0,18 показывает, что с ростом дохода на 1 ден. ед. потребление в тот же период времени увеличивается на 0,18 ден. ед. Долгосрочный мультипликатор 0,95 означает, что в долгосрочной перспективе увеличение дохода на 1 ден. ед. приведет к росту потребления на 0,95 ден. ед.

В модели Койка (8.7) случайная ошибка $U_t = \xi_t - \lambda\xi_{t-1}$ коррелирована с переменной y_{t-1} . Поэтому оценивание параметров ее модели традиционным МНК дает смещенные и несостоительные оценки. Вместо МНК могут быть применены инструментальные переменные (п. 8.3.2) или метод максимального правдоподобия.

Поскольку уравнение (8.7) является моделью авторегрессии, то остатки U_t могут быть автокоррелированы. Для их анализа неприменим рассмотренный ранее критерий Дарбина–Уотсона (DW). Вместо него необходимо использовать h -статистику Дарбина (8.15).

8.3. Модели авторегрессии

Преобразование Койка сворачивает модель с распределенными лагами к модели авторегрессии, т. е. к модели, в правой части которой используется лаговая зависимая переменная. Это не единственный вид авторегрессионных моделей, но все же достаточно распространенный:

$$y_t = a + b_0x_t + c_1y_{t-1} + \xi_t. \quad (8.8)$$

Между тем интерпретация параметров данной модели имеет свою специфику, что и будет рассмотрено ниже.

8.3.1. Интерпретация параметров модели авторегрессии

Для модели (8.8), как и в модели с распределенными лагами, параметр b_0 характеризует краткосрочное изменение y_t под воздействием изменения x_t на 1 единицу. Параметр c_1 по существу представляет собой величину λ из преобразования Койка, т. е. $|c_1| < 1$ и показывает коэффициент снижения лаговых коэффициентов при увеличении значения лага в соответствии с концепцией их геометрического убывания.

* См.: Маленво Э. Статистические методы эконометрии: пер. с франц. М.: Статистика, 1975. С.137.

Следовательно, к моменту времени $t + 1$ результат y изменится дополнительно на $b_0 c_1$ единиц, а к моменту времени $t + 2$ дополнительное изменение y составит $b_0 c_1^2$ единиц, к моменту времени $t + 3 - b_0 c_1^3$ и т. д. Соответственно долгосрочный мультипликатор окажется равным

$b = b_0 + b_0 c_1 + b_0 c_1^2 + b_0 c_1^3 + \dots$ (в предположении бесконечного числа лагов).

Учитывая геометрическую прогрессию лаговых коэффициентов,

$$b = b_0(1 + c_1 + c_1^2 + c_1^3 + \dots) = \frac{b_0}{1 - c_1} \quad \text{— долгосрочный мультипликатор}$$

изменения y .

Трактовка данного мультипликатора была показана на примере зависимости потребления от доходов в п. 8.2.2.2.

8.3.2. Инструментальные переменные как метод оценивания параметров модели авторегрессии

В силу того что в модели авторегрессии в правой части содержатся лаговые эндогенные переменные, принято считать, что оценка параметров традиционным МНК дает неудовлетворительные результаты.

Предположим, что рассматривается модель авторегрессии вида

$$y_t = a + b_0 x_t + c_1 y_{t-1} + \xi_t. \quad (8.8)$$

Применение для оценивания параметров уравнения (8.8) традиционного МНК возможно, если выполняется предпосылка МНК относительно отсутствия автокорреляции остатков. Между тем при наличии в правой части лаговой зависимой переменной может иметь место автокорреляция остатков. Кроме того, может иметь место и зависимость объясняющей переменной y_{t-1} с остатками ξ_t , т. е. нарушается предпосылка о гомоскедастичности остатков. В силу этого классический метод наименьших квадратов в случае малых выборок даст смещенные оценки параметров.

Одним из возможных методов оценивания параметров модели (8.8) является *метод инструментальных переменных*. Суть метода состоит в том, что вместо лаговой зависимой переменной y_{t-1} , для которой нарушается предпосылка МНК, используется другая переменная, называемая инструментальной. При этом инструментальная переменная должна обладать двумя свойствами:

- она должна быть тесно коррелирована с лаговой переменной y_{t-1} ;
- она не должна коррелировать с остатками U_t (случайными ошибками).

Иными словами, от модели авторегрессии (8.8) необходимо перейти к модели вида

$$y_t = a + b x_t + c z_t + \xi_t. \quad (8.9)$$

Результаты регрессии по модели (8.9), естественно, зависят от того, насколько удачно подобрана инструментальная переменная. В качест-

ве инструментальной переменной можно, например, взять оценку \hat{y}_{t-1} , т. е. \hat{y}_{t-1} , полученную по регрессии y_{t-1} от x_{t-1} .

Поскольку в модели (8.8) предполагается наличие зависимости y_t от x_t , то можно предположить, что также имеет место зависимость y_{t-1} от x_{t-1} , т. е. найдем регрессию

$$\hat{y}_{t-1} = A + Bx_{t-1}. \quad (8.10)$$

Используя для оценки параметров уравнения (8.10) обычный МНК, что возможно ввиду отсутствия в правой части модели лаговой зависимой переменной, найдем теоретические значения \hat{y}_{t-1} , которые и будут рассматриваться как значения инструментальной переменной z в модели (8.9). Далее вновь применяем МНК уже к модели (8.9), т. е. по существу оценка параметров модели авторегрессии (8.8) будет найдена исходя из модели вида

$$y_t = a + b_0 x_t + c_1 \hat{y}_{t-1} + \xi_t. \quad (8.11)$$

Если вместо оценки \hat{y}_{t-1} подставить выражение (8.10), то получим следующую модель:

$$y_t = (a + c_1 A) + b_0 x_t + Bc_1 x_{t-1} + \xi_t. \quad (8.12)$$

Она представляет собой модель с распределенным лагом, оценка параметров которой может быть дана МНК.

Таким образом, используя в качестве инструментальной переменной оценки \hat{y}_{t-1} исходя из регрессии от x_{t-1} (8.10), модель авторегрессии (8.8) заменяется на модель с распределенным лагом (8.12).

Вместе с тем следует отметить, что применение рассмотренной инструментальной переменной может привести при практической реализации модели (8.8) к появлению коллинеарности факторов. Объясняется это тем, что в модель (8.8) одновременно вводятся в качестве объясняющих переменных линейно связанные и высоко коррелируемые между собой \hat{y}_{t-1} и x_t , ибо $\hat{y}_{t-1} = A + Bx_{t-1}$ и $r_{\hat{y}_{t-1}, x_{t-1}} = 1$, а соответственно и $\hat{y}_{t-1} x_t$ будет близок к 1. Однако если коллинеарность факторов не повлекла за собой неверные знаки у коэффициентов регрессии и не привела к большим стандартным ошибкам оценок, то применение инструментальной переменной можно считать возможным.

Пример. Применим метод инструментальных переменных к модели авторегрессии (8.8) по данным фирмы об импорте сырья (y , т) и величине производства (x , тыс. ед.) за январь – декабрь 2005–2006 гг.

	Месяцы											
	1	2	3	4	5	6	7	8	9	10	11	12
2005 г.												
y	164	162	165	168	172	177	182	186	187	191	196	201
x	78	81	89	76	105	101	93	94	107	103	116	170
2006 г.												
y	213	211	219	228	232	239	244	249	255	264	265	267
x	101	110	138	145	180	165	144	130	155	142	151	305

Рассмотрим модель (8.8):

$$y_t = a + b_0 x_t + c_1 y_{t-1} + \xi_t.$$

Для оценивания параметров этой модели введем инструментальную переменную $z = \hat{y}_{t-1} = A + Bx_{t-1}$. Используя МНК, получим

$$\hat{y}_{t-1} = 102,4025 + 0,8705x_{t-1} \quad R^2 = 0,6396.$$

$$(t) \quad (5,77) \quad (6,10) \quad F = 37,26$$

Уравнение регрессии значимо, как и его параметры. Подставляя в это уравнение значения x_{t-1} , получим расчетные значения \hat{y}_{t-1} : 170,3; 172,9; 179,9 ... 233,8. Далее вновь применяем МНК к модели (8.8), в которой вместо фактических значений y_{t-1} используются расчетные величины, т. е. \hat{y}_{t-1} . Результаты оказались следующими:

$$y_t = 12,6808 + 0,2636x_t + 0,7946\hat{y}_{t-1} + \xi_t, \quad R^2 = 0,7819.$$

$$(t) \quad (0,4) \quad (2,9) \quad (4,86) \quad F = 35,8$$

Уравнение авторегрессии в целом значимо, значимыми являются и коэффициенты регрессии.

Если к модели (8.8) сразу же применить МНК, т. е. без введения инструментальной переменной, то результаты окажутся следующими:

$$y_t = -0,4948 - 0,012x_t + 1,0315y_{t-1} + \xi_t, \quad R^2 = 0,9915.$$

$$(t) \quad -0,1 \quad -0,5 \quad 33,1$$

Хотя коэффициент детерминации для модели, оцененной по обычному МНК, выше, чем для модели с инструментальной переменной, но коэффициент регрессии при x_t не только статистически не значим, но и имеет неверный знак, ибо увеличение объема продукции, на производство которой требуется ввоз сырья, ведет к росту величины импорта, что и показывает модель авторегрессии, оцененная с помощью метода инструментальных переменных.

8.3.3. Оценка автокорреляции остатков по модели авторегрессии

Рассмотренный ранее критерий Дарбина—Уотсона неприменим для моделей авторегрессии, содержащих в составе объясняющих переменных лаговые значения зависимой переменной. Связано это с тем, что критерий Дарбина—Уотсона для модели авторегрессии может принимать значение, близкое к 2, как при отсутствии, так и при наличии автокорреляции остатков.

Предположим, что в модели авторегрессии (8.8) имеет место автокорреляция остатков, т. е. случайное отклонение ξ_t , можно рассматривать как авторегрессию вида

$$\xi_t = \rho \xi_{t-1} + U_t, \quad (8.13)$$

где ρ — коэффициент автокорреляции первого порядка;

U_t — случайная составляющая.

Тогда уравнение (8.8) можно представить как

$$y_t = a + b_0 x_t + c_1 y_{t-1} + \rho \xi_{t-1} + U_t. \quad (8.14)$$

В уравнении (8.14) y_{t-1} связан с ξ_{t-1} , как и по уравнению (8.8) y_t связан с ξ_t . Таким образом, имеется систематическая связь лаговой зависимой переменной со случайной компонентой. Применение теста Дарбина—Уотсона к модели (8.14) может показать отсутствие автокорреляции в остатках U_t , при наличии ее для остатков ξ_t . Как указано в названной работе Э. Маленво, критерий Дарбина—Уотсона теряет мощность в авторегрессионных моделях. Дж. Дарбин предложил в моделях авторегрессии для оценки существенности автокорреляции остатков использовать другой критерий, который получил название в литературе «*h*-статастика Дарбина»

$$h = \rho \sqrt{\frac{n}{1 - nV}}, \quad (8.15)$$

где ρ — коэффициент автокорреляции в остатках первого порядка, который практически используется при расчете критерия Дарбина—Уотсона, т. е.

$$\rho = \frac{\sum e_t e_{t-1}}{\sum e_t^2} = 1 - \frac{DW}{2},$$

n — число наблюдений в модели;

V — выборочная дисперсия коэффициента при лаговой зависимой переменной y_{t-1} .

При большом числе наблюдений и при отсутствии в остатках автокорреляции первого порядка *h*-статастика Дарбина подчиняется стандартизированному нормальному распределению. Поэтому фактическое значение h сравнивается с табличным по заданному уровню значимости α . Если $|h|$ больше критического значения, то нулевая гипотеза об отсутствии автокорреляции ошибок отклоняется. При практических расчетах чаще всего α берется как 0,05, и если $|h| > 1,96$, то гипотеза об отсутствии автокорреляции остатков отвергается.

Из уравнения (8.15) следует, что *h*-статастика не применима, если величина $(nV) \geq 1$. Кроме того, данный критерий предназначен для больших выборок (например, для $n > 30$). *h*-статастика зависит от квадрата стандартной ошибки параметра только при лаговой зависимой переменной $y_{t-1}(V)$ и не зависит от числа лагов, используемых в модели авторегрессии. Так, для модели $y_t = a + b_0 x_1 + c_1 y_{t-1} + c_2 y_{t-2} + \xi_t$, оценка значимости автокорреляции остатков также проводится с помощью *h*-статастики Дарбина.

Аналогично данный критерий используется и для модели авторегрессии с несколькими экзогенными переменными:

$$y_t = a + b_0 x_{1t} + b_1 x_{2t} + b_2 x_{3t} + c_1 y_{t-1} + \xi_t.$$

В рассматриваемом примере автокорреляция остатков не устранена, чем свидетельствует *h*-статастика Дарбина: коэффициент автокорреляции в остатках (ρ) составил 0,440; стандартная ошибка коэффици-

ента регрессии при переменной y_{t-1} оказалась равной 0,1635 (0,7946/4,86), соответственно $V = 0,1635^2 = 0,026732$ и при $n = 23$

$$h = 0,440 \sqrt{\frac{23}{1 - 23 \cdot 0,026732}} = 3,4, \text{ что больше необходимого } 1,96.$$

Автокорреляция в остатках по авторегрессионным моделям может быть устранена с помощью авторегрессионных преобразований, используя модели *ARMA* и *ARIMA*.

8.4. Авторегрессионные процессы и их моделирование

8.4.1. Авторегрессионные процессы

Рассмотренные ранее модели авторегрессии содержали в правой части наряду с лаговыми зависимыми переменными (y_{t-1}, y_{t-2} и т. п.) независимые переменные (x). Авторегрессионная модель, в которой отсутствуют независимые переменные и y_t , рассматривается как линейная функция только предыдущих своих значений, представляет собой *авторегрессионный процесс*:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \xi_t. \quad (8.16)$$

В зависимости от того, сколько предыдущих уровней временного ряда включено в уравнение (8.16), авторегрессионный процесс может быть разного порядка. Если текущее значение уровня динамического ряда (y_t) рассматривается как линейная функция от одного предыдущего значения, то имеем дело с авторегрессионным процессом первого порядка, что обычно в англоязычной литературе обозначается как AR (1):

$$y_t = a_0 + a_1 y_{t-1} + \xi_t. \quad (8.17)$$

Увеличивая число лаговых переменных в модели (8.17), получим авторегрессионный процесс более высокого порядка. Например, процесс AR (3) сводится к уравнению:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \xi_t, \quad (8.18)$$

и отражает авторегрессионный процесс третьего порядка.

Процессы AR могут быть стационарными и нестационарными. Чтобы процесс был стационарным, коэффициенты a_1, a_2, \dots, a_p в модели (8.16) должны образовывать сходящийся ряд и все корни характеристического уравнения $1 - a_1 z - a_2 z^2 - \dots - a_p z^p = 0$ (вещественные и комплексные) должны лежать вне единичного круга, т. е. $|z| > 1$.

Рассмотренное условие стационарности для процесса AR (1) означает, что в уравнении (8.17) параметр a_1 должен соответствовать величине $|a_1| < 1$, так как характеристическое уравнение $1 - a_1 z = 0$ имеет корень $z = \frac{1}{a_1}$ и $|z| > 1$ составит при $|a_1| < 1$. Так, для ряда $y_t = 3 + 0,8 y_{t-1} + \xi_t$, при $y_0 = 2$ (начальный уровень динамического ряда) характеристиче-

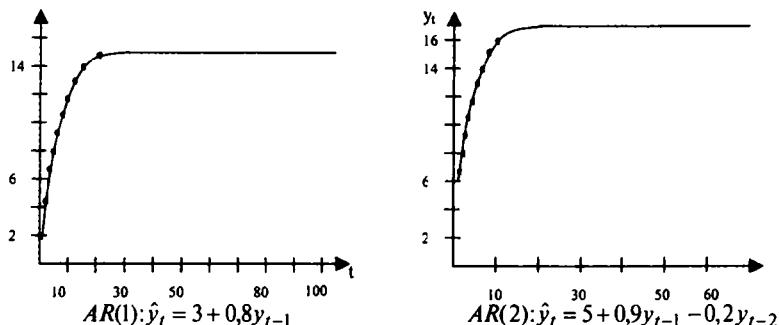


Рис. 8.4. Асимптотические стационарные временные ряды

ское уравнение имеет вид $1 - 0,8z = 0$. Соответственно $z = 1,25$ и рассматриваемый процесс является стационарным. Его асимптота (μ) окажется равной $\mu = \frac{a_0}{1 - a_1}$, т. е. имеем $\mu = \frac{3}{1 - 0,8} = 15$, т. е. траектория процесса флюктуирует и не превышает 15. Так, при $t = 100$ \hat{y}_t принимает значения: 2; 4,6; 6,7; 8,3; 9,7; 10,7; 11,6; 12,3; 12,8; 13,3; 13,6; 13,8; 14,1,

далее возрастающие до 15, а начиная с $t = 27$ не превышающие 15.

Предположим, что рассматривается процесс AR (2), а именно:

$$y_t = 5 + 0,9y_{t-1} - 0,2y_{t-2} + \xi_t.$$

Для него характеристическое уравнение составит

$$1 - 0,9z + 0,2z^2 = 0.$$

Корни этого уравнения составят $z_1 = 2,5$ и $z_2 = 2$, что больше 1, и, следовательно, процесс является стационарным. Асимптота данного ряда окажется равной $\mu = \frac{5}{1 - 0,9 + 0,2} = 16,(6)$, т. е. начиная с $t = 14$ \hat{y}_t ,

варьирует вокруг величины 16,(6): уровни ряда принимали значения 7; 6; 9; 11,9; 13,9; 15,1 и т. д. В рассмотренных примерах AR (1) и AR (2) динамические ряды обнаруживают вначале некоторую тенденцию, которая постепенно затухает и ряд становится стационарным (см. графики на рис. 8.4).

Авторегрессионный процесс с большим числом лагов предполагает очень длинные динамические ряды, которые далеко не всегда имеются в эконометрических исследованиях. При наличии коротких временных рядов стационарные AR-процессы могут иметь место после удаления из уровней ряда тенденции и сезонных колебаний. Это означает, что исследователь должен вычленить эти компоненты динамического ряда и подвергать дальнейшей обработке остаточные величины. В этом случае авторегрессионный процесс первого порядка AR (1) примет выражение

$$\xi_t = a\xi_{t-1} + V_t, \quad (8.19)$$

где ξ_t — остатки после устранения из уровней ряда y_t тенденции и периодической составляющей;
 V_t — белый шум.

8.4.2. Модели скользящей средней

Среди моделей для стационарных временных рядов широкое распространение имеют *модели скользящей средней*.

Для стационарного ряда моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т. е. разностей между прошлыми фактическими и теоретическими уровнями:

$$y_t = \mu + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2} - \dots - \theta_q \xi_{t-q}, \quad (8.20)$$

где μ — константа;
 $\xi_t, \xi_{t-1}, \dots, \xi_{t-q}$ — белый шум в текущий и предыдущий период времени;
 $\xi_t = y_t - \hat{y}_t$.

Термин «скользящая средняя», используемый здесь, не синоним скользящей средней как методу слаживания уровней динамического ряда.

В модели (8.20) уровень динамического ряда рассматривается как сумма константы (μ) и скользящей средней между текущими и предыдущими значениями белого шума (случайных отклонений).

Обозначим скользящую среднюю модели (8.20) через x_t :

$$x_t = \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2} - \dots - \theta_q \xi_{t-q}. \quad (8.21)$$

Уравнение (8.21) принято называть *процессом скользящего среднего порядка q* и обозначать как *MA(q)* (от англ. *moving average*). Порядок скользящей средней определяется числом учитываемых в модели предыдущих значений случайных отклонений. Так, MA (2) можно записать как $x_t = \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2}$, а модель уровня динамического ряда с использованием MA (2) будет иметь вид

$$y_t = \mu + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2}.$$

Соответственно модель уровня ряда с использованием MA (1) примет выражение

$$y_t = \mu + \xi_t - \theta_1 \xi_{t-1}.$$

При $q = 0$, $\mu = 0$ получаем процесс белого шума.

