Министерство образования и науки Российской Федерации

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

ФАКУЛЬТЕТ ДИСТАНЦИОННОГО ОБУЧЕНИЯ (ФДО)

Е. Б. Грибанова

ЭКОНОМЕТРИКА

Учебное пособие

Томск
2014

УДК 330.43(085.8)
ББК 65в6я73
Г 820

Рецензенты:
Мицель А. А., проф. кафедры автоматизированных систем управления ТУСУРа;
Крицкий О. Л., канд. физ.-мат. наук, доцент кафедры высшей математики и математической физики Национального исследовательского Томского политехнического университета.

Грибанова Е. Б.
Г 820 Эконометрика : учебное пособие / Е. Б. Грибанова. — Томск : факультет дистанционного обучения ТУСУРа, 2014. — 156 с.

В пособии рассматриваются основные понятия эконометрического моделирования, характеристики и виды случайных величин, выборок и оценок. Представлены методы нахождения оценок неизвестных параметров регрессионной модели. Рассмотрена классическая линейная модель множественной регрессии, а также её модификации: нелинейные модели, модели с гетероскедастичными и автокоррелированными остатками.

УДК 330.43(085.8)
ББК 65в6я73

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Сегодня деятельность в любой области экономики (управлении, финансово-кредитной сфере, маркетинге, учете, аудите) требует от специалиста применения современных методов работы, знания достижений мировой экономической мысли и понимания научного языка. Большинство новых методов основано на эконометрических моделях, концепциях и приемах. Без глубоких знаний эконометрики научиться использовать их невозможно. Чтение современной экономической литературы также предполагает хорошую эконометрическую подготовку.

Специфической особенностью деятельности экономиста является работа в условиях недостатка информации и неполноты исходных данных. Анализ такой информации требует специальных методов, которые составляют один из аспектов эконометрики. Центральной проблемой эконометрики являются построение эконометрической модели и определение возможностей ее использования для описания, анализа и прогнозирования реальных экономических процессов.

Последние десятилетия эконометрика как научная дисциплина стремительно развивается. Растет число научных публикаций и исследований с применением эконометрических методов.

Свидетельством всемирного признания эконометрики является присуждение за наиболее выдающиеся разработки в этой области Нобелевских премий по экономике Р. Фришу и Я. Тинбергу (1969), Л. Клейну (1980), Т. Хаавельмо (1989), Дж. Хекману и Д. Макфаддену (2000), Р. Инглу и К. Грэнджеру (2003).

Достижения современной экономической науки предъявляют новые требования к высшему профессиональному образованию экономистов. Современное экономическое образование, — утверждает директор ЦЭМИРАН академик В. Л. Макаров, — держится на трех китах: макроэкономике, микроэкономике и эконометрике [1].

Курс «Эконометрика» разбит на семь частей.

В первой главе определяется понятие эконометрики и эконометрической модели.

Во второй главе Вы познакомитесь с понятием случайной величины, выборки, закона распределения, оценки, статистической проверки гипотез.

Третья глава посвящена описанию методов и моделей регрессионного анализа.

Глава четвертая посвящена нелинейным моделям регрессии и линеаризации. Здесь также рассмотрены основные способы выбора типа зависимости между переменными регрессионной модели.

В пятой главе рассказывается о природе гетероскедастичности. Описаны способы определения и устранения гетероскедастичности.

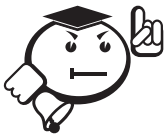
Шестая глава посвящена автокорреляции. Рассматриваются способы определения автокорреляции и её устранения.

В седьмой главе описаны некоторые вопросы практического использования регрессионных моделей. Приводится расчет эластичности, рассматриваются способы определения и устранения мультиколлинеарности, описывается тест на возможность объединения выборки и т. д.

Соглашения, принятые в книге

Для улучшения восприятия материала в данной книге используются пиктограммы и специальное выделение важной информации.

Этот блок означает определение или новое понятие.



В блоке «На заметку» автор может указать дополнительные сведения или другой взгляд на изучаемый предмет, чтобы помочь читателю лучше понять основные идеи.

Пример

Эта пиктограмма означает пример. В данном блоке автор может привести практический пример для пояснения и разбора основных моментов, отраженных в теоретическом материале.

Контрольные вопросы по главе

Глава 1

ЭКОНОМЕТРИКА И ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ: ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

1.1 Вероятностно-статистическая (эконометрическая) модель как частный случай математической модели

Математическая модель *— это абстракция реального мира, в которой интересующие исследователя отношения между реальными элементами заменены подходящими отношениями между математическими категориями.*

Эти отношения, как правило, представлены в форме уравнений или неравенств между переменными, характеризующими функционирование моделируемой системы.

Вероятностно-статистическая модель *— это вероятностная модель, значения отдельных характеристик (параметров) которой оцениваются по результатам наблюдений, характеризующих функционирование моделируемого конкретного явления.*

Вероятностно-статистическая модель, описывающая механизм функционирования экономической или социально-экономической системы, называется ***эконометрической****.*

. . . . . .

Если же речь идет о любой математической модели, описывающей механизм функционирования гипотетической экономической или социально-экономической системы, то такую модель называют экономической [2].

В качестве примера экономической модели рассмотрим простейший (идеализированный) вариант так называемой «паутинной модели», которая описывает процесс формирования спроса и предложения определенного товара на рынке. Речь идет о формализации экономического закона спроса и предложения, гласящего: количество товара, которое можно продать на рынке (то есть спрос), изменяется в направлении, противоположном изменению его цены; количество товара, которое продавцы доставляют на рынок (то есть предложение), изменяется в том же направлении, что и цена; реальная рыночная цена складывается на уровне, при котором спрос и предложение равны друг другу.

Займемся математической формализацией этих положений. Пусть x_t (ден. ед.) — цена в «момент времени» t. И пусть $y_t^{(n)}$ и $y_t^{(c)}$ — количество товара, соответственно предложенного и купленного («спрошенного») на рынке в тот же момент времени t. Тогда, с учетом одного такта времени, необходимого продавцам на то, чтобы «среагировать» на цену x, можно математически сформулировать приведенные выше закономерности в виде:

$$\begin{cases} y_t^{(n)} = f\left(x_{t-1}\right), \\ y_t^{(c)} = g\left(x_t\right), \\ \lim\limits_{t\to\infty} f\left(x_{t-1}\right) = \lim\limits_{t\to\infty} g\left(x_t\right), \\ \lim\limits_{t\to\infty} x_t = \overline{x}, \end{cases}$$

где $f(x)$ — некоторая монотонно возрастающая, а $g(x)$ — монотонно убывающая функция от аргумента x (от цены).

Математические соотношения, отражающие закон спроса-предложения, могут быть проиллюстрированы рисунком 1.1.

Из рисунка 1.1 видно, что процесс формирования цены начался с назначения в 1-й (начальный) момент времени цены на уровне x_1. Продавец отреагировал на это в следующий (2-й) момент времени величиной предложения, равной $y_2^{(n)} = f(x_1)$, в то время как спрос на этот товар сформировался всего на уровне $y_1^{(c)} = g(x_1)$. Заметное превышение предложения над спросом привело к понижению цены в следующий (2-й) момент времени до уровня x_2. Это сразу отразилось на предложении в следующий (3-й) момент времени: оно снизилось до $y_3^{(n)} = f(x_2)$. Зато спрос резко подскочил и составил во второй момент времени величину $y_2^{(c)} = g(x_2)$ и т. д. В результате этого процесса траектория сходится паутинообразно к точке равновесия, к точке пересечения кривых $g(x)$ и $f(x)$.

Реалистическая модель закона спроса-предложения, конечно, сложнее. В частности, $y^{(n)}$ и $y^{(c)}$ зависят не только от цены x, поскольку связь между $y^{(n)}$ и $y^{(c)}$,

с одной стороны, и ценой x — с другой, носит не детерминированный, а стохастический характер.

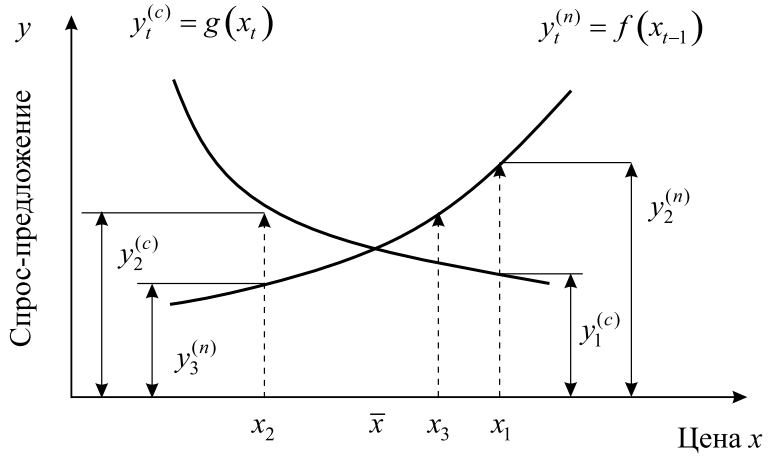


Рис. 1.1 – График процесса формирования спроса-предложения («паутинная» модель)

Для того, чтобы эта модель превратилась из экономической в эконометрическую, следует говорить не вообще о законе спроса-предложения, а о конкретном его действии в данном месте, данное время и применительно к данному конкретному товару.

1.2 Эконометрика и ее место в ряду математико-статистических и экономических дисциплин

Название «эконометрика» было введена в 1926 г. норвежским экономистом и статистиком Р. Фришем. В буквальном переводе этот термин означает «измерения в экономике».

Эконометрика — самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, методов и моделей, предназначенных для того, чтобы на базе экономической теории, экономической статистики, математико-статистического инструментария придавать конкретное количественное выражение общим закономерностям, обусловленным экономической теорией.

В соответствии с этим определением суть эконометрики — именно в синтезе экономики, экономической статистики и математики. На рисунке 1.2 приведена диаграмма, иллюстрирующая место эконометрики в сопредельных научных областях.

Говоря об экономической теории в рамках эконометрики, будем интересоваться не просто выявлением экономических законов и связей между экономическими показателями, но и подходами к их формализации, включающими в себя методы спецификации соответствующих моделей с учетом проблемы их идентифицируемости. Под математико-статистическим инструментарием эконометрики подразумевается не вся математическая статистика, а лишь отдельные ее разделы (такие

как: классическая и обобщенная линейные модели регрессионного анализа, анализ временных рядов, построение и анализ систем одновременных уравнений).

Рис. 1.2 – Область эконометрики в сопредельных научных областях

Именно «приземление» экономической теории на базу конкретной экономической статистики и извлечение из этого «приземления» с помощью подходящего математического аппарата определенных количественных взаимосвязей являются ключевыми моментами в эконометрике, обеспечивают разграничение эконометрики с дисциплинами: математической экономикой, описательной экономической статистикой и математической статистикой. Так, математическая экономика, которая на самом деле является математически сформулированной экономической теорией, изучает взаимосвязи между экономическими переменными на общем (неколичественном) уровне. Она становится эконометрикой, когда символически представленные в этих взаимосвязях коэффициенты заменяются конкретными численными оценками, полученными на базе соответствующих экономических данных.

Из определения эконометрики следует, что ее происхождение и главное назначение — это экономические и социально-экономические приложения, *а именно модельное описание* ***конкретных количественных взаимосвязей***, существующих между анализируемыми показателями. Задачи, решаемые с помощью эконометрики, можно классифицировать по трем параметрам: по конечным прикладным целям, по уровню иерархии и по профилю анализируемой экономической системы.

По конечным прикладным целям выделим две основные:

- *прогноз* экономических и социально-экономических показателей (переменных), характеризующих состояние и развитие анализируемой системы;
- *имитация* различных возможных сценариев социально-экономического развития анализируемой системы, когда статистически выявленные взаимосвязи между характеристиками производства, потребления, социальной и финансовой политики и т. п. используются для прослеживания того, как планируемые (возможные) изменения тех или иных поддающихся управлению параметров производства или распределения скажутся на значениях интересующих нас «выходных» характеристик.

По уровню иерархии анализируемой экономической системы выделяются *макроуровень* (т. е. страны в целом), *мезоуровень* (регионы, отрасли, корпорации) и *микроуровень* (семьи, предприятия, фирмы).

В некоторых случаях должен быть определен ***профиль*** эконометрического моделирования: исследование может быть сконцентрировано на проблемах рынка, инвестиционной, финансовой или социальной политики, ценообразования, распределительных отношений, спроса и потребления или на определенном *комплексе* проблем.

1.3 От простых взаимосвязей между переменными к эконометрической модели

Рассмотрим идею о взаимосвязях между экономическими переменными. Формирующийся на рынке спрос на некоторый товар рассматривается как функция его цены; затраты, связанные с изготовлением какого-либо продукта, предполагаются зависящими от объема производства; потребительские расходы могут быть функцией дохода и т. д. Это примеры связей между двумя переменными, одна из которых (спрос на товар, производственные затраты, потребительские расходы) играет роль объясняемой переменной (или результирующего показателя), а другие интерпретируются как объясняющие переменные (или факторы-аргументы). Но для большей реалистичности в каждое такое соотношение приходится вводить несколько объясняющих переменных и остаточную случайную составляющую, отражающую влияние на результирующий показатель всех неучтенных факторов. Спрос на товар можно рассматривать как функцию его цены, потребительского дохода и цен на конкурирующие и дополняющие товары; производственные затраты будут зависеть от объема производства, от его динамики и от цен на основные производственные ресурсы; потребительские расходы можно определить как функцию дохода, ликвидных активов и предыдущего уровня потребления. При этом участвующая в каждом из этих соотношений случайная составляющая, отражающая влияние на анализируемый результирующий показатель всех неучтенных факторов, обусловливает стохастический характер зависимости, а именно: даже зафиксировав на определенных уровнях значения объясняющих переменных, скажем, цены на сам товар и на конкурирующие с ним или дополняющие товары, а также потребительский доход, мы не можем ожидать, что тем самым однозначно определяется спрос на этот товар. Другими словами, переходя в своих наблюдениях спроса от одного временного или пространственного такта к другому, мы обнаружим случайное варьирование величины спроса около некоторого уровня даже при сохранении значений всех объясняющих переменных неизменными.

В прикладном статистическом анализе анализируются различные варианты формализации понятия стохастической зависимости между результирующим показателем y и объясняющими переменными $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$.

Распространенной в эконометрических приложениях формой представления стохастической зависимости является *аддитивная линейная форма*, которая и будет главным предметом исследования:

$$y_t = \theta_0 + \theta_1 x_t^{(1)} + \ldots + \theta_p x_t^{(p)} + \varepsilon_t. \qquad (1.1)$$

Здесь y_t — значение результирующей (объясняемой) переменной, измеренное в t-м временном (или пространственном) такте; $x_1^{(1)}, x_2^{(2)}, \ldots, x_t^{(p)}$ — значения объясняющих переменных, полученные в том же t-м измерении; $\theta_0, \theta_1, \ldots, \theta_p$ — некоторые параметры (как правило, не известные до проведения соответствующего статистического анализа); ε_t — случайная составляющая, характеризующая разницу между модельным и наблюденным значениями анализируемой результирующей переменной, зафиксированную в t-м измерении.

Под модельным значением результирующей переменной $\hat{y}_t$ мы будем понимать ее значение, восстановленное по заданным величинам объясняющих переменных при условии, что коэффициенты $\theta_0, \theta_1, \ldots, \theta_p$ известны, т. е.

$$\hat{y}_t = \theta_0 + \theta_1 x_t^{(1)} + \ldots + \theta_p x_t^{(p)}. \tag{1.2}$$

Случайную составляющую ε можно интерпретировать как случайную ошибку прогноза y по заданным значениям $x_1^{(1)}, x_2^{(2)}, \ldots, x_t^{(p)}$, причем, чтобы исключить систематическую ошибку в оценке y_t по $\hat{y}_t$, полагают, что среднее значение случайной составляющей ε_t при всех значениях t равно нулю (т. е. $M\varepsilon_t = 0$).

Следующий шаг в развитии экономических теорий состоит в группировке отдельных соотношений в модель. Всякая математическая модель является лишь упрощенным формализованным представлением реального объекта (явления, процесса), и искусство ее построения состоит в том, чтобы совместить как можно большую лаконичность параметризации модели с достаточной адекватностью описания именно тех сторон моделируемой реальности, которые интересуют исследователя. Количество связей, включаемых в экономическую модель, зависит от условий, при которых эта модель конструируется, и от подробности объяснения, к которой мы стремимся. Например, традиционная модель спроса и предложения должна объяснять соотношения между ценой и объемом выпуска, характерные для некоторого определенного рынка. Она содержит три уравнения, а именно: уравнение спроса, уравнение предложения и уравнение реакции рынка. В эти уравнения, помимо интересующих нас объема выпуска и цены, будут входить и другие переменные; так, например, в уравнение спроса войдет потребительский доход, а в уравнение предложения — цена. Объяснение, достигнутое с помощью такой модели, обусловлено значениями некоторых «внешних» по отношению к модели переменных, и в этом смысле модель является неполной, или условной. Более претенциозные модели содержат гораздо больше уравнений и с их помощью пытаются отразить поведение существенно большего числа переменных; однако и они остаются условными, поскольку тоже содержат переменные, не определяемые или не объясняемые моделью.

Все экономические модели, независимо от того, относятся они ко всему хозяйству или к его элементам (т. е. к макроэкономике, отрасли, фирме или рынку), имеют некоторые общие особенности. Во-первых, они основаны на предположении, что поведение экономических переменных определяется с помощью совместных и одновременных операций с некоторым числом экономических соотношений. Во-вторых, принимается гипотеза, в силу которой модель, допуская упрощение сложной действительности, тем не менее улавливает главные характеристики изучаемого объекта. В-третьих, создатель модели полагает, что на основе достигнутого с ее

помощью понимания реальной системы удастся предсказать ее будущее движение и, возможно, управлять им в целях улучшения экономического благосостояния.

Для пояснения общих черт одного из важнейших этапов эконометрического моделирования, в процессе которого исследователь математически формализует отдельные положения экономической теории и объединяет их в систему, рассмотрим пример.

Пример 1.1

Предположим, что экономист-теоретик сформулировал следующие положения:

- потребление есть возрастающая функция от имеющегося в наличии дохода, но возрастающая, видимо, медленнее, чем рост дохода;
- объем инвестиций есть возрастающая функция национального дохода и убывающая функция характеристики государственного регулирования (например, нормы процента);
- национальный доход есть сумма потребительских, инвестиционных и государственных закупок товаров и услуг.

Первая задача — перевести эти положения на математический язык. И тут мы немедленно сталкиваемся с многообразием открывающихся перед нами возможных способов удовлетворения сформулированным априорным требованиям теоретика. Какие соотношения выбрать между переменными — линейные или нелинейные? Если остановиться на нелинейных, то какими они должны быть — логарифмическими, полиномиальными или какими-либо еще? Даже определив форму конкретного соотношения, мы оставляем еще нерешенной проблему выбора для различных уравнений запаздываний по времени. Будут ли, например, инвестиции текущего периода реагировать только на национальный доход, произведенный в последнем периоде, или же на них скажется динамика нескольких предыдущих периодов? Обычный выход из этих трудностей состоит в выборе при первоначальном анализе наиболее простой из возможных форм этих соотношений. Тогда появляется возможность записать на основе указанных выше положений следующую линейную относительно анализируемых переменных и аддитивную относительно случайных составляющих модель:

$$y_t^{(1)} = \alpha_0 + \alpha_1 \left(y_t^{(3)} - x_t^{(1)} \right) + \varepsilon_t^{(1)}, \tag{1.3}$$

$$y_t^{(2)} = \beta_1 y_{t-1}^{(3)} + \beta_2 x_t^{(2)} + \varepsilon_t^{(2)}, \tag{1.4}$$

$$y_t^{(3)} = y_t^{(1)} + y_t^{(2)} + x_t^{(3)}, \tag{1.5}$$

где априорные ограничения выражены неравенствами

$$0 < \alpha_1 < 1; \quad \beta_1 > 0; \quad \beta_2 < 0.$$

Эти три соотношения вместе с ограничениями образуют модель. В ней $y_t^{(1)}$ обозначает потребление, $y_t^{(2)}$ — инвестиции, $y_t^{(3)}$ — национальный доход, $x_t^{(1)}$ — подоходный налог, $x_t^{(2)}$ — норму процента как инструмент государственного регули-

рования, $x_t^{(3)}$ — государственные закупки товаров и услуг, измеренные в «момент времени» t.