Временные ряды с использованием процесса скользящего среднего могут иметь место, когда уровни динамического ряда характеризуются случайной колеблемостью.

8.4.3. Модели ARMA

Соединение в одной модели авторегрессионного процесса AR и модели скользящего среднего MA приводит к модели *авторегрессии-*

онного процесса со скользящими средними в остатках (от англ. ARMA – Auto Regressive – Moving Average):

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2} - \dots - \theta_q \xi_{t-q}. \quad (8.22)$$

В модели (8.22) в качестве объясняющих переменных рассматриваются лаговые значения зависимой переменной с p интервалами сдвига и скользящие средние порядка q для остатков авторегрессии. Иными словами, модель включает в себя AR (p) и MA (q). Ее принято обозначать ARMA (p, q). Например, ARMA (3,2) имеет вид

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \xi_t - \theta_1 \xi_{t-1} - \theta_2 \xi_{t-2}. \quad (8.23)$$

При практической реализации моделей ARMA наиболее сложным является выбор числа лагов p и q .

Инструментом идентификации модели ARMA является изучение частной автокорреляционной функции по моделям с разным числом лагов. Частная автокорреляционная функция (от англ. PACF – partial autocorrelation function) представляет собой серию частных коэффициентов автокорреляции (PAC), которые измеряют связь между текущим уровнем динамического ряда (y_t) и предыдущими значениями $y_{t-1}, y_{t-2}, \dots, y_{t-k}$ в условиях, когда влияние других промежуточных временных лагов устранено. Так, частный коэффициент автокорреляции при лаге k будет представлять собой корреляцию y_t и y_{t-k} , очищенную от влияния $y_{t-1}, y_{t-2}, \dots, y_{t-k-1}$.

Обозначим частный коэффициент автокорреляции с лагом k через $\rho(k)$. При $k=0 \rho(0)=1$ (уровни ряда коррелируют сами с собой); при $k=1 \rho(1)=v_{a_1}$, где v_{a_1} – коэффициент автокорреляции первого порядка. Это равенство связано с тем, что при расчете $\rho(1)$ отсутствуют промежуточные лаги. Вычисление ρ более высокого порядка можно производить по формулам

$$\rho(2) = \frac{\begin{vmatrix} 1 & v_{a_1} \\ v_{a_1} & v_{a_2} \end{vmatrix}}{\begin{vmatrix} 1 & v_{a_1} \\ v_{a_1} & 1 \end{vmatrix}} = \frac{(v_{a_2} - v_{a_1}^2)}{(1 - v_{a_1}^2)},$$

$$\rho(3) = \frac{\begin{vmatrix} 1 & v_{a_1} & v_{a_1} \\ v_{a_1} & 1 & v_{a_2} \\ v_{a_2} & v_{a_1} & v_{a_3} \end{vmatrix}}{\begin{vmatrix} 1 & v_{a_1} & v_{a_2} \\ v_{a_1} & 1 & v_{a_1} \\ v_{a_2} & v_{a_1} & 1 \end{vmatrix}},$$

$$\rho(k) = \frac{\begin{vmatrix} 1 & v_{a_1} & v_{a_2} & v_{a_1} \\ v_{a_1} & 1 & v_{a_1} & v_{a_2} \\ v_{a_2} & v_{a_1} & 1 & v_{a_3} \\ \dots & & & \\ v_{a_{k-1}} & v_{a_{k-2}} & v_{a_{k-3}} & \dots & v_{a_k} \end{vmatrix}}{\begin{vmatrix} 1 & v_{a_1} & v_{a_2} & v_{a_{k-1}} \\ v_{a_1} & 1 & v_{a_1} & v_{a_{k-2}} \\ v_{a_2} & v_{a_1} & 1 & v_{a_{k-3}} \\ \dots & & & \\ v_{a_{k-1}} & v_{a_{k-2}} & v_{a_{k-3}} & & 1 \end{vmatrix}}$$

Для авторегрессионного процесса порядка p частная автокорреляционная функция отлична от нуля при $k \leq p$ и равна нулю при $k > p$. Это позволяет определять порядок p процесса AR. Так, для модели AR (1): $y_t = a_1 y_{t-1} + \xi_t$, $\rho(2)$ близко к нулю.

Пример. За 50 месяцев темпы прироста объема продукции k характеризовались авторегрессией вида $y_t = 0,616 y_{t-1} + \xi_t$, $R^2 = 0,662$ и $F = 93,9$. Автокорреляционная функция составила убывающие значения автокорреляции:

Лаг	1	2	3	4	5	6	7	...	13	14
r_a	0,813	0,747	0,617	0,474	0,408	0,457	0,353	...	-0,107	-0,051

Частная автокорреляционная функция начиная с лага 2 достаточно близка к нулю:

Лаг	1	2	3	4	5
ρ	0,813	0,252	0,134	0,202	0,122

Для модели типа MA (q) порядок q определяется по поведению автокорреляционной функции: при $k \geq p r_a$ стремится к нулю. Для модели ARMA (p, q) автокорреляционная функция характеризуется убыванием, начинающимся с лага q , а частная автокорреляционная функция убывает начиная с лага p . Так, для модели ARMA (1, 1) при $a_1 > 0$ ACF наблюдает экспоненциальное затухание с лага 1, а PACF – осциллирующее убывание с лага 1. При $a_1 < 0$ ACF для модели ARMA (1, 1) наблюдает осциллирующее убывание с лага 1, а PACF – экспоненциальное затухание с лага 1.

Выбор типа модели ARMA не ограничивается обычно исследованием автокорреляционных функций. С этой целью может использо-

ваться, например, информационный критерий Акаике*, рассмотрение которого не входит в задачу данного учебника.

8.4.4. Модели ARIMA

Для получения стационарного ряда могут рассчитываться разности уровней временного ряда (Δ) разного порядка (d). Модель, в которой соединены нахождение последовательных разностей временного ряда порядка d и ARMA – модель порядка (p, q), получила название *авторегрессионной интегрированной модели скользящего среднего – ARIMA (Autoregressive Integrated Moving Average)*.

Модель ARIMA обладает тремя параметрами: p – порядок авторегрессии (AR); d – порядок последовательных разностей уровней временных рядов, обеспечивающий стационарность ряда; q – порядок скользящей средней (MA).

В общем виде модель ARIMA (p, d, q) выражается формулой

$$\Delta^k y_t = a_1 \Delta^k y_{t-1} + \dots + a_p \Delta^k y_{t-p} + \xi_t - \theta_1 \xi_{t-1} - \dots - \theta_q \xi_{t-q}, \quad (8.24)$$

где $\Delta^k y_t$ – k -я последовательная разность уровней y_t , т. е. $\Delta^k y_t = y_t - y_{t-1} - \dots - y_{t-k}$; $\xi_t, \xi_{t-1}, \dots, \xi_{t-q}$ – нормально распределенные случайные величины с нулевым математическим ожиданием и постоянной дисперсией.

Из модели (8.24) для $\Delta^k y_t$ можно получить модель для исходного динамического ряда с помощью выражения

$$y_t = y_{t-1} + \Delta y_t. \quad (8.25)$$

Так, если модель ARIMA (1, 1, 1) имеет вид

$$\Delta^1 y_t = 0,2 \Delta^1 y_{t-1} + \xi_t - 0,1 \xi_{t-1},$$

то динамический ряд описывается моделью

$$y_t = y_{t-1} + 0,2 y_{t-1} - 0,2 y_{t-2} + \xi_t - 0,1 \xi_{t-1},$$

так как $\Delta^1 y_t = y_t - y_{t-1}$, $\Delta^1 y_t = y_{t-1} - y_{t-2}$

Модель ARIMA практически пригодна для большинства временных рядов. При $p = 1, 2, \dots, k$, $d = 0$ и $q = 0$ модель ARIMA превращается в AR-процесс:

$$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \dots + b_k y_{t-k} + \xi_t.$$

Если $p = 0$, $d = 0$ и $q = 1, 2, \dots, k$, то имеем модель MA:

$$y_t = \mu + \xi_t - \theta_1 \xi_{t-1} - \dots - \theta_k \xi_{t-k}.$$

Наиболее распространены модели ARIMA с параметрами p, d и q , не превышающими 2. Современные компьютерные программы пред-

*См, например: Носко В. П. Эконометрика. Элементарные методы и введение в регрессионный анализ временных рядов. М.: ИЭПП, 2004. С. 251.

лагают разные варианты оценивания параметров модели ARIMA, среди которых преобладает оценка методом максимального правдоподобия. Такой подход можно видеть при реализации модели ARIMA в системе SPSS*.

Контрольные вопросы

1. Что такое модель с распределенными лагами?
2. Как интерпретируются параметры модели с распределенными лагами?
3. В каких случаях оценка параметров модели с распределенными лагами может быть дана методом наименьших квадратов?
4. Объясните, что такое структура лага и как она используется при построении модели с распределенными лагами?
5. Опишите методику построения модели с использованием лагов Алмон.
6. В чем состоит суть преобразования Койка?
7. Что такое модели авторегрессии?
8. Как интерпретируются параметры модели авторегрессии?
9. Для чего используются инструментальные переменные?
10. Что такое авторегressive процессы?
11. Как строятся модели ARMA?
12. В чем состоят особенности модели ARIMA?

*Прогнозирование с помощью моделей ARIMA см.: *Дубрава Т. А. Статистические методы прогнозирования*. М.: ЮНИТИ, 2003. С. 178–184.

ГЛАВА 9. СИСТЕМА ЭКОНОМЕТРИЧЕСКИХ УРАВНЕНИЙ

9.1. Общая характеристика системы эконометрических уравнений

Ввиду многогранности экономические процессы наиболее полно могут быть описаны с помощью системы эконометрических уравнений, которая в матричном виде может быть представлена как

$$BY + \Gamma X = E, \quad (9.1)$$

где B — матрица коэффициентов при зависимых переменных;

Y — вектор зависимых переменных;

Γ — матрица коэффициентов при объясняющих переменных;

X — вектор объясняющих переменных;

E — вектор ошибок.

Зависимые переменные, которые определяются моделью, принято называть *эндогенными* переменными. Часто они обозначаются через y , т. е. вектор эндогенных переменных системы может быть записан как

$$\vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}.$$

Экономические переменные, которые влияют на значения эндогенных переменных и не зависят от них, называются *экзогенными* переменными, обычно обозначаемыми через x . Вектор экзогенных переменных составит

$$\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix}.$$

Экономические переменные могут в зависимости от теоретической концепции модели выступать в одной ситуации как эндогенные, а в другой — как экзогенные. В качестве объясняющих переменных могут рассматриваться эндогенные переменные с временным запаздыванием (*лагом*). Так, спрос на товар текущего года (y_t) может зависеть не только от дохода, цен и других экономических переменных, но и от уровня достигнутого спроса в предыдущем году (y_{t-1}). *Лаговые* переменные вместе с экзогенными переменными принято называть *предопределенные* переменными. К предопределенным переменным относятся и лаговые экзогенные переменные, а также включаемый в модель фактор времени t . В дальнейшем в системе (9.1) под вектором \vec{X} будут пониматься предопределенные переменные.

В зависимости от того, что представляет собой матрица коэффициентов при эндогенных переменных, т. е. матрица B , различают следующие виды систем эконометрических уравнений:

- система независимых уравнений;
- система рекурсивных уравнений;
- система взаимозависимых уравнений.

Система независимых уравнений представляет собой систему, в которой эндогенные переменные (y_1, y_2, \dots, y_n) рассматриваются как функции объясняющих переменных (x_1, x_2, \dots, x_m). В результате матрица коэффициентов B является диагональной. Так, при трех эндогенных и трех экзогенных переменных система имеет вид

$$\begin{cases} y_1 = a_{01} + a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \xi_1 \\ y_2 = a_{02} + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \xi_2 \\ y_3 = a_{03} + a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \xi_3 \end{cases} \quad (9.2)$$

Соответственно матрица коэффициентов при эндогенных переменных является диагональной:

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

При этом набор объясняющих переменных в каждом уравнении системы может меняться в зависимости от реальной связи экономических переменных, что не оказывает влияния на вид системы и способ оценивания ее параметров. Каждое уравнение такой системы является уравнением регрессии, оценка параметров которого может быть дана традиционным МНК. Это значит, что каждое уравнение системы независимых уравнений рассматривается самостоятельно и в системе из n уравнений n раз применяется МНК.

Система рекурсивных уравнений представляет собой систему, в которой эндогенная переменная y в одном уравнении, являющаяся функцией ряда объясняющих переменных x , используется как экзогенная переменная в другом уравнении системы. Например,

$$\begin{cases} y_1 = a_{01} + a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \xi_1 \\ y_2 = a_{02} + b_{21}y_1 + a_{22}x_2 + \xi_2 \\ y_3 = a_{03} + b_{31}y_1 + b_{32}y_2 + a_{34}x_4 + a_{35}x_5 + \xi_3 \end{cases} \quad (9.3)$$

где y_1 — производительность труда;

y_2 — фондоотдача;

y_3 — себестоимость единицы продукции;

x_1 — фондоооруженность труда;

x_2 — энерговооруженность труда;

x_3 — квалификация рабочих;

x_4 — зарплаталяемость продукции;

x_5 — материалоемкость продукции.

Матрица коэффициентов B при зависимых переменных y в рекурсивной системе является треугольной. В нашем примере она составит

$$B = \begin{bmatrix} 1 & 0 & 0 \\ -b_{21} & 1 & 0 \\ -b_{31} & -b_{32} & 1 \end{bmatrix}$$

В рекурсивной модели в качестве объясняющих переменных в правой части модели могут участвовать только те эндогенные переменные, которые ранее были определены в предыдущих уравнениях системы. Иными словами, в модели отсутствуют обратные связи между эндогенными переменными, т. е.

$$\begin{aligned} y_2 &= f(y_1), \text{ но } y_1 \neq f(y_2); \\ y_3 &= f(y_1, y_2), \text{ но } y_2 \neq f(y_3). \end{aligned}$$

Поэтому каждое уравнение рекурсивной модели можно рассматривать самостоятельно, а оценку параметров производить с помощью обычного МНК.

Система взаимозависимых уравнений в отличие от предыдущей системы содержит обратные связи между эндогенными переменными, т. е. одни и те же переменные одновременно рассматриваются как зависимые в одних уравнениях и как объясняющие в других. Поэтому система взаимозависимых уравнений получила название системы одновременных уравнений. Матрица коэффициентов B при эндогенных переменных в данной модели не является ни диагональной, ни треугольной.

Рассмотрим модель следующего вида:

$$\begin{cases} y_1 = a_{01} + b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \xi_1 \\ y_2 = a_{02} + b_{21}y_1 + b_{23}y_3 + a_{23}x_3 + \xi_2, \\ y_3 = a_{03} + b_{32}y_2 + a_{33}x_3 + \xi_3 \end{cases} \quad (9.4)$$

где y_1 — объем продукции;

y_2 — производительность труда;

y_3 — фондоотдача;

x_1 — основные производственные фонды;

x_2 — затраты на рабочую силу;

x_3 — энерговооруженность труда.

Матрица коэффициентов при эндогенных переменных имеет вид

$$B = \begin{bmatrix} 1 & -b_{12} & 0 \\ -b_{21} & 1 & -b_{23} \\ 0 & -b_{32} & 1 \end{bmatrix}.$$

Она не является ни диагональной, ни треугольной, что и характерно для системы одновременных уравнений. Ввиду обратной связи между зависимыми переменными в системе (9.4) оценка ее параметров не может быть дана МНК и требуются специальные методы оценивания.

Необходимость специальных методов оценивания параметров системы одновременных уравнений можно видеть, рассматривая известную кейнсианскую модель:

$$\begin{cases} C_t = a + bY_t + \xi_t, \\ Y_t = C_t + I_t, \end{cases}, \quad (9.5)$$

где C_t — личное потребление в момент времени t ;

Y_t — национальный доход в момент времени t ;

I_t — инвестиции в момент времени t .

В модели (9.5) C_t и Y_t — эндогенные переменные; I_t — экзогенная переменная. На первый взгляд, первое уравнение системы (9.5) — уравнение регрессии, параметры которого (a и b) могут быть оценены МНК. Однако в силу тождества $Y_t = C_t + I_t$, модель (9.5) представляет собой систему одновременных уравнений. Матрица коэффициентов при эндогенных переменных имеет вид

$$B = \begin{bmatrix} 1 & -b \\ -1 & 1 \end{bmatrix}.$$

Если мы подставим первое уравнение системы (9.5) во второе, то получим

$$Y_t = a + bY_t + \xi_t + I_t = a + bY_t + I_t + \xi_t.$$

Как видим, Y_t линейно связан с ошибкой. Иными словами, в уравнении регрессии для C_t , объясняющая переменная Y_t и ошибки ξ связаны. Между тем одной из предпосылок МНК является независимость ошибок от значений объясняющей переменной. Следовательно, для оценки параметров первого уравнения системы (9.5) МНК неприменим и требуются иные подходы статистического решения.

9.2. Структурная и приведенная формы модели

Система одновременных уравнений получила название также *структурной формы модели*. Структурная модель из n эндогенных переменных (y_1, y_2, \dots, y_n) содержит n уравнений в системе, каждое из которых рассматривается как функция предопределенных и ряда эндогенных переменных. Так, структурная модель (9.4) содержит три эндогенные переменные, которым соответствуют три уравнения, и три экзогенные (предопределенные) переменные (x_1, x_2, \dots, x_3), коэффициенты при эндогенных ($b_{12}, b_{21}, \dots, b_{32}$) и предопределенных переменных ($a_{11}, a_{12}, a_{23}, a_{33}$) называются структурными коэффициентами модели. Параметры a_{01}, a_{02} , и a_{03} — свободные члены уравнений регрессии. Они могут отсутствовать в модели, если переменные выражены в отклонениях от среднего уровня, т. е. под x_i подразумевается $(x_j - \bar{x}_j)$, а под y_j соответственно $(y_j - \bar{y}_j)$.

Для оценки параметров структурной модели используется *приведенная форма модели*, которая представляет собой систему линейных функций эндогенных переменных от экзогенных, имеющихся в системе. Для структурной модели (9.4) приведенная модель составит

$$\begin{cases} y_1 = A_1 + \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3 + U_1 \\ y_2 = A_2 + \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 + U_2 \\ ry_3 = A_3 + \delta_{31}x_1 + \delta_{32}x_2 + \delta_{33}x_3 + U_3 \end{cases} \quad (9.6)$$

где U_1, U_2, U_3 — ошибки приведенной модели;

$\delta_{11}, \delta_{12}, \dots, \delta_{33}$ — коэффициенты приведенной модели.

По внешнему виду приведенная модель похожа на систему независимых уравнений (9.2). Отличие между моделями (9.2) и (9.6) состоит в том, что в модели (9.2) в каждом уравнении системы набор экзогенных переменных может быть свой в зависимости от содержательной концепции и существенности связи y_j и x_j . В модели же (9.6) в каждом уравнении используется полный набор экзогенных переменных, имеющихся в системе (9.4).

Коэффициенты структурной и приведенной модели связаны нелинейными соотношениями. Покажем это для простейшей структурной модели:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{01} + \xi_1 \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{02} + \xi_2 \end{cases} \quad (9.7)$$

для которой приведенная модель составит

$$\begin{cases} y_1 = A_1 + \delta_{11}x_1 + \delta_{12}x_2 + U_1 \\ y_2 = A_2 + \delta_{21}x_1 + \delta_{22}x_2 + U_2 \end{cases} \quad (9.8)$$

Выразим из первого структурного уравнения системы (9.7) y_2 . Тогда система (9.7) примет вид

$$\begin{cases} y_2 = \frac{y_1 - a_{11}x_1 - a_{01} - \xi_1}{b_{12}} \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{02} + \xi_2 \end{cases} \quad (9.9)$$

Приравняв правые части системы (9.9), будем иметь после соответствующих преобразований:

$$y_1 = \frac{a_{11}}{1 - b_{21}b_{12}}x_1 + \frac{a_{22}b_{12}}{1 - b_{21}b_{12}}x_2 + \frac{a_{01} + a_{02}b_{12}}{1 - b_{21}b_{12}} + \frac{\xi_1 + \xi_2b_{12}}{1 - b_{21}b_{12}}. \quad (9.10)$$

Нетрудно видеть, что уравнение (9.10) представляет собой первое уравнение приведенной модели (9.8), где

$$A_1 = \frac{a_{01} + a_{02}b_{12}}{1 - b_{21}b_{12}}; \quad \delta_{11} = \frac{a_{11}}{1 - b_{21}b_{12}}; \quad \delta_{12} = \frac{a_{22} \cdot b_{12}}{1 - b_{21}b_{12}}; \quad U_1 = \frac{\xi_1 + \xi_2b_{12}}{1 - b_{21}b_{12}}.$$

Аналогично можно показать, что коэффициенты второго уравнения приведенной модели (9.8) также нелинейно связаны с коэффициентами структурной модели (9.7), а именно:

$$A_2 = \frac{a_{01}b_{21} + a_{02}}{1 - b_{21}b_{12}}; \quad \delta_{21} = \frac{a_{11} \cdot b_{21}}{1 - b_{21}b_{12}}; \quad \delta_{22} = \frac{a_{22}}{1 - b_{21}b_{12}}; \quad U_2 = \frac{\xi_1b_{21} + \xi_2}{1 - b_{21}b_{12}}.$$

Взаимосвязь коэффициентов приведенной и структурной модели имеет вполне реальное практическое применение. Коэффициенты приведенной модели как коэффициенты системы независимых уравнений могут быть оценены обычным МНК, и на их основе может быть произведена оценка структурных коэффициентов модели.

9.3. Идентификация структурной модели

Оценке коэффициентов структурной модели предшествует исследование модели на идентифицируемость. Если модель идентифицируема, то она имеет статистическое решение, т. е. на основе приведенной модели могут быть найдены коэффициенты структурной модели. *Идентификация модели* — это соответствие между приведенной и структурной формами модели, позволяющее однозначно оценить структурные коэффициенты по приведенным коэффициентам модели.

Рассмотрим проблему идентификации для модели с тремя эндогенными и тремя экзогенными переменными:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \xi_1 \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \xi_2 \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \xi_3 \end{cases} \quad (9.11)$$

В модели (9.11) для упрощения все переменные отцентрированы, т. е. выражены в отклонениях от среднего уровня и параметры $a_{01}, a_{02}, a_{03} = 0$, что не влияет на решение вопроса об идентификации модели.

Модель (9.11) содержит 15 структурных коэффициентов: 6 при эндогенных и 9 при экзогенных переменных. В структурной модели из n эндогенных и m экзогенных переменных в каждом уравнении системы общее число структурных коэффициентов составит $n(n-1+m)$, что и имеет место в модели (9.11): $n = 3$ и $m = 3$ и $3(3 - 1 + 3) = 15$. Приведенная форма модели из n эндогенных и m экзогенных переменных содержит nm параметров, что составит в примере $3 \cdot 3 = 9$ коэффициентов. Это можно видеть, написав приведенную форму модели:

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3 + U_1 \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 + U_2 \\ y_3 = \delta_{31}x_1 + \delta_{32}x_2 + \delta_{33}x_3 + U_3 \end{cases} \quad (9.12)$$

Как видим, на основе 9 коэффициентов приведенной модели необходимо найти 15 коэффициентов структурной модели (9.11), что, естественно, не может привести к единственности решения.

Чтобы получить единственно возможное решение, необходимо предположить, что некоторые из структурных коэффициентов модели ввиду слабой связи признаков равны нулю. Так, в модели (9.11), предполагая $b_{13} = b_{31} = a_{12} = a_{21} = a_{23} = a_{32} = 0$, уменьшим число структурных коэффициентов до 9:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{13}x_3 + \xi_1 \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{22}x_2 + \xi_2 \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + \xi_3 \end{cases} \quad (9.13)$$

Для модели (9.13) приведенная форма примет вид модели (9.12), тот же, что и для модели (9.11), ибо число экзогенных переменных не изменилось.

Найдя приведенные коэффициенты модели (9.12), можно оценить и структурные коэффициенты модели (9.13).

Уменьшение числа структурных коэффициентов модели возможно и другим путем, например приравниванием некоторых коэффициентов друг к другу. Так, если в первом уравнении модели (9.13) предположим, что $b_{12} = a_{13}$, то уравнение примет вид

$$y_1 = b_{12}(y_2 + x_3) + a_{11}x_1 + \xi_1. \quad (9.14)$$

В результате число структурных коэффициентов модели (9.13) уменьшится до 8. На структурные коэффициенты модели могут накладываться и другие ограничения, например вида $b_{ij} + a_{ij} = 0$.

По идентифицируемости структурные модели подразделяются на три класса:

- идентифицируемые;
- неидентифицируемые;
- сверхидентифицируемые.

Модель *идентифицируема*, если число коэффициентов структурной модели равно числу коэффициентов приведенной модели и структурные коэффициенты однозначно определяются по приведенным коэффициентам. Идентифицируемой, например, является модель (9.13).

Модель *неидентифицируема*, если число структурных коэффициентов больше числа приведенных коэффициентов. Модели, в которых в каждом уравнении системы участвуют все эндогенные и экзогенные переменные, имеющиеся в системе, всегда неидентифицируемы (например, модель (9.11)), и структурные коэффициенты модели не могут быть оценены.

Модель *сверхидентифицируема*, если число приведенных коэффициентов превышает число структурных коэффициентов. В результате на основе коэффициентов приведенной модели можно получить несколько значений одного структурного коэффициента. Так, модель (9.13) становится сверхидентифицируемой, если первое уравнение системы примет вид уравнения (9.14). Тогда 8 структурных коэффициентов будут определяться по 9 коэффициентам приведенной модели (9.11). Сверхидентифицируемая модель имеет статистическое решение, но требует специальных методов оценки параметров.

На идентификацию проверяется каждое уравнение системы. Чтобы уравнение было идентифицировано, необходимо, чтобы число предопределенных переменных, отсутствующих в данном уравнении, но присутствующих в системе, было равно числу эндогенных переменных в данном уравнении без одного.

Если обозначить число эндогенных переменных в j -м уравнении системы через H , а число экзогенных (предопределенных) переменных, которые содержатся в системе, но не входят в данное уравнение, — через D , то условие идентифицированности модели может быть записано в виде следующего счетного правила:

- $D + 1 = H$ — уравнение идентифицируемо;
- $D + 1 < H$ — уравнение неидентифицируемо;
- $D + 1 > H$ — уравнение сверхидентифицируемо.

Предположим, рассматривается следующая система стохастических уравнений:

$$\begin{cases} y_1 = a_1 + b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \xi_1 \\ y_2 = a_2 + b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \xi_2 \\ y_3 = a_3 + b_{31}y_1 + b_{32}y_2 + a_{33}x_3 + a_{34}x_4 + \xi_3 \end{cases} \quad (9.15)$$

В системе (9.15) три эндогенные переменные (y_1, y_2 и y_3) и четыре экзогенные переменные (x_1, x_2, x_3, x_4). Используем счетное правило для проверки на идентификацию каждого уравнения системы (9.15). В первом уравнении $H = 3$, так как в нем присутствуют 3 эндогенные переменные, а $D = 2$, так как отсутствуют переменные x_3 и x_4 . Тогда имеем равенство $D + 1 = H(2 + 1 = 3)$, что означает, что первое уравнение идентифицировано. К аналогичным выводам приедем, рассматривая второе и третье уравнение системы (9.15):

- во втором уравнении $H = 2(y_1$ и $y_2)$ и $D = 1(x_4)$, т. е. $D + 1 = H$;
- в третьем уравнении $H = 3(y_1, y_2, y_3)$ и $D = 2(x_1$ и $x_2)$, т. е. $D + 1 = H$.

Таким образом, система (9.15) по счетному правилу является точно идентифицируемой, ибо каждое уравнение в ней идентифицируемо.

Предположим, что в модели (9.15) $a_{21} = a_{33} = 0$. Тогда система примет вид

$$\begin{cases} y_1 = a_1 + b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \xi_1 \\ y_2 = a_2 + b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \xi_2 \\ y_3 = a_3 + b_{31}y_1 + b_{32}y_2 + a_{34}x_4 + \xi_3 \end{cases} \quad (9.16)$$

Так как система (9.16) также содержит три эндогенные и четыре экзогенные переменные, то первое уравнение системы точно идентифицировано (оно не изменилось). Второе уравнение системы (9.16) теперь имеет $H = 2$ и $D = 2$ (отсутствуют x_1 и x_4), и так как $D + 1 > H(2 + 1 > 2)$, то данное уравнение оказывается сверхидентифицированным. Третье уравнение системы (9.16) также является сверхидентифицированным: $H = 3(y_1, y_2$ и $y_3)$ и $D = 3$ (отсутствуют x_1, x_2 и x_3), т. е. $D + 1 > H$.

Если предположим, что последнее уравнение системы (9.16) примет вид

$$y_3 = a_3 + b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + a_{34}x_4 + \xi_3,$$

то оно становится неидентифицированным, ибо в нем $H = 3$ и $D = 1$ (отсутствует только x_3), т. е. $D + 1 < H$. Соответственно его параметры не имеют статистического решения. Рассмотренное счетное правило отражает необходимое, но недостаточное условие идентификации.

Чтобы уравнение, входящее в систему одновременных уравнений, было идентифицировано, необходимо и достаточно, чтобы ранг матрицы коэффициентов по отсутствующим в нем переменным был на единицу меньше числа эндогенных переменных в системе и определяль этой матрицы не был равен нулю.

Обратимся к следующей структурной модели:

$$\begin{cases} y_1 = A_1 + b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \xi_1 \\ y_2 = A_2 + b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 + \xi_2 \\ y_3 = A_3 + b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \xi_3 \end{cases} \quad (9.17)$$

Проверим каждое уравнение системы на необходимое и достаточное условия идентификации:

Первое уравнение $H = 3$ (y_1, y_2, y_3) и $D = 2$ (x_3 и x_4 отсутствуют), $D + 1 = H$ и по счетному правилу уравнение точно идентифицируемо. Для проверки на достаточное условие идентификации заполним следующую таблицу:

Таблица
Коэффициенты при отсутствующих в первом уравнении переменных

Уравнения	Переменные	
	x_3	x_4
I I	a_{23}	a_{24}
I I I	0	0

Определитель матрицы ($\det A$) коэффициентов равен 0. Следовательно, достаточное условие идентификации не выполняется и первое уравнение нельзя считать идентифицированным.

Второе уравнение — $H = 2$ (y_1 и y_2), $D = 1$ (отсутствует x_1). Счетное правило дает утвердительный ответ: уравнение идентифицированное ($D + 1 = H$).

Достаточное условие идентификации также дает подтверждение наличия идентификации для данного уравнения.

Таблица
Коэффициенты при отсутствующих во втором уравнении переменных

Уравнения	Переменные	
	y_3	x_1
I	b_{13}	a_{11}
II	-1	a_{31}

$\det A \neq 0$, ранг матрицы равен 2, что соответствует критерию: ранг матрицы коэффициентов должен быть не менее чем число эндогенных переменных в системе без одного. Итак, второе уравнение точно идентифицировано.

Третье уравнение системы содержит $H = 3$ и $D = 2$, т. е. по необходимому условию идентификации оно точно идентифицируемо. Противоположный вывод имеем, проверив его на достаточное условие идентификации:

Таблица

Коэффициенты при отсутствующих в третьем уравнении переменных

Уравнения	Переменные	
	x_3	x_4
I	0	0
II	a_{23}	a_{24}

$\det A = 0$. Достаточное условие идентификации не выполняется. Уравнение не идентифицировано. Следовательно, рассматриваемая в целом структурная модель, идентифицированная по счетному правилу, не может считаться идентифицированной исходя из достаточного условия идентификации.

В эконометрических моделях часто наряду с уравнениями, параметры которых должны быть статистически оценены, используются балансовые тождества переменных, коэффициенты при которых равны плюс или минус единице. В этом случае само тождество не требует проверки на идентификацию, ибо коэффициенты при его переменных известны. Вместе с тем при проверке на идентификацию собственно структурных уравнений системы тождества участвуют.

Так, например, рассмотрим упрощенный вариант модели Клейна:

$$\begin{aligned} \text{функция потребления}^* &= CN_t = a_0 + a_1 y_t + a_2 T + u_1; \\ \text{уравнение инвестиций} &= I_t = b_0 + b_1 y_t + b_2 K_{t-1} + u_2; \\ \text{тождество} &= y_t = CN_t + I_t, \end{aligned} \quad (9.18)$$

где CN_t — конечное потребление в период времени t ;

y_t — национальный доход в период времени t ;

T — фактор времени, измеряемый в годах;

I_t — инвестиции в период времени t ;

K_{t-1} — капитал в период времени t .

В данной модели три эндогенные переменные (CN_t , I_t , y_t) и две предопределенные переменные (T и K_{t-1}). Первое и второе уравнения модели (9.18) точно идентифицированы по счетному правилу и по достаточному условию идентификации (ранг матрицы равен 2).

* Бернхт Э. Практика эконометрики: классика и современность: пер. с англ. М.: ЮНИТИ, 2005. С. 680.

Матрица коэффициентов при отсутствующих переменных

		/	\		
		в первом уравнении	во втором уравнении		
Уравнение II	I_t	K_{t-1}		Уравнение I	CN_t
Тождество	1	$-b_2$		Тождество	1
	-1	0			- a_2
	$\det A \neq 0$			$\det A \neq 0$	

Исследуем на идентификацию классическую модель спроса и предложения:

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + U_t, \\ Q_t^S = b_0 + b_1 P_t + V_t, \\ Q_t^S = Q_t^d \end{cases} \quad (9.19)$$

где Q_t^d – спрашиваемое количество благ (объем спроса) в момент времени t ;

Q_t^S – предлагаемое количество благ (объем предложения) в момент времени t ;

P_t – цена в момент времени t .

В системе (9.19) три эндогенные переменные – Q_t^d , Q_t^S и P_t . При этом если Q_t^d и Q_t^S расположены в левой части системы и, стало быть, являются эндогенными исходя из структуры системы, то P_t является эндогенной по экономическому содержанию (цена зависит от спроса и предложения), также в результате наличия тождества $Q_t^d = Q_t^S$.

Приравняв первое и второе уравнения системы (9.19), можно показать, что P_t является зависимой (эндогенной) переменной:

$$a_0 + a_1 P_t + U_t = b_0 + b_1 P_t + V_t.$$

$$\text{Отсюда } P_t = \frac{b_0 - a_0}{a_1 - b_1} + \frac{V_t - U_t}{a_1 - b_1}.$$

Модель спроса и предложения (9.19) не содержит ни одной экзогенной переменной и является неидентифицируемой: не выполняется порядковое условие идентификации – в первом и втором уравнениях системы $D+1$ Н, ибо $H = 2$, а $D = 0$.

Чтобы модель имела статистическое решение, в нее вводятся экзогенные переменные.

Предположим, что в правую часть уравнения спроса включен совокупный располагаемый доход y_t , тогда система примет вид

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + a_2 y_t + U_t, \\ Q_t^S = b_0 + b_1 P_t + V_t, \\ Q_t^d = Q_t^S \end{cases} \quad (9.20)$$

В системе (9.20) уравнение спроса неидентифицируемо: не соблюдено порядковое условие идентификации ($H = 2$; $D = 0$ и $D + 1$ Н),

а уравнение предложения идентифицируемо, исходя из необходимого и достаточного условия идентификации: $H = 2$; $D = 1$ и $D + 1 = H$ ранг матрицы равен 2 при трех эндогенных переменных в системе.

Чтобы уравнение спроса в системе (9.20) было идентифицированным, дополним уравнение предложения, включив в него экзогенную переменную W_t . Так, при изучении спроса на сельскохозяйственную продукцию в качестве экзогенной переменной может выступать размер осадков. Экзогенной переменной может быть и спрос на взаимозаменяемый товар. В этом случае система уравнений примет вид

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + a_2 y_t + U_t, \\ Q_t^S = b_0 + b_1 P_t + b_2 W_t + V_t, \\ Q_t^d = Q_t^S \end{cases} \quad (9.21)$$

где W_t — например, погодные условия.

Модель (9.21) становится точно идентифицируемой как по достаточному, так и по необходимому условиям идентификации: $D + 1 = H$ и ранг матрицы равен 2 при трех эндогенных переменных в системе.

9.4. Оценивание параметров системы одновременных уравнений

Структурные коэффициенты системы одновременных уравнений могут быть оценены разными способами в зависимости от вида модели с точки зрения ее идентификации.

По точно идентифицированным моделям, когда каждое уравнение системы идентифицировано, оценка параметров может быть дана *косвенным методом наименьших квадратов (КМНК)*. По сверхидентифицированным моделям, в которых каждое уравнение системы сверхидентифицировано, обычно используются:

- двухшаговый метод наименьших квадратов (ДМНК);
- трехшаговый метод наименьших квадратов (ТМНК);
- метод максимального правдоподобия (ММП).

По моделям, содержащим одновременно как точно идентифицированные, так и сверхидентифицированные уравнения, по каждому виду уравнений соответственно могут использоваться перечисленные выше методы КМНК — к точно идентифицированным уравнениям; ДМНК (ТМНК) — к сверхидентифицированным уравнениям.

9.4.1. Косвенный метод наименьших квадратов (КМНК)

Основная идея метода заключается в использовании коэффициентов приведенной модели для получения оценок параметров структурной модели. Применение КМНК предполагает выполнение следующих этапов работы:

1. Строится приведенная форма модели.
2. Для каждого уравнения приведенной модели традиционным МНК оцениваются параметры модели.

3. Коэффициенты приведенной модели трансформируются в параметры структурной модели.

Пример. Исследуется зависимость спроса и предложения некоторого товара от его цены (P_t), дохода на душу населения (y_t) и инвестиций в производство (I_t). Модель спроса и предложения имеет вид

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + a_2 y_t + \xi_1 \\ Q_t^S = b_0 + b_1 P_t + b_2 I_t + \xi_2 \\ Q_t^d = Q_t^S \end{cases}$$

где Q_t^d — спрос в момент времени t ,

Q_t^S — предложение в момент времени t .

Примем, что $Q_t^d = Q_t^S = Q_t$.

В данной модели Q_t и P_t — эндогенные переменные, а y_t и I_t — экзогенные переменные. Каждое уравнение системы точно идентифицировано. Применим КМНК, используя следующую информацию:

Q_t	20	33	28	41	40	36	42	38	51
P_t	3	3	5	4	5	6	6	7	7
y_t	34	43	51	49	55	62	70	68	78
I_t	5	6	6	7	7	6	8	8	12

Приведенная форма модели составит

$$\begin{cases} Q_t = A_1 + \delta_1 y_t + \delta_2 I_t + U_t \\ P_t = A_2 + \delta_3 y_t + \delta_4 I_t + V_t \end{cases}$$

где U_t и V_t случайные ошибки приведенной модели.