Присутствие в уравнениях «остаточных» случайных составляющих $e_t^{(1)}$ и $e_t^{(2)}$ обусловлено необходимостью учесть влияние соответственно на $y_t^{(1)}$ и $y_t^{(2)}$ ряда неучтенных факторов. Действительно, нереалистично ожидать, что величина потребления $y_t^{(1)}$ будет однозначно определяться уровнями национального дохода $y_t^{(3)}$ и подоходного налога $x_t^{(1)}$; величина инвестиций $y_t^{(2)}$ зависит не только от достигнутого в предыдущий год уровня национального дохода $y_{t-1}^{(3)}$ и от величины нормы процента $x_t^{(2)}$, но и от ряда не учтенных в уравнении факторов.

Полученная модель содержит два уравнения, объясняющие поведение потребителей и инвесторов, и одно тождество. Мы сформулировали ее для дискретных периодов времени и выбрали запаздывание (лаг) в один период для отражения воздействия национального дохода на инвестиции.

1.4 Основные понятия эконометрического моделирования

В любой эконометрической модели в зависимости от конечных прикладных целей ее использования все участвующие в ней переменные подразделяются на *экзогенные*, *эндогенные* и *предопределенные*.

***Экзогенные**, т. е. задаваемые как бы «извне», автономно, в определенной степени управляемые (планируемые).*

***Эндогенные**, т. е. такие переменные, значения которых формируются в процессе и внутри функционирования анализируемой социально-экономической системы в существенной мере под воздействием экзогенных переменных и, конечно, во взаимодействии друг с другом; в эконометрической модели они являются предметом объяснения.*

***Предопределенные**, т. е. выступающие в системе в роли факторов-аргументов, или объясняющих переменных.*

Множество предопределенных переменных формируется из всех экзогенных переменных (которые могут быть «привязаны» к прошлым, текущему или будущим моментам времени) и так называемых лаговых эндогенных переменных, т. е. таких эндогенных переменных, значения которых входят в уравнения анализируемой эконометрической системы измеренными в прошлые (по отношению к текущему) моменты времени, а следовательно, являются уже известными, заданными.

В формулах (1.3)–(1.5) потребление $\left(y_t^{(1)}\right)$, инвестиции $\left(y_t^{(2)}\right)$ и национальный доход $\left(y_t^{(3)}\right)$ в текущий момент времени t являются эндогенными переменными;

подоходный налог $\left(x_t^{(1)}\right)$, характеристика государственного регулирования $\left(x_t^{(2)}\right)$ и государственные закупки товаров и услуг $\left(x_t^{(3)}\right)$ — экзогенными переменными, которые вместе с национальным доходом в предшествующий момент времени $\left(y_{t-1}^{(3)}\right)$ образуют множество предопределенных переменных.

Таким образом, можно сказать, что эконометрическая модель служит ***для объяснения поведения эндогенных переменных в зависимости от значений экзогенных и лаговых эндогенных переменных***.

При построении и анализе эконометрической модели следует различать ее структурную и приведенную формы. Обозначим буквой X все предопределенные переменные, т. е. все экзогенные переменные и все участвующие в модели лаговые эндогенные переменные. Пусть общее число эндогенных переменных равно m, а общее число предопределенных переменных — p. Примем, что общее число уравнений и тождеств в эконометрической модели равно числу эндогенных переменных, т. е. равно m. И пусть из общего числа m соотношений модели мы имеем m_1 уравнений, включающих случайные остаточные компоненты, и m_2 тождеств ($m_1 + m_2 = m$). Разобьем вектор эндогенных переменных $Y_t = \left(y_t^{(1)} \quad y_t^{(2)} \quad \ldots \quad y_t^{(m)}\right)^T$ на два подвектора $Y_t^{(1)} = \left(y_t^{(1)} \quad y_t^{(2)} \quad \ldots \quad y_t^{(m_1)}\right)^T$ и $Y_t^{(2)} = \left(y_t^{(m+1)} \quad \ldots \quad y_t^{(m_1+m_2)}\right)^T$, при этом порядок, в котором перенумерованы эндогенные переменные, не имеет значения.

Тогда общий вид линейной эконометрической модели может быть представлен в форме

$$\begin{cases} B_1 Y_t^{(1)} + B_2 Y_t^{(2)} + C_1 X_t = \Delta t, \\ B_3 Y_t^{(1)} + B_4 Y_t^{(2)} + C_2 X_t = 0, \qquad t = 1, 2, \ldots, n, \end{cases} \tag{1.6}$$

где $B_1 = (\beta_{ij})$, $i,j = 1,\ldots,m_1$ — матрица размерности $(m_1 \times m_1)$ из коэффициентов при $y_t^{(1)},\ldots,y_t^{(m_1)}$ в m_1 первых уравнениях; $B_2 = (\beta_{ij})$, $i = 1,\ldots,m_1$, $j = m_1 + 1,\ldots,m_1 + m_2$ — матрица размерности $m_1 \times m_2$ из коэффициентов при $y_t^{(m_1+1)},\ldots,y_t^{(m_1+m_2)}$ в m_1 первых уравнениях; $X_t = \left(x_t^{(0)} \quad x_t^{(1)} \quad x_t^{(2)} \quad \ldots \quad x_t^{(p)}\right)^T$ — вектор-столбец предопределенных переменных (в нем $x_t^{(0)} \equiv 1$); $C_1 = (C_{ij})$, $i = 1,\ldots,m_1$, $j = 0,\ldots,p$ — матрица размерности $m_1 \times (p + 1)$ из коэффициентов при предопределенных переменных в первых m_1 уравнениях (очевидно, коэффициенты c_{i0} играют роль свободных членов уравнений); $B_3 = (\beta_{ij})$, $i = m_1 + 1,\ldots,m_1 + m_2$, $j = 1,\ldots,m_1$ — матрица размерности $m_2 \times m_1$ из коэффициентов при $y_t^{(1)},\ldots,y_t^{(m_1)}$ в m_2 тождествах системы; $C_2 = (C_{ij})$, $i = m_1 + 1,\ldots,m_1 + m_2$, $j = 0,\ldots,p$ — матрица размерности $m_2 \times (p + 1)$ из коэффициентов при предопределенных переменных в m_2 тождествах системы; $\Delta t = \left(\delta_t^{(1)} \quad \delta_t^{(2)} \quad \ldots \quad \delta_t^{m_1}\right)^T$ — вектор-столбец размерности m_1 случайных остаточных составляющих m_1 первых уравнений системы и $O_{m_2} = \left(0 \quad 0 \quad \ldots \quad 0\right)^T$ — вектор-столбец размерности m_2, состоящий из нулей; $B_4 = (\beta_{ij})$, $i,j = m_1 + 1,\ldots,m_1 + m_2$ — матрица размерности $m_2 \times m_2$ из коэффициентов при $y_t^{(m_1+1)},\ldots,y_t^{(m_1+m_2)}$.

Исходными статистическими данными, необходимыми для проведения статистического анализа системы (а именно для оценки неизвестных коэффициентов β_{ij} и c_{ij}, проверки статистических гипотез, например о линейном характере исследуемых зависимостей и т. п.), являются матрицы

$$Y = \begin{pmatrix} Y_1^T \\ \dots \\ Y_n^T \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ \dots \\ X_n^T \end{pmatrix}$$

соответственно размерностей $n \times m$ и $n \times (p+1)$, а все элементы матриц B_3, B_4, C_2 являются известными.

Система (1.6) может быть записана также в виде

$$B \cdot Y_t + C \cdot X_t = \overline{\Delta}_t, \quad t = 1, 2, \dots, n.$$

Система уравнений и тождеств вида (1.6) (или эквивалентных ей записей) называется структурной формой линейной эконометрической модели. При этом предполагается, что коэффициент при i-й эндогенной переменной в i-м структурном стохастическом уравнении равен единице (правило нормировки системы), а матрицы B_4 и B невырождены.

Поскольку при реализации конечных прикладных целей эконометрического моделирования главный интерес представляют соотношения, позволяющие явно выразить все эндогенные переменные Y_t через предопределенные X_t, то одновременно со структурной формой имеет смысл рассмотреть *приведенную (редуцированную)* форму линейной эконометрической модели. Требуемый результат мы получим, домножив слева обе части соотношений (1.6) на матрицу B^{-1} и уединив затем Y_t:

$$Y_t = -B^{-1}CX_t + B^{-1}\overline{\Delta}_t, \quad t = 1, 2, \dots, n.$$

1.5 Этапы эконометрического моделирования

Для пояснения сущности именно *эконометрической модели* и описания основных возникающих при ее построении и анализе проблем нам будет удобно разбить весь процесс моделирования на ***шесть основных этапов***.

1-й этап (постановочный) — определение конечных целей моделирования, набора участвующих в модели факторов и показателей, их роли.

2-й этап (априорный) — предмодельный анализ экономической сущности изучаемого явления, формирование и формализация априорной информации, в частности, относящейся к природе и генезису исходных статистических данных и случайных остаточных составляющих.

3-й этап (параметризация) — собственно моделирование, т. е. выбор общего вида модели, в том числе состава и формы входящих в нее связей.

4-й этап (информационный) — сбор необходимой статистической информации, т. е. регистрация значений участвующих в модели факторов и показателей на различных временных или пространственных тактах функционирования изучаемого явления.

5-й этап (идентификация модели) — статистический анализ модели и в первую очередь статистическое оценивание неизвестных параметров модели.

6-й этап (верификация модели) — сопоставление реальных и модельных данных, проверка адекватности модели, оценка точности модельных данных.

Последние три этапа (4-й, 5-й и 6-й) сопровождаются крайне трудоемкой ***процедурой калибровки модели***. Дело в том, что при построении эконометрической

модели исследователь, как правило, находится в ситуации, когда, с одной стороны, действует большое число «нормативных» (т. е. определенных содержательным смыслом анализируемых связей) ограничений на коэффициенты матриц B и C, а с другой стороны, ему приходится действовать в условиях определенной нечеткости (или неполноты) исходной статистической информации. Процедура калибровки модели заключается в переборе большого числа различных вариантов «нормативные ограничения — значения отдельных переменных» (что связано с многократными «вычислительными прогонами» модели) с целью получения совместной, непротиворечивой и идентифицируемой модели.

Математическая модель, в том числе математическая модель экономического явления или процесса, может быть сформулирована на общем (качественном) уровне, без настройки на конкретные статистические данные, т. е. она может иметь смысл и без 4-го и 5-го этапов. Тогда она не является эконометрической. Суть именно эконометрической модели заключается в том, что она, будучи представленной в виде набора математических соотношений, описывает функционирование конкретной экономической системы, а не системы вообще (именно экономики России или процесса «спрос — предложение» в данном конкретном месте и в данное время). Поэтому она обязательно «настраивается» на конкретных статистических данных, а значит, предусматривает обязательную реализацию 4-го и 5-го этапов моделирования.

Рассмотрим основные проблемы, которые приходится решать в процессе эконометрического моделирования.

Проблема спецификации модели. Эта проблема решается на первых трех этапах моделирования и включает в себя:

1) определение конечных целей моделирования (прогноз, имитация различных сценариев социально-экономического развития анализируемой системы, управление);
2) определение списка экзогенных и эндогенных переменных;
3) определение состава анализируемой системы уравнений и тождеств, их структуры и соответственно списка предопределенных переменных;
4) формулировка исходных предпосылок и априорных ограничений относительно:

   - стохастической природы остатков Δt (в классических вариантах моделей постулируются их взаимная статистическая независимость или некоррелированность, нулевые значения их средних величин и, иногда, сохранение постоянными в процессе наблюдения значений их дисперсий — *гомоскедастичность*);
   - числовых значений отдельных элементов матриц B и C в структурной форме модели (1.6).

Итак, спецификация модели — это первый и, быть может, *важнейший* шаг эконометрического исследования. Спецификация опирается на имеющиеся экономические теории, специальные знания или на интуитивные представления исследователя об анализируемой экономической системе. Эти априорные сведения определяют, в частности, природу матриц B и C. Например, информация о том, что

определенные переменные непосредственно не участвуют в том или ином уравнении, означает равенство нулю соответствующих элементов в строках матриц B и C. Дополнительные сведения о системе могут иметь вид ограничений на комбинации элементов матриц B и C.

Проблема идентификации. Решение этой проблемы предусматривает «настройку» записанной в общей структурной форме модели на реальные статистические данные. Другими словами, речь идет о выборе и реализации методов статистического оценивания неизвестных параметров модели (т. е. той части элементов матриц B и C, значения которых не являются априори известными) по исходным статистическим данным.

Проблема верификации модели. Эта проблема, так же как и проблема идентификации, является специфичной, связанной с построением именно эконометрической модели. Собственно построение эконометрической модели завершается ее идентификацией, т. е. статистическим оцениванием участвующих в ней неизвестных коэффициентов (параметров). После этого, однако, возникают вопросы:

а) можно ли рассчитывать на то, что использование построенной модели в целях прогноза эндогенных переменных и имитационных расчетов, определяющих варианты социально-экономического развития анализируемой системы, даст результаты, достаточно адекватные реальной действительности?

б) какова точность (абсолютная, относительная) прогнозных и имитационных расчетов, основанных на построенной модели?

Получение ответов на эти вопросы с помощью тех или иных математико-статистических методов и составляет содержание проблемы верификации эконометрической модели.

Методы верификации модели основаны на процедурах статистической проверки гипотез (при ответе на вопрос (а)) и на статистическом анализе характеристик точности различных приемов статистического оценивания параметров (при ответе на вопрос (б)). Отметим также, что наиболее распространенным и эффективным подходом к верификации эконометрической модели можно признать принцип так называемых ретроспективных расчетов. Сущность его заключается в следующем.

Предположим, мы строим эконометрическую модель с целью прогноза эндогенных переменных или имитационных расчетов на τ временных тактов вперед. Тогда исходные статистические данные делятся на две части:

- обучающую выборку:

$$\left\{\begin{pmatrix} Y_1^T \\ \cdots \\ Y_{n-\tau}^T \end{pmatrix}, \begin{pmatrix} X_1^T \\ \cdots \\ X_{n-\tau}^T \end{pmatrix}\right\},$$

- экзаменационную выборку:

$$\left\{\begin{pmatrix} Y_{n-\tau+1}^T \\ \cdots \\ Y_n^T \end{pmatrix}, \begin{pmatrix} X_{n-\tau+1}^T \\ \cdots \\ X_n^T \end{pmatrix}\right\}.$$

Далее все ранее принятые решения по проблемам спецификации, идентифицируемости и идентификации модели применяются только к наблюдениям обуча-

ющей выборки. Из полученной таким образом модели подстановкой в её приведенную форму значений экзогенных переменных $X_{n-\tau+1}, X_{n-\tau+2}, \ldots, X_n$ получают *модельные (ретроспективно прогнозные)* значения соответственно $\overline{Y}_{n-\tau+1}, \overline{Y}_{n-\tau+2}, \ldots, \overline{Y}_n$. Сравнение этих модельных значений с соответствующими реальными значениями экзаменующей выборки $Y_{n-\tau+1}, Y_{n-\tau+2}, \ldots, Y_n$ позволяет проанализировать и адекватность модельных выводов реальной действительности, и их точность.

Контрольные вопросы по главе 1

1. Что понимается под математической моделью?
2. Что понимается под вероятностно-статистической моделью?
3. При каких условиях «паутинная» модель из экономической переходит в эконометрическую?
4. Как в «паутинной» модели происходит формирование цены?
5. В чем суть эконометрики?
6. Каково главное назначение эконометрики?
7. Как можно классифицировать задачи, которые решаются с помощью эконометрики?
8. Каковы цели эконометрики?
9. Что представляет собой аддитивная линейная форма представления стохастической зависимости в эконометрике?
10. Что понимают под модельным значением переменной $\hat{y}_t$?
11. Как можно интерпретировать случайную составляющую ε?
12. В чем отличие макроуровня от микроуровня?
13. Что понимается под экзогенными, эндогенными, предопределенными переменными?
14. Какие можно привести примеры экзогенных, эндогенных и предопределенных переменных?
15. Записать общий вид линейной эконометрической модели.
16. Что является исходными статистическими данными, необходимыми для проведения статистического анализа системы?
17. Какие существуют формы линейной эконометрической модели?
18. Что называется структурной формой линейной эконометрической модели?
19. На какие этапы можно разбить процесс эконометрического моделирования?
20. Какие основные проблемы приходится решать в процессе эконометрического моделирования?
21. В чем суть проблемы спецификации модели?
22. Что предусматривает решение проблемы идентификации модели?
23. В чем состоит проблема верификации модели?

Глава 2

СЛУЧАЙНЫЕ ПЕРЕМЕННЫЕ, ВЫБОРКИ ОЦЕНКИ

2.1 Характеристики случайных величин

Событие *— это любой исход или совокупность исходов какого-либо вероятностного эксперимента* [3].

Получение прибыли можно рассматривать как результат строительства завода.

Событие, которое может произойти или не произойти в условиях данного эксперимента, называется ***случайным*** *(прибыль может быть, а может и не быть). Если событие происходит всегда в условиях данного эксперимента, то называется* ***достоверным*** *(спрос на автомобили упадет при резком снижении доходов населения).*

Событие называется ***невозможным****, если оно не происходит никогда в условиях данного эксперимента (при прочих равных условиях рост спроса на автомобили приводит к снижению их цены).*

События, которые не могут происходить одновременно, называются ***несовместимыми*** *(увеличение налогов — рост располагаемого дохода). В противном случае они называются* ***совместимыми*** *(увеличение объема продаж — увеличение прибыли).*

Два события называются ***противоположными****, если одно из них происходит тогда и только тогда, когда не происходит другое (товар реализован — товар не реализован). Событие, которое нельзя разбить на более простые, называется* ***элементарным*** *(продажа автомобиля). Событие, представимое в виде совокупности (суммы) нескольких элементарных событий, называется* ***составным*** *(предприятие не потерпело убытки — прибыль может быть положительной либо равной нулю).*

Вероятность события — это количественная мера, которая вводится для сравнения событий по степени возможности их появления.

Вероятностью события A *называется отношение числа* ***m*** *элементарных событий (исходов), благоприятствующих появлению события* ***A****, к числу* ***n*** *всех элементарных событий в условиях данного вероятностного эксперимента:*

$$P(A) = \frac{m}{n}.$$

Из определения вытекают следующие *свойства вероятности*:

1) $0 \leqslant P(A) \leqslant 1$;
2) вероятность достоверного события $P(A) = 1$;
3) вероятность невозможного события $P(A) = 0$;
4) если события A и B несовместимы, то $P(A + B) = P(A) + P(B)$;
5) если A и $\overline{A}$ — противоположные события, то $P(A) = 1 - P\left(\overline{A}\right)$.

Случайной величиной (СВ) *называют величину, которая в результате наблюдения принимает то или иное значение, заранее неизвестное и зависящее от случайных обстоятельств.*

Объем ВНП, количество реализованной продукции, прибыль фирмы, размер чистого экспорта за год и т. д. являются случайными величинами.

Различают дискретные и непрерывные СВ.

Дискретной *называют такую СВ, которая принимает отдельные, изолированные значения с определенными вероятностями (такая СВ имеет счетное количество значений).*

Непрерывной *называют такую СВ, которая может принимать любое значение из некоторого конечного или бесконечного промежутка (т. е. число возможных значений непрерывной СВ бесконечно).*

. . . . . .

Например, можно считать, что число покупателей в магазине, побывавших там в течение дня; число автомобилей, ремонтируемых еженедельно в данной мастерской; число находящихся в аэропорту самолетов являются дискретными СВ. Однако большинство СВ, рассматриваемых в экономике, имеют настолько большое число возможных значений, что их удобнее представлять в виде непрерывных СВ. Например, курсы валют, доход, объемы ВНП, ВВП и т. п. обычно рассматриваются как непрерывные СВ.

. . . . . .

Для описания дискретной СВ необходимо установить соответствие между всеми возможными значениями СВ и их вероятностями. Такое соответствие называется ***законом распределения дискретной СВ****. Его можно задать таблично, аналитически (в виде формулы) либо графически.*

. . . . . .

При табличном задании закона распределения дискретной СВ X первая строка таблицы содержит ее возможные значения, а вторая — их вероятности (табл. 2.1).

Таблица 2.1 – Закон распределения случайной величины

X	x_1	x_2	$\ldots$	x_k
p_i	p_1	p_2	$\cdots$	p_k

$p_1 + p_2 + \ldots + p_k = 1$

Непрерывные случайные величины могут принимать бесконечные количества значений. Проиллюстрируем наши рассуждения на примере температуры в комнате. Для определенности предположим, что она меняется в пределах от 55 до 75°F, и вначале предположим, что все значения в этом диапазоне равновероятны.