Оценку параметров каждого уравнения приведенной модели произведем обычным МНК. Результаты окажутся следующими:

$$\begin{cases} Q_t = 6,022 + 0,234 y_t + 2,394 I_t + U_t \\ P_t = -0,692 + 0,127 y_t - 0,189 I_t + V_t \end{cases}$$

Переходим от приведенной модели к структурной модели. Найдем уравнение спроса $Q_t^d = a_0 + a_1 P_t + a_2 y_t + \xi_1$, используя коэффициенты приведенной модели. Для этой цели из второго уравнения приведенной модели выразим I_t и подставим его в первое уравнение:

$$I_t = \frac{-P_t - 0,692 + 0,127 y_t}{0,189};$$

$$Q_t^d = 6,022 + 0,234 y_t + 2,394 \cdot \frac{-P_t - 0,692 + 0,127 y_t}{0,189} = -2,743 - 12,667 P_t + 1,842 y_t -$$

первое уравнение структурной модели. Аналогично найдем и второе уравнение структурной модели, характеризующее функцию предложения:

$$Q_t^S = b_0 + b_1 P_t + b_2 I_t + \xi_2.$$

Для этого надо из первого уравнения приведенной модели исключить y_t , выразив его через второе уравнение и подставив в первое:

$$y_t = \frac{P_t + 0,692 + 0,189I_t}{0,127};$$

$$Q_t^S = 6,022 + 0,234 \cdot \frac{P_t + 0,692 + 0,189I_t}{0,127} + 2,394I_t = 7,297 + 1,843P_t + 2,742I_t -$$

второе уравнение структурной модели. Таким образом, модель спроса и предложения имеет вид

$$\begin{cases} Q_t^d = -2,743 - 12,667P_t + 1,842y_t + \xi_1 \\ Q_t^S = 7,297 + 1,843P_t + 2,742I_t + \xi_2 \\ Q_t^S = Q_t^d \end{cases}$$

Если к исходным данным применить обычный МНК, то результаты окажутся следующими:

$$Q_t^d = 1,424 - 7,238P_t + 1,273y_t + \xi_1,$$

$$Q_t^S = 8,824 + 0,663P_t + 3,371I_t + \xi_2.$$

Как видим, оценки, полученные КМНК и обычным МНК, различаются. Эти различия могут быть как более, так и менее существенными в зависимости от характера исследуемой информации. В рассматриваемом примере знаки при переменных согласуются с теоретической концепцией функций спроса и предложения. Так, в модели спроса рост цены способствует снижению спроса, а рост доходов — увеличению спроса. В модели предложения рост инвестиций и цен ведет к росту объема выпуска продукции. Эти выводы характерны как при использовании КМНК, так и обычного МНК. Вместе с тем необходимо учитывать, что оценки, полученные применением обычного МНК к каждому уравнению структурной модели, несостоительны. Впервые это показал Т. Хаавельмо, рассматривая две взаимосвязанные регрессии:

$$\begin{cases} y = ax + \xi_1 \\ x = by + \xi_2 \end{cases}$$

где ξ_1 и ξ_2 — нормально распределенные случайные ошибки: $\bar{\xi}_1 = \bar{\xi}_2 = 0$, а дисперсии ошибок равны σ_1^2 и σ_2^2 . Коэффициент регрессии a отличается от структурного коэффициента (когда рассматриваются оба уравнения в системе) и совпадает с ним только в одном частном случае, когда переменная y не содержит ошибок, т. е. $\xi_1 = 0$, а ошибки переменной x имеют дисперсию, равную единице.

Как показано в работе Э. Кейна*, при исследовании потребительской функции предельная склонность к сбережению, полученная

*См.: Кейн Э. Экономическая статистика и эконометрия. Введение в количественный экономический анализ. Вып. 2: пер. с англ. М.: Статистика, 1971. С. 127—128.

обычным (b) и косвенным МНК (b^*), различаются: в большинстве случаев наблюдается

$$|b| < |b^*|$$

Подобное соотношение имеется и в нашем примере: структурные коэффициенты функции спроса по абсолютной величине превышают коэффициенты регрессии уравнения регрессии спроса от цены и дохода.

Несостоятельность оценок структурных коэффициентов при использовании к каждому уравнению системы одновременных уравнений обычного МНК возрастает при увеличении числа эндогенных переменных в правой части системы. Так, предположим, что рассматривается модель вида

$$\begin{cases} y_1 = A_1 + b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + \xi_1 \\ y_2 = A_2 + b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \xi_2 \\ y_3 = A_3 + b_{31}y_1 + b_{32}y_2 + a_{33}x_3 + \xi_3 \end{cases}$$

Применяя МНК к первому уравнению системы, коэффициент регрессии b_{13} покажет изолированное воздействие y_3 на y_1 , когда y_2 закреплен на постоянном уровне. Между тем в соответствии с системой одновременных уравнений y_2 не может быть в данном уравнении неизменным, ибо y_2 влияет на y_3 в третьем уравнении системы. Аналогично невозможно расщепить совместное влияние и других эндогенных переменных во втором и третьем уравнениях системы.

Оценка значимости каждого уравнения структурной модели дается через F -критерий Фишера и коэффициент детерминации R^2 на основе результатов приведенной модели. Так, для уравнения приведенной модели $Q_t = 6,022 + 0,234y_t + 2,394I_t + U_t$, $F = 10,54$ и $R^2 = 0,778$; это означает, что уравнение статистически значимо: F табличное при $\alpha = 0,05$ составило 5,14 и вариация спроса на 78% обусловлена вариацией дохода и инвестиций. Для структурного уравнения $Q_t^d = -2,743 - 12,667P_t + 1,842y_t + \xi_1$ оценка значимости та же, т. е. $F = 10,54$ и $R^2 = 0,778$.

Поскольку $Q_t^d = Q_t^S$, то и для уравнения

$$Q_t^S = 7,297 + 1,843P_t + 2,742I_t + \xi_2 \quad F = 10,54 \text{ и } R^2 = 0,778.$$

При применении КМНК затруднена оценка значимости структурных коэффициентов модели по t -критерию Стьюдента из-за нелинейных соотношений между структурными и приведенными коэффициентами моделей (см. п. 9.2). Вместе с тем оценка значимости структурных коэффициентов модели может быть дана, если применить двухшаговый метод наименьших квадратов (ДМНК). Как будет показано далее, для точно идентифицируемого уравнения оценка структурных коэффициентов по КМНК совпадает с использованием для нее ДМНК. В рассматриваемом примере ввиду малого числа наблюдений не все структурные коэффициенты статистически значимы: в функции

спроса влияние дохода значимо при $\alpha = 0,067$, а влияние цен не значимо, ибо вероятность ошибки 0,16; в функции предложения влияние инвестиций значимо с вероятностью ошибки в 0,067, а влияние цены не значимо.

9.4.2. Двухшаговый метод наименьших квадратов

Двухшаговый метод наименьших квадратов (ДМНК) применяется для нахождения параметров систем совместных (одновременных) уравнений. В отличие от косвенного метода наименьших квадратов двухшаговый применяется как к точно идентифицированной, так и к сверхидентифицированной системе.

Применение метода предусматривает следующие шаги.

Первый шаг, как и в КМНК, — построение приведенной формы модели (ПФМ). Сначала составляют ПФМ в символьном виде, затем с помощью МНК находят числовые параметры каждого уравнения ПФМ.

Второй шаг предполагает работу с каждым уравнением исходной (структурной) формы модели (СФМ). А именно — для каждого уравнения выполняют следующие действия:

- находят эндогенные переменные, являющиеся факторными признаками (стоят в правой части уравнения);
- для этих переменных определяют их выровненные (теоретические) значения, используя соответствующие уравнения ПФМ;
- находят параметры рассматриваемого уравнения СФМ обычным МНК, заменяя исходные значения эндогенных переменных-факторов их выровненными значениями.

Рассмотрим двухшаговый метод наименьших квадратов на примере.

Имеются квартальные данные по Российской Федерации об объемах валового внутреннего продукта ($ВВП_t$, трлн руб.), расходов на конечное потребление ($КП_t$, трлн руб.), валового накопления ($ВН_t$, трлн руб.) и чистого экспорта ($Э_t$, трлн руб.) в среднегодовых ценах 1995 г. (табл. 9.1)*

Таблица 9.1

Динамика валового внутреннего продукта Российской Федерации по кварталам (в 1995–2005 гг. трлн руб.).

Номер наблюдения	Год, квартал	$ВВП_t$	$КП_t$	$ВН_t$	$Э_t$
1	1995, I квартал	330,10	260,15	43,40	26,55
2	1995, II квартал	341,60	248,27	73,10	20,22
3	1995, III квартал	395,70	266,59	124,23	4,89
4	1995, IV квартал	361,10	251,42	105,49	4,19
5	1996, I квартал	322,80	257,22	55,71	9,87

* По данным сайта <http://www.gks.ru>

Окончание табл. 9.1

Номер наблюдения	Год, квартал	ВВП _t	КП _t	ВН _t	Э _t
6	1996, II квартал	330,10	254,48	64,34	11,28
7	1996, III квартал	374,00	244,54	116,24	13,22
8	1996, IV квартал	350,10	240,90	86,66	22,55
9	1997, I квартал	321,40	255,54	52,31	13,55
10	1997, II квартал	327,30	257,72	63,56	6,01
11	1997, III квартал	384,70	266,72	112,83	5,16
12	1997, IV квартал	362,60	278,34	77,86	6,39
13	1998, I квартал	316,70	247,37	73,12	-3,79
14	1998, II квартал	324,20	237,30	87,60	-0,70
15	1998, III квартал	350,80	270,15	62,16	18,49
16	1998, IV квартал	329,70	275,55	-6,67	60,82
17	1999, I квартал	310,80	244,14	23,93	42,73
18	1999, II квартал	334,20	232,50	51,61	50,09
19	1999, III квартал	390,90	242,69	89,33	58,87
20	1999, IV квартал	369,40	244,74	41,35	83,31
21	2000, I квартал	346,30	221,76	40,13	84,41
22	2000, II квартал	368,40	222,53	61,49	84,38
23	2000, III квартал	432,00	253,44	99,66	78,90
24	2000, IV квартал	399,80	250,08	85,17	64,55
25	2001, I квартал	347,10	235,57	46,33	65,21
26	2001, II квартал	364,15	242,31	70,86	50,98
27	2001, III квартал	424,67	261,28	116,53	46,86
28	2001, IV квартал	399,80	265,24	100,71	33,85
29	2002, I квартал	363,31	271,79	53,62	37,90
30	2002, II квартал	382,50	269,36	69,12	44,02
31	2002, III квартал	450,21	289,35	111,53	49,33
32	2002, IV квартал	417,90	287,06	87,79	43,05
33	2003, I квартал	377,14	271,76	50,60	54,77
34	2003, II квартал	399,38	279,64	74,57	45,17
35	2003, III квартал	470,24	295,94	127,63	46,66
36	2003, IV квартал	443,71	300,05	97,98	45,68
37	2004, I квартал	405,83	298,01	58,24	49,58
38	2004, II квартал	431,29	296,59	80,61	54,09
39	2004, III квартал	499,50	316,51	122,58	60,42
40	2004, IV квартал	478,06	306,31	112,89	58,86
41	2005, I квартал	435,47	314,10	60,92	60,45
42	2005, II квартал	463,66	309,69	84,13	69,83
43	2005, III квартал	535,01	336,24	127,61	71,15
44	2005, IV квартал	510,18	318,43	128,26	63,49

На основе этих данных была построена следующая система эконо-
метрических уравнений:

$$\begin{cases} \text{КП}_t = a_1 + b_{11} \text{ВВП}_{t-4} + \varepsilon_1 \\ \text{ВН}_t = a_2 + b_{21} \text{ВВП}_{t-4} + \varepsilon_2, \\ \text{ВВП}_t = \text{КП}_t + \text{ВН}_t + \vartheta, \end{cases} \quad (9.22)$$

где ВВП_{t-4} — объем ВВП за аналогичный квартал предыдущего года.

В этой системе имеются три эндогенные переменные — КП_t , ВН_t , и ВВП_t , одна экзогенная переменная — ϑ , и одна лаговая переменная — ВВП_{t-4} . Таким образом, в системе всего две предопределенных переменных — ϑ , и ВВП_{t-4} .

Проверим систему на идентификацию с помощью необходимого и достаточного условия.

Для первого уравнения имеем:

- количество эндогенных переменных, входящих в это уравнение, — две (КП_t , и ВВП_t), $H = 2$;
- количество предопределенных переменных, не входящих в это уравнение, — две (ϑ , и ВВП_{t-4}), $D = 2$.

$$H < D + 1,$$

следовательно, уравнение сверхидентифицировано.

Для второго уравнения:

- количество эндогенных переменных, входящих в это уравнение, одна (ВН_t), $H = 1$;
- количество предопределенных переменных, не входящих в это уравнение, одна (ϑ), $D = 1$.

$$H < D + 1,$$

следовательно, уравнение сверхидентифицировано.

Третье выражение является тождеством и на идентификацию не проверяется.

Таким образом, после проверки необходимого условия идентификации можно утверждать, что система сверхидентифицирована.

Для проверки достаточного условия построим для первых двух уравнений матрицы коэффициентов при переменных, отсутствующих в данном уравнении, но присутствующих в системе, и найдем их ранг.

Для первого уравнения столбцы коэффициентов при отсутствующих в нем переменных ВН_t , ВВП_{t-4} , ϑ , взятые из второго и третьего выражений системы, образуют матрицу

$$\begin{pmatrix} -1 & b_{21} & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Очевидно, что из этой матрицы можно выделить три квадратные матрицы 2×2 , причем для выполнения достаточного условия иденти-

фикации необходимо, чтобы определитель хотя бы одной из них был не равен нулю. Это действительно так. Например, определитель матрицы

$$\begin{vmatrix} -1 & b_{21} \\ 1 & 0 \end{vmatrix} = -1 \cdot 0 - b_{21} \cdot 1 = -b_{21} \neq 0.$$

Ранг матрицы равен 2, и достаточное условие для первого уравнения соблюдается.

Для второго уравнения матрица содержит коэффициенты при переменных КП_t, ВВП_t и Э_t, взятые из первого и третьего выражений системы:

$$\begin{pmatrix} -1 & b_{11} & 0 \\ 1 & -1 & 1 \end{pmatrix}.$$

Из этой матрицы также можно получить три квадратные, например:

$$\begin{vmatrix} -1 & b_{11} \\ 1 & -1 \end{vmatrix} = -1 \cdot (-1) - b_{11} \cdot 1 = 1 - b_{11} \neq 0.$$

Определитель матрицы порядка 2 × 2, выделенной из исходной, не равен нулю, следовательно, ранг исходной матрицы равен 2 и достаточное условие идентификации для второго уравнения соблюдается.

Таким образом, после проверки системы по достаточному условию идентификации можно утверждать, что система идентифицирована.

Обобщая выводы относительно идентификации системы, получаем, что система имеет решение и является сверхидентифицированной. Для нахождения ее параметров необходимо применить двухшаговый метод наименьших квадратов.

Построим приведенную форму модели и найдем ее параметры. В символьном виде ПФМ имеет вид

$$\begin{cases} \text{ВВП}_t = A_1 + B_{11} \text{ВВП}_{t-4} + B_{12} \mathcal{E}_t + u_1 \\ \text{КП}_t = A_2 + B_{21} \text{ВВП}_{t-4} + B_{22} \mathcal{E}_t + u_2, \\ \text{ВН}_t = A_3 + B_{31} \text{ВВП}_{t-4} + B_{32} \mathcal{E}_t + u_3 \end{cases} \quad (9.23)$$

где A , B — неизвестные параметры; u — случайные остатки.

В качестве исходной информации возьмем данные из табл. 9.1. В системе используются две переменные, характеризующие валовой внутренний продукт: за текущий момент времени (ВВП_t) и за аналогичный момент времени прошлого года (ВВП_{t-4}). Поэтому для проведения анализа следует взять укороченные динамические ряды: все показатели, кроме ВВП_{t-4}, берутся с пятого (I квартал 1996 г.) по последнее наблюдение (IV квартал 2005 г.), а показатель ВВП_{t-4} — с первого (I квартал 1995 г.) по сороковое наблюдение (IV квартал 2004 г.).

После применения метода наименьших квадратов к каждому уравнению ПФМ в отдельности получим следующую систему:

$$\begin{cases} \text{ВВП}_t = -31,53 + 1,06\text{ВВП}_{t-4} + 0,51\mathbb{E}_t + u_1 \\ \text{КП}_t = 79,06 + 0,52\text{ВВП}_{t-4} - 0,11\mathbb{E}_t + u_2 \\ \text{ВН}_t = -110,59 + 0,54\text{ВВП}_{t-4} - 0,38\mathbb{E}_t + u_3 \end{cases} \quad (9.24)$$

Рассмотрим уравнения структурной формы модели.

В первом уравнении СФМ эндогенной переменной-фактором является переменная ВВП_t. Найдем ее выровненные значения по первому уравнению приведенной формы (табл. 9.2).

Таблица 9.2

Расчет выровненных значений переменной ВВП_t

Номер наблюдения	ВВП _{t-4}	\mathbb{E}_t	
5	330,10	9,87	323
6	341,60	11,28	336
7	395,70	13,22	394
8	361,10	22,55	362
9	322,80	13,55	317
10	330,10	6,01	321
11	374,00	5,16	367
12	350,10	6,39	342
13	321,40	-3,79	307
14	327,30	-0,70	315
15	384,70	18,49	385
16	362,60	60,82	383
17	316,70	42,73	325
18	324,20	50,09	337
19	350,80	58,87	370
20	329,70	83,31	360
21	310,80	84,41	340
22	334,20	84,38	365
23	390,90	78,90	422
24	369,40	64,55	392
25	346,30	65,21	368
26	368,40	50,98	384
27	432,00	46,86	449
28	399,80	33,85	409
29	347,10	37,90	355
30	364,15	44,02	376
31	424,67	49,33	443
32	399,80	43,05	413

Окончание табл. 9.2

Номер наблюдения	ВВП _{t-4}	Э _t	
33	363,31	54,77	381
34	382,50	45,17	396
35	450,21	46,66	469
36	417,90	45,68	434
37	377,14	49,58	393
38	399,38	54,09	419
39	470,24	60,42	497
40	443,71	58,86	468
41	405,83	60,45	429
42	431,29	69,83	460
43	499,50	71,15	533
44	478,06	63,49	507

После этого применим к первому уравнению СФМ метод наименьших квадратов, используя в качестве исходной информации фактические значения эндогенной переменной-результата ($K\hat{I}_t$) и выровненные значения эндогенной переменной-фактора ($\hat{V}\hat{B}P_t$). Система нормальных уравнений будет иметь вид:

- в символьном виде:

$$\begin{cases} \sum K\hat{I}_t = a_1 + b_{11} \sum \hat{V}\hat{B}P_t, \\ \sum \hat{V}\hat{B}P_t \cdot K\hat{I}_t = a_1 \sum \hat{V}\hat{B}P_t + b_{11} \sum \hat{V}\hat{B}P_t^2; \end{cases}$$

- после расчета соответствующих сумм:

$$\begin{cases} 10762,95 = 40a_1 + 15646,32b_{11} \\ 4259441 = 15645,32a_1 + 6241957b_{11} \end{cases}$$

Решив эту систему, получим следующее уравнение регрессии:

$$K\hat{I}_t = 110,48 + 0,41 \hat{V}\hat{B}P_t + e_1.$$

Фактическое значение t -критерия для коэффициента регрессии в этом уравнении равно 7,96, табличное — 2,0244 ($df = 40 - 2 = 38$; $\alpha = 0,05$), следовательно, коэффициент при переменной $\hat{V}\hat{B}P_t$, значим. Значимо также все уравнение ($R^2 = 0,62$; $F = 63,3$; $F_{табл} = 4,1$).

Второе уравнение СФМ не содержит эндогенных переменных в качестве факторов. Поэтому его параметры можно найти, применяя к нему обычный МНК:

$$\begin{cases} \sum BH_t = a_2 + b_{21} \sum VBP_{t-4} \\ \sum BH_t \cdot VBP_{t-4} = a_2 \sum VBP_{t-4} + b_{21} \sum VBP_{t-4}^2 \end{cases}$$

$$\begin{cases} 3120,942 = 40a_2 + 15129,49b_{21} \\ 1223653 = 15129,49a_2 + 5812350b_{21} \end{cases}$$

Получим следующее уравнение регрессии:

$$ВН_1 = -103,89 + 0,48 ВВП_{t-4} + \epsilon_2; R^2 = 0,57; F = 49,7; (t_{\text{факт}}) (7,05)$$

Это уравнение, так же как и первое, имеет значимый коэффициент регрессии и значимо в целом по F -критерию.

Третье выражение СФМ не является уравнением регрессии, это тождество. У него нет неизвестных параметров (все параметры равны единице).

Таким образом, мы получили параметры для всех уравнений нашей системы:

$$\begin{cases} КП_1 = 110,48 + 0,41 ВВП_1 + \epsilon_1 \\ ВН_1 = -103,89 + 0,48 ВВП_{t-4} + \epsilon_2 \\ ВВП_1 = КП_1 + ВН_1 + \Theta_1 \end{cases}$$

Отметим также, что полученные уравнения регрессии имели значимые коэффициенты регрессии и были значимы в целом, что подтверждает правильность предложенной теоретической модели.

Для точно идентифицированной системы применение к ней КМНК или ДМНК дает одинаковые результаты.

Покажем справедливость этого утверждения на простом примере. Пусть имеется следующая система уравнений:

$$\begin{cases} y_1 = a_1 + b_{11}x_1 + c_{12}y_2 + e_1 \\ y_2 = a_2 + b_{22}x_2 + c_{21}y_1 + e_2 \end{cases}$$

Нетрудно убедиться, что эта система точно идентифицирована. Эндогенными переменными в ней являются y_1 и y_2 , а предопределенными — x_1 и x_2 . Приведенная форма модели для этой системы имеет вид

$$\begin{cases} y_1 = A_1 + B_{11}x_1 + B_{12}x_2 + \delta_1 \\ y_2 = A_2 + B_{21}x_1 + B_{22}x_2 + \delta_2 \end{cases}$$

Применим *косвенный метод наименьших квадратов*. Предположим, что уже найдены параметры приведенной формы модели $A_1, A_2, B_{11}, B_{12}, B_{21}, B_{22}$. Найдем параметры первого уравнения структурной формы модели.

Из второго уравнения ПФМ выразим фактор x_2 , подставим это выражение в первое уравнение ПФМ и приведем подобные члены:

$$x_2 = \frac{y_2 - A_2 - B_{21}x_1}{B_{22}};$$

$$\hat{y}_1 = A_1 - \frac{A_2}{B_{22}}B_{12} + (B_{11} - B_{12}\frac{B_{21}}{B_{22}})x_1 + \frac{B_{12}}{B_{22}}y_2.$$

Таким образом, параметры первого уравнения структурной формы модели равны:

$$a_1 = A_1 - \frac{A_2}{B_{22}}B_{12}; b_{11} = B_{11} - B_{12}\frac{B_{21}}{B_{22}}; c_{12} = \frac{B_{12}}{B_{22}}.$$

Применим теперь двухшаговый метод наименьших квадратов. По-прежнему будем предполагать, что параметры приведенной формы модели нам известны.

Для того чтобы найти параметры первого уравнения структурной формы, необходимо выявить в нем эндогенные переменные факторы. В данном случае это переменная y_2 . В соответствии с ДМНК подставим выровненные значения переменной y_2 , найденные по приведенной форме, в первое уравнение структурной формы. В символьном виде такая подстановка означает замену переменной y_2 на соответствующую функцию из ПФМ:

$$y_1 = a_1 + b_{11}x_1 + c_{12}(A_2 + B_{21}x_1 + B_{22}x_2) + e_1.$$

Раскроем скобки и приведем подобные члены:

$$y_1 = (a_1 + c_{12}A_2) + (b_{11} + c_{12}B_{21})x_1 + c_{12}B_{22}x_2 + e_1.$$

Полученное уравнение по своей структуре соответствует первому уравнению ПФМ. Следовательно, их параметры равны между собой:

$$a_1 + c_{12}A_2 = A_1; \quad b_{11} + c_{12}B_{21} = B_{11}; \quad c_{12}B_{22} = B_{12}.$$

Выразим структурные параметры через приведенные:

$$c_{12} = \frac{B_{12}}{B_{22}}; \quad b_{11} = B_{11} - B_{21} \frac{B_{12}}{B_{22}}; \quad a_1 = A_1 - A_2 \frac{B_{12}}{B_{22}}.$$

Полученные результаты не отличаются от результатов применения КМНК.

Аналогично для второго уравнения СФМ имеем следующие результаты:

- при применении КМНК:

$$x_1 = \frac{y_1 - A_1 - B_{12}x_2}{B_{11}};$$

$$\hat{y}_2 = (A_2 - B_{21} \frac{A_1}{B_{11}}) + (B_{22} - B_{21} \frac{B_{12}}{B_{11}})x_2 + \frac{B_{21}}{B_{11}}y_1;$$

$$a_2 = A_2 - B_{21} \frac{A_1}{B_{11}}; \quad b_{22} = B_{22} - B_{21} \frac{B_{12}}{B_{11}}; \quad c_{21} = \frac{B_{21}}{B_{11}};$$

- при применении ДМНК:

$$y_2 = a_2 + b_{22}x_2 + c_{21}(A_1 + B_{11}x_1 + B_{12}x_2) + e_2;$$

$$y_2 = (a_2 + c_{21}A_1) + c_{21}B_{11}x_1 + (b_{22} + c_{21}B_{12})x_2 + e_2;$$

$$c_{21}B_{11} = B_{21} \Rightarrow c_{21} = \frac{B_{21}}{B_{11}};$$

$$a_2 + c_{21}A_1 = A_2 \Rightarrow a_2 = A_2 - A_1 \frac{B_{21}}{B_{11}};$$

$$b_{22} + c_{21}B_{12} = B_{22} \Rightarrow b_{22} = B_{22} - B_{12} \frac{B_{21}}{B_{11}}.$$

Таким образом, для второго уравнения СФМ также получаем равенство параметров, определенных с помощью КМНК и ДМНК.

9.4.3. Трехшаговый метод наименьших квадратов

Рассмотренный выше двухшаговый метод наименьших квадратов позволяет получить удовлетворительные оценки параметров уравнений системы тогда, когда случайные остатки разных уравнений не коррелированы между собой. Пусть, например, система состоит из трех уравнений регрессии. Обозначим случайные остатки первого уравнения как e_{i1} , второго — как e_{i2} , третьего — как e_{i3} . Тогда некоррелированность остатков уравнений означает, что

$$r_{e_{i1} e_{i2}} = r_{e_{i1} e_{i3}} = r_{e_{i2} e_{i3}} = 0.$$

Если это предположение не выполняется, то рассчитанные оценки параметров уравнений системы теряют свойство эффективности. В этом случае необходимо применять *трехшаговый метод наименьших квадратов (ТМНК)*.

- Первые два шага трехшагового МНК совпадают с шагами ДМНК находят параметры уравнений приведенной формы модели;
- находят параметры уравнений структурной формы модели, используя выровненные значения эндогенных переменных-факторов.

Третий шаг заключается в применении обобщенного метода наименьших квадратов (ОМНК) ко всей системе в целом (тождества в расчет не включаются). Для этого все уравнения регрессии как бы объединяются в одно:

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{Z}}\mathbf{B}^{3MK} + \tilde{\mathbf{E}}. \quad (9.25)$$

Рассмотрим отдельные элементы уравнения (9.25).

Переменная — результат $\tilde{\mathbf{Y}}$ имеет вид

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{X}^T \mathbf{Y}_1 \\ \mathbf{X}^T \mathbf{Y}_2 \\ \vdots \\ \mathbf{X}^T \mathbf{Y}_k \end{pmatrix},$$

где \mathbf{X}^T — транспонированная матрица предопределенных переменных модели;
 $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k$ — матрицы-столбцы эндогенных переменных, являющихся результативными признаками в первом, втором, ..., k -м уравнениях системы:

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & & 1 \\ x_{11} & x_{12} & & x_{1n} \\ x_{21} & x_{22} & & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & & x_{pn} \end{pmatrix}; \quad \mathbf{Y}_1 = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \end{pmatrix}; \quad \mathbf{Y}_2 = \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \end{pmatrix}; \quad \mathbf{Y}_k = \begin{pmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kn} \end{pmatrix}.$$

Величина n обозначает количество наблюдений, величина p — количество предопределенных переменных в системе. Стока элементов, равных единице, в матрице $\tilde{\mathbf{X}}^T$ приводится, если в уравнениях модели есть свободный член.

Матрицы $\tilde{\mathbf{Y}}$ и $\tilde{\mathbf{Z}}$ для удобства восприятия можно разделить на блоки, соответствующие различным уравнениям системы. Каждый блок содержит $(p+1)$ строк. Будем называть блок первых $(p+1)$ строк блоком первого уравнения, блок вторых $(p+1)$ строк (с $(p+2)$ по $2(p+1)$) — блоком второго уравнения и т. д.

Матрица $\tilde{\mathbf{Z}}$ содержит преобразованные значения факторов из всех уравнений системы. Общее количество столбцов в ней равно количеству неизвестных параметров в структурной форме модели. Если обозначить количество неизвестных параметров в s -ном уравнении как q_s , то первые q_1 столбцов матрицы $\tilde{\mathbf{Z}}$ соответствуют значениям преобразованных факторных признаков — сомножителей параметров первого уравнения, следующие столбцы — с (q_1+1) по (q_1+q_2) — значениям преобразованных факторных признаков — сомножителей параметров второго уравнения и т. д.

Обозначим столбец, содержащий значения переменной при j -м параметре s -го уравнения, как $\tilde{\mathbf{Z}}_{sj}$.

Элементы столбца $\tilde{\mathbf{Z}}_{sj}$, соответствующие блоку s -го уравнения, равны элементам вектора, рассчитанным по формуле

$$\mathbf{X}^T \mathbf{Z}_{sj}, \quad (9.26)$$

где \mathbf{Z}_{sj} — столбец значений j -го фактора из s -го уравнения. Отметим, что в данном случае используются фактические значения как предопределенных, так и эндогенных переменных. Свободный член уравнения также рассматривается как коэффициент регрессии, умноженный на фактор, все значения которого равны единице. Таким образом, столбец \mathbf{Z}_{sj} для свободного члена равен*

$$\mathbf{Z}_{s1} = \begin{pmatrix} 1 \\ & \ddots \\ & & 1 \end{pmatrix}.$$

Элементы столбца \mathbf{Z}_{sj} , соответствующие блокам других уравнений, равны нулю. Так как размерность произведения $\mathbf{X}^T \mathbf{Z}_{sj}$ равна $(p+1) \times 1$ и всего имеется k уравнений регрессии, общее количество строк в матрице $\tilde{\mathbf{Z}}$ (и в каждом ее столбце \mathbf{Z}_{sj}) равно $(p+1) \cdot k$.

С учетом приведенных выше соображений матрица в общем виде может быть представлена как

*Как правило, свободный член является первым по счету параметром уравнения регрессии. Тогда все столбцы Z_{sj} соответствуют свободным членам разных уравнений системы, а нумерация коэффициентов при факторах начинается с $j=2$.

$$\tilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{X}^T \mathbf{Z}_{11} & \mathbf{X}^T \mathbf{Z}_{12} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}^T \mathbf{Z}_{21} & \mathbf{X}^T \mathbf{Z}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}^T \mathbf{Z}_{k1} & \mathbf{X}^T \mathbf{Z}_{k2} \\ \vdots & & & & & \end{pmatrix},$$

где $\mathbf{0}$ — векторы, все элементы которых равны нулю.

Матрица \mathbf{B}^{3MK} содержит параметры всех уравнений структурной формы модели. Для единобразия записи свободные члены в этой матрице обозначены как $b_{11}, b_{21}, \dots, b_{k1}$.

$$\mathbf{B}^{3MHK} = \begin{pmatrix} b_{11} \\ b_{12} \\ b_{21} \\ b_{22} \\ \vdots \\ b_{k1} \\ b_{k2} \end{pmatrix}.$$

Матрица случайных остатков $\tilde{\mathbf{E}}$ имеет ту же структуру, что и матрица $\tilde{\mathbf{Y}}$: векторы остатков \mathbf{E} , рассчитанных по каждому уравнению структурной формы модели, умножены слева на матрицу \mathbf{X}^T :

$$\tilde{\mathbf{E}} = \begin{pmatrix} \mathbf{X}^T \mathbf{E}_1 \\ \mathbf{X}^T \mathbf{E}_2 \\ \vdots \\ \mathbf{X}^T \mathbf{E}_k \end{pmatrix}.$$

Для нахождения параметров уравнения (9.25) применяют обобщенный метод наименьших квадратов:

$$\mathbf{B}^{3MHK} = (\tilde{\mathbf{Z}}^T \Omega^{-1} \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}^T \Omega^{-1} \tilde{\mathbf{Y}}, \quad (9.27)$$

где Ω — ковариационная матрица случайных остатков $\tilde{\mathbf{E}}$:

$$\Omega = \begin{pmatrix} \text{cov}_{11} \mathbf{X}^T \mathbf{X} & \text{cov}_{12} \mathbf{X}^T \mathbf{X} & \text{cov}_{1k} \mathbf{X}^T \mathbf{X} \\ \text{cov}_{21} \mathbf{X}^T \mathbf{X} & \text{cov}_{22} \mathbf{X}^T \mathbf{X} & \text{cov}_{2k} \mathbf{X}^T \mathbf{X} \\ \text{cov}_{k1} \mathbf{X}^T \mathbf{X} & \text{cov}_{k2} \mathbf{X}^T \mathbf{X} & \text{cov}_{kk} \mathbf{X}^T \mathbf{X} \end{pmatrix}, \quad (9.28)$$

В матрице Ω величина cov_{sl} — ковариация случайных остатков уравнения s и уравнения l . Для ее оценки используются случайные ос-

татки, рассчитанные по уравнениям структурной формы модели, полученным на втором шаге ДМНК:

$$\hat{c}ov_{sl} = \frac{1}{n} \sum (y_s - \hat{y}_s)(y_l - \hat{y}_l). \quad (9.29)$$

Трехшаговый метод наименьших квадратов позволяет улучшить оценки параметров модели с точки зрения их эффективности, если остатки разных уравнений связаны друг с другом. Чем менее тесная эта связь, тем ближе оценки по ТМНК к оценкам по ДМНК.

Рассмотрим в качестве примера модель, приведенную для иллюстрации двухшагового МНК. После второго шага была получена модель:

$$\begin{cases} КП_t = 110,48 + 0,41 ВВП_t + \varepsilon_1 \\ ВН_t = -103,89 + 0,48 ВВП_{t-4} + \varepsilon_2. \\ ВВП_t = КП_t + ВН_t + \vartheta_t \end{cases}$$

В системе три выражения, но только два из них являются уравнениями регрессии, поэтому далее будем говорить о двух уравнениях (первом и втором).

Найдем случайные остатки по каждому из уравнений. Отметим, что для расчета выровненных значений переменных КП_t и ВН_t используются фактические значения переменных-факторов (в частности, фактическое значение эндогенной переменной ВВП_t).

Таблица 9.3

Расчет остатков по уравнениям структурной формы модели

Номер наблюдения	КП _t	КП _t	$\ell_{1,t}$	ВН _t	ВН _t	$\ell_{2,t}$
5	257,22	241,37	15,86	55,71	54,87	0,84
6	254,48	244,33	10,16	64,34	60,40	3,94
7	244,54	262,13	-17,59	116,24	86,42	29,82
8	240,90	252,44	-11,54	86,66	69,78	16,88
9	255,54	240,80	14,74	52,31	51,36	0,95
10	257,72	243,19	14,53	63,56	54,87	8,69
11	266,72	266,47	0,25	112,83	75,99	36,84
12	278,34	257,50	20,84	77,86	64,49	13,37
13	247,37	238,89	8,48	73,12	50,69	22,44
14	237,30	241,93	-4,64	87,60	53,52	34,08
15	270,15	252,72	17,43	62,16	81,13	-18,97
16	275,55	244,16	31,39	-6,67	70,50	-77,17
17	244,14	236,50	7,64	23,93	48,43	-24,50
18	232,50	245,99	-13,49	51,61	52,03	-0,42
19	242,69	268,98	-26,29	89,33	64,83	24,51
20	244,74	260,26	-15,52	41,35	54,68	-13,33

Окончание табл. 9.3

Номер наблюдения	$K\Pi_t$	$\hat{K}\Pi_t$	ℓ_{1t}	BH_t	\hat{BH}_t	ℓ_{2t}
21	221,76	250,89	-29,13	40,13	45,59	-5,46
22	222,53	259,86	-37,33	61,49	56,84	4,65
23	253,44	285,64	-32,21	99,66	84,11	15,54
24	250,08	272,59	-22,51	85,17	73,77	11,40
25	235,57	251,22	-15,65	46,33	62,66	-16,34
26	242,31	258,13	-15,82	70,86	73,29	-2,43
27	261,28	282,67	-21,39	116,53	103,88	12,65
28	265,24	272,59	-7,34	100,71	88,39	12,31
29	271,79	257,79	13,99	53,62	63,05	-9,42
30	269,36	265,57	3,78	69,12	71,25	-2,13
31	289,35	293,03	-3,68	111,53	100,35	11,18
32	287,06	279,93	7,13	87,79	88,39	-0,60
33	271,76	263,40	8,36	50,60	70,84	-20,24
34	279,64	272,42	7,22	74,57	80,07	-5,50
35	295,94	301,15	-5,21	127,63	112,64	14,99
36	300,05	290,39	9,66	97,98	97,10	0,88
37	298,01	275,03	22,98	58,24	77,50	-19,25
38	296,59	285,36	11,24	80,61	88,19	-7,58
39	316,51	313,02	3,49	122,58	122,27	0,30
40	306,31	304,32	1,99	112,89	109,51	3,38
41	314,10	287,05	27,05	60,92	91,30	-30,37
42	309,69	298,48	11,21	84,13	103,54	-19,41
43	336,24	327,41	8,83	127,61	136,35	-8,73
44	318,43	317,35	1,08	128,26	126,03	2,22

Найдем коэффициент корреляции между случайными остатками первого и второго уравнений. Он равен -0,49, т. е. имеет место умеренная обратная связь.

Проведем необходимые преобразования данных для применения ТМНК. Для облегчения чтения в приведенных ниже матрицах указаны только первые и последние элементы.

Матрица X предопределенных переменных $B\Pi_{t-4}$ и \mathcal{E}_t :

$$X = \begin{pmatrix} 1 & 330,10 & 9,87 \\ 1 & 341,60 & 11,28 \\ 1 & 395,70 & 13,22 \\ 1 & 478,06 & 63,49 \end{pmatrix}$$

Соответственно транспонированная матрица \mathbf{X}^T имеет вид

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 330,10 & 341,60 & 395,70 & 478,06 \\ 9,87 & 11,28 & 13,22 & 63,49 \end{pmatrix}$$

Векторы значений результативных переменных отдельных уравнений системы:

$$\mathbf{Y}_1 = \mathbf{K}\Pi_t = \begin{pmatrix} 257,22 \\ 254,48 \\ 244,54 \\ 318,42 \end{pmatrix}; \quad \mathbf{Y}_2 = \mathbf{B}\mathbf{H}_t = \begin{pmatrix} 55,71 \\ 64,34 \\ 116,24 \\ 128,26 \end{pmatrix}.$$

Векторы значений факторных переменных, включая единичные множители для свободных членов:

$$\mathbf{Z}_{11} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}; \quad Z_{12} = \mathbf{B}\mathbf{V}\Pi_t = \begin{pmatrix} 322,80 \\ 330,10 \\ 374,00 \\ 510,18 \end{pmatrix}; \quad \mathbf{Z}_{21} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}; \quad Z_{22} = \mathbf{B}\mathbf{V}\Pi_{t-4} = \begin{pmatrix} 330,10 \\ 341,60 \\ 395,70 \\ 478,06 \end{pmatrix}.$$

На основе исходных данных получим преобразованные переменные.