Поскольку число различных значений, принимаемых показателем температуры, бесконечно, здесь бессмысленно пытаться придать определенный «вес» каждому ее значению. Вместо этого мы будем говорить о вероятности того, что случайная величина лежит в пределах заданного интервала, и вероятность представлена графически площадью в пределах данного интервала. Например, в представленном случае вероятность того, что случайная величина X лежит в интервале от 59 до 60, равна 0.05, поскольку данный промежуток составляет $1/20$ от всего промежутка между 55 и 75. На рисунке 2.1 вероятность того, что величина A лежит внутри данного интервала, показана в виде прямоугольника. Поскольку его площадь равна 0.05, а его основание равно единице, его высота должна равняться 0.05. То же правило верно для всех интервалов из диапазона значений, которые может принимать X.

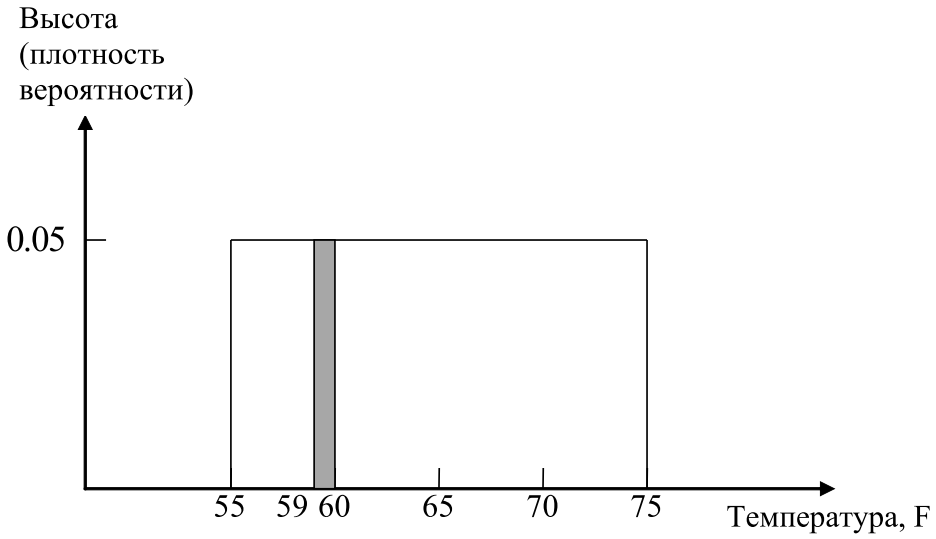


Рис. 2.1 – Графическая интерпретация плотности вероятности

Найдя высоты для всех точек в данном диапазоне, мы сможем ответить на такой, например, вопрос: с какой вероятностью температура будет находиться в диапазоне от 65 до 70°F? Ответ определяется площадью в диапазоне от 65 до 70°F. Основание прямоугольника равно 5, его высота равна 0.05, и, соответственно, площадь равна 0.25. Искомая вероятность равна $1/4$, что в любом случае очевидно, поскольку промежуток от 65 до 70°F составляет $1/4$ всего рассматриваемого диапазона. Высота заштрихованной площади в конкретной точке формально называется плотностью вероятности в этой точке, и если эта высота может быть записана как функция значений случайной переменной, то эта функция называется *функцией плотности вероятности*. В нашем примере это $f(x)$, где X — температура, и

$$f(X) = 0.05 \text{ для } 55 \leqslant X \leqslant 75,$$

$$f(X) = 0 \text{ для } X < 55 \text{ или } X > 75.$$

Свойства плотности вероятности

1) $f(x) > 0$;

2) $P(a \leqslant X \leqslant b) = \int_a^b f(x)\,dx$;

3) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ (условие нормировки).

Математическое ожидание характеризует среднее ожидаемое значение СВ, т. е. приближенно равно ее среднему значению. Для решения многих задач достаточно знать эту величину. Например, при оценивании покупательной способности населения вполне может хватить знания среднего дохода. При анализе выгодности двух видов деятельности можно ограничиться сравнением их средних прибыльностей. Знание того, что выпускники данного университета зарабатывают в среднем

больше выпускников другого университета, может послужить основанием для принятия решения о поступлении в высшее учебное заведение и т. д.

Математическое ожидание $M(X)$ определяется следующим образом.

Для дискретной СВ:

$$M(X) = \sum_{i=1}^{k} x_i p_i,$$

где k — число всех возможных значений СВ X.

Для непрерывной СВ:

$$M(X) = \int_{-\infty}^{+\infty} xf(x)\,dx.$$

Свойства математического ожидания:

1. $M(C) = C$, где C = const.
2. $M(CX) = CM(X)$.
3. $M(X \pm Y) = M(X) \pm M(Y)$.
4. $M(aX + b) = aM(X) + b$.
5. Для независимых СВ $M(XY) = M(X)M(Y)$.

Однако для детального анализа поведения СВ знания лишь среднего значения явно недостаточно. Существуют отличные друг от друга случайные величины, имеющие одинаковые математические ожидания. Например, средний уровень жизни в Швеции и США приблизительно одинаков, однако разброс в доходах в этих странах существенно отличается. Акции двух компаний могут приносить в среднем одинаковые дивиденды, однако вложение денег в одну из них может быть гораздо более рискованной операцией, чем в другую. Следовательно, нужна числовая характеристика, которая будет оценивать разброс возможных значений СВ относительно ее среднего значения (математического ожидания). Такой характеристикой является дисперсия.

Дисперсией $D(X)$ *СВ* X *называется математическое ожидание квадрата отклонения СВ от ее математического ожидания. Она рассчитывается по формуле:*

$$D(X) = M\left(X - M(X)\right)^2 = M\left(X^2\right) - M^2(X).$$

При этом для дискретных СВ

$$D(X) = \sum_{i=1}^{k} \left(x_i - M(X)\right)^2 p_i = \sum_{i=1}^{k} x_i^2 \cdot p_i - M^2(X).$$

Для непрерывных СВ

$$D(X) = \int_{-\infty}^{+\infty} \left(x - M(X)\right)^2 \cdot f(x)\,dx = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - M^2(X).$$

Свойства дисперсии:

1. $D(C) = 0$, где $C = \text{const}$.
2. $D(CX) = C^2 D(X)$.
3. $D(X + Y) = D(X) + D(Y)$, где X и Y — независимые СВ.
4. $D(aX + b) = a^2 D(X)$.

Дисперсия имеет размерность, равную квадрату размерности СВ. Для того чтобы представить разброс значений СВ в тех же единицах, что и саму СВ, вводится другая числовая характеристика — среднее квадратическое отклонение.

Средним квадратическим отклонением $\sigma(X)$ *СВ* X *называют квадратный корень из дисперсии* $D(X)$*:* $\sigma(X) = \sqrt{D}$.

Для описания связи между СВ X и Y применяют *ковариацию* СВ X и Y:

$$\operatorname{cov}(X, Y) = M\left[\left(X - M(X)\right)\left(Y - M(Y)\right)\right] = M(XY) - M(X)M(Y).$$

2.2 Закон распределения

Большинство СВ подчиняется определенному закону распределения, на основании знания которого можно предвидеть вероятности попадания исследуемой СВ в определенные интервалы. Такое предсказание весьма желательно при анализе экономических показателей, ведь в этом случае появляется возможность осуществлять продуманную политику с учетом возможности возникновения той или иной ситуации. Законов распределений достаточно много. Рассмотрим те, которые наиболее активно используются в эконометрическом анализе.

Нормальное распределение (распределение Гаусса). Нормальное распределение является предельным случаем почти всех реальных распределений вероятности. Поэтому оно используется в очень большом числе реальных приложений теории вероятностей. Говорят, что СВ X имеет *нормальное распределение*, если ее плотность вероятности имеет вид (рис. 2.2)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Нормальное распределение зависит от параметров m и σ и полностью определяется ими. При этом $m = M(X)$, $\sigma = \sigma(X)$, т. е. $D(X) = \sigma^2$, $\pi = 3.14159\ldots$, $e = 2.71828\ldots$ Если СВ X имеет нормальное распределение с параметрами $M(X) = m$ и $\sigma(X) = \sigma$, то символически это можно записать так: $X \sim N(m, \sigma^2)$. Важным частным случаем нормального распределения является ситуация, когда $m = 0$ и $\sigma = 1$. В этом случае говорят о *стандартизированном (стандартном) нормальном распределении* (см. Приложение А).

В дальнейшем стандартизированную нормальную СВ будем обозначать через $U \sim N(0, 1)$. Для практических расчетов специально разработаны таблицы функций $f(u)$, $F(u)$ стандартизированного нормального распределения.

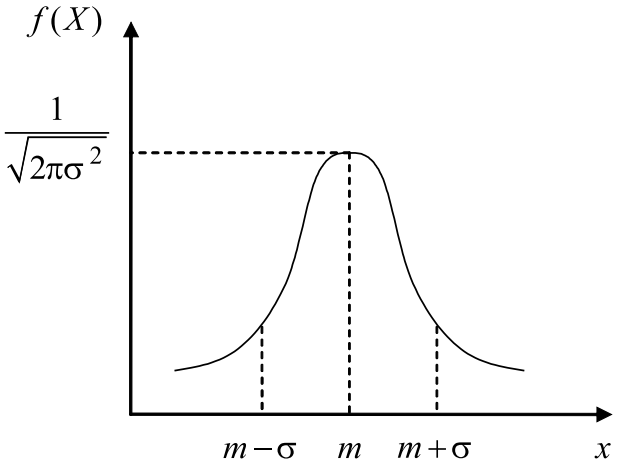


Рис. 2.2 – Функция плотности распределения

Распределение χ^2 (хи-квадрат). Пусть X_i, $i = 1,\ldots,n$ — независимые нормально распределенные СВ с математическими ожиданиями m_i и средними квадратическими отклонениями σ_i соответственно, т. е. $X_i \sim N(m_i, \sigma_i)$.

Тогда СВ $U_i = (X_i - m_i)/\sigma_i$, $i = 1,\ldots,n$ являются независимыми СВ, имеющими стандартизированное нормальное распределение, $U_i \sim N(0,1)$.

СВ z имеет хи-квадрат распределение с n степенями свободы, если

$$z = \sum_{i=1}^{n} U_i^2 = U_1^2 + U_2^2 + \ldots + U_n^2. \qquad (2.1)$$

Отметим, что *число степеней свободы* исследуемой СВ определяется числом случайных величин, ее составляющих, уменьшенным на число линейных связей между ними. Например, число степеней свободы исследуемой СВ, являющейся композицией n случайных величин, которые в свою очередь связаны m линейными уравнениями, определяется числом $\nu = m - n$.

Из определения (2.1) следует, что распределение χ^2 определяется одним параметром — числом степеней свободы (см. Приложение Б).

График плотности вероятности СВ, имеющей χ^2-распределение, лежит только в первой четверти декартовой системы координат и имеет асимметричный вид с вытянутым правым «хвостом». Однако с увеличением числа степеней свободы распределение χ^2 постепенно приближается к нормальному (сравните графики на рис. 2.3).

Среднее значение и дисперсия: $M(\chi^2) = \nu = n - m$, $D(\chi^2) = 2\nu = 2(n - m)$.

Если X и Y — две независимые χ^2-распределенные СВ с числами степеней свободы n и k, то их сумма $(X + Y)$ также является χ^2-распределенной СВ с числом степеней свободы $\nu = n + k$.

Распределение χ^2 применяется для нахождения интервальных оценок, а также при проверке статистических гипотез. При этом активно используется таблица критических точек χ^2-распределения.

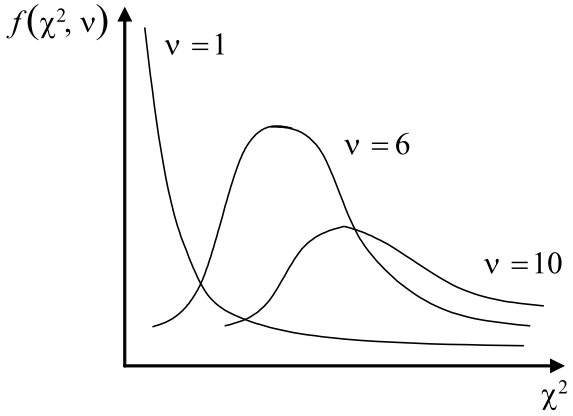


Рис. 2.3 – Плотность распределения

Распределение Стьюдента. Пусть СВ $U \sim N(0,1)$, СВ V — независимая от U величина, распределенная по закону χ^2 с n степенями свободы. Тогда величина

$$T = \frac{U}{\sqrt{\frac{V}{n}}} \tag{2.2}$$

имеет распределение Стьюдента (t-распределение) с n степенями свободы.

Из формулы (2.2) очевидно, что распределение Стьюдента определяется только одним параметром n — числом степеней свободы. График функции плотности вероятности СВ, имеющей распределение Стьюдента, является симметричной кривой (линия симметрии — ось ординат) (рис. 2.4).

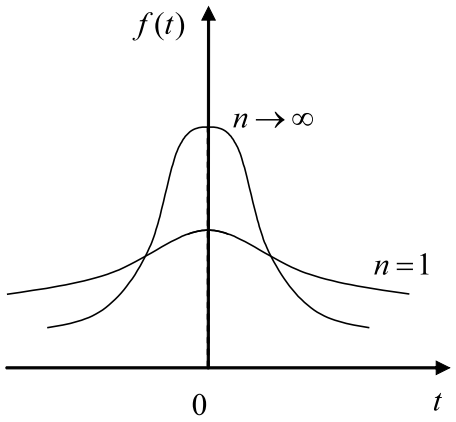


Рис. 2.4 – Плотность распределения Стьюдента

Среднее значение и дисперсия: $M(T) = 0$, $D(T) = n/(n-2)$.

При этом с увеличением числа степеней свободы распределение Стьюдента приближается к стандартизированному нормальному, причем при $n > 30$ распределение Стьюдента практически можно заменить нормальным распределением.

Распределение Стьюдента применяется для нахождения интервальных оценок, а также при проверке статистических гипотез. При этом активно используется таблица критических точек распределения Стьюдента (см. Приложение В).

Распределение Фишера. Пусть V и W — независимые СВ, распределение по закону χ^2 со степенями свободы $\nu_1 = m$ и $\nu_2 = n$ соответственно. Тогда величина

$$F = \frac{V}{m} \Big/ \frac{W}{n}$$

имеет распределение Фишера со степенями свободы $\nu_1 = m$ и $\nu_2 = n$. Таким образом, распределение Фишера F определяется двумя параметрами — числами степеней свободы m и n (рис. 2.5) (см. Приложение Г).

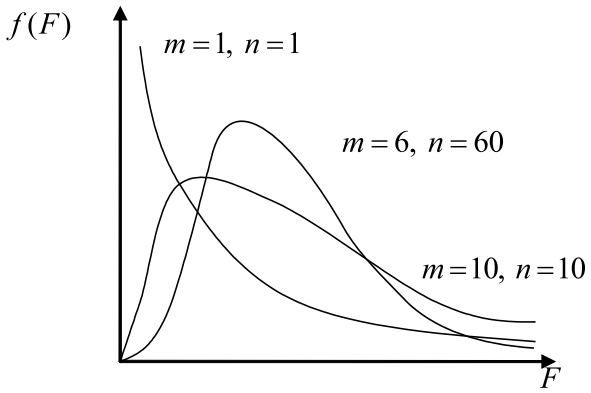


Рис. 2.5 – Плотность распределения Фишера

При больших m и n это распределение приближается к нормальному. Нетрудно заметить, что $T_n^2 = F_{1,n}$, где T_n — СВ, имеющая распределение Стьюдента с числом степеней свободы $\nu = n$; $F_{1,n}$ — СВ, имеющая распределение Фишера с числами степеней свободы $\nu_1 = 1$ и $\nu_2 = n$.

Распределение Фишера используется при проверке статистических гипотез, в дисперсионном и регрессионном анализах. При этом активно используется таблица критических точек распределения Стьюдента.

Для практического применения приведенных выше СВ к осуществлению статистических расчетов служат таблицы распределений. Перед их рассмотрением введем понятие квантиля (критической точки) распределения.

Пусть Y — СВ, имеющая одно из вышеперечисленных распределений. α-квантилем (критической точкой уровня α) называется значение y_α СВ Y, такое, что $P(Y > y_\alpha) = \int_{y_\alpha}^{+\infty} f(y)\,dy = \alpha$.

Квантили y_α и $y_{1-\alpha}$ называются симметричными. Если распределение симметрично относительно оси ординат, то $y_{1-\alpha} = -y_\alpha$.

С геометрической точки зрения нахождение квантиля y_α заключается в таком выборе значения $Y = y_\alpha$, при котором площадь заштрихованной криволинейной трапеции была бы равна α (рис. 2.6).

Нетрудно заметить, что нахождение α-квантиля (критической точки) для вышеперечисленных законов распределений определяется величиной (уровнем значимости) самого α и числом (числами) степеней свободы рассматриваемого закона распределения.

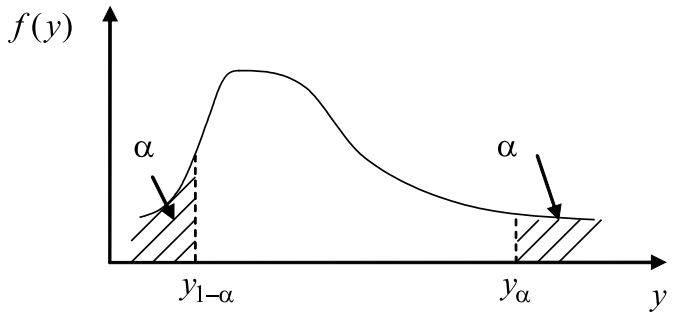


Рис. 2.6 – Представление α-квантилей

2.3 Генеральная совокупность и выборка

Пусть изучается совокупность однородных объектов относительно некоторого количественного признака, характеризующего эти объекты. Например, доход населения, количество покупателей в магазине в течение дня, количество качественных товаров в исследуемой партии и т. д.

Генеральной совокупностью *называется множество всех возможных значений или реализаций исследуемой случайной величины X при данном реальном комплексе условий.*

Выборкой *(выборочной совокупностью) называют часть генеральной совокупности, отобранную для изучения.*

Число элементов рассматриваемой совокупности называется ее *объемом*.

Изучение всей генеральной совокупности во многих случаях либо невозможно, либо нецелесообразно в силу больших материальных затрат или в силу уничтожения или порчи исследуемых объектов. Например, анализ среднего дохода населения г. Минска формально предполагает наличие достоверной информации о каждом жителе города в конкретный момент времени. Получение такой информации просто невозможно. Проверка качества обуви связана с воздействием на нее различных экстремальных факторов: растяжения, сжатия, влажности, температуры, солнечных лучей, химического воздействия, что приведет к потере товарного вида исследуемой обуви. Поэтому на практике вся генеральная совокупность почти никогда не анализируется. Для осуществления выводов о генеральной совокупности в большинстве случаев используется выборка ограниченного объема. В силу этого задача математической статистики состоит в исследовании свойств выборки и обобщении этих свойств на генеральную совокупность. Полученный при этом вывод называется *статистическим*.

Информация о генеральной совокупности, полученная на основании выборочного наблюдения, практически всегда будет обладать некоторой погрешностью, так как она основывается на изучении только части элементов. Вряд ли средний доход и разброс в доходах, полученных по выборке объема $n = 1000$, будет в точности

таким же, что и во всем городе. Это определяет две проблемы, составляющие содержание математической теории выборки:

- как организовать выборочное наблюдение, чтобы полученная информация достаточно полно отражала пропорции генеральной совокупности (*проблема репрезентативности выборки*);
- как использовать результаты выборки для суждения по ним с наибольшей надежностью о свойствах и параметрах генеральной совокупности (*проблема оценки*).

В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если отбор будет носить случайный характер.

Различают ***повторную*** *и* ***бесповторную выборки****. В первом случае отобранный объект перед отбором следующего возвращается в генеральную совокупность. Во втором — отобранный в выборку объект не возвращается в генеральную совокупность. Если выборка составляет незначительную часть генеральной совокупности, то различие между повторной и бесповторной выборками стирается.*

Случайный отбор может проводиться с помощью датчика таблицы случайных чисел либо обычной жеребьевкой. Однако строгое соблюдение правил случайного отбора не всегда осуществимо, так как оно требует четко ограниченной базы статистического анализа, каковой является генеральная совокупность, перенумеровки всех ее элементов или непосредственного их извлечения при жеребьевке. Так, при проведении обследований дохода населения в масштабах города практически невозможно составить список всех его жителей или семей с последующей организацией выборки с помощью датчика случайных чисел. Аналогично невозможно организовать опросы по изучению покупательного спроса, потребностей населения и т. д. путем образования строго случайной выборки. Поэтому прибегают к различным приемам *неслучайного* отбора, стремясь, однако, приблизиться к условиям случайного. К этим приемам относится *механический* отбор, при котором элементы генеральной совокупности, предварительно упорядоченные, отбираются по заранее установленному правилу, не связанному с вариацией исследуемого признака. Например, можно фиксировать доход каждого сотого, входящего в метро. *Серийным* называют отбор, при котором объекты выбираются из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Например, о продукции предприятия можно судить по продукции, выпущенной в какой-то конкретный день. При *типическом* отборе объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части. Например, население города можно предварительно классифицировать по социальному статусу (бизнесмены, чиновники, служащие, рабочие и т. д.). Нередко на практике применяется комбинированный отбор, при котором сочетаются описанные выше способы.