Вектор результативной переменной $\tilde{\mathbf{Y}}$ равен

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{X}^T \mathbf{Y}_1 \\ \mathbf{X}^T \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} 10762,95 \\ 4115601,19 \\ 478757,01 \\ 3120,94 \\ 1223652,66 \\ 136099,95 \end{pmatrix}.$$

Матрица $\tilde{\mathbf{Z}}$ имеет четыре столбца. Найдем ненулевые элементы каждого из них:

$$\mathbf{X}^T \mathbf{Z}_{11} = \begin{pmatrix} 40,00 \\ 15129,49 \\ 1761,42 \end{pmatrix}; \quad \mathbf{X}^T \mathbf{Z}_{12} = \begin{pmatrix} 15645,31 \\ 6020116,88 \\ 716766,01 \end{pmatrix};$$

$$\mathbf{X}^T \mathbf{Z}_{21} = \begin{pmatrix} 40,00 \\ 15129,49 \\ 1761,42 \end{pmatrix}; \quad \mathbf{X}^T \mathbf{Z}_{22} = \begin{pmatrix} 15129,49 \\ 5812350,05 \\ 680863,04 \end{pmatrix}.$$

Тогда матрица \tilde{Z} будет равна

$$\tilde{Z} = \begin{pmatrix} 40,00 & 15645,31 & 0 & 0 \\ 15129,49 & 6020116,88 & 0 & 0 \\ 1761,42 & 716766,01 & 0 & 0 \\ 0 & 0 & 40,00 & 15129,49 \\ 0 & 0 & 15129,49 & 5812350,05 \\ 0 & 0 & 1761,42 & 680863,04 \end{pmatrix}.$$

Для вычисления элементов матрицы Ω найдем ковариации остатков рассматриваемых уравнений, а также произведение матриц $X^T X$:

$$\hat{cov}_{11} = 278,2932; \hat{cov}_{22} = 396,8266; \hat{cov}_{12} = \hat{cov}_{21} = -162,837.$$

$$X^T X = \begin{pmatrix} 40,00 & 15129,49 & 1761,42 \\ 15129,49 & 5812350,05 & 680863,04 \\ 1761,42 & 680863,04 & 101909,06 \end{pmatrix}.$$

Рассчитаем элементы матрицы Ω :

$$\Omega = \begin{pmatrix} 11131,7 & 4210434,1 & 490191,1 & & & \\ 4210434,1 & 1617537282 & 189479528,6 & & & \\ 490191,1 & 189479528,6 & 283605953 & & & \\ -6513,5 & -2463633,9 & -286823,5 & 15873,1 & 6003784,3 & 698978,1 \\ -2463633,9 & -946462907,3 & -110869373,7 & 6003784,3 & 2306494920 & 270184542,3 \\ -286823,5 & -110869373,7 & -16594518,0 & 698978,1 & 270184542,3 & 40440223,4 \end{pmatrix}.$$

Правый верхний угол матрицы Ω не заполнен, так как его элементы равны элементам левого нижнего угла. При выполнении вычислений следует указать значения всех элементов матрицы Ω .

Теперь мы имеем все исходные матрицы, чтобы применить формулу (9.27). Из-за громоздкости записи опустим промежуточные результаты расчетов. Окончательно получаем

$$B^{3MHK} = \begin{pmatrix} 121,19 \\ 0,38 \\ -118,45 \\ 0,52 \end{pmatrix}.$$

В матрице B^{3MHK} первые два элемента представляют собой параметры первого уравнения структурной формы модели, а вторые два — второго. Таким образом, после применения трехшагового метода наименьших квадратов система имеет вид:

$$\begin{cases} КП_1 = 121,19 + 0,38 ВВП_1 + e_1 \\ ВН_1 = -118,45 + 0,52 ВВП_{1-4} + e_2 \\ ВВП_1 = КП_1 + ВН_1 + \mathcal{E}_1 \end{cases}$$

Сравнивая результаты ТМНК и ДМНК, можно отметить, что параметры системы изменились. Так как связь между случайными остатками первого и второго уравнений была умеренной, эти изменения не значительны.

Контрольные вопросы

1. Дайте описание системы эконометрических уравнений в общем виде.
2. Какие типы переменных принято выделять в системах эконометрических уравнений?
3. Назовите основные виды систем эконометрических уравнений.
4. Поясните, почему нельзя использовать МНК для нахождения параметров системы одновременных уравнений.
5. Что называют структурной формой модели?
6. Для чего необходима приведенная форма модели? Какой вид она имеет?
7. Что такое идентификация модели?
8. Какие классы моделей можно выделить с точки зрения их идентификации?
9. В чем состоят необходимое и достаточное условия идентификации?
10. Какие методы могут быть использованы для нахождения параметров системы эконометрических уравнений? Какова область их применения?
11. Кратко опишите методику применения косвенного метода наименьших квадратов.
12. Сопоставьте косвенный и двухшаговый методы наименьших квадратов. В чем их сходство и в чем различие?
13. В каких случаях применяется трехшаговый метод наименьших квадратов? В чем состоит его отличие от двухшагового МНК?
14. Приведите примеры применения систем эконометрических уравнений в экономической теории.

ГЛАВА 10. МОДЕЛИ С АВТОРЕГРЕССИОННОЙ УСЛОВНОЙ ГЕТЕРОСКЕДАСТИЧНОСТЬЮ

10.1. Оценка порядка интегрируемости. Тест Дикки–Фуллера

Коинтеграция является статистическим выражением концепции долгосрочной связи между экономическими переменными. В основе понятия коинтеграции лежит идея о том, что в некоторых случаях отсутствие стационарности* у многомерного процесса вызывается общим стохастическим трендом, который может быть устранен путем взятия определенной линейной комбинации компонент процесса, в результате чего эта линейная комбинация будет стационарной. Необходимым условием коинтеграционного анализа является одинаковый порядок интегрируемости двух временных рядов. В большинстве случаев при анализе макроэкономических данных исследователь сталкивается с проблемой нестационарных рядов и (или) рядов, имеющих тренд, поэтому на первом этапе необходимо осуществить проверку исследуемых рядов на порядок интегрируемости.

Метод определения порядка интегрируемости временного ряда был предложен Д. Дикки и У. Фуллером в 1979 г. Суть метода состоит в проверке гипотезы о стационарности временного ряда и последовательно его разностей повышающегося порядка.

Тест Дикки–Фуллера основан на оценке параметра $\delta = \alpha_1 - 1$ уравнения

$$\Delta y_t = \delta y_{t-1} + \varepsilon_t, \quad (10.1)$$

эквивалентного уравнению регрессии

$$y_t = \alpha_1 y_{t-1} + \varepsilon_t. \quad (10.2)$$

Нулевая гипотеза состоит в равенстве $\delta = 0$, альтернативная гипотеза: $\delta < 0$:

$$H_0 \quad \delta = 0;$$

$$H_1 \quad \delta < 0;$$

Отклонение нулевой гипотезы в пользу альтернативной означает, что $|\alpha_1| < 1$ и временной ряд y_t – стационарный или интегрируемый нулевого порядка ($Y_t \sim I(0)$).

*Ряд y_t называется строго стационарным (или стационарным в узком смысле), если совместное распределение вероятностей m наблюдений $y_{11}, y_{12}, \dots, y_{1m}$ такое же, как и для m наблюдений $y_{(1+1)}, y_{(2+1)}, \dots, y_{(m+1)}$ для любых m, t, t^1, \dots, t^m . Так как в исследовании временных рядов обычно интересует не все распределение, а средние значения, дисперсии и ковариации, часто используется еще и понятие слабой стационарности (или стационарности в широком смысле), которое состоит в том, что среднее, дисперсия и ковариации y_t не зависят от момента времени t :

$$E(y_t) = \mu < \infty, \quad \text{Var}(y_t) = \sigma^2, \quad \text{Cov}(y_t, y_{t-k}) = \gamma_k, \quad \text{для любых } t \text{ и } k.$$

Значение μ определяет постоянный уровень, относительно которого колебляется анализируемый временной ряд y_t , а σ характеризует размах этих колебаний, см.: Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс. 2005. С. 246.

Для проверки временного ряда y_t на порядок интегрируемости необходимо рассчитать значение t -статистики Стьюдента для параметра δ , которое совпадает с расчетным значением статистики Дикки–Фуллера, и сравнить его с критическими значениями статистики Дикки–Фуллера. Если значение расчетной t -статистики меньше, чем критическое значение для соответствующего числа наблюдений, нулевую гипотезу о наличии единичного корня следует отклонить и принять альтернативную о стационарности временного ряда y_t . Если расчетное значение t -статистики превышает критическое значение, тогда нулевая гипотеза не может быть отклонена. В случае когда нулевая гипотеза о равенстве $\delta = 0$ не отклоняется, можно утверждать, что временной ряд y_t нестационарен. Следовательно, можно сделать вывод: либо временной ряд y_t , интегрируем более высокого порядка, либо не интегрируем вообще, поэтому переходим ко второму шагу, проверке гипотезы о том, что y_t – временной ряд, интегрируемый первого порядка, $Y_t \sim I(1)$, т. е. тест Дикки–Фуллера далее применяется к первым разностям Δy_t , вместо y_t . Анализируемое уравнение примет вид

$$\begin{aligned} \Delta \Delta y_t &= \delta \Delta y_{t-1} + \epsilon_t; \\ H_0 \quad \delta &= 0; \\ H_1 \quad \delta &< 0; \end{aligned} \quad (10.3)$$

Если по результатам DF -теста нулевая гипотеза отклоняется и принимается альтернативная, следовательно, ряд Δy_t – стационарный, а исходный временной ряд y_t , интегрируемый первого порядка, т. е. $Y_t \sim I(1)$. Если нулевая гипотеза не может быть отклонена, тогда следует проверить ряд y_t на интегрируемость второго порядка.

Проверка временного ряда на стационарность может продолжаться до тех пор, пока или не определится порядок интегрируемости ряда, или, с другой стороны, не будет установлена неинтегрируемость исследуемого временного ряда.

На практике наибольшую популярность получил расширенный тест Дикки–Фуллера (*Augmented Dickey–Fuller test, ADF-test*), который учитывает возможную автокорреляцию в остатках. Для этого в правую часть уравнения (10.1) было предложено включить дополнительные объясняющие переменные, а именно лаговые значения переменной из левой части, и, следовательно, уравнение примет вид

$$\Delta y_t = \alpha_1 y_{t-1} + \sum_{i=1}^k \alpha_{i+1} \Delta y_{t-i} + \epsilon_t. \quad (10.4)$$

Процедура тестирования аналогична простому тесту Дикки–Фуллера: оценивается значение t -статистики Стьюдента для параметра α_1 , критические значения для ADF-теста те же самые, что и для обычного DF -теста. При проверке ряда на стационарность как в уравнение (10.1), так и в уравнение (10.4) могут быть добавлены константа и тренд.

Следующим шагом после проверки временных рядов на стационарность и определения порядка их интегрируемости является провер-

ка временных рядов на наличие коинтеграции. В исследовании применяется метод Энгла—Грэндера. Алгоритм проведения данной процедуры следующий.

Пусть существуют два временных ряда y_t и x_t , для которых строится уравнение регрессии вида

$$y_t = a + bx_t + \varepsilon_t. \quad (10.5)$$

На первом шаге выдвигается нулевая гипотеза о нестационарности остатков и отсутствии коинтеграции между временными рядами и оценивается коинтеграционное соотношение (10.5) обычным методом наименьших квадратов.

Так как необходимым условием является одинаковый порядок интегрируемости временных рядов, то:

- если $y_t \sim I(1)$, $x_t \sim I(0)$, строить регрессию в этом случае бессмысленно, так как для любых a и b в такой ситуации $y_t - a - bx_t \sim I(1)$;
- если, наоборот, $y_t \sim I(0)$, $x_t \sim I(1)$. Для любых a и $b \neq 0$ здесь опять $y_t - a - bx_t \sim I(1)$, и только при $b = 0$ получаем $y_t - a - bx_t \sim I(0)$, следовательно, и в таком сочетании строить регрессию одного ряда на другой не имеет смысла;
- если $y_t \sim I(0)$, $x_t \sim I(0)$, проверка временных рядов на коинтеграцию также бессмысленна.

На втором шаге остатки, полученные при построении уравнения регрессии (10.5), тестируются на стационарность. Таким образом, если остатки стационарны, то временные ряды y_t и x_t , коинтегрируемы. Для этого рассчитываются параметры уравнения вида

$$\Delta \varepsilon_t = a + b \varepsilon_{t-1}. \quad (10.6)$$

Далее фактическое значение t -критерия для коэффициента регрессии b в уравнении (10.6), которое совпадает с расчетным значением статистики Дикки—Фуллера, сравнивается с критическим значением статистики Дикки—Фуллера из специальной таблицы критических значений. Если вычисленное значение статистики меньше критического, то нулевая гипотеза отклоняется и принимается альтернативная о стационарности остатков коинтеграционного уравнения (10.5) и делается вывод о том, что временные ряды y_t и x_t , коинтегрируемы.

Рассмотрим в качестве примера российский индекс ММВБ и зарубежные индексы, такие как Nikkei 225, Dow Jones Composite, FTSE 100, за период с 1 августа 2003 г. по 31 августа 2006 г. В исследовании были использованы ежедневные значения индексов на момент закрытия за три года, всего 806 значений. Для проведения более точного анализа информация по всем индексам была синхронизирована, так как различные праздники и выходные дни в России, Японии, США и Великобритании сдвигают значения в рядах данных. Изучим долгосрочные взаимодействия между рядами значений ежедневных доходностей российского индекса ММВБ и доходностями индексов других стран. Введем условные обозначения:

r_{dj} — показатель доходности индекса Dow Jones Composite;

r_{ftse} — показатель доходности индекса FTSE 100;

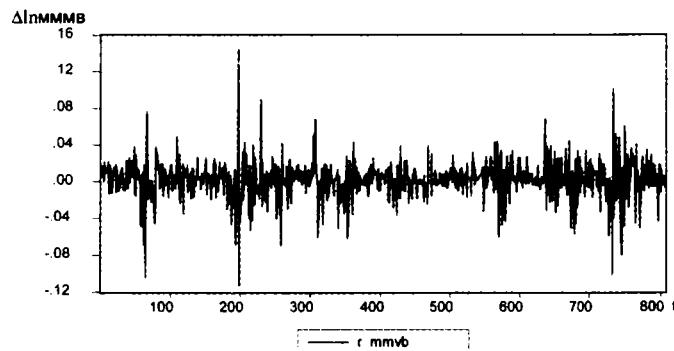


Рис. 10.1. Доходность индекса ММВБ с 1 августа 2003 г. по 31 августа 2006 г.

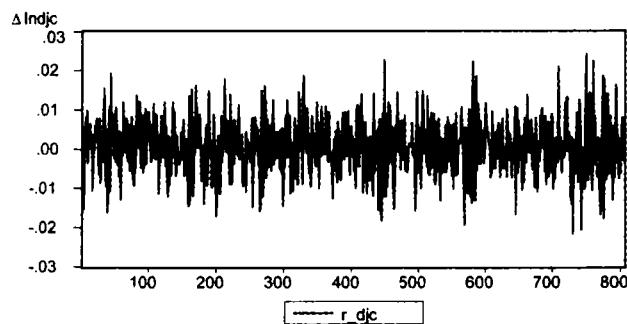


Рис 10.2. Доходность индекса Dow Jones Composite с 1 августа 2003 г. по 31 августа 2006 г.

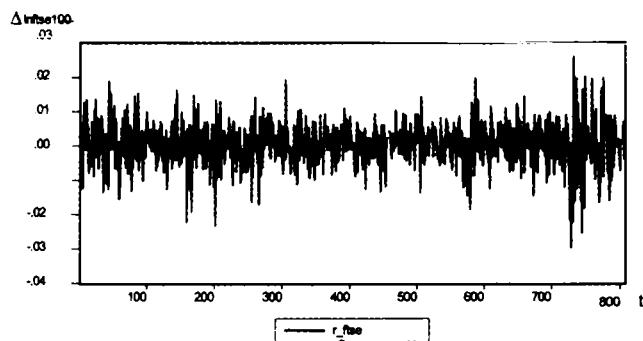


Рис. 10.3. Доходность индекса FTSE 100 с 1 августа 2003 г. по 31 августа 2006 г.

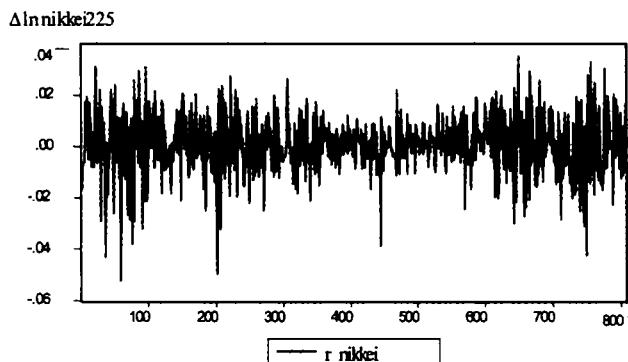


Рис. 10.4. Доходность индекса Nikkei 225
с 1 августа 2003 г. по 31 августа 2006 г.

$r_{\text{ттвб}}$ – показатель доходности индекса ММВБ;
 r_{nikkei} – показатель доходности индекса Nikkei 225.

Первый шаг при проверке исследуемого ряда на стационарность заключается в его визуализации. Графики ежедневных значений доходностей индексов с 1 августа 2003 г. по 31 августа 2006 г. (см. рис 10.1–10.4) позволяют предположить, что исследуемые ряды являются стационарными. Проверим данное предположение, применив ряд эконометрических методик.

Рассмотрим подробно процесс проверки ряда на стационарность на примере ряда ежедневных доходностей индекса ММВБ. Проверка на наличие или отсутствие автокорреляции в уровнях ряда доходностей показала, что автокорреляция отсутствует (табл. 10.1). Следующим этапом проверки временного ряда на стационарность является тест на наличие единичных корней – тест Дикки–Фуллера (DF) или расширенный тест Дикки–Фуллера (ADF).

Таблица 10.1
Автокорреляционная (AC) и частная автокорреляционная (PAC)
функции переменной $r_{\text{ттвб}}$

Autocorrelation	Partial	Correlation	AC	PAC	Q-Stat	Probability
..	..	1	-0,048	-0,048	1,8390	0,175
..	..	2	-0,015	-0,017	2,0132	0,365
..	..	3	0,002	0,001	2,0186	0,569
..	..	4	-0,011	-0,011	2,1255	0,713
.*	.*	5	0,066	0,066	5,7450	0,332
..	..	6	0,028	0,035	6,4000	0,380
..	..	7	-0,040	-0,035	7,7114	0,359
..	..	8	-0,053	-0,057	10,038	0,262

Окончание табл. 10.1

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Probability
..	..	9	0,013	0,007	10,193	0,335
..	..	10	-0,050	-0,055	12,191	0,272
..	..	11	0,005	0,004	12,212	0,348
..	..	12	-0,050	-0,049	14,247	0,285
..	..	13	0,034	0,040	15,198	0,295
..	..	14	0,013	0,015	15,331	0,356
..	..	15	-0,017	-0,012	15,566	0,411
..	..	16	0,040	0,039	16,859	0,395
..	..	17	0,014	0,022	17,025	0,453
..	..	18	-0,004	-0,009	17,039	0,520
..	..	19	-0,016	-0,024	17,276	0,571
..	..	20	-0,048	-0,055	19,177	0,510
..	..	21	0,054	0,052	21,588	0,424
..	..	22	0,011	0,004	21,693	0,478
..	..	23	-0,016	-0,009	21,909	0,526
..	..	24	-0,041	-0,036	23,325	0,501
.*	.*	25	0,096	0,108	30,940	0,191
..	..	26	-0,015	-0,009	31,122	0,224
..	..	27	-0,004	-0,013	31,132	0,266
..	..	28	-0,001	-0,001	31,134	0,311
..	..	29	-0,055	-0,044	33,685	0,251
..	..	30	0,037	0,012	34,851	0,248
* .	* .	31	-0,082	-0,092	40,626	0,116
.*	..	32	0,066	0,065	44,241	0,073
..	..	33	-0,018	0,004	44,481	0,088
..	..	34	-0,004	-0,004	44,494	0,107
..	..	35	-0,029	-0,027	45,132	0,117
..	..	36	-0,031	-0,024	45,938	0,124

Тест Дикки—Фуллера — это многошаговая процедура. На первом этапе построим модель

$$r_mmvb_t = \alpha_0 + \alpha_1 r_mmvb_{t-1} + \varepsilon_t \sim iid \quad (10.7)$$

и проверим гипотезу о наличии единичного корня в уровнях ряда ежедневных доходностей индекса ММВБ:

$$H_0: \alpha_1 = 1;$$

$$H_1: |\alpha_1| < 1.$$

Если нулевая гипотеза будет отвергнута и в модели будет отсутствовать автокорреляция в остатках, то можно утверждать, что ряд еже-

дневных доходностей индекса ММВБ стационарен. Если не будет оснований отвергнуть нулевую гипотезу, то перейдем к следующему шагу, на котором проведем исследование первых разностей ежедневных доходностей индекса ММВБ. Данная процедура прекратится, если на каком-либо из шагов мы отвергнем нулевую гипотезу и примем альтернативную. Полученные оценки модели (10.7) приведены в табл. 10.2.

Таблица 10.2

Оценка модели (с константой)

Dependent Variable: r_mmvbt				
Method: Least Squares				
Sample(adjusted): 3 805				
Included observations: 803 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
α_0	0,001620	0,000768	2,107516	0,0354
r_mmvbt _{t-1} -1	-0,047617	0,035293	-1,349187	0,1777
R-squared	0,002267	Mean dependent var		0,001546
Adjusted R-squared	0,001022	S.D. Dependent var		0,021733
S.E. Of regression	0,021722	Akaike info criterion		-4,818494
Sum squared resid	0,377948	Schwarz criterion		-4,806817
Log likelihood	1936,625	F-statistic		1,820306
Durbin-Watson stat	2,001616	Prob(F-statistic)		0,177658

Согласно табл. 10.2 коэффициент α_1 статистически незначим. Исключим из модели (10.7) константу. Оценка модели приведена в табл. 10.3.

Таблица 10.3

Оценка модели (без константы)

Dependent Variable: r_mmvbt				
Method: Least Squares				
Sample(adjusted): 3 806				
Included observations: 804 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
r_mmvbt _{t-1} -1	-0,042445	0,035258	-1,203835	0,2290
R-squared	-0,003190	Mean dependent var		0,001535
Adjusted R-squared	-0,003190	S.D. Dependent var		0,021722
S.E. Of regression	0,021756	Akaike info criterion		-4,816576
Sum squared resid	0,380092	Schwarz criterion		-4,810743
Log likelihood	1937,264	Durbin-Watson stat		2,000930

После исключения константы коэффициент α_1 также остается незначимым, и, несмотря на отсутствие автокорреляции в остатках, мы не можем утверждать, что ряд ежедневных доходностей индекса

ММВБ является стационарным, поэтому перейдем к расчету первых разностей ежедневных доходностей индекса ММВБ и проведем тест Дикки—Фуллера на порядок интегрируемости ряда. Для того чтобы учесть возможную автокорреляцию в остатках, в исследовании будем использовать расширенный тест Дикки—Фуллера.

При проведении расширенного теста Дикки—Фуллера необходимо решить проблему включения в тестовое уравнение константы, тренда и числа добавок, а именно лаговых значений зависимой переменной. Так как в ряду ежедневных доходностей индекса ММВБ за анализируемый период тренд отсутствует, то не имеет смысла включать его в уравнение. При построении тестового уравнения будем использовать метод исключений: выберем максимальное число лагов и будем постепенно исключать из уравнения незначимые переменные и добиваться отсутствия автокорреляции в остатках.

Таким образом, рассмотрим тестовое уравнение с константой и числом добавок, равным 6:

$$\begin{aligned} dr_mmvb_t = & c + \delta_0 r_mmvb_{t-1} + \delta_1 dr_mmvb_{t-1} + \\ & + \delta_2 dr_mmvb_{t-2} + \dots + \delta_6 dr_mmvb_{t-6} + \\ & \epsilon_t \sim iid, \end{aligned} \quad (10.8)$$

где dr_mmvb_t — первая разность доходностей индекса ММВБ.

Сформулируем рабочие гипотезы:

$H_0: \delta = 0 \Rightarrow$ ряд r_mmvb нестационарный, $r_mmvb \sim I(1)$;

$H_1: \delta < 0 \Rightarrow$ ряд r_mmvb стационарный, $r_mmvb \sim I(0)$.

Если нулевая гипотеза будет принята, то перейдем к следующему шагу и рассчитаем вторые разности доходностей индекса ММВБ, после чего аналогичным образом проведем расширенный тест Дикки — Фуллера на порядок интегрируемости на ряду вторых разностей ежедневных доходностей индекса ММВБ. Результаты оценки тестового уравнения (10.8) приведены в табл. 10.4.

Таблица 10.4

Оценка модели

Dependent Variable: dr_mmvbt				
Method: Least Squares				
Sample(adjusted): 9 806				
Included observations: 798 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
c	0,001484	0,000785	1,890470	0,0591
r_mmvbt _{t-1}	-1,008432	0,097003	-10,39590	0,0000
dr_mmvbt _{t-1}	-0,041218	0,090626	-0,454814	0,6494
dr_mmvbt _{t-2}	-0,056056	0,083320	-0,672779	0,5013

Окончание табл. 10.4

dr_mmvb _{t-3}	-0,055647	0,074154	-0,750432	0,4532
dr_mmvb _{t-4}	-0,064148	0,063836	-1,004879	0,3153
dr_mmvb _{t-5}	0,002419	0,051501	0,046973	0,9625
dr_mmvb _{t-6}	0,035223	0,035536	0,991177	0,3219
R-squared	0,527683	Mean dependent var	-2,12E-05	
Adjusted R-squared	0,523498	S.D. Dependent var	0,031542	
S.E. Of regression	0,021773	Akaike info criterion	-4,806275	
Sum squared resid	0,374526	Schwarz criterion	-4,759336	
Log likelihood	1925,704	F-statistic	126,0863	
Durbin-Watson stat	2,004088	Prob(F-statistic)	0,000000	

Таблица 10.4 показывает, что все добавки и константа незначимы, но чтобы полностью удостовериться в этом, проверим совместную незначимость данных показателей с помощью *теста Вальда* (*Wald test*).

Проверим следующие гипотезы:

$$H_0: c = \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0;$$

$$H_1: c \neq \delta_1 \neq \delta_2 \neq \delta_3 \neq \delta_4 \neq \delta_5 \neq \delta_6 \neq 0.$$

Таблица 10.5

Тест Вальда

Wald Test:	
Null Hypothesis:	c = 0
	$\delta_1 = 0$
	$\delta_2 = 0$
	$\delta_3 = 0$
	$\delta_4 = 0$
	$\delta_5 = 0$
	$\delta_6 = 0$
F-statistic	1,386898
Chi-square	9,708289
	Probability 0,207469
	Probability 0,205716

Согласно табл. 10.5 так как prob > 0,05, то на 5%-ном уровне значимости мы не можем отклонить нулевую гипотезу о совместном равенстве нулю анализируемых переменных, следовательно, они могут быть исключены из тестового уравнения.

Оценим тестовое уравнение (10.8) без добавок и константы:

$$dr_mmvb_t = \delta_0 r_mmvb_{t-1} + \varepsilon_t; \\ \varepsilon_t \sim iid. \quad (10.9)$$

Таблица 10.6

Оценка модели (без константы и добавок)

ADF Test Statistic	-29.56597	1% Critical Value*	-2,5683	
		5% Critical Value	-1,9398	
		10% Critical Value	-1,6158	
*MacKinnon critical values for rejection of hypothesis of a unit root.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: dr_mmvtb				
Method: Least Squares				
Date: 03/13/07 Time: 20:50				
Sample(adjusted): 3 806				
Included observations: 804 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
r_mmvtb_{t-1}	-1,042445	0.035258	-29.56597	0.0000
R-squared	0,521211	Mean dependent var	-1,61E-05	
Adjusted R-squared	0,521211	S.D. Dependent var	0,031442	
S.E. Of regression	0,021756	Akaike info criterion	-4,816576	
Sum squared resid	0,380092	Schwarz criterion	-4,810743	
Log likelihood	1937,264	Durbin-Watson stat	2,000930	

Табл. 10.6 показывает, что параметр $\delta_0 < 0$ и статистически значим. С помощью *LM-теста Брайша–Годфри* (*Breusch-Godfrey Serial Correlation LM Test*), суть которого состоит в построении уравнения регрессии остатков, проверим, присутствует ли в модели автокорреляция в остатках:

$$e_t = \rho_1 e_{t-1} + \dots + \rho_m e_{t-m} + u_t;$$

$H_0 : \rho_1 = \rho_2 = \dots = \rho_m = 0$, \Rightarrow автокорреляции в остатках нет;

$H_1 : \rho_1 \neq \dots \neq \rho_2 \neq \dots \neq \rho_m \neq 0$, \Rightarrow автокорреляция в остатках присутствует.

Результаты *LM*-теста Брайша–Годфри (см. табл. 10.7) свидетельствуют об отсутствии автокорреляции остатков, так как на всех этапах проведения теста и взятии разного числа лагов проб $> 0,05$. Таким образом, можно сделать вывод, что модель (10,9) построена и оценена корректно, значение *ADF*-статистики на 5%-ном уровне значимости, равное -29,56597, меньше критического значения, что позволяет сделать вывод о том, что ряд ежедневных доходностей индекса ММВБ стационарный и интегрируемый нулевого порядка ($r_mmvtb \sim I(0)$).

Таблица 10.7

LM-тест Бройша—Годфри

Breusch-Godfrey Serial Correlation LM Test:			
Лаг = 2			
F-statistic	0,099934	Probability	0,904908
Obs*R-squared	0,000000	Probability	1,000000
Лаг = 4			
F-statistic	0,056626	Probability	0,994039
Obs*R-squared	0,000000	Probability	1,000000
Лаг = 6			
F-statistic	0,902105	Probability	0,492663
Obs*R-squared	1,069986	Probability	0,982820
Лаг = 12			
F-statistic	0,983941	Probability	0,462370
Obs*R-squared	7,506374	Probability	0,822419
Лаг = 16			
F-statistic	0,976684	Probability	0,480426
Obs*R-squared	11,35609	Probability	0,786991
Лаг = 18			
F-statistic	0,896812	Probability	0,582700
Obs*R-squared	11,90560	Probability	0,852073

Аналогичная проверка на стационарность была проведена по остальным рядам ежедневных значений доходностей индексов. Ряды ежедневных значений доходностей индексов Dow Jones Composite, Nikkei 225, FTSE 100 являются стационарными, следовательно, не имеет смысла проводить коинтеграционный анализ.

10.2. Модель ARCH

В эконометрическом анализе предполагается, что автокорреляция ошибок в модели равна нулю. Однако в последнее десятилетие значительно возрос интерес исследователей к систематически изменяющимся ошибкам и дисперсиям ошибок, так как во многих временных рядах, таких как обменный курс, доходности фондового рынка, было обнаружено чередование периодов малых значений ошибок с периодами больших значений ошибок или соответственно низкой и высокой волатильностью. Волатильность измеряется дисперсией σ^2 временного ряда. Гомоскедастичность модели означает, что остатки модели имеют постоянную дисперсию, соответственно гетероскедастичность модели подразумевает, что остатки модели имеют непостоянную дисперсию.

Условная гетероскедастичность означает, что условная дисперсия ошибок, т. е. дисперсия при условии известной информации зависит

от времени, и она может проявляться, несмотря на общую гомоскедастичность.

Модель *ARCH*, т. е. модель с авторегрессионной условной гетероскедастичностью, была предложена Р. Энглом в 1982 году для моделирования кластеризации волатильности. Процесс *ARCH* q -го порядка $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ задается следующими соотношениями:

$$\begin{aligned}\varepsilon_t | \Omega_{t-1} &\sim N(0, \sigma_t^2); \\ \sigma_t^2 &= \omega + \gamma_1 \varepsilon_{t-1}^2 + \dots + \gamma_q \varepsilon_{t-q}^2.\end{aligned}\quad (10.10)$$

Здесь $\Omega_{t-1} = (\varepsilon_{t-1}, \varepsilon_{t-2}, \dots)$ – предыстория процесса ε_t , а σ_t^2 – условная по предыстории дисперсия ε_t , т. е. $\sigma_t^2 = \text{Var}(\varepsilon_t | \Omega_{t-1}) = E(\varepsilon_t^2 | \Omega_{t-1})$.

Условную дисперсию часто называют *волатильностью* процесса. Для того чтобы условная дисперсия оставалась положительной, требуется выполнение соотношений $\omega > 0$, и $\gamma_1, \dots, \gamma_q \geq 0$.

Данный процесс может быть записан и по-другому:

$$\begin{aligned}\xi_t &\sim NID(0, 1), \\ \varepsilon_t &= \xi_t \sigma_t, \\ \sigma_t^2 &= \omega + \gamma_1 \varepsilon_{t-1}^2 + \dots + \gamma_q \varepsilon_{t-q}^2.\end{aligned}\quad (10.11)$$

Аббревиатура *NID* означает, что ξ_t нормально распределены и независимы, т. е. нормированный случайный процесс ξ_t , не зависит от предыстории.

Смысл модели *ARCH* состоит в том, что если абсолютная величина ε_t оказывается большой, то это приводит к повышению условной дисперсии в последующие периоды. В свою очередь, при высокой условной дисперсии более вероятно появление больших (по абсолютной величине) значений ε_t . Наоборот, если значения ε_t в течение нескольких периодов близки к 0, то это приводит к понижению условной дисперсии в последующие периоды практически до уровня ω . В свою очередь, при низкой условной дисперсии более вероятно появление малых (по абсолютной величине) значений ε_t . Таким образом, *ARCH*-процесс характеризуется инерционностью условной дисперсии (кластеризации волатильности).

Несложно показать, что процесс *ARCH* не автокоррелирован:

$$E(\varepsilon_t \varepsilon_{t-j}) = E(E(\varepsilon_t \varepsilon_{t-j} | \Omega_{t-1})) = E(\varepsilon_{t-j} E(\varepsilon_t | \Omega_{t-1})) = 0.$$

Поскольку процесс имеет постоянное (нулевое) математическое ожидание и не автокоррелирован, то он является слабо стационарным в случае, если у него есть дисперсия. Если обозначить разницу между величиной ε_t и ее условным математическим ожиданием, σ_t^2 , через η_t , то получим следующую эквивалентную запись процесса *ARCH*:

$$\varepsilon_t^2 = \omega + \gamma_1 \varepsilon_{t-1}^2 + \dots + \gamma_q \varepsilon_{t-q}^2 + \eta_t. \quad (10.12)$$

Поскольку условное математическое ожидание η_t равно 0, то безусловное математическое ожидание также равно 0. Кроме того, η_t не автокоррелирован. Следовательно, квадраты процесса *ARCH*(q) следуют авторегрессионному процессу q -го порядка.

Если все корни характеристического уравнения

$$1 - \gamma_1 x + \dots + \gamma_q x^q = 0$$

лежат за пределами единичного круга, то у процесса ARCH (q) существует безусловная дисперсия и он является слабо стационарным. Поскольку коэффициенты γ_j неотрицательны, то это условие эквивалентно условию $\sum_{j=1}^q \gamma_j < 1$.

В качестве примера изучим динамику российского индекса ММВБ, используя ежедневные значения индекса на момент закрытия за три года, за период с 01 августа 2003 г. по 31 августа 2006 г. Подбор модели будем осуществлять на трансформированном ряду значений индекса ММВБ, а именно на ряду доходностей российского индекса.

В итоге получаем три модели, адекватные данным ежедневных доходностей индекса ММВБ, а именно: модель MA (5) с ограничениями ($MA(1) = MA(2) = MA(3) = MA(4) = 0$), модель AR (5) с ограничениями ($AR(1) = AR(2) = AR(3) = AR(4) = 0$, в силу незначимости коэффициентов) и модель ARMA (2,2). В случае когда несколько моделей являются значимыми, следует выбрать модель с наименьшим количеством параметров. Так как число параметров в моделях MA (5) с ограничениями и AR (5) с ограничениями равны, сравним значения информационных критерии и выберем модель с меньшим значением данных показателей.

Таблица 10.8

	Модель AR (5) с ограничениями	Модель MA (5) с ограничениями	Модель ARMA (2,2)
Модель	$r_{\text{mmvb}_t} = 0,07 r_{\text{mmvb}_{t-5}}$ $t\text{-крит.} = 2,022$, $p\text{rob} = 0,0434$	$r_{\text{mmvb}_t} = 0,078 \epsilon_{t-5}$ $t\text{-крит.} = 2,241$, $p\text{rob} = 0,0253$	$r_{\text{mmvb}_t} = 1,474 r_{\text{mmvb}_{t-1}} - 0,953$ Все коэффициенты значимы
Условие стационарности	Выполняется (0,07)	Выполняется (при любых значениях коэффициента при регрессоре ϵ_{t-5})	Выполняется ($ -0,953 $; $1,474 + (-0,953) = 0,521$ и меньше 1; $-0,953 - 1,474 = -2,426$ и меньше 1)
Автокорреляция остатков	Отсутствует	Отсутствует	Отсутствует
AIC (Критерий Акайке)	-4,825828	-4,821572	-4,815647
BIC (Критерий Шварца)	-4,819972	-4,815745	-4,805124

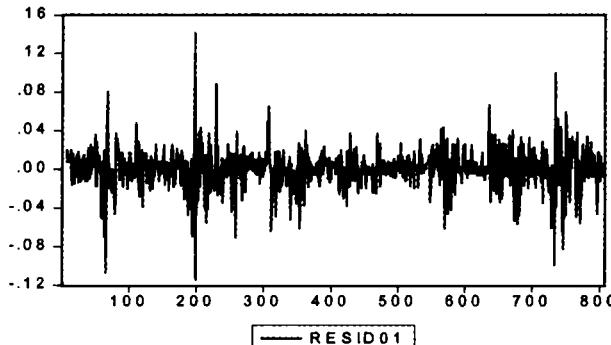


Рис. 10.5. График остатков модели AR(5) с ограничениями

Согласно табл. 10.8 модель AR(5) с ограничениями наиболее точно описывает изменение ряда ежедневных доходностей индекса ММВБ за рассматриваемый период.

Одна из характерных черт финансовых рынков – это то, что присущая рынку неопределенность изменяется во времени, и как следствие – наблюдается «кластеризация волатильности» и традиционные модели временных рядов не всегда могут адекватно учесть все характеристики, которыми обладают финансовые временные ряды, в данном случае ряд ежедневных доходностей индекса ММВБ, и требуют расширения. Построенный график остатков модели AR(5) с ограничениями для ряда доходностей индекса ММВБ за период с 01 августа 2003 г. по 31 августа 2006 г. (рис. 10.5) позволяет судить о наличии кластеров, когда происходит чередование областей низкой и высокой волатильности, то есть для ряда доходностей индекса ММВБ характерны периоды, когда показатель ведет себя непостоянно, и относительно спокойные периоды.

Для моделирования кластеризации волатильности была построена модель ARCH. На первом этапе проведем ARCH-LM-тест на статистическое выявление условной гетероскедастичности в остатках модели и оценим тестовое уравнение:

$$u_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \dots + \alpha_p u_{t-p}^2,$$

где u_t – остатки модели.

Выдвинем гипотезы:

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$, следовательно, принятие нулевой гипотезы означает отсутствие ARCH-эффекта в модели;

$$H_1: \text{не } H_0.$$

ARCH-LM тест уже при лаге, равном 1, уверенно отвергает гипотезу отсутствия условной гетероскедастичности (табл. 10.9). Для того чтобы описать процесс поведения остатков, не учтенный в выбранной модели AR(5) с ограничениями, оценим полученную модель с коррекцией на условную гетероскедастичность.

Таблица 10.9

ARCH-LM-тест для модели AR(5) с ограничениями

ARCH Test: лаг равен 1			
F-statistic	74,11121	Probability	0,000000
Obs*R-squared	67,97623	Probability	0,000000
ARCH Test: лаг равен 6			
F-statistic	16,55849	Probability	0,000000
Obs*R-squared	88,99934	Probability	0,000000
ARCH Test: лаг равен 12			
F-statistic	8,547836	Probability	0,000000
Obs*R-squared	92,10429	Probability	0,000000

Оценка модели AR (5) с ограничениями + ARCH (1) приведена в табл. 10.10.