2.4 Вычисление выборочных характеристик

Для любой СВ X кроме определения ее функции распределения желательно указать ее числовые характеристики, важнейшими из которых являются математическое ожидание, дисперсия, среднее квадратическое отклонение. Пусть объем генеральной совокупности равен N. Тогда математическим ожиданием СВ X является *генеральное среднее*:

$$\overline{x}_{\text{г}} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Дисперсией СВ X является *генеральная дисперсия*:

$$D_{\text{г}} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Корень квадратный из генеральной дисперсии называется *генеральным средним квадратическим отклонением*:

$$\sigma_{\text{г}} = \sqrt{D_{\text{г}}}.$$

Таким образом, для нахождения генеральных числовых характеристик необходим анализ всей генеральной совокупности. В силу того, что в реальности практически всегда имеют дело с выборками, приходится находить оценки указанных выше генеральных характеристик — выборочные числовые характеристики: выборочное среднее, выборочную дисперсию, выборочное среднее квадратическое отклонение.

Выборочное среднее — это среднее арифметическое наблюдаемых значений выборки:

$$\overline{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Дисперсия:

$$D_{\text{в}} = \overline{x^2} - \overline{x}^2.$$

2.5 Точечные и интервальные оценки

Пусть оценивается некоторый параметр θ наблюдаемой СВ X генеральной совокупности. Пусть из генеральной совокупности извлечена выборка объема n: $x_1, x_2, \ldots, x_n$, по которой может быть найдена оценка $\hat{\theta}$ параметра θ. Например, для нормального закона распределения с плотностью параметрами являются математическое ожидание m и среднее квадратическое отклонение σ.

Точечной оценкой $\hat{\theta}$ *параметра* θ *называется числовое значение этого параметра, полученное по выборке объема* n.

Например, оценками m и σ могут быть $\hat{m} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ и $\hat{\sigma} = \sigma_{\text{в}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2}$ соответственно.

После получения точечной оценки $\hat{\theta}$ желательно иметь данные о надежности такой оценки. Особенно важно иметь сведения о точности оценок для небольших выборок. Поэтому точечная оценка может быть дополнена *интервальной оценкой* — интервалом (θ_1, θ_2), внутри которого с наперед заданной вероятностью γ находится точное значение оцениваемого параметра θ. Задачу определения такого интервала называют *интервальным оцениванием*, а сам интервал — *доверительным интервалом*. При этом γ называют *доверительной вероятностью*, или *надежностью*, с которой оцениваемый параметр θ попадает в интервал (θ_1, θ_2).

Зачастую для определения доверительного интервала заранее выбирают число $\alpha = 1 - \gamma$, $0 < \alpha < 1$, называемое *уровнем значимости*, и находят два числа θ_1, θ_2, зависящих от точечной оценки $\hat{\theta}$ такие, что

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha = \gamma.$$

В этом случае говорят, что интервал (θ_1, θ_2) накрывает неизвестный параметр θ с вероятностью $(1-\alpha)$ или в $100(1-\alpha)\%$ случаев. Границы интервала θ_1 и θ_2 называются доверительными, и они обычно находятся из условия $P(\theta < \theta_1) = P(\theta > \theta_2) = \alpha/2$ (рис. 2.7).

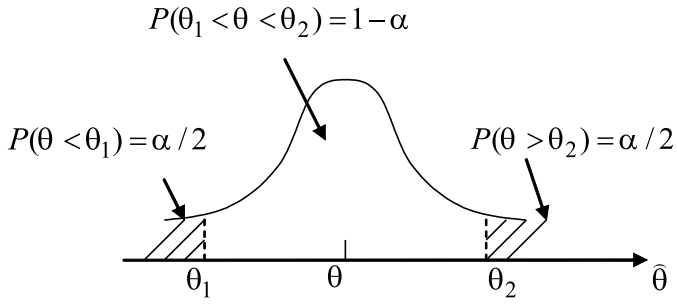


Рис. 2.7 – Интервальная оценка

Длина доверительного интервала, характеризующая точность интервальной оценки, зависит от объема выборки n и надежности γ (уровня значимости $\alpha = 1-\gamma$). При увеличении величины n длина доверительного интервала уменьшается, а с приближением надежности γ к единице — увеличивается. Выбор α (или $\gamma = 1 - \alpha$) определяется конкретными условиями. Обычно используется $\alpha = 0.1;\ 0.05;\ 0.01$, что соответствует 90, 95, 99%-ным доверительным интервалам.

2.6 Статистическая проверка гипотез

Большинство эконометрических моделей требуют многократного улучшения и уточнения. Для этого требуется проведение соответствующих расчетов, связанных с установлением выполнимости или невыполнимости тех или иных предпосылок, анализом качества найденных оценок, достоверностью полученных выводов. Обычно эти расчеты проводятся по схеме статистической проверки гипотез. Поэтому знание основных принципов проверки гипотез является обязательным для эконометрики.

Во многих случаях необходимо знать закон распределения генеральной совокупности. Если закон распределения неизвестен, но есть основания предположить, что он имеет определенный вид (назовем его A), выдвигают гипотезу: генеральная совокупность (СВ X) распределена по закону A. Например, можно выдвинуть предположение, что доход населения, ежедневное количество покупателей в магазине, размер выпускаемых деталей имеют нормальный закон распределения.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр θ равен ожидаемому числу θ_0, выдвигают гипотезу: $\theta = \theta_0$. Например, можно выдвинуть предположение о величине среднего дохода населения, среднего ожидаемого дохода по акциям, о разбросе в доходах и т. д.

Статистической *называют гипотезу о виде закона распределения или о параметрах известного распределения. В первом случае гипотеза называется* ***непараметрической****, а во втором —* ***параметрической****.*

Гипотеза H_0, подлежащая проверке, называется *нулевой (основной)*. Наряду с нулевой рассматривают гипотезу H_1, которая будет приниматься, если отклоняется H_0. Такая гипотеза называется *альтернативной* (*конкурирующей*). Например, если проверяется гипотеза о равенстве параметра θ некоторому значению θ_0, т. е. H_0: $\theta = \theta_0$, то в качестве альтернативной могут рассматриваться следующие гипотезы:

$$H_1^{(1)}: \theta \neq \theta_0; \quad H_1^{(2)}: \theta > \theta_0; \quad H_1^{(3)}: \theta < \theta_0; \quad H_1^{(4)}: \theta = \theta_1 \ (\theta_1 \neq \theta_0).$$

Выбор альтернативной гипотезы определяется конкретной формулировкой задачи, а нулевая гипотеза часто специально подбирается так, чтобы отвергнуть ее и принять тем самым альтернативную гипотезу. Для того чтобы принять гипотезу о наличии корреляции между двумя экономическими показателями (например, между инфляцией и безработицей), можно опровергнуть гипотезу об отсутствии такой корреляции, взяв ее в качестве нулевой гипотезы.

Гипотезу называют ***простой****, если она содержит одно конкретное предположение* $\left(H_1^{(1)}: \theta = \theta_0;\ H_1^{(4)}: \theta = \theta_1\right)$.

Гипотезу называют ***сложной****, если она состоит из конечного или бесконечного числа простых гипотез* $\left(H_1^{(1)}: \theta \neq \theta_0;\ H_1^{(2)}: \theta > \theta_0;\ H_1^{(3)}: \theta < \theta_0\right)$.

Сущность проверки статистической гипотезы заключается в том, чтобы установить, согласуются или нет данные наблюдений и выдвинутая гипотеза. Можно ли расхождение между гипотезой и результатом выборочных наблюдений отнести за счет случайной погрешности, обусловленной механизмом случайного отбора? Эта

задача решается с помощью специальных методов математической статистики — методов статистической проверки гипотез.

При проверке гипотезы выборочные данные могут противоречить гипотезе H_0. Тогда она *отклоняется*. Если же статистические данные согласуются с выдвинутой гипотезой, то она *не отклоняется*. На практике часто в таких случаях говорят, что нулевая гипотеза принимается (такая формулировка не совсем точна, однако она широко распространена). Статистическая проверка гипотез на основании выборочных данных неизбежно связана с риском принятия ложного решения. При этом возможны ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная нулевая гипотеза.

Ошибка второго рода состоит в том, что будет принята нулевая гипотеза, в то время как в действительности верна альтернативная гипотеза.

Возможные результаты статистических выводов представлены следующей таблицей 2.2.

Таблица 2.2 – Возможные результаты статистических выводов

Результаты проверки гипотезы	**Возможные состояния гипотезы**	
	Верна H_0	**Верна** H_1
Гипотеза H_0 отклоняется	Ошибка первого рода	Правильный вывод
Гипотеза H_0 не отклоняется	Правильный вывод	Ошибка второго рода

В большинстве случаев последствия указанных ошибок неравнозначны. Первая приводит к более осторожному, консервативному решению, вторая — к неоправданному риску. Что лучше или хуже — зависит от конкретной постановки задачи и содержания нулевой гипотезы. Например, если H_0 состоит в признании продукции предприятия качественной и допущена ошибка первого рода, то будет забракована годная продукция. Допустив ошибку второго рода, мы отправим потребителю брак. Очевидно, последствия второй ошибки более серьезны с точки зрения имиджа фирмы и ее долгосрочных перспектив.

Исключить ошибки первого и второго рода невозможно в силу ограниченности выборки. Поэтому стремятся минимизировать потери от этих ошибок. Отметим, что одновременное уменьшение вероятностей данных ошибок невозможно, так как задачи их уменьшения являются конкурирующими и уменьшение вероятности допустить одну из них влечет за собой увеличение вероятности допустить другую. В большинстве случаев единственный способ уменьшения вероятности ошибок состоит в увеличении объема выборки.

Вероятность совершить ошибку первого рода принято обозначать буквой α, и ее называют *уровнем значимости*. Вероятность совершить ошибку второго рода обозначают β. Тогда вероятность несовершения ошибки второго рода $(1 - \beta)$ называется *мощностью критерия*.

Обычно значения α задают заранее круглыми числами (например, 0.1; 0.05; 0.01 и т. п.), а затем стремятся построить критерий наибольшей мощности. Таким образом, если $\alpha = 0.05$, то это означает, что исследователь не хочет совершить ошибку первого рода более чем в 5 случаях из 100.

Проверку статистической гипотезы осуществляют на основании данных выборки. Для этого используют специально подобранную СВ (статистику, критерий), точное или приближенное значение которой известно. Обозначим такую СВ через K.

Таким образом, *статистическим критерием* называют СВ K, которая служит для проверки нулевой гипотезы. После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отклоняется, другое — при которых она не отклоняется. Совокупность значений критерия, при которых нулевую гипотезу отклоняют, называют *критической областью*. Совокупность значений критерия, при которых нулевую гипотезу не отклоняют, называют *областью принятия гипотезы*.

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия K (вычисленное по выборке) принадлежит критической области, то нулевую гипотезу отклоняют. Если же наблюдаемое значение критерия K принадлежит области принятия гипотезы, то нулевую гипотезу не отклоняют (принимают).

Точки, разделяющие критическую область и область принятия гипотезы, называют *критическими*.

Перейдем к определению критических точек, а следовательно, и критической области. В основу этого определения положен принцип практической невозможности маловероятных событий.

Пусть для проверки нулевой гипотезы H_0 служит критерий K. Предположим, что плотность распределения вероятности СВ K в случае справедливости H_0 имеет вид $f\left(k|H_0\right)$, а математическое ожидание K равно k_0. Тогда вероятность того, что СВ K попадет в произвольный интервал $\left(k_{1-\alpha/2}, k_{\alpha/2}\right)$, можно найти по формуле:

$$P\left(k_{1-\alpha/2} < K < k_{\alpha/2}\right) = \int\limits_{k_{1-\alpha/2}}^{k_{\alpha/2}} f\left(k|H_0\right)\, dk.$$

Зададим эту вероятность равной $(1-\alpha)$ и вычислим критические точки (квантили) K-распределения $k_{1-\alpha/2}, k_{\alpha/2}$ из условий:

$$P\left(K \leqslant k_{1-\alpha/2}\right) = \int\limits_{-\infty}^{k_{1-\alpha/2}} f\left(k|H_0\right)\, dk = \frac{\alpha}{2},$$

$$P\left(K \geqslant k_{\alpha/2}\right) = \int\limits_{k_{\alpha/2}}^{+\infty} f\left(k|H_0\right)\, dk = \frac{\alpha}{2}.$$

Следовательно, $P\left(k_{1-\alpha/2} < K < k_{\alpha/2}\right) = 1-\alpha$, а $P\left(\left(K \leqslant k_{1-\alpha/2}\right) \cup \left(K \geqslant k_{\alpha/2}\right)\right) = \alpha$.

Зададим вероятность α настолько малой (0.05; 0.01), чтобы попадание СВ K за пределы интервала $\left(k_{1-\alpha/2}, k_{\alpha/2}\right)$ можно было бы считать маловероятным событием. Тогда исходя из принципа практической невозможности маловероятных событий можно считать, что если H_0 справедлива, то при ее проверке с помощью критерия K по данным одной выборки наблюдаемое значение K должно наверняка попасть

в интервал $(k_{1-\alpha/2}, k_{\alpha/2})$. Если же наблюдаемое значение K попадает за пределы указанного интервала, то произойдет маловероятное, практически невозможное событие. Это дает основание считать, что с вероятностью $(1-\alpha)$ нулевая гипотеза H_0 несправедлива (рис. 2.8).

Точки $k_{1-\alpha/2}, k_{\alpha/2}$ являются *критическими*.

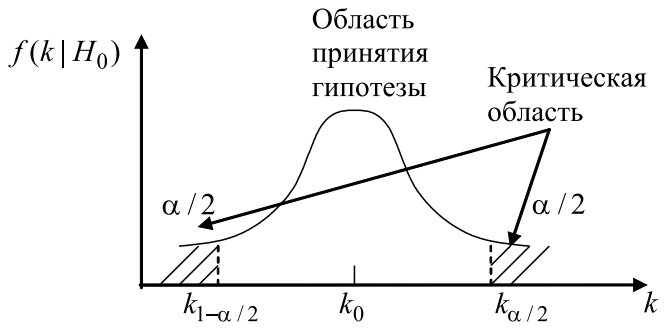


Рис. 2.8 – Графическое представление проверки гипотезы

Контрольные вопросы по главе 2

1. Что понимается под событием?
2. Что понимается под вероятностью события?
3. Каковы свойства вероятности?
4. Что понимается под случайной величиной?
5. Какую случайную величину называют дискретной?
6. Какую случайную величину называют непрерывной?
7. Что понимается под законом распределения случайной величины?
8. Как с помощью таблицы можно задать закон распределения дискретной случайной величины?
9. Как можно интерпретировать плотность вероятности с помощью графика?
10. Что понимается под математическим ожиданием?
11. Как вычисляется математическое ожидание дискретной и непрерывной случайной величины?
12. Перечислите свойства математического ожидания.
13. Что понимается под дисперсией?
14. Как вычисляется дисперсия дискретной и непрерывной случайной величины?
15. Перечислите свойства дисперсии.
16. Какова связь между дисперсией и средним квадратическим отклонением?
17. Записать функцию плотности нормального распределения.

18. Какие параметры определяют нормальное распределение?
19. В каком случае нормальное распределение считается стандартным?
20. Записать функцию плотности распределения хи-квадрат.
21. Какие параметры определяют распределение хи-квадрат?
22. Записать функцию плотности распределения Стьюдента.
23. Какие параметры определяют распределение Стьюдента?
24. Записать функцию плотности распределения Фишера.
25. Какие параметры определяют распределение Фишера?
26. Что понимается под квантилем распределения?
27. Что понимается под генеральной совокупностью?
28. Что понимается под выборкой?
29. В чем отличие повторной выборки от бесповторной?
30. Что понимается под серийным, типическим, механическим отбором?
31. Как вычисляются генеральное среднее и генеральная дисперсия?
32. Как вычисляются выборочное среднее и выборочная дисперсия?
33. Что понимается под точечной оценкой?
34. Какая гипотеза называется нулевой?
35. В чем отличие простой гипотезы от сложной?
36. Что представляют собой ошибки первого и второго рода?

Глава 3

МЕТОДЫ И МОДЕЛИ РЕГРЕССИОННОГО АНАЛИЗА

3.1 Введение в регрессионный анализ

Результирующая (зависимая, эндогенная) переменная y. Переменная, характеризующая результат или эффективность функционирования анализируемой экономической системы. Ее значения формируются в процессе и внутри функционирования этой системы под воздействием ряда других переменных и факторов, часть из которых поддается регистрации и, в определенной степени, управлению и планированию (эту часть принято называть объясняющими переменными). В регрессионном анализе результирующая переменная выступает в роли функции, значения которой определяются (с некоторой случайной погрешностью) значениями упомянутых выше объясняющих переменных, выступающих в роли аргументов. Поэтому по *природе своей результирующая переменная y всегда стохастична*.

Объясняющие (предикторные, экзогенные) переменные $X=\left(x^{(1)}\ x^{(2)}\ \ldots\ x^{(p)}\right)^T$. Переменные (или признаки), поддающиеся регистрации, описывающие условия функционирования изучаемой реальной экономической системы и в существенной мере определяющие процесс формирования значений результирующих переменных. Как правило, часть из них поддается хотя бы частичному регулированию и управлению. Значения ряда объясняющих переменных могут задаваться как бы «извне» анализируемой системы. В этом случае их принято называть экзогенными. В регрессионном анализе они играют роль аргументов той функции, в качестве которой рассматривается анализируемый результирующий показатель y. По своей природе объясняющие переменные могут быть как случайными, так и неслучайными.

Функция регрессии *y* по *X* = $\left(x^{(1)} \quad x^{(2)} \quad \ldots \quad x^{(p)}\right)^T$.

Функция $\boldsymbol{f}\left(\boldsymbol{X}^*\right)$ *называется* ***функцией регрессии y по X****, если она описывает изменение условного среднего значения результирующей переменной* ***y*** *(при условии, что значения объясняющих переменных* ***X*** *зафиксированы на уровнях* $\boldsymbol{X}^*$*) в зависимости от изменения значений* $\boldsymbol{X}^*$ *объясняющих переменных. Соответственно математически это определение может быть записано в виде*

$$f\left(X^*\right) = M\left(y|X = X^*\right).$$

Далее в целях упрощения обозначений правую часть мы будем записывать просто $M\left(y|X^*\right)$ или $M(y|X)$. Поэтому сокращенно функция регрессии может быть определена также соотношением

$$f(X) = M(y|X).$$

Объясняющие переменные X могут быть как случайными величинами, так и неслучайными параметрами, от значений которых зависит закон распределения вероятностей случайной результирующей переменной y.

Уравнения регрессионной связи между *y* и *X*. Уже отмечено, что в регрессионном анализе результирующая переменная y выступает в роли функции, значения которой определяются (с некоторой случайной погрешностью) значениями объясняющих переменных $X = \left(x^{(1)} \quad x^{(2)} \quad \ldots \quad x^{(p)}\right)^T$, выступающих в роли аргументов этой функции. Математически это может быть выражено в виде уравнений регрессионной связи

$$\begin{cases} y(X) = f(X) + \varepsilon(X), \\ M\varepsilon(X) = 0. \end{cases} \tag{3.1}$$

Если в модели присутствует только один регрессор, то модель называется моделью парной регрессии. Если в модели присутствует два или более регрессора, то она называется моделью множественной регрессии.

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

В качестве примера рассмотрим парную линейную модель вида

$$y_i = \Theta_0 + \Theta_1 x_i + \varepsilon_i.$$

Величина y_i, значение зависимой переменной в наблюдении i, состоит из двух составляющих: 1) неслучайной составляющей $\Theta_0 + \Theta_1 x_i$, где Θ_0 и Θ_1 — это постоянные величины, называемые параметрами уравнения, а x — значение объясняющей переменной в наблюдении i, и 2) случайного члена ε_i.

На рисунке 3.1 показано, как комбинация этих двух составляющих определяет Y. Показатели X_1, X_2, X_3, X_4 — это четыре гипотетических значения переменной. Если бы соотношение между Y и X было точным, то соответствующие значения Y были бы представлены точками $Q_1 - Q_4$ на прямой. Наличие случайного члена приводит к тому, что в действительности значение Y получается другим. Предполагалось, что случайный член положителен в первом и четвертом наблюдениях и отрицателен в двух других, поэтому если отметить на графике реальные значения Y при соответствующих значениях X, то мы получим точки $P_1 - P_4$.

Следует подчеркнуть, что точки P — это все, что вы можете видеть на рисунке 3.1 на практике. Фактические значения Θ_0 и Θ_1 и, следовательно, положения точек Q неизвестны, так же как и фактические значения случайного члена.

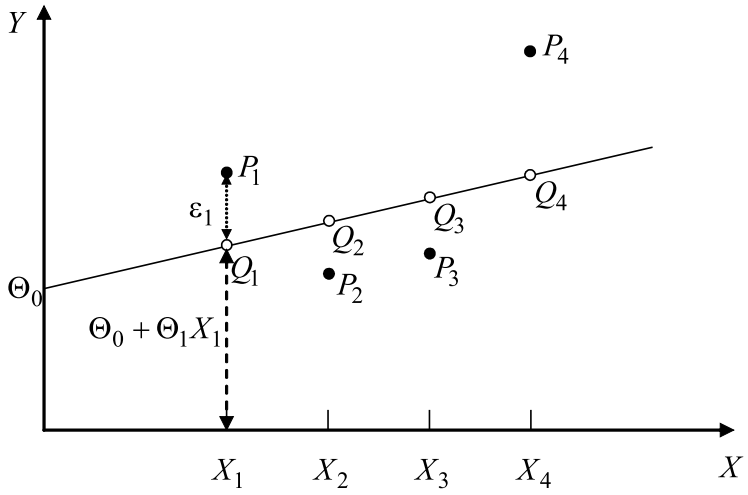


Рис. 3.1 – Истинная зависимость между Y и X

Задача регрессионного анализа состоит в получении оценок Θ_0 и Θ_1 и, следовательно, в определении положения прямой по точкам P.

Почему же существует случайный член? Имеется несколько причин.

1. Невключение объясняющих переменных. Соотношение между Y и X почти наверняка является очень большим упрощением. В действительности существуют другие факторы, влияющие на Y, которые не учитываются в уравнении, влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой. Часто происходит так, что имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем этого сделать потому, что не знаем, как их измерить. Например, необходимо оценить функцию заработка, связывающую часовые заработки с продолжительностью образования. Известно, что срок обучения является не единственным фактором, влияющим на заработки, и в конечном счете нужно будет усовершенствовать модель, включив в нее и другие переменные, такие как, например, трудовой стаж. Тем не менее даже наилучшим образом специфицированная функция заработка объясняет не более половины разброса уровня заработков. Многие другие факторы влияют на возможность получения хорошей работы, такие как, например, неизмеримые характеристики индивида или даже чистый фактор удачи в смысле нахождения

данным индивидом работы, наилучшим образом соответствующей его индивидуальным способностям. Все эти прочие факторы вносят свой вклад в случайный член.

2. Агрегирование переменных. Во многих случаях рассматриваемая зависимость — это попытка объединить вместе некоторое количество микроэкономических соотношений. Например, функция суммарного потребления — это попытка суммарного выражения совокупности решений отдельных индивидуумов о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между суммарными расходами и совокупным доходом является лишь аппроксимацией. Наблюдаемое расхождение при этом приписывается наличию случайного члена.
3. Неправильное описание структуры модели. Структура модели может быть описана неправильно или не вполне правильно. Здесь можно привести один из многих возможных примеров. Если зависимость относится к данным о временном ряде, то значение Y может зависеть не от фактического значения X, а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между Y и X существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайного члена.
4. Неправильная функциональная спецификация. Функциональное соотношение между Y и X математически может быть определено неправильно. Например, истинная зависимость может не являться линейной, а быть более сложной. Нелинейные зависимости будут рассмотрены в следующих главах. Безусловно, надо постараться избежать возникновения этой проблемы, используя подходящую математическую формулу, но любая самая изощренная формула является лишь приближением, и существующее расхождение вносит вклад в остаточный член.
5. Ошибки измерения. Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не будут соответствовать точному соотношению, и существующее расхождение будет вносить вклад в остаточный член.

Случайный член является суммарным проявлением всех этих факторов. Очевидно, что если бы исследователя интересовало только измерение влияния X на Y, то было бы значительно удобнее, если бы случайного члена не было. Если бы он отсутствовал, то точки P на рисунке 3.1 совпадали бы с точками Q и любое изменение в Y от наблюдения к наблюдению было вызвано изменением X и можно было бы точно вычислить Θ_0 и Θ_1.

Однако в действительности каждое изменение Y отчасти вызвано изменением ε, и это значительно усложняет жизнь. По этой причине ε иногда описывается как «шум».

Второе соотношение в уравнении (3.1) непосредственно следует из смысла функции регрессии $f(X) = M(y|X)$, поскольку усреднение (вычисление математического ожидания) левых и правых частей первого из соотношений при любом фиксированном значении X дает

$$M\big(y(X)|X\big) = M\big(f(X)\big) + M\big(\varepsilon(X)\big).$$

А так как $M\big(y(X)|X\big) = f(X)$ по определению и $M\big(f(X)\big) = f(X)$ (поскольку величина $f(X)$ при фиксированных значениях X не является случайной), то $M\big(\varepsilon(X)\big) = 0$ при любом фиксированном значении X.

Исходные статистические данные. Все выводы в регрессионном анализе, так же как и в любом статистическом исследовании, строятся на основании имеющихся *исходных статистических данных*.

Далее будем полагать, что мы располагаем результатами регистрации значений анализируемых объясняющих $\big(X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})\big)$ и результирующей (y) переменных на n статистически обследованных объектах. Так что, если i — номер обследованного объекта, то имеющиеся исходные статистические данные состоят из n строк вида

$$\big(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; y_i\big), \quad i = 1, 2, \ldots, n, \tag{3.2}$$

где $x_i^{(j)}$ и y_i — значения соответственно j-й объясняющей переменной ($j = 1, 2, \ldots, p$) и результирующего показателя, зарегистрированные на i-м обследованном объекте.

Данные (3.2) в регрессионном анализе обычно представляют в виде двух матриц:

$$X = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_1^{(p)} \\ 1 & x_2^{(1)} & \ldots & x_2^{(p)} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_n^{(1)} & \ldots & x_n^{(p)} \end{pmatrix} — \tag{3.3}$$

матрица размера $n \times (p+1)$, составленная из наблюденных значений объясняющих переменных, и

$$Y = \begin{pmatrix} y_1 & y_2 & \ldots & y_n \end{pmatrix}^T — \tag{3.4}$$

матрица размера $n \times 1$ (вектор-столбец высоты n), составленная из наблюденных значений результирующей переменной.

Возможны ситуации, когда данные регистрируются на одном и том же объекте, но в разные периоды («такты») времени. Тогда i будет означать номер периода времени, к которому «привязаны» соответствующие данные, а n — общее число тактов времени, в течение которых собирались исходные данные (случай «временной» выборки в отличие от предыдущей — «пространственной» или «перекрестной»).

Возможна ситуация, когда «отслеживается» каждый из n объектов в течение N тактов времени («*пространственно-временная*» выборка, или «*панельные данные*»). В любой из упомянутых ситуаций исходные данные могут быть представлены в конечном счете в форме (3.3)–(3.4).

3.2 Основные задачи прикладного регрессионного анализа

Анализ регрессионных зависимостей вида (3.1), базирующийся на исходных статистических данных (3.3)–(3.4), нацелен на решение следующих основных задач:

Задача 1. Для любых заданных значений объясняющих переменных $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ построить наилучшие в определенном смысле точечные и интервальные (с доверительной вероятностью P) оценки $\hat{f}(X)$ и $\Delta[f(X)^T]_P$ для неизвестной функции регрессии $f(X)$.

Задача 2. По заданным значениям объясняющих переменных $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ построить наилучший в определенном смысле точечный и интервальный (с доверительной вероятностью P) прогноз соответственно $\hat{y}(X)$ для неизвестного значения результирующей переменной $y(X)$.

Задача 3. Пусть известно, что искомая функция регрессии принадлежит некоторому параметрическому семейству функций $\{\hat{f}(X;\Theta)\}$, где $\Theta = (\Theta_0\ \Theta_1\ \ldots\ \Theta_p)^T$ — векторный параметр, все или некоторые компоненты которого допускают определенную экономическую интерпретацию. Требуется построить наилучшие в определенном смысле точечные и интервальные оценки для неизвестных значений этих параметров.

Задача 4. Оценить удельный вес влияния каждой из объясняющих переменных $X = (x^{(1)}\ \ x^{(2)}\ \ \ldots\ \ x^{(p)})$ на результирующий показатель $y(X)$ и, в частности, определить, какие из объясняющих переменных можно исключить из модели как практически не влияющие на процесс формирования значений результирующего показателя.

Итак, регрессионный анализ начинается с решения задачи 1 и, в частности, с конструирования по исходным данным вида (3.3)–(3.4) оценки $\hat{f}(X)$ для неизвестной функции регрессии $f(X)$. Исходным этапом в решении этой задачи следует признать выбор параметрического семейства функций $F = \{f(X;\Theta)\}$ — класса допустимых решений, в рамках которого предполагается вести поиск наилучшей (в определенном смысле) аппроксимации $\hat{f}(X)$ для $f(X)$. При этом необходимо помнить, что не следует гнаться за чрезмерной сложностью функции, описывающей поведение искомой функции регрессии. При подборе общего вида функции регрессии идут от простого к сложному, т. е. начинают с анализа возможности использовать простейшую линейную модель вида

$$f(X;\Theta) = \Theta_0 + \sum_{k=1}^{p} \Theta_k \cdot x^{(k)}.$$

3.3 Классическая линейная модель множественной регрессии (КЛММР)

Классическая линейная модель множественной регрессии (КЛММР) представляет собой простейшую версию конкретизации требований к общему виду функции регрессии $f(X)$, природе объясняющих переменных X и статистических регрессионных остатков $\varepsilon(X)$ в общих уравнениях регрессионной связи (3.1). В рамках КЛММР эти требования формулируются следующим образом:

$$\begin{cases} y_i = \Theta_0 + \Theta_1 x_1^{(1)} + \ldots + \Theta_p x_i^{(p)} + \varepsilon_i, \ i = 1, 2, \ldots, n; \\ M\varepsilon_i = 0, \ i = 1, 2, \ldots, n; \\ M\left(\varepsilon_i \varepsilon_j\right) = \begin{cases} \sigma^2 & \text{при } i = j, \\ 0 & \text{при } i \neq j; \end{cases} \\ \left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right) \text{ — неслучайные переменные;} \\ \text{ранг}(X) = P + 1 < n; \\ \text{матрица } X \text{ определена соотношением (3.3).} \end{cases} \tag{3.5}$$

Из (3.5) следует, что в рамках КЛММР рассматриваются только *линейные* функции регрессии, т. е.

$$f(X) = M(y|X) = \Theta_0 + \Theta_1 x^{(1)} + \ldots + \Theta_p x^{(p)},$$

где объясняющие переменные $X = \left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right)$ играют роль неслучайных параметров, от которых зависит закон распределения вероятностей результирующей переменной y. Это означает, что в повторяющихся выборочных наблюдениях $\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; y_i\right)$ единственным источником случайных возмущений значений y_i являются случайные возмущения регрессионных остатков ε_i.

Кроме того, постулируется взаимная некоррелированность случайных регрессионных остатков $M\left(\varepsilon_i \varepsilon_j\right) = 0$ для $i \neq j$. В случае, когда речь идет о пространственных выборках (3.3)–(3.4), т. е. когда значения анализируемых переменных регистрируются на различных объектах (индивидуумах, семьях, предприятиях, банках, регионах и т. п.), данное предположение означает, что «возмущения» (регрессионные остатки), получающиеся при наблюдении одного какого-либо обследуемого объекта, не влияют на «возмущения», характеризующие наблюдения над другими объектами, и наоборот.

Тот факт, что для всех остатков $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ выполняется соотношение $M\varepsilon_i^2 = \sigma^2$, где величина σ^2 от номера наблюдения i не зависит, означает неизменность (постоянство, независимость от того, при каких значениях объясняющих переменных производятся наблюдения) дисперсий регрессионных остатков. Последнее свойство принято называть гомоскедастичностью регрессионных остатков.

Наконец, требуется, чтобы ранг матрицы X, составленной из наблюденных значений объясняющих переменных, был бы максимальным, т. е. равнялся бы числу столбцов этой матрицы, которое в свою очередь должно быть меньше числа ее строк (т. е. общего числа имеющихся наблюдений). Случаи $p + 1 \geqslant n$ не рассматриваются, поскольку при этом число n имеющихся в нашем распоряжении исходных статистических данных оказывается меньшим или равным числу оцениваемых параметров модели $(p + 1)$, что исключает принципиальную возможность получения сколько-нибудь надежных статистических выводов.

Требование к рангу матрицы X означает, что не должно существовать строгой линейной зависимости между объясняющими переменными. Так, если, например, одна объясняющая переменная может быть линейно выражена через какое-то количество других, то ранг матрицы X окажется меньше $p + 1$, а следовательно, и ранг матрицы $X^T X$ будет тоже меньше $p + 1$. А это означает вырождение симметрической матрицы $X^T X$ (т. е. $\det\left(X^T X\right) = 0$), что исключает существование матрицы

$\left(X^T X\right)^{-1}$, которая играет важную роль в процедуре оценивания параметров анализируемой модели.

Далее нам удобнее будет оперировать с матричной записью модели (3.5). При этом, кроме обозначений (3.3)–(3.4), введем также матрицы (векторы):

- $I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$ — единичная матрица размерности $n \times n$;
- $\Theta = \begin{pmatrix} \Theta_0 & \Theta_1 & \dots & \Theta_p \end{pmatrix}^T$ — вектор-столбец неизвестных значений параметров;
- $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{pmatrix}^T$ — вектор-столбец регрессионных остатков;
- $0_n = \begin{pmatrix} 0 & 0 & \dots & 0 \end{pmatrix}^T$ — вектор-столбец высоты n, состоящий из одних нулей;
- $\Sigma_\varepsilon = M\left(\varepsilon\varepsilon^T\right) = \begin{pmatrix} M\left(\varepsilon_1^2\right) & M\left(\varepsilon_1\varepsilon_2\right) & \dots & M\left(\varepsilon_1\varepsilon_n\right) \\ M\left(\varepsilon_2\varepsilon_1\right) & M\left(\varepsilon_2^2\right) & \dots & M\left(\varepsilon_2\varepsilon_n\right) \\ \dots & \dots & \dots & \dots \\ M\left(\varepsilon_n\varepsilon_1\right) & M\left(\varepsilon_n\varepsilon_2\right) & \dots & M\left(\varepsilon_n^2\right) \end{pmatrix}$ — ковариационная матрица размерности $n \times n$ вектора остатков;
- $\hat{\Theta} = \begin{pmatrix} \hat{\Theta}_0 & \hat{\Theta}_1 & \dots & \hat{\Theta}_p \end{pmatrix}^T$ — вектор-столбец оценок неизвестных значений параметров.

Тогда матричная форма записи КЛММР имеет вид:

$$\begin{cases} Y = X\Theta + \varepsilon, \\ M\varepsilon = 0_n, \\ \Sigma_\varepsilon = \sigma^2 \cdot I_n, \\ x^{(1)}, x^{(2)}, \dots, x^{(p)} \text{ — неслучайные переменные}, \\ \text{ранг}\left(X\right) = p + 1 < n. \end{cases} \tag{3.6}$$

Когда дополнительно к условиям (3.5) постулируют *нормальный* характер распределения регрессионных остатков $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{pmatrix}^T \left(\varepsilon \in N\left(0; \sigma^2 I_n\right)\right)$, говорят, что Y и X связаны нормальной КЛММР.

Пример 3.1

Исследуется зависимость урожайности зерновых культур (y ц/га) от ряда переменных, характеризующих различные факторы сельскохозяйственного производства, а именно:

- $x^{(1)}$ — число тракторов (приведенной мощности) на 100 га;
- $x^{(2)}$ — число зерноуборочных комбайнов на 100 га;
- $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га;
- $x^{(4)}$ — количество удобрений, расходуемых на гектар (т/га);
- $x^{(5)}$ — количество химических средств защиты растений, расходуемых на гектар (ц/га).

Исходные данные для двадцати сельскохозяйственных районов области приведены в таблице 3.1.

Таблица 3.1 – Исходные данные

***I* (номер района)**	y_i	$x_i^{(1)}$	$x_i^{(2)}$	$x_i^{(3)}$	$x_i^{(4)}$	$x_i^{(5)}$
1	9.70	1.59	0.26	2.05	0.32	0.14
2	8.40	0.34	0.28	0.46	0.59	0.66
3	9.00	2.53	0.31	2.46	0.30	0.31
4	9.90	4.63	0.40	6.44	0.43	0.59
5	9.60	2.16	0.26	2.16	0.39	0.16
6	8.60	2.16	0.30	2.69	0.32	0.17
7	12.50	0.68	0.29	0.73	0.42	0.23
8	7.60	0.35	0.26	0.42	0.21	0.08
9	6.90	0.52	0.24	0.49	0.20	0.08
10	13.50	3.42	0.31	3.02	1.37	0.73
11	9.70	1.78	0.30	3.16	0.73	0.17
12	10.70	2.40	0.32	3.30	0.25	0.14
13	12.10	9.36	0.40	11.51	0.39	0.38
14	9.70	1.72	0.28	2.26	0.82	0.17
15	7.00	0.59	0.29	0.60	0.13	0.35
16	7.20	0.28	0.26	0.30	0.09	0.15
17	8.20	1.64	0.29	1.44	0.20	0.08
18	8.40	0.09	0.22	0.05	0.43	0.20
19	13.10	0.08	0.25	0.03	0.73	0.20
20	8.70	1.36	0.26	0.17	0.99	0.42

В данном примере мы располагаем пространственной выборкой объема $n = 20$; число объясняющих переменных $p = 5$. Матрица X будет составлена из шести столбцов размерности 20 каждый, причем в качестве первого столбца используется вектор, состоящий из одних единиц, а столбцы со 2-го по 6-й представлены соответственно 3–7-м столбцами таблицы 3.1. Вектор-столбец Y определяется 2-м столбцом таблицы 3.1. Анализ технологии сбора исходных статистических данных показал, что допущение о взаимной некоррелированности и гомоскедастичности регрессионных остатков ε может быть принято в качестве рабочей гипотезы. Поэтому мы можем записать уравнения статистической связи между y_i и $X_i = \left(x_i^{(1)} \quad x_i^{(2)} \quad x_i^{(3)} \quad x_i^{(4)} \quad x_i^{(5)}\right)^T$ в виде (3.5).

. . . . . .

3.4 Оценивание неизвестных параметров КЛММР: метод наименьших квадратов и метод максимального правдоподобия

Соотношения (3.5) и (3.6) определяют специфицированные уравнения статистической связи, существующей между результирующей переменной y и объясняющими переменными X. Но значения участвующих в этих уравнениях параметров $\Theta = \begin{pmatrix}\Theta_0 & \Theta_1 & \dots & \Theta_p\end{pmatrix}^T$ и σ^2 нам не известны; их требуется определить (статистически оценить) по имеющимся исходным статистическим данным вида (3.3)–(3.4).

Далее рассмотрим способы статистического оценивания параметров $\Theta = \begin{pmatrix}\Theta_0 & \Theta_1 & \dots & \Theta_p\end{pmatrix}^T$ и σ^2 в рамках КЛММР (метод наименьших квадратов) и в рамках нормальной КЛММР (метод максимального правдоподобия).

Рассмотрим **метод наименьших квадратов (МНК)**.

В основе метода наименьших квадратов лежит стремление исследователя подобрать такие оценки $\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p$ для неизвестных значений параметров функции регрессии соответственно $\Theta_0, \Theta_1, \dots, \Theta_p$, при которых сглаженные (регрессионные) значения $\hat{\Theta}_0 + \hat{\Theta}_1 x_i^{(1)} + \dots + \hat{\Theta}_p x_i^{(p)}$ результирующего показателя как можно меньше отличались бы от соответствующих наблюденных значений y_i. Сформулируем этот принцип математически. Введем в качестве меры расхождения сглаженного и наблюденного (в i-м наблюдении) значений результирующего показателя разность

$$\hat{\varepsilon}_i = y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i^{(1)} - \dots - \hat{\Theta}_p x_i^{(p)} \tag{3.7}$$

(будем в дальнейшем называть $\hat{\varepsilon}_i$ «невязками»). Значения $\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p$ следует подбирать таким образом, чтобы минимизировать некоторую интегральную (по всем имеющимся наблюдениям) характеристику невязок. Примем за такую интегральную характеристику подгонки (выравнивания) значений y_i с помощью линейной функции от $x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}$ $(i = 1, 2, \dots, n)$ величину

$$Q\left(\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p\right) = \sum_{i=1}^{n} \left(y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i^{(1)} - \dots - \hat{\Theta}_p x_i^{(p)}\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{3.8}$$

Величина Q будет определяться при заданной системе наблюдений (3.3)–(3.4) конкретным выбором значений оценок параметров $\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p$. Оценки по методу наименьших квадратов (МНК-оценки) $\hat{\Theta}_{0.\text{МНК}}, \hat{\Theta}_{1.\text{МНК}}, \dots, \hat{\Theta}_{p.\text{МНК}}$ подбираются таким образом, чтобы минимизировать величину Q, определенную соотношением (3.8), т. е.

$$Q\left(\hat{\Theta}_{0.\text{МНК}}, \hat{\Theta}_{1.\text{МНК}}, \dots, \hat{\Theta}_{p.\text{МНК}}\right) = \min_{\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p} Q(\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_p) \tag{3.9}$$

или

$$\hat{\Theta}_{\text{МНК}} = \arg\min_{\hat{\Theta}} Q(\hat{\Theta}).$$

Например, для рисунка 3.2 справедлива формула

$$Q\left(\hat{\Theta}_0, \hat{\Theta}_1\right) = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2.$$

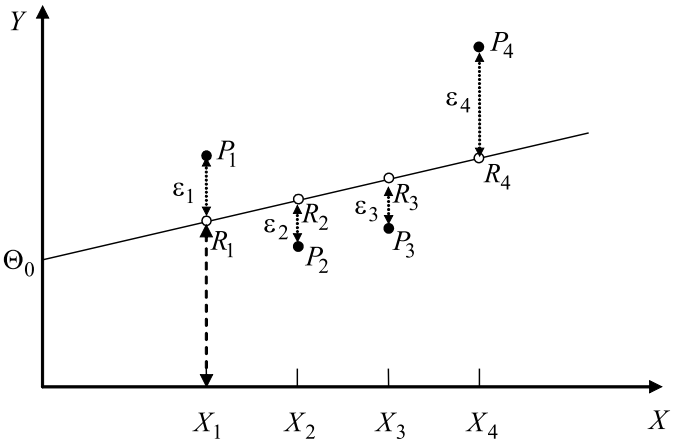


Рис. 3.2 – Оцененная по точкам наблюдения линия регрессии

Опишем процедуру решения оптимизационной задачи (3.9). Начнем с простейшего частного случая, когда рассматривается зависимость y от *единственной* объясняющей переменной x (т. е. $p = 1$). Этот случай в литературе обычно называют моделью парной линейной регрессии.

Первое из уравнений связи (3.5) в данном случае имеет вид: $y_i = \Theta_0 + \Theta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$.

Критерий Q метода наименьших квадратов:

$$Q\left(\widehat{\Theta}_0, \widehat{\Theta}_1\right) = \sum_{i=1}^{n} \left(y_i - \widehat{\Theta}_0 - \widehat{\Theta}_1 x_i\right)^2.$$

Необходимые условия экстремума по $\widehat{\Theta}_0$ и $\widehat{\Theta}_1$ функции $Q\left(\widehat{\Theta}_0, \widehat{\Theta}_1\right)$:

$$\begin{cases} \dfrac{\partial Q}{\partial \widehat{\Theta}_0} = -2\sum\limits_{i=1}^{n} \left(y_i - \widehat{\Theta}_0 - \widehat{\Theta}_1 x_i\right) = 0, \\ \dfrac{\partial Q}{\partial \widehat{\Theta}_1} = -2\sum\limits_{i=1}^{n} x_i \left(y_i - \widehat{\Theta}_0 - \widehat{\Theta}_1 x_i\right) = 0 \end{cases}$$

или, после раскрытия скобок и очевидных тождественных преобразований:

$$\begin{cases} n\widehat{\Theta}_0 + \widehat{\Theta}_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ \left(\sum\limits_{i=1}^{n} x_i\right) \widehat{\Theta}_0 + \widehat{\Theta}_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{3.10}$$

Система (3.10) из двух линейных уравнений относительно $\widehat{\Theta}_0$ и $\widehat{\Theta}_1$ представляет так называемую ***стандартную форму нормальных уравнений*** (для случая $p = 1$). Ее решения легко выписываются в явном виде:

$$\widehat{\Theta}_{1.\text{МНК}} = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}, \tag{3.11}$$

$$\hat{\Theta}_{0.\text{МНК}} = \overline{y} - \hat{\Theta}_{1.\text{МНК}} \cdot \overline{x},$$

где $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ и $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Пример 3.2

Приведем простой пример всего с двумя наблюдениями для того, чтобы продемонстрировать механику процесса: как показано на рисунке 3.3, — наблюдаемое значение $Y = 3$, когда $X = 1$; и $Y = 5$ при $X = 2$ [4].

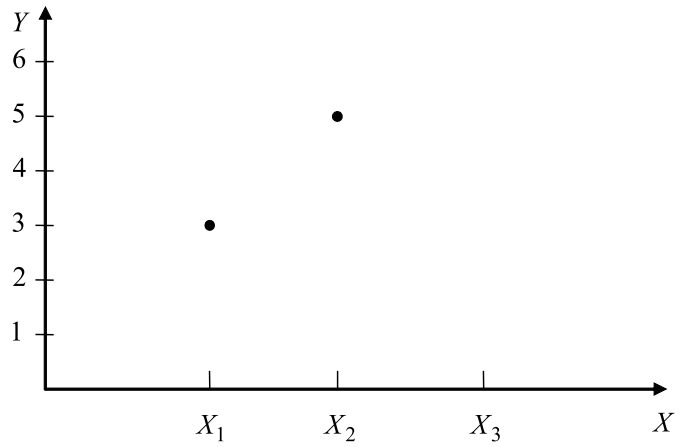


Рис. 3.3 – Пример с двумя наблюдениями

Предположим, что истинная модель имеет вид

$$y_i = \Theta_0 + \Theta_1 x_i + \varepsilon_i,$$

оценим коэффициенты Θ_0 и Θ_1 уравнения

$$\hat{y}_i = \hat{\Theta}_0 + \hat{\Theta}_1 x_i.$$

Очевидно, что при наличии всего двух наблюдений мы можем получить точное соответствие, проведя линию регрессии через две точки. Придем к тому же выводу, используя метод регрессии.

Если $X = 1$, то $Y = (\hat{\Theta}_0 + \hat{\Theta}_1)$ в соответствии с уравнением регрессии. Если $X = 2$, то $Y - (\hat{\Theta}_0 + 2\hat{\Theta}_1)$. Следовательно, мы можем сформировать таблицу 3.2. Таким образом, остаток для первого наблюдения $(\hat{\varepsilon}_1)$, который задается выражением $(y_1 - \hat{y}_1)$, равен $(3 - \hat{\Theta}_0 - \hat{\Theta}_1)$, и $\hat{\varepsilon}_2$, заданный выражением $(y_2 - \hat{y}_2)$, равен $(5 - \hat{\Theta}_0 - 2\hat{\Theta}_1)$. Следовательно,

$$Q(\hat{\Theta}_0, \hat{\Theta}_1) = (3 - \hat{\Theta}_0 - \hat{\Theta}_1)^2 + (5 - \hat{\Theta}_0 - 2\hat{\Theta}_1)^2 = 9 + \hat{\Theta}_0^2 + \hat{\Theta}_1^2 - 6\hat{\Theta}_0 - 6\hat{\Theta}_1 + 2\hat{\Theta}_0\hat{\Theta}_1 +$$
$$+ 25 + \hat{\Theta}_0^2 + 4\hat{\Theta}_1^2 - 10\hat{\Theta}_0 - 20\hat{\Theta}_1 + 4\hat{\Theta}_0\hat{\Theta}_1 = 34 + 2\hat{\Theta}_0^2 + 5\hat{\Theta}_1^2 - 16\hat{\Theta}_0 - 26\hat{\Theta}_1 + 6\hat{\Theta}_0\hat{\Theta}_1.$$

Таблица 3.2 – Расчет характеристик

x	y	$\hat{y}$	$\hat{\varepsilon}$
1	3	$\hat{\Theta}_0 + \hat{\Theta}_1$	$3 - \hat{\Theta}_0 - \hat{\Theta}_1$
2	5	$\hat{\Theta}_0 + 2\hat{\Theta}_1$	$5 - \hat{\Theta}_0 - 2\hat{\Theta}_1$

Теперь нужно выбрать такие значения $\hat{\Theta}_0$ и $\hat{\Theta}_1$, чтобы значение Q было минимальным. Для этого мы используем дифференциальное исчисление и находим значения $\hat{\Theta}_0$ и $\hat{\Theta}_1$, удовлетворяющие следующим соотношениям:

$$\frac{\partial Q}{\partial \hat{\Theta}_0} = 0 \text{ и } \frac{\partial Q}{\partial \hat{\Theta}_1} = 0.$$

Взяв частные производные, получаем

$$\frac{\partial Q}{\partial \hat{\Theta}_0} = 4\hat{\Theta}_0 + 6\hat{\Theta}_1 - 16,$$

$$\frac{\partial Q}{\partial \hat{\Theta}_1} = 10\hat{\Theta}_1 + 6\hat{\Theta}_0 - 26.$$

Таким образом, имеем

$$2\hat{\Theta}_0 + 3\hat{\Theta}_1 - 8 = 0,$$

$$3\hat{\Theta}_0 + 5\hat{\Theta}_1 - 13 = 0.$$

Решив эти два уравнения, получим $\hat{\Theta}_0 = 1$ и $\hat{\Theta}_1 = 2$, следовательно, уравнение регрессии будет иметь следующий вид:

$$\hat{y}_i = 1 + 2x_i.$$

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

$$\hat{\varepsilon}_1 = 3 - \hat{\Theta}_0 - \hat{\Theta}_1 = 3 - 1 - 2 = 0;$$

$$\hat{\varepsilon}_2 = 5 - \hat{\Theta}_0 - 2\hat{\Theta}_1 = 5 - 1 - 4 = 0.$$

Таким образом, оба остатка равны нулю, что означает, что линия проходит точно через обе точки.

Пример 3.3

Используем предыдущий пример и добавим третье наблюдение: $Y = 6$ при $X = 3$. Три наблюдения, показанные на рисунке 3.4, не лежат на одной прямой, поэтому точное соответствие получить невозможно. В этом случае для вычисления положения прямой мы используем регрессию по методу наименьших квадратов. Начнем с задания стандартного уравнения

$$\hat{y}_i = \hat{\Theta}_0 + \hat{\Theta}_1 x_i.$$

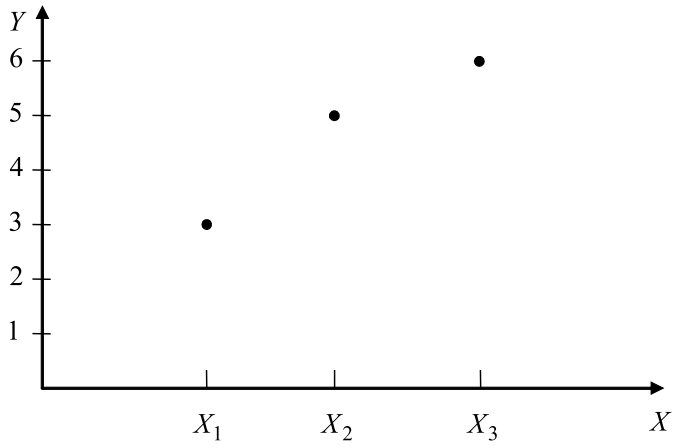


Рис. 3.4 – Пример с тремя наблюдениями

Для значений X, равных 1, 2 и 3, расчетные значения Y равны соответственно $(\Theta_0 + \Theta_1)$, $(\Theta_0 + 2\Theta_1)$ и $(\Theta_0 + 3\Theta_1)$, они приведены в таблице 3.3.

Таблица 3.3 – Расчет характеристик

x	y	$\hat{y}$	$\hat{\varepsilon}$
1	3	$\hat{\Theta}_0 + \hat{\Theta}_1$	$3 - \hat{\Theta}_0 - \hat{\Theta}_1$
2	5	$\hat{\Theta}_0 + 2\hat{\Theta}_1$	$5 - \hat{\Theta}_0 - 2\hat{\Theta}_1$
3	6	$\hat{\Theta}_0 + 3\hat{\Theta}_1$	$6 - \hat{\Theta}_0 - 3\hat{\Theta}_1$

Следовательно,

$$Q\left(\hat{\Theta}_0, \hat{\Theta}_1\right) = \left(3 - \hat{\Theta}_0 - \hat{\Theta}_1\right)^2 + \left(5 - \hat{\Theta}_0 - 2\hat{\Theta}_1\right)^2 + \left(6 - \hat{\Theta}_0 - 3\hat{\Theta}_1\right)^2 =$$
$$= 9 + \hat{\Theta}_0^2 + \hat{\Theta}_1^2 - 6\hat{\Theta}_0 - 6\hat{\Theta}_1 + 2\hat{\Theta}_0\hat{\Theta}_1 + 25 + \hat{\Theta}_0^2 + 4\hat{\Theta}_1^2 - 10\hat{\Theta}_0 - 20\hat{\Theta}_1 + 4\hat{\Theta}_0\hat{\Theta}_1 +$$
$$+ 36 + \hat{\Theta}_0^2 + 9\hat{\Theta}_1^2 - 2\hat{\Theta}_0 - 360\hat{\Theta}_1 + 6\hat{\Theta}_0\hat{\Theta}_1 = 70 + 3\hat{\Theta}_0^2 + 14\hat{\Theta}_1^2 - 28\hat{\Theta}_0 - 620\hat{\Theta}_1 + 12\hat{\Theta}_0\hat{\Theta}_1.$$

Условия первого порядка $\frac{\partial Q}{\partial \hat{\Theta}_0} = 0$ и $\frac{\partial Q}{\partial \hat{\Theta}_1} = 0$ дают $6\hat{\Theta}_0 + 12\hat{\Theta}_1 - 28 = 0$, и $12\hat{\Theta}_0 + 28\hat{\Theta}_1 - 62 = 0$.

Решая систему этих двух уравнений, получим $\hat{\Theta}_0 = 1.67$ и $\hat{\Theta}_1 = 1.5$. Следовательно, уравнение регрессии имеет следующий вид:

$$\hat{y}_i = 1.67 + 1.5x_i.$$

Три наблюдения и линия регрессии представлены на рисунке 3.5.

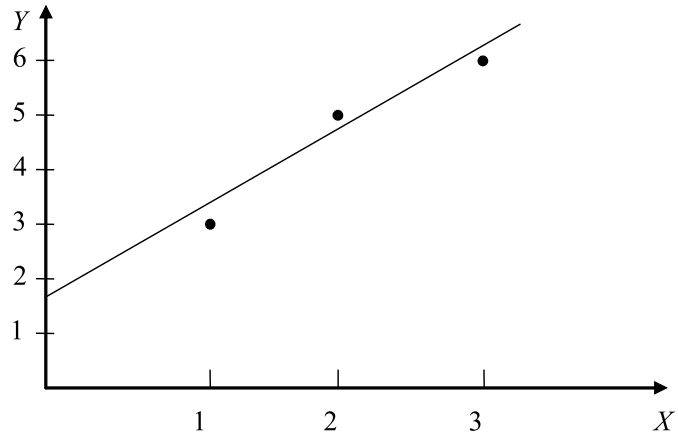


Рис. 3.5 – Пример построения линии регрессии с тремя наблюдениями

Перейдем к случаю многих объясняющих переменных ($p > 1$). В этом случае более удобной оказывается матричная форма записи всех необходимых в данной задаче условий и соотношений:

$\hat{\varepsilon} = Y - X\hat{\Theta} = \left(y_1 - \hat{\Theta}_0 - \hat{\Theta}_1 x_1^{(1)} - \ldots - \hat{\Theta}_p x_1^{(p)}, \ldots, y_n - \hat{\Theta}_0 - \hat{\Theta}_1 x_n^{(1)} - \ldots - \hat{\Theta}_p x_n^{(p)}\right)^T$ — вектор-столбец невязок;

$$Q(\hat{\Theta}) = \sum_{i=1}^{n} \left(y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i^{(1)} - \ldots - \hat{\Theta}_p x_i^{(p)}\right)^2 = \left(Y - X\hat{\Theta}\right)^T \left(Y - X\hat{\Theta}\right) — \qquad (3.12)$$

оптимизируемый (по $\hat{\Theta}$) критерий метода наименьших квадратов.

Перед тем, как выписать необходимые условия экстремума функции $Q(\hat{\Theta})$ по $\hat{\Theta}$, преобразуем правую часть (3.12):

$$Q(\hat{\Theta}) = Y^T Y - 2\hat{\Theta}^T X^T Y + \hat{\Theta}^T X^T X \hat{\Theta}. \qquad (3.13)$$

В этом преобразовании мы воспользовались правилом транспонирования произведения матриц, а также тем, что $\hat{\Theta}^T X^T Y$ — число, а потому оно совпадает со своим транспонированным выражением $Y^T X \hat{\Theta}$.

Необходимые условия, которым удовлетворяют решения оптимизационной задачи, получаются дифференцированием правой части (3.13) по $\hat{\Theta}_0, \hat{\Theta}_1, \ldots, \hat{\Theta}_p$. При выписывании получающейся при этом системы уравнений относительно $\hat{\Theta}_0, \hat{\Theta}_1, \ldots, \hat{\Theta}_p$ мы воспользуемся матричным обозначением производной

$$\frac{\partial Q(\hat{\Theta})}{\partial \hat{\Theta}} = \left(\frac{\partial Q(\hat{\Theta})}{\partial \hat{\Theta}_0}, \frac{\partial Q(\hat{\Theta})}{\partial \hat{\Theta}_1}, \ldots, \frac{\partial Q(\hat{\Theta})}{\partial \hat{\Theta}_p}\right)^T,$$

а также правилами записи матричного дифференцирования линейных и квадратичных функций от $\hat{\Theta}$:

$$\frac{\partial Q(\hat{\Theta})}{\partial \hat{\Theta}} = -2X^T Y + 2X^T X \hat{\Theta} = O_{p+1}, \qquad (3.14)$$

где O_{p+1} — вектор-столбец, имеющий размерность $p+1$, состоящий из одних нулей.

Разрешая систему уравнений (3.14) относительно $\widehat{\Theta}$, получаем:

$$X^T X\widehat{\Theta} = X^T Y, \tag{3.15}$$

и, следовательно,

$$\widehat{\Theta}_{\text{МНК}} = (X^T X)^{-1} X^T Y. \tag{3.16}$$

В основной формуле метода наименьших квадратов (3.16) мы воспользовались невырожденностью матрицы $X^T X$, которая следует из требования максимального ранга для матрицы X, входящего в описание КЛММР.

Применим общую формулу (3.15) к ранее рассмотренному частному случаю парной регрессии ($p = 1$). В этом случае:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \quad X^T X = \begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix}.$$

Подставляя эти выражения в (3.14), получаем:

$$\begin{cases} n\widehat{\Theta}_0 + \widehat{\Theta}_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ \left(\sum\limits_{i=1}^{n} x_i\right)\widehat{\Theta}_0 + \widehat{\Theta}_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i, \end{cases}$$

т. е. ту же стандартную форму нормальных уравнений, с которой мы встретились (3.11), анализируя эту же модель покомпонентно.

Пример 3.4

Применение формулы (3.16) к данным таблицы 3.1 позволяет получить следующие МНК-оценки для параметров $\Theta = (\Theta_0, \Theta_1, \dots, \Theta_5)$:

$$\widehat{\Theta}_{0.\text{МНК}} = 3.515; \quad \widehat{\Theta}_{1.\text{МНК}} = -0.006; \quad \widehat{\Theta}_{2.\text{МНК}} = 15.542; \quad \widehat{\Theta}_{3.\text{МНК}} = 0.110;$$
$$\widehat{\Theta}_{4.\text{МНК}} = 4.475; \quad \widehat{\Theta}_{5.\text{МНК}} = -2.932.$$

Таким образом, оценка $\widehat{f}(X)$ неизвестной функции регрессии $f(X)$ в данном случае имеет вид:

$$\widehat{f}(X) = 3.515 - 0.006x^{(1)} + 15.542x^{(2)} + 0.110x^{(3)} + 4.475x^{(4)} - 2.932x^{(5)}.$$

Далее рассмотрим **метод максимального правдоподобия (ММП)**. Метод максимального правдоподобия может быть применен в тех случаях, когда с точностью до неизвестных значений параметров известен общий вид закона распределения

вероятностей имеющихся выборочных данных. Поэтому если мы проводим регрессионный анализ в рамках нормальной КЛММР, т. е. если дополнительно к условиям (3.5) постулируется нормальность регрессионных остатков $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, то, учитывая их взаимную некоррелированность, можно выписать функцию правдоподобия в терминах остатков:

$$L\left(\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n | \Theta; \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(y_i - \Theta_0 - \Theta_1 x_i^{(1)} - \ldots - \Theta_p x_i^{(p)}\right)^2} =$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\Theta)^T (Y - X\Theta)\right).$$

Оценки $\hat{\Theta}_{\text{ММП}}$ и $\hat{\sigma}^2_{\text{ММП}}$ максимального правдоподобия определяются как такие значения Θ и σ^2, при которых функция правдоподобия L (или логарифмическая функция правдоподобия $l = \ln L$) достигает своей максимальной величины. Соответствующие уравнения ММП получаются приравниванием к нулю производных функции l по Θ и σ^2:

$$\begin{aligned} l\left(\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n | \Theta; \sigma^2\right) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\Theta)^T (Y - X\Theta); \\ \frac{\partial l}{\partial \Theta} &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \Theta}\left[(Y - X\Theta)^T (Y - X\Theta)\right] = O_{p+1}, \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\Theta)^T (Y - X\Theta) = O. \end{aligned} \tag{3.17}$$

Первая строка (3.17) после сокращения левой и правой частей на $-1/(2\sigma^2)$ повторяет систему уравнений (3.14) метода наименьших квадратов. Следовательно,

$$\hat{\Theta}_{\text{ММП}} = \hat{\Theta}_{\text{МНК}} = (X^T X)^{-1} X^T Y. \tag{3.18}$$

Вторая строка системы (3.17) позволяет вычислить ММП-оценку для σ^2:

$$\hat{\sigma}^2_{\text{ММП}} = \frac{1}{n}\left(Y - X\hat{\Theta}\right)^T (Y - X\hat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i^{(1)} - \ldots - \hat{\Theta}_p x_i^{(p)}\right)^2,$$

где $\hat{\Theta} = \left(\hat{\Theta}_0, \hat{\Theta}_1, \ldots, \hat{\Theta}_p\right)$ — оценки но методу наименьшие квадратов (они же — оценки по методу максимального правдоподобия) неизвестных коэффициентов регрессии Θ. Далее оценки, полученные по формулам (3.18), мы будем обозначать просто $\hat{\Theta}$ (без индексирования метода).

3.5 Статистические свойства оценок параметров КЛММР

Оценки $\hat{\Theta}$ и $\hat{\sigma}^2$ «ведут себя» как случайные величины. Поэтому представляет интерес исследовать статистические свойства этих оценок.

Состоятельность оценок $\hat{\Theta}$ и $\hat{\sigma}^2$. В данном случае свойство состоятельности оценок определяется структурой матрицы X. Существуют различные формулировки условий (в терминах элементов матрицы X), при которых оценки $\hat{\Theta}$ и $\hat{\sigma}^2$

являются состоятельными. Рассмотрим наиболее удобное для приложений условие состоятельности оценок $\widehat{\Theta}$ и $\widehat{\sigma}^2$.

*Оценки $\widehat{\Theta}$ и $\widehat{\sigma}^2$ являются **состоятельными** тогда и только тогда, когда наименьшее собственное значение матрицы X^TX стремится к бесконечности при $n \to \infty$.*

Наименьшее собственное значение $\lambda_{\min}$ матрицы X^TX определяется как минимальный по величине корень уравнения $\left|X^TX - \lambda I_{p+1}\right| = O$ (матрица X^TX как симметричная и положительно определенная имеет $p + 1$ действительных положительных собственных значений $\lambda_1, \lambda_2, \ldots, \lambda_p$. Сформулированное условие состоятельности оценок $\widehat{\Theta}$ и $\widehat{\sigma}^2$ для случая парной линейной регрессионной зависимости (т. е. при $p = 1$) равносильно требованию

$$\frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}{n} \to \text{const при } n \to \infty.$$

Несмещенность оценок.

*Оценка $\widehat{\Theta}$ параметра Θ называется **несмещенной**, если $M\widehat{\Theta} = \Theta$.*

Чтобы подсчитать среднее значение оценки (3.18), подставим в формулу (3.18) вместо Y его выражение из основного (первого) соотношения системы (3.6):

$$\widehat{\Theta} = \left(X^TX\right)^{-1}X^T(X\Theta + \varepsilon) = \Theta + \left(X^TX\right)^{-1}X^T\varepsilon. \qquad (3.19)$$

Здесь оценка представлена как сумма истинного (неизвестного нам) значения Θ и линейной комбинации случайных остатков $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Беря математические ожидания от левой и правой частей (3.19) с учетом того, что величины $\widehat{\Theta}$ и $\left(X^TX\right)^{-1}X^T$ неслучайны, а средние значения остатков равны нулю (т. е. $M\varepsilon = O_n$), получаем:

$$M\widehat{\Theta} = M\Theta + \left(X^TX\right)^{-1}X^TM\varepsilon = \Theta.$$

Тем самым показано, что МНК-оценки (они же ММП-оценки) $\widehat{\Theta}$ неизвестных параметров КЛММР являются несмещенными.

В отличие от $\widehat{\Theta}$ оценка $\widehat{\sigma}^2_{\text{ММП}}$ параметра σ^2 оказывается смещенной и

$$M\widehat{\sigma}^2_{\text{ММП}} = M\left[\frac{1}{n}\left(Y - X\widehat{\Theta}\right)^T\left(Y - X\widehat{\Theta}\right)\right] = \sigma^2\left(1 - \frac{p+1}{n}\right).$$

Воспользуемся вместо $\widehat{\sigma}^2_{\text{ММП}}$ оценкой

$$\widehat{\sigma}^2 = \frac{n}{n-p-1}\widehat{\sigma}^2_{\text{ММП}} = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \widehat{\Theta}_0 - \widehat{\Theta}_1x_i^{(1)} - \ldots - \widehat{\Theta}_px_i^{(p)}\right)^2 =$$

$$= \frac{1}{n-p-1}\left(Y - X\widehat{\Theta}\right)^T\left(Y - X\widehat{\Theta}\right).$$

Такой способ оценивания неизвестной дисперсии остатков ε_i (так называемой *остаточной дисперсии* σ^2) уже будет *несмещенным*.

Оптимальность МНК-оценок. При сравнении различных способов оценивания решающей характеристикой качества оценки $\widehat{\Theta}$ неизвестного *числового* параметра Θ оказывается средний квадрат ошибки $M\left(\widehat{\Theta}-\Theta\right)^2$.

Оценка $\widehat{\Theta}_1$ точнее (лучше, эффективнее), чем оценка $\widehat{\Theta}_2$, если $M\left(\widehat{\Theta}_1-\Theta\right)^2 < M\left(\widehat{\Theta}_2-\Theta\right)^2$.

*Оценка $\widehat{\Theta}$ является **оптимальной в классе оценок** $\widetilde{M}$, если*

$$M\left(\widehat{\Theta}_{\text{опт}}-\Theta\right)^2=\min M\left(\widehat{\Theta}-\Theta\right)^2,\quad \widehat{\Theta}\in\widetilde{M}.$$

Как определить качество и оптимальность векторной оценки $\widehat{\Theta}=\left(\widehat{\Theta}_0\ \widehat{\Theta}_1\ \ldots\ \widehat{\Theta}_p\right)^T$? Необходимо показать, что для любой линейной функции $C^T\Theta=\Theta_c$ от неизвестных параметров $\widehat{\Theta}_0,\widehat{\Theta}_1,\ldots,\widehat{\Theta}_p$ оценка $C^T\Theta_{\text{МНК}}$ является оптимальной для параметра Θ в некотором достаточно широком классе оценок $\widetilde{M}$ (здесь $C=\left(C_0\ \ C_1\ \ \ldots\ \ C_p\right)^T$ — $(p+1)$-мерный вектор-столбец произвольных констант C_j). Действительно, полагая $C^T=(0,0,\ldots,0,1,0,\ldots,0)$, где единица стоит на j-м месте, получаем оптимальность $\widehat{\Theta}$ с точки зрения точности оценивания параметра $\widehat{\Theta}_j$ $(j=0,1,\ldots,p)$. Полагая $C^T=\left(1,x^{(1)},\ldots,x^{(p)}\right)$, получаем оптимальность $\widehat{\Theta}$ с точки зрения точности оценивания неизвестной функции регрессии $f(X)=\Theta_0+\Theta_1x^{(1)}+\ldots+\Theta_px^{(p)}$ при любых заданных значениях объясняющих переменных X.

3.6 Определение доверительных интервалов для коэффициентов и функции регрессии

Перейдем теперь к оценке значимости коэффициентов регрессии $\widehat{\Theta}_j$ и построению доверительного интервала для параметров регрессионной модели Θ_j.

Оценка $\widehat{\sigma}^2_{\widehat{\Theta}_j}$ дисперсии $\sigma^2_{\widehat{\Theta}_j}$ коэффициента регрессии $\widehat{\Theta}_j$, определяется по формуле:

$$\widehat{\sigma}^2_{\widehat{\Theta}_j}=\widehat{\sigma}^2\left[\left(X^TX\right)^{-1}\right]_{jj},$$

где $\widehat{\sigma}^2$ — несмещенная оценка параметра σ^2; $\left[\left(X^TX\right)^{-1}\right]_{jj}$ — диагональный элемент матрицы $\left(X^TX\right)^{-1}$.

Значимость коэффициента регрессии $\widehat{\Theta}_j$ можно проверить, если учесть, что статистика $\left(\widehat{\Theta}_j-\Theta_j\right)/\widehat{\sigma}_{\widehat{\Theta}_j}$ имеет t-распределение Стьюдента с $k=n-p-1$ степенями свободы. Поэтому $\widehat{\Theta}_j$ значимо отличается от нуля (иначе — гипотеза H_0 о равенстве параметра Θ_j нулю, т. е. H_0: $\Theta_j=0$, отвергается) на уровне значимости α, если

$|t| = \left|\widehat{\Theta}_j\right| / \widehat{\sigma}_{\widehat{\Theta}_j} > t_{1-\alpha;n-p-1}$, где $t_{1-\alpha;n-p-1}$ — табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $k = n - p - 1$.

В общей постановке гипотеза H_0 о равенстве параметра Θ_j заданному числу Θ_{j0}, т. е. H_0: $\Theta_j = \Theta_{j0}$, отвергается, если

$$|t| = \frac{\left|\widehat{\Theta}_j - \Theta_{j0}\right|}{\widehat{\sigma}_{\widehat{\Theta}_j}} > t_{1-\alpha;n-p-1}.$$

Поэтому доверительный интервал для параметра Θ_j есть

$$\widehat{\Theta}_j - t_{1-\alpha;n-p-1}\widehat{\sigma}_{\widehat{\Theta}_j} \leqslant \Theta_j \leqslant \widehat{\Theta}_j + t_{1-\alpha;n-p-1}\widehat{\sigma}_{\widehat{\Theta}_j}.$$

Наряду с интервальным оцениванием коэффициентов регрессии весьма важным для оценки точности определения зависимой переменной (прогноза) является построение доверительного интервала для функции регрессии или для условного математического ожидания зависимой переменной $M_x(Y)$, найденного в предположении, что объясняющие переменные $X^{(1)}, X^{(2)}, \ldots, X^{(p)}$ приняли значения, задаваемые вектором $\left(X^{(0)}\right)^T = \begin{pmatrix} 1 & x_1^{(0)} & x_2^{(0)} & \ldots & x_p^{(0)} \end{pmatrix}$. Доверительный интервал для $M_x(Y)$:

$$\hat{y} - t_{1-\alpha;k}\sigma_{\hat{y}} \leqslant M(Y) \leqslant \hat{y} + t_{1-\alpha;k}\sigma_{\hat{y}},$$

где $\hat{y}$ — групповая средняя, определяемая по уравнению регрессии; $\sigma_{\hat{y}} = \sigma\sqrt{\left(X^{(0)}\right)^T\left(X^TX\right)^{-1}X^{(0)}}$ — ее стандартная ошибка.

Аналогичный доверительный интервал для индивидуальных значений зависимой переменной y_0^* примет вид:

$$\hat{y}_0 - t_{1-\alpha;n-p-1}\sigma_{\hat{y}_0} \leqslant y_0^* \leqslant \hat{y}_0 + t_{1-\alpha;n-p-1}\sigma_{\hat{y}_0},$$

где $\sigma_{\hat{y}_0} = \sigma\sqrt{1 + \left(X^{(0)}\right)^T\left(X^TX\right)^{-1}X^{(0)}}$.

3.7 Обобщенная линейная модель

При моделировании реальных экономических процессов встречаются ситуации, в которых условия классической линейной модели регрессии оказываются нарушенными.

В частности, могут не выполняться предпосылки о том, что случайные возмущения (ошибки) модели имеют постоянную дисперсию и не коррелированы между собой. Для линейной множественной модели эти предпосылки означают, что ковариационная матрица вектора возмущений (ошибок) ε имеет вид: $\sum_\varepsilon = \sigma^2 I_n$.

В тех случаях, когда имеющиеся статистические данные достаточно однородны, допущение $\sum_\varepsilon = \sigma^2 I_n$ вполне оправдано.

Однако в других ситуациях оно может оказаться неприемлемым. Так, например, при использовании зависимости расходов на потребление от уровня доходов семей можно ожидать, что в более обеспеченных семьях вариация расходов выше, чем в малообеспеченных, т. е. дисперсии возмущений неодинаковы. При рассмотрении временных рядов мы, как правило, сталкиваемся с ситуацией, когда наблюдаемые в данный момент значения зависимой переменной коррелируют с их значениями

в предыдущие моменты времени, т. е. наблюдается корреляция между возмущениями в разные моменты времени.

Обобщенная линейная модель множественной регрессии описывается следующей системой соотношений и условий:

1) ε — случайный вектор; X — неслучайная (детерминированная) матрица;
2) $M(\varepsilon) = 0_n$;
3) $\sum_\varepsilon = M(\varepsilon\varepsilon^T) = \Omega$, где Ω — положительно определенная матрица;
4) ранг$(X) = p + 1 < n$, где p — число объясняющих переменных; n — число наблюдений.

Сравнивая обобщенную модель с классической, видим, что она отличается от классической только видом ковариационной матрицы: вместо $\sum_\varepsilon = \sigma^2 I_n$ для классической модели имеем $\sum_\varepsilon = M(\varepsilon\varepsilon^T) = \Omega$ для обобщенной. Это означает, что в отличие от классической, в обобщенной модели ковариации и дисперсии объясняющих переменных могут быть произвольными. В этом состоит суть обобщения регрессионной модели.

Обычный метод наименьших квадратов в обобщенной линейной регрессионной модели дает смещенную оценку ковариационной матрицы $\sum_\Theta$ вектора оценок Θ.

Теорема Айткена. В классе линейных несмещенных оценок вектора Θ для обобщенной регрессионной модели оценка

$$\hat{\Theta} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} Y$$

имеет наименьшую ковариационную матрицу.

В случае классической модели, т. е. при выполнении предпосылки $\sum_\varepsilon = \Omega = \sigma^2 I_n$, оценка обобщенного метода наименьших квадратов совпадает с оценкой «обычного» метода Θ.

При выполнении предпосылки о нормальном законе распределения вектора возмущений ε можно убедиться в том, что оценка $\hat{\Theta}$ обобщенного метода наименьших квадратов для параметра Θ при известной матрице Ω совпадает с его оценкой, полученной методом максимального правдоподобия.

Для применения обобщенного метода наименьших квадратов необходимо знание ковариационной матрицы вектора возмущений Ω, что встречается крайне редко в практике эконометрического моделирования. Поэтому для практической реализации обобщенного метода наименьших квадратов необходимо вводить дополнительные условия на структуру матрицы Ω.

Например, для модели с гетероскедастичными остатками:

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}.$$

Контрольные вопросы по главе 3

1. Что понимается под результирующей переменной?
2. Что понимается под объясняющей переменной?
3. Что называется функцией регрессией?
4. Что представляют собой уравнения регрессионной связи?
5. Каковы причины наличия случайного члена в функции регрессии?
6. Как представляют в регрессионном анализе исходные статистические связи?
7. Каковы основные задачи прикладного регрессионного анализа?
8. Какие функции регрессии рассматриваются в рамках КЛММР?
9. Сформулировать требования, предъявляемые к КЛММР.
10. Какой вид имеет матричная форма записи КЛММР?
11. Какие условия должны выполняться для нормальной КЛММР?
12. Какие существуют методы статистического оценивания параметров в рамках КЛММР и в рамках нормальной КЛММР?
13. Записать критерий метода наименьших квадратов для случая парной линейной регрессии.
14. Что представляет собой стандартная форма нормальных уравнений для случая парной линейной регрессии?
15. Записать критерий метода наименьших квадратов для случая многих объясняющих переменных.
16. По какой формуле вычисляется оценка неизвестных параметров при использовании метода наименьших квадратов?
17. Записать функцию правдоподобия в случае нормальной КЛММР.
18. Каким образом определяются оценки по методу максимального правдоподобия?
19. При каком условии оценки являются состоятельными?
20. Какие оценки называются несмещенными?
21. Что понимают под оптимальностью оценок?
22. С помощью какой системы соотношений можно описать обобщенную линейную модель множественной регрессии?
23. Записать теорему Айткена.

Глава 4

НЕЛИНЕЙНЫЕ МОДЕЛИ РЕГРЕССИИ И ЛИНЕАРИЗАЦИЯ

4.1 Нелинейные связи в экономике. Линеаризация модели

До сих пор мы рассматривали лишь *линейные* модели регрессионной зависимости y от $X = \left(x^{(1)} \quad x^{(2)} \quad \ldots \quad x^{(p)}\right)^T$. В то же время многие важные связи в экономике являются *нелинейными*. Примеры такого рода регрессионных моделей доставляет нам изучение так называемых *производственных функций* (зависимостей, существующих между объемом произведенной продукции и основными факторами производства — трудом, капиталом и т. п.), функций *спроса* (зависимостей, существующих между спросом на какой-либо вид товаров или услуг, с одной стороны, и доходом и ценами на этот и другие товары — с другой). *Этап параметризации регрессионной модели*, т. е. выбора параметрического семейства функций $\{f(X;\Theta)\}$, в рамках которого производится дальнейший поиск неизвестной функции регрессии $f(X) = M(y|X)$, является одновременно наиболее важным и наименее формализованным и теоретически обоснованным этапом регрессионного анализа. Если же в результате реализации этого этапа мы приходим к выводу, что функция $f(X;\Theta)$ нелинейна, то далее действуем следующим образом:

1) вначале пытаемся подобрать такие преобразования к анализируемым переменным $y, x^{(1)}, \ldots, x^{(p)}$, которые позволили бы представить искомую зависимость в виде линейного соотношения между преобразованными переменными; другими словами, если $\varphi_0, \varphi_1, \ldots, \varphi_p$ — те самые искомые функции, которые определяют переход к преобразованным переменным, т. е. $\widetilde{y} = \varphi_0(y)$, $\widetilde{x}^{(1)} = \varphi_1\left(x^{(1)}\right), \ldots, \widetilde{x}^{(p)} = \varphi_p\left(x^{(p)}\right)$, то связь между y и $X = \left(x^{(1)} \quad x^{(2)} \quad \ldots \quad x^{(p)}\right)^T$ может быть представлена в виде линейной функции регрессии $\widetilde{y}$ по $\widetilde{X}$, а именно:

$$\widetilde{y}_i = \Theta_0 + \Theta_1 \widetilde{x}_i^{(1)} + \ldots + \Theta_p \widetilde{x}_i^{(p)} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Эту часть исследования обычно называют ***процедурой линеаризации модели***;

2) в случае невозможности линеаризации модели приходится исследовать искомую регрессионную зависимость в терминах *исходных переменных*, а именно:

$$y_i = f(X_i; \Theta) + \varepsilon_i,$$

если спецификация регрессионных остатков ε_i соответствует условиям *классической* модели, то для вычисления МНК-оценок $\widehat{\Theta}_{\text{МНК}}$ векторного параметра Θ решается оптимизационная задача вида

$$\widehat{\Theta}_{\text{МНК}} = \arg\min_{\Theta} \sum_{i=1}^{n} \left(y_i - f(X_i; \Theta) \right)^2.$$

4.2 Использование априорной информации о содержательной сущности анализируемой зависимости

Анализируя содержательную сущность изучаемой зависимости, исследователь еще до обращения к исходным статистическим данным может попытаться ответить на ряд вопросов по поводу характера искомой регрессионной связи:

а) будет ли искомая функция $f(X)$ монотонной или она должна иметь один экстремум (может быть, несколько)?

б) следует ли ожидать стремления $\left(\text{в процессе } x^{(k)} \to \infty\right) f(X)$ к асимптомам (по одной или нескольким предикторным переменным) и какова их содержательная интерпретация? Так, например, если $f(X)$ — средний объем благ определенного вида, потребляемых семьями группы X по доходам, то, очевидно, при $x^{(k)} \to \infty$ следует ожидать «насыщения», т. е. $f(X)$ будет стремиться (снизу) к горизонтальной асимптоте;

в) какова принципиальная природа воздействия объясняющих переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ на формирование результирующего показателя y — аддитивная или мультипликативная? Так, например, многие схемы зависимостей в экономике характеризуются мультипликативной природой воздействия предикторов на y;

г) не диктует ли содержательный смысл анализируемой зависимости обязательное прохождение графика искомой функции $f(X)$ через одну или несколько априори заданных точек в исследуемом факторном пространстве?

При выборе общего вида искомой функции регрессии $f(X)$, помимо соображений и приемов, описанных выше, полезно учитывать следующие общие рекомендации.

1. Не следует гнаться за чрезмерной сложностью функции, описывающей поведение искомой функции регрессии, руководствуясь исключительно соображениями

оптимизации критерия качества аппроксимации. Дело в том, что если и оценки $\widehat{\Theta}$ неизвестных параметров модели, и значение критерия качества вычисляются на основании одной и той же выборки, то за счет увеличения размерности k оцениваемого векторного параметра $\widehat{\Theta}$ можно добиться, на первый взгляд, идеального результата. Возможность эта основана на известном в математическом анализе результате, в соответствии с которым для любой заданной системы точек $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (с неповторяющимися абсциссами) можно подобрать алгебраический полином степени $n - 1$, проходящий в точности через все точки этой системы. А это значит, что все «невязки» $y_i - f(x_i)$, на основании которых строится критерий качества, равны нулю, т. е. «лучше» модели для описания поведения функции регрессии подобрать невозможно. На самом деле при таком подходе мы как бы заставляем функцию $f(x)$ реагировать на случайные флюктуации, объясняемые наличием остаточной случайной компоненты $\varepsilon(X)$. Поэтому если мы попробуем применить полученный таким образом результат к другой выборке из той же самой генеральной совокупности, то увидим явное рассогласование модельных $\left(f(X_i)\right)$ и наблюденных (y_i) значений результирующего показателя. Поэтому при подборе общего вида функции регрессии, как правило, идут от простого к сложному, т. е. начиная с анализа возможности использовать простейшую линейную модель.

2. Следует добиваться компромисса между сложностью регрессионной модели и точностью ее оценивания. Из общих результатов математической статистики, относящихся к анализу точности оценивания исследуемой модели при ограниченных объемах выборки, следует, что с увеличением сложности модели (выраженной, например, размерностью k векторного параметра Θ, участвующего в ее уравнении) точность оценивания падает. Например, ширина доверительного интервала для неизвестного значения $y(X)$ при прочих равных характеристиках анализируемой схемы, увеличивается с ростом размерности параметра Θ, участвующего в вычислении функции регрессии $f(X;\Theta)$. Именно поэтому в ситуациях, когда исследователь располагает ограниченной исходной выборочной информацией вида, он вынужден искать компромисс между степенью общности привлекаемого класса допустимых решений и точностью оценивания, которой возможно при этом добиться.

3. При обнаружении нелинейности в парных статистических связях анализируемых переменных $x^{(j)}$ и y $(j = 1, 2, \ldots, p)$ следует попытаться применить к этим переменным линеаризующие преобразования. Простейший пример такого приема мы имеем, когда вместо анализа степенной зависимости вида

$$y = \Theta_0 x^{\Theta_1}$$

исследователь рассматривает линейную зависимость между логарифмами исходных переменных, а именно:

$$\widetilde{y} = \widetilde{\Theta}_0 + \Theta_1 \widetilde{x},$$

где $\widetilde{y} = \ln y$, $\widetilde{x} = \ln x$, $\widetilde{\Theta}_0 = \ln \Theta_0$. В зависимости от типа нелинейной связи, существующей между исходными переменными, подбираются и другие линеаризующие преобразования.

4. Анализ регрессионных остатков. Ряд статистических критериев проверки адекватности используемой аппроксимирующей модели регрессии основан на ана-

лизе регрессионных остатков (невязок). В основе их конструирования — положение, в соответствии с которым правильный выбор модели $f_a(X)$ предопределяет асимптотическую (по $n \to \infty$) независимость остатков $\hat{\varepsilon}(X_i)$. Поэтому статистическая проверка правильности выбора общего вида функции регрессии сводится к проверке статистической независимости остатков, для чего могут быть использованы, например, критерии, описанные в п. 6.4.

5. Поиск модели, наиболее устойчивой к варьированию состава выборочных данных, на основании которых она оценивается. Идея этого подхода к выбору общего вида исследуемой регрессионной зависимости основана на следующем простом соображении. Если общий параметрический вид зависимости $f\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \Theta\right)$ «угадан» правильно, то результаты оценивания $\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots$ параметра Θ по различным подвыборкам выборки $\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\right)$, $i = 1, \ldots, n$ будут мало отличаться друг от друга (а следовательно, не сильно будут различаться между собой и соответствующие значения $f\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \widehat{\Theta}_1\right)$, $f\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \widehat{\Theta}_2\right), \ldots$). И наоборот, при неудачном выборе общего вида искомой зависимости результаты ее восстановления по различным выборкам, как правило, будут сильно отличаться один от другого.

4.3 Некоторые виды нелинейных зависимостей, поддающиеся линеаризации. Зависимости гиперболического типа

Пусть y и $X = \left(x^{(1)} \quad x^{(2)} \quad \ldots \quad x^{(p)}\right)^T$ — исходные анализируемые переменные (соответственно результирующая и объясняющие), а ε — случайная остаточная компонента, участвующая в записи регрессионной зависимости, связывающей между собой y и X. За редким исключением вопросы линеаризации анализируемых связей решаются на основе рассмотрения *парных* зависимостей (и графически представляющих их парных корреляционных полей) типа $\left(x_i^{(j)}, y_i\right)$ и $\left(x_i^{(j)}, y_i^l\right)$, $i = 1, 2, \ldots, n$. Поэтому ниже будут представлены и проанализированы именно *парные* регрессионные зависимости, поддающиеся линеаризации.

Зависимости гиперболического типа.

1. Предположим, что анализируемые переменные и случайные регрессионные остатки соответственно x, y и ε связаны между собой статистической зависимостью вида

$$y = \Theta_0 + \Theta_1 \frac{1}{x} + \varepsilon, \quad (0 < x < \infty).$$

Соответствующая кривая регрессии $f(x; \Theta) = f\left(x; \Theta_0; \Theta_1\right) = \Theta_0 + \Theta_1 / x$ (рис. 4.1) характеризуется двумя асимптотами (т. е. прямыми, к которым график функции неограниченно приближается, не достигая их) — горизонтальной ($y = \Theta_0$) и вертикальной ($x = 0$).

С помощью преобразования объясняющей переменной $\widetilde{x} = 1/x$ (т. е. при переходе к новой объясняющей переменной $\widetilde{x}$) эта зависимость приводится к линейному виду $y = \Theta_0 + \Theta_1 \widetilde{x} + \varepsilon$. Соответственно при вычислении МНК-оценок пара-

метров Θ_0 и Θ_1 второй столбец матрицы X должен быть сформирован из чисел $1/x_1, 1/x_2, \ldots, 1/x_n$.

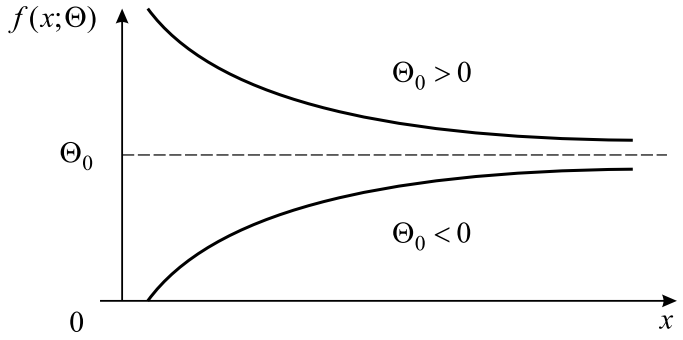


Рис. 4.1 – График гиперболической зависимости вида $f(x;\Theta) = \Theta_0 + \Theta_1/x$

2. Пусть переменные x, y и случайные регрессионные остатки ε связаны между собой статистической зависимостью вида

$$y = \frac{1}{\Theta_0 + \Theta_1 x + \varepsilon}, \quad \left(-\frac{\Theta_0}{\Theta_1} < x < \infty\right).$$

Мы придем к линейной модели $y = \Theta_0 + \Theta_1 x + \varepsilon$, если в качестве результирующего признака рассмотрим переменную $\widetilde{y} = 1/y$. Следует не забыть только, что при вычислении МНК-оценок Θ_0 и Θ_1 надо использовать в качестве вектора наблюденных значений зависимой переменной вектор $\widetilde{Y} = \left(1/y_1 \quad 1/y_2 \quad \ldots \quad 1/y_n\right)^T$.

3. Если этап параметризации модели регрессии приводит нас к зависимости вида

$$y = \frac{x}{\Theta_0 x + \Theta_1 + x\varepsilon}, \quad \left(-\frac{\Theta_1}{\Theta_0} < x < \infty\right),$$

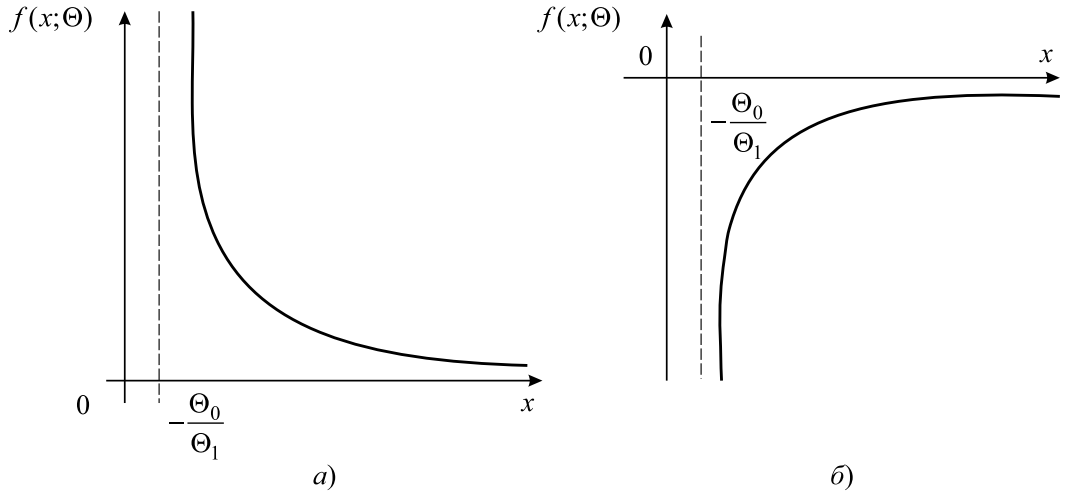
то линеаризацию исследуемой связи обеспечит переход к новым переменным $\widetilde{y} = 1/y$ и $\widetilde{x} = 1/x$.

Легко видеть, что эти переменные будут связаны между собой зависимостью вида

$$\widetilde{y} = \Theta_0 + \Theta_1 \widetilde{x} + \varepsilon.$$

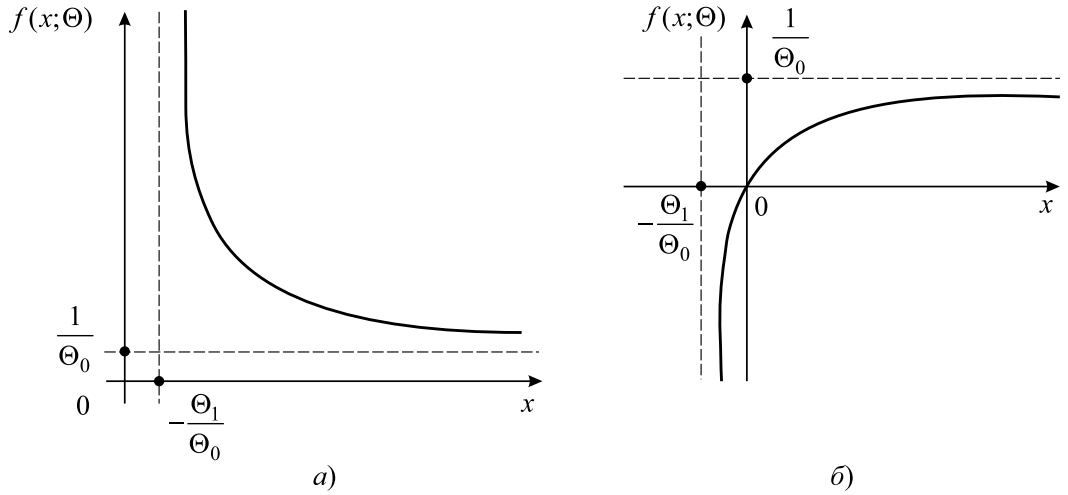
Матрицы $\widetilde{X}$ и $\widetilde{Y}$, используемые в формулах метода наименьших квадратов при вычислении оценок Θ_0 и Θ_1, должны формироваться не из наблюденных значений, соответственно x_i и y_i, а из обратных к ним величин $\widetilde{y}_i = 1/y_i$ и $\widetilde{x}_i = 1/x_i$.

Функции, изображенные на рисунках 4.1 (вариант $\Theta_1 < 0$) и 4.3 (вариант *б*), используются в определенных ситуациях при построении так называемых кривых *Энгеля*, которые описывают зависимость спроса на определенный вид товаров или услуг (y) от уровня доходов (x) потребителей. При этом спрос определяется либо абсолютными, либо относительными расходами на данный вид товаров или услуг. Функции, изображенные на рисунках 4.1 (вариант $\Theta_1 > 0$), 4.2, *а* и 4.3, *а*, могут оказаться полезными при изучении спроса на товар (y) в зависимости от его цены (x).



а) *б*)

Рис. 4.2 – График гиперболической зависимости вида $f(x;\Theta) = 1/(\Theta_0 + \Theta_1 x)$: *а*) случай $\Theta < 0$; $\Theta_1 > 0$ (для $x > -\Theta_0/\Theta_1$); *б*) случай $\Theta_0 > 0$; $\Theta_1 < 0$ (для $x > -\Theta_0/\Theta_1$)



а) *б*)

Рис. 4.3 – График гиперболической зависимости вида $f(x;\Theta) = x/(\Theta_0 x + \Theta_1)$: *а*) случай $\Theta_0 > 0$; $\Theta_1 < 0$ (для $x > -\Theta_1/\Theta_0$); *б*) случай $\Theta_0 > 0$; $\Theta_1 > 0$ (для $x > -\Theta_1/\Theta_0$)

4.4 Зависимости показательного (экспоненциального) типа

Достаточно широкий класс экономических показателей характеризуется приблизительно постоянным темпом относительного прироста во времени. Этому соответствует следующая форма зависимости этого показателя (y) от времени (x):

$$y = \Theta_0 e^{\Theta_1 x + \varepsilon}.$$

Если пренебречь влиянием случайной остаточной компоненты ε (т. е. положить $\varepsilon = 0$, см. рис. 4.4, *а*), то непосредственные расчеты дают:

$$\frac{dy}{dx} = \Theta_1 \Theta_0 e^{\Theta_1 x} = \Theta_1 y,$$

так что относительный прирост y за единицу времени (т. е. за единицу «количества» x) определяется выражением

$$\frac{dy}{dx}\Big/y = \Theta_1 \ \text{(в долях } y\text{)}.$$

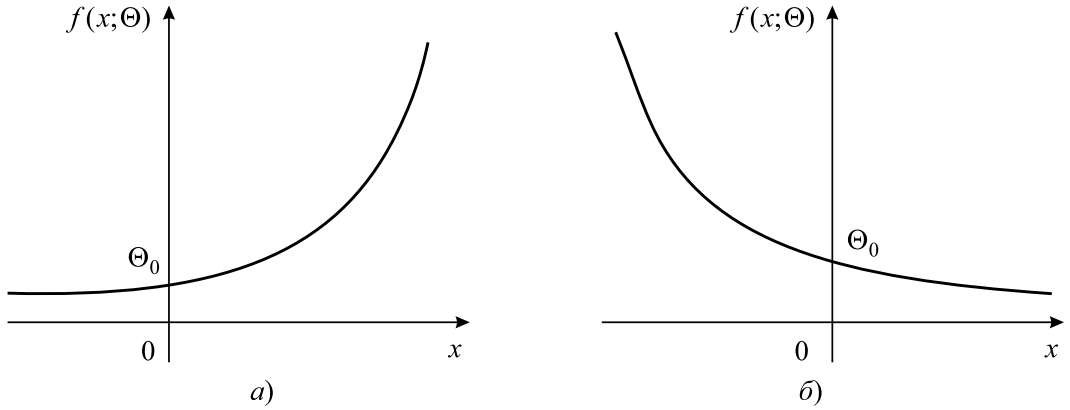


Рис. 4.4 – График показательной (экспоненциальной) зависимости вида $f(x;\Theta) = \Theta_0 e^{\Theta_1 x}$: *а*) случай $\Theta_1 > 0$; *б*) случай $\Theta_1 < 0$

Переход к новой переменной $\widetilde{y} = \ln y$ позволяет свести исследуемую зависимость к линейному виду:

$$\widetilde{y} = \widetilde{\Theta}_0 + \Theta_1 x + \varepsilon,$$

где $\widetilde{\Theta}_0 = \ln \Theta_0$. Располагая наблюдениями $(x_1;y_1),(x_2;y_2),\ldots,(x_n;y_n)$ и формируя вектор-столбец $\widetilde{Y}$ из $\ln y_1, \ln y_2, \ldots, \ln y_n$, мы с помощью МНК можем построить оценки $\widehat{\widetilde{\Theta}}_0$ и $\widehat{\Theta}_1$ параметров $\widetilde{\Theta}_0$ и Θ_1, а затем получить оценку $\widehat{\Theta}_0 = e^{\widehat{\widetilde{\Theta}}}$ для параметра Θ_0 *исходного* уравнения.

Если в результате параметризации модели мы пришли к необходимости исследовать экспоненциальную статистическую зависимость вида

$$y = \Theta_0 e^{\frac{\Theta_1}{x} + \varepsilon},$$

то линеаризация искомой зависимости достигается с помощью следующих преобразований переменных: $\widetilde{y} = \ln y$, $\widetilde{x} = 1/x$.

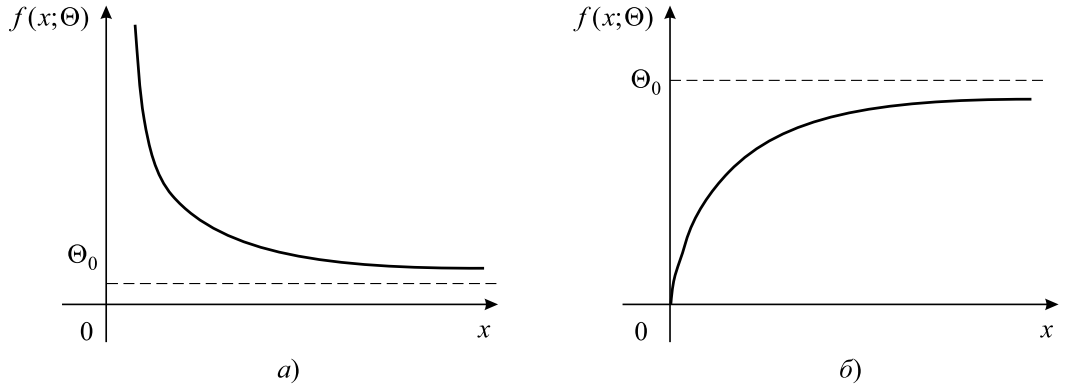
В терминах переменных $(\widetilde{x},\widetilde{y})$ исследуемая зависимость будет иметь вид

$$\widetilde{y} = \widetilde{\Theta}_0 + \Theta_1 \widetilde{x} + \varepsilon,$$

где $\widetilde{\Theta}_0 = \ln \Theta_0$. Соответственно вектор-столбец $\widetilde{Y}$ и матрица $\widetilde{X}$, участвующие в формулах МНК, определяются по исходным наблюдениям $\{(x_i,y_i)\}$, $i = 1,2,\ldots,n$ следующим образом:

$$\widetilde{Y} = \begin{pmatrix} \ln y_1 & \ln y_2 & \ldots & \ln y_n \end{pmatrix}^T; \quad \widetilde{X} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1/x_1 & 1/x_2 & \ldots & 1/x_n \end{pmatrix}^T.$$

Графики показательной зависимости представлены на рисунке 4.5.



а) *б*)

Рис