Таблица 10.10

Оценка модели AR(5) с ограничениями + ARCH (1)

Dependent Variable: r_mmvbt				
Method: ML - ARCH (Marquardt)				
Sample(adjusted): 7 806				
Included observations: 800 after adjusting endpoints				
Convergence achieved after 17 iterations				
Variance backcast: ON				
	Coefficient	Std. Error	z-Statistic	Prob.
r_mmvbt-5	0,039809	0,032026	1,243045	0,2139
Variance Equation				
C	0,000384	1,10E-05	34,99505	0,0000
ARCH(1)	0,174890	0,036132	4,840285	0,0000
R-squared	0,000689	Mean dependent var	0,001508	
Adjusted R-squared	0,003200	S.D. Dependent var	0,021768	
S.E. Of regression	0,021803	Akaike info criterion	-4,872440	
Sum squared resid	0,378858	Schwarz criterion	-4,854873	
Log likelihood	1951,976	Durbin-Watson stat	2,087477	

Коэффициент при регрессоре AR (5) оказался незначимым на 5%-ном уровне, следовательно, данная модель не подходит для описания поведения остатков в модели AR(5) с ограничениями. Построим модель AR (5) + ARCH (2).

Тестовое уравнение для ARCH (2) имеет вид:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2,$$

где σ_t^2 — дисперсия;

u_{t-1}^2, u_{t-2}^2 — ошибки модели AR (5) с ограничениями.

В модели AR(5) +ARCH (2) коэффициент при регрессоре AR(5) также незначим (см. табл. 10.10), поэтому необходимо построить модель ARCH более высокого порядка, а именно модель, в которой остатки подчиняются процессу GARCH (1,1).

Таблица 10.11

Оценка модели AR(5) с ограничениями + ARCH(2)

Dependent Variable: r_mmvt				
Method: ML - ARCH (Marquardt)				
Sample(adjusted): 7 806				
Included observations: 800 after adjusting endpoints				
Convergence achieved after 11 iterations				
Variance backcast: ON				
	Coefficient	Std. Error	z-Statistic	Prob.
r_mmvt _{t-5}	0,069169	0,035880	1,927807	0,0539
Variance Equation				
C	0,000325	9,61E-06	33,77854	0,0000
ARCH(1)	0,132405	0,034344	3,855281	0,0001
ARCH(2)	0,164004	0,039508	4,151176	0,0000
R-squared	0,000307	Mean dependent var		0,001508
Adjusted R-squared	0,003461	S.D. Dependent var		0,021768
S.E. Of regression	0,021805	Akaike info criterion		-4,910816
Sum squared resid	0,378481	Schwarz criterion		-4,887393
Log likelihood	1968,326	Durbin-Watson stat		2,089446

Рассмотрим теоретические основы модели GARCH.

10.3. Модель GARCH

Модель GARCH (*generalized ARCH* — обобщенная модель ARCH), предложенная Т. Боллерслевом (1986), является альтернативной модификацией модели ARCH, позволяющей получить более длинные кластеры при малом числе параметров.

Известно, что модель ARMA зачастую позволяет получить более «сжатое» описание временных зависимостей для условного математического ожидания, чем модель AR. Подобным же образом модель GARCH дает возможность обойтись меньшим количеством параметров по сравнению с моделью ARCH, если речь идет об условной дис-

персии. Для того чтобы вывести модель GARCH, следует использовать в модели ARCH бесконечный геометрический лаг:

$$\sigma_t^2 = \omega + \gamma \sum_{j=1}^{\infty} \delta^{j-1} \varepsilon_{t-j}^2 = \omega + (\gamma / (1 - \delta)) \varepsilon_{t-1}^2.$$

В общем виде модель GARCH (p, q) имеет вид

$$\sigma_t^2 = \omega + \sum_{j=1}^p \delta_j \sigma_{t-j}^2 + \sum_{j=1}^q \gamma_j \varepsilon_{t-j}^2. \quad (10.13)$$

При этом предполагается, что $\omega > 0$, $\delta_1, \dots, \delta_p \geq 0$ и $\gamma_1, \dots, \gamma_q \geq 0$. Как и в модели ARCH, σ_t^2 служит условной дисперсией процесса $\varepsilon_t | \Omega_t \sim N(0, \sigma_t^2)$.

Рассчитаем безусловную дисперсию GARCH-процесса. Для этого возьмем математическое ожидание от обеих частей уравнения для условной дисперсии:

$$E(\sigma_t^2) = \sum_{j=1}^p \delta_j E(\sigma_{t-j}^2) + \sum_{j=1}^q \gamma_j E(\varepsilon_{t-j}^2), \text{ откуда } \sigma^2 = \sum_{j=1}^p \delta_j \sigma^2 + \sum_{j=1}^q \gamma_j \sigma^2$$

и

$$\sigma^2 = 1 / (1 - \sum_{j=1}^p \delta_j - \sum_{j=1}^q \lambda_j).$$

Таким образом, с точки зрения безусловной дисперсии GARCH-процесс гомоскедастичен. Для того чтобы дисперсия была конечной, требуется $\sum_{j=1}^p \delta_j + \sum_{j=1}^q \lambda_j < 1$.

Процесс GARCH можно записать в эквивалентной форме, если $\varepsilon_t^2 - \sigma_t^2 = \eta_t$:

$$\varepsilon_t^2 = \omega + \sum_{j=1}^m (\delta_j + \gamma_j) \varepsilon_{t-j}^2 + \eta_t - \sum_{j=1}^p \delta_j \eta_{t-j}, \quad (10.14)$$

где $m = \max(p, q)$.

Такая форма позволяет увидеть, что квадраты GARCH-процесса подчиняются модели ARMA(m, p).

Вернемся к примеру, рассмотренному в предыдущем параграфе. Спецификация дисперсии процесса GARCH(1,1) имеет вид

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Дисперсия должна быть неотрицательной, т. е. $\alpha_0, \alpha_1, \beta_1 \geq 0$, и стационарной:

$$\alpha_1 > 0 \quad \beta_1 > 0 \quad \alpha_1 + \beta_1 < 1.$$

Оценка модели AR(5) с ограничениями + GARCH(1,1) (табл. 10.12) показывает, что коэффициенты в модели статистически значимы.

Таблица 10.12
Оценка модели AR(5) с ограничениями + GARCH(1,1)

Dependent Variable: r_mmvbt				
Method: ML - ARCH (Marquardt)				
Sample(adjusted): 7 806				
Included observations: 800 after adjusting endpoints				
Convergence achieved after 27 iterations				
Variance backcast: ON				
	Coefficient	Std. Error	z-Statistic	Prob.
R_MMVB(-5)	0,080153	0,039666	2,020714	0,0433
Variance Equation				
C	2,72E-05	4,59E-06	5,929044	0,0000
ARCH(1)	0,149817	0,019745	7,587488	0,0000
GARCH(1)	0,796637	0,025477	31,26863	0,0000
R-squared	0,000234	Mean dependent var		0,001508
Adjusted R-squared	0,003534	S.D. Dependent var		0,021768
S.E. Of regression	0,021806	Akaike info criterion		-5,030766
Sum squared resid	0,378509	Schwarz criterion		-5,007343
Log likelihood	2016,306	Durbin-Watson stat		2,090181

Дисперсия положительна и стационарна:

$$\alpha_1 = 0,149817, \beta_1 = 0,796637,$$

$$\alpha_1 > 0 \beta_1 > 0, \alpha_1 + \beta_1 = 0,946454 \quad \alpha_1 + \beta_1 < 1.$$

Автокорреляция в модели отсутствует (табл. 10.13).

Таблица 10.13

**Проверка на наличие автокорреляции в остатках
с помощью статистики Льюнга—Бокса в модели AR(5)
с ограничениями + GARCH(1,1)**

Sample: 7 806							
Included observations: 800							
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob	
. .	. .	1	0,005	0,005	0,0195	0,889	
. .	. .	2	-0,005	-0,005	0,0362	0,982	
. .	. .	3	0,014	0,014	0,1948	0,978	
. .	. .	4	0,026	0,026	0,7464	0,945	
. .	. .	5	0,014	0,014	0,9074	0,970	
. .	. .	6	0,048	0,048	2,8044	0,833	
. .	. .	7	-0,018	-0,020	3,0803	0,877	
. .	. .	8	-0,041	-0,041	4,4253	0,817	

Окончание табл. 10.13

Sample: 7 806						
Included observations: 800						
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. .	. .	9	0,010	0,008	4,5011	0,875
. .	. .	10	-0,013	-0,016	4,6404	0,914
. .	. .	11	-0,005	-0,004	4,6628	0,946
. .	. .	12	-0,028	-0,028	5,2933	0,947
. .	. .	13	0,028	0,031	5,9370	0,948
. .	. .	14	0,007	0,010	5,9727	0,967
. .	. .	15	-0,021	-0,022	6,3287	0,974
. .	. .	16	-0,018	-0,017	6,6043	0,980
. .	. .	17	-0,007	-0,008	6,6481	0,988
. .	. .	18	0,016	0,017	6,8604	0,991
. .	. .	19	-0,016	-0,019	7,0694	0,994
. .	. .	20	-0,048	-0,049	8,9757	0,983
. .	. *	21	0,064	0,070	12,303	0,931
. .	. .	22	0,019	0,019	12,606	0,944
. .	. .	23	0,005	0,006	12,626	0,960
. .	. .	24	-0,024	-0,027	13,121	0,964
. *	. *	25	0,082	0,084	18,650	0,814
. .	. .	26	-0,013	-0,013	18,788	0,845
. .	. .	27	-0,007	-0,020	18,825	0,876
. .	. .	28	-0,002	-0,006	18,830	0,903
. .	. .	29	-0,036	-0,031	19,893	0,896
. .	. .	30	0,047	0,050	21,726	0,864
. .	. .	31	-0,043	-0,056	23,286	0,839
. .	. .	32	0,045	0,050	24,992	0,806
. .	. .	33	-0,006	0,011	25,018	0,839
. .	. .	34	-0,016	-0,022	25,243	0,861
. .	. .	35	-0,029	-0,031	25,972	0,866
. .	. .	36	-0,025	-0,035	26,493	0,876

Тест ARCH-LM не отвергает нулевую гипотезу отсутствия условной гетероскедастичности в остатках после построения GARCH(1,1) (табл. 10.14). Следовательно, модель AR(5) с ограничениями + GARCH(1,1) адекватно описывает изменение ряда ежедневных доходностей индекса ММВБ за рассматриваемый период.

Таблица 10.14

ARCH-LM-тест для модели AR(5) с ограничениями + GARCH(1,1)

ARCH Test: лаг равен 1			
F-statistic	0,195677	Probability	0,658353
Obs*R-squared	0,196120	Probability	0,657871
ARCH Test: лаг равен 8			
F-statistic	1,099139	Probability	0,361409
Obs*R-squared	8,795408	Probability	0,359848
ARCH Test: лаг равен 12			
F-statistic	0,790748	Probability	0,660427
Obs*R-squared	9,531448	Probability	0,656992

Контрольные вопросы

1. Определите понятие «коинтеграция».
2. Что является необходимым условием коинтеграционного анализа?
3. Для чего используется тест Дикки–Фуллера?
4. В чем сущность теста Дикки–Фуллера?
5. Раскройте методику применения расширенного теста Дикки–Фуллера.
6. Каков алгоритм применения метода Энгла–Грэнджера?
7. В чем суть LM-теста Броайша–Годфри?
8. Опишите модель ARCH.
9. Что такое волатильность процесса?
10. Назовите свойства ARCH-процесса.
11. Для описания каких экономических явлений и процессов может быть использована модель ARCH?
12. Опишите модель GARCH.
13. Сопоставьте модели ARCH и GARCH.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики: учебник. 2-е изд. – М. ЮНИТИ-ДАНА, 2001.
2. *Бернхт Э. Р.* Практика эконометрики: классика и современность: Учебник для студентов вузов, обучающихся по специальностям 060000 экономика и управление / пер. с англ. под ред. проф. С. А. Айвазяна. М. ЮНИТИ-ДАНА, 2005.
3. *Джонстон Дж.* Эконометрические методы. М. Статистика, 1980.
4. *Доугерти К.* Введение в эконометрику. М. Финансы и статистика, 1999.
5. *Кремер Н. Ш., Путко Б. А.* Эконометрика. М. ЮНИТИ, 2002.
6. *Магнус Я. Р., Катышев П. К., Пересецкий А. А.* Эконометрика. Начальный курс: учебник. 7-е изд., испр. – М. Дело, 2005.
7. *Тихомиров Н. П., Дорохина Е. Ю.* Эконометрика учебник. М. Экзамен, 2003.
8. Эконометрика, учебник / под ред. И. И. Елисеевой: – 2-е изд., перераб. и доп. М. Финансы и статистика, 2007.

СТАТИСТИКО-МАТЕМАТИЧЕСКИЕ ТАБЛИЦЫ*Приложение 1*

**Таблица значений F-критерия Фишера
на уровне значимости $\alpha = 0,05$**

df_1 df_2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	243,91	249,05	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,85	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,69	2,51	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,37	2,20	2,01	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,31	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,27	2,87	2,64	2,49	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,20	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44

Окончание табл.

df_1	1	2	3	4	5	6	8	12	24	∞
df_2										
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,31	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,63	2,40	2,24	2,13	1,97	1,78	1,55	1,10
400	3,86	3,02	2,63	2,39	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,12	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,11	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

**Критические значения *t*-критерия Стьюдента
на уровне значимости 0,10; 0,05; 0,01 (двусторонний)**

Число степеней свободы <i>df</i>	Уровень значимости			Число степеней свободы <i>df</i>	Уровень значимости		
	0,01	0,05	0,1		0,01	0,05	0,1
1	63,6567	12,7062	6,3138	24	2,7969	2,0639	1,7109
2	9,9248	4,3027	2,9200	25	2,7874	2,0595	1,7081
3	5,8409	3,1824	2,3534	26	2,7787	2,0555	1,7056
4	4,6041	2,7764	2,1318	27	2,7707	2,0518	1,7033
5	4,0321	2,5706	2,0150	28	2,7633	2,0484	1,7011
6	3,7074	2,4469	1,9432	29	2,7564	2,0452	1,6991
7	3,4995	2,3646	1,8946	30	2,7500	2,0423	1,6973
8	3,3554	2,3060	1,8595	35	2,7238	2,0301	1,6896
9	3,2498	2,2622	1,8331	40	2,7045	2,0211	1,6839
10	3,1693	2,2281	1,8125	45	2,6896	2,0141	1,6794
11	3,1058	2,2010	1,7959	50	2,6778	2,0086	1,6759
12	3,0545	2,1788	1,7823	60	2,6603	2,0003	1,6706
13	3,0123	2,1604	1,7709	70	2,6479	1,9944	1,6669
14	2,9768	2,1448	1,7613	80	2,6387	1,9901	1,6641
15	2,9467	2,1314	1,7531	90	2,6316	1,9867	1,6620
16	2,9208	2,1199	1,7459	100	2,6259	1,9840	1,6602
17	2,8982	2,1098	1,7396	125	2,6157	1,9791	1,6571
18	2,8784	2,1009	1,7341	150	2,6090	1,9759	1,6551
19	2,8609	2,0930	1,7291	200	2,6006	1,9719	1,6525
20	2,8453	2,0860	1,7247	300	2,5923	1,9679	1,6499
21	2,8314	2,0796	1,7207	400	2,5882	1,9659	1,6487
22	2,8188	2,0739	1,7171	500	2,5857	1,9647	1,6479
23	2,8073	2,0687	1,7139	∞	2,5758	1,9600	1,6449

Приложение 3

**Критические значения χ^2
на уровне значимости 0,10; 0,05; 0,01**

Число степеней свободы <i>df</i>	Уровень значимости			Число степеней свободы <i>df</i>	Уровень значимости		
	0,01	0,05	0,1		0,01	0,05	0,1
1	6,6349	3,8415	2,7055	24	42,9798	36,4150	33,1962
2	9,2103	5,9915	4,6052	25	44,3141	37,6525	34,3816
3	11,3449	7,8147	6,2514	26	45,6417	38,8851	35,5632
4	13,2767	9,4877	7,7794	27	46,9629	40,1133	36,7412
5	15,0863	11,0705	9,2364	28	48,2782	41,3371	37,9159
6	16,8119	12,5916	10,6446	29	49,5879	42,5570	39,0875
7	18,4753	14,0671	12,0170	30	50,8922	43,7730	40,2560
8	20,0902	15,5073	13,3616	35	57,3421	49,8018	46,0588
9	21,6660	16,9190	14,6837	40	63,6907	55,7585	51,8051
10	23,2093	18,3070	15,9872	45	69,9568	61,6562	57,5053
11	24,7250	19,6751	17,2750	50	76,1539	67,5048	63,1671
12	26,2170	21,0261	18,5493	60	88,3794	79,0819	74,3970
13	27,6882	22,3620	19,8119	70	100,4252	90,5312	85,5270
14	29,1412	23,6848	21,0641	80	112,3288	101,8795	96,5782
15	30,5779	24,9958	22,3071	90	124,1163	113,1453	107,5650
16	31,9999	26,2962	23,5418	100	135,8067	124,3421	118,4980
17	33,4087	27,5871	24,7690	125	164,6940	152,0939	145,6430
18	34,8053	28,8693	25,9894	150	193,2077	179,5806	172,5812
19	36,1909	30,1435	27,2036	200	249,4451	233,9943	226,0210
20	37,5662	31,4104	28,4120	300	359,9064	341,3951	331,7885
21	38,9322	32,6706	29,6151	400	468,7245	447,6325	436,6490
22	40,2894	33,9244	30,8133	500	576,4928	553,1268	540,9303
23	41,6384	35,1725	32,0069	1000	1106,9690	1074,6794	1057,7239

**Значения статистик Дарбина–Уотсона
на уровне значимости 0,05**

n	k = 1		k = 2		k = 3		k = 4		k = 5	
	dl	du								
6	0,61	1,40	—	—	—	—				
7	0,70	1,36	0,47	1,90	—	—				
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1,05	1,35	0,91	1,55	0,77	1,78				
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,67	1,80
200	1,76	1,78	1,75	1,79	1,74	1,80	1,73	1,81	1,72	1,82

Источники: Эконометрика: учебник / под ред. И. И. Елисеевой. 2-е изд., перераб. и доп. М. Финансы и статистика, 2005. С. 566; Дугерти К. Введение в эконометрику: пер. с англ. М: ИНФРА-М, 1997. С. 372.

Приложение 5

$$\text{Значение интеграла вероятностей } F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+t} e^{-\frac{t^2}{2}} dt$$

t	Сотые доли t									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,10	0,0797	0,0876	0,0955	0,1034	0,1113	0,1192	0,1271	0,1350	0,1428	0,1507
0,20	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,30	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2961	0,3035
0,40	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,50	0,3829	0,3899	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,60	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4907	0,4971	0,5035	0,5098
0,70	0,5161	0,5223	0,5285	0,5346	0,5407	0,5467	0,5527	0,5587	0,5646	0,5705
0,80	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,90	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6680	0,6729	0,6778
1,00	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,10	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,20	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7995	0,8029
1,30	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,40	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,50	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,60	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,70	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,80	0,9281	0,9297	0,9312	0,9328	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,90	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,00	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9615	0,9625	0,9634
2,10	0,9643	0,9651	0,9660	0,9668	0,9676	0,9684	0,9692	0,9700	0,9707	0,9715
2,20	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,30	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,40	0,9836	0,9840	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,50	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904
2,60	0,9907	0,9909	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9929

**Односторонние критические
значения статистики Дикки–Фуллера (*DF*)**

Доверительный уровень	Размер выборки			
	25	50	100	?
<i>AR</i> = модель вида $y_t = b_1 y_{t-1} + e_{1t}$				
0,010	-2,66	-2,62	-2,60	-2,58
0,025	-2,26	-2,25	-2,24	-2,23
0,050	-1,95	-1,95	-1,95	-1,95
<i>AR</i> = модель вида $y_t = a_2 + b_2 y_{t-1} + e_{2t}$				
0,010	-3,75	-3,58	-3,51	-3,43
0,025	-3,33	-3,22	-3,17	-3,12
0,050	-3,00	-2,93	-2,89	-2,86
<i>AR</i> = модель вида $y_t = a_3 + b_3 y_{t-1} + c_3 t + e_{3t}$				
0,010	-4,38	-4,15	-4,04	-3,96
0,025	-3,95	-3,80	-3,69	-3,66
0,050	-3,60	-3,50	-3,45	-3,41

Источник: Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: учебник. 6-е изд., перераб. и доп. М.: Дело, 2004. С. 280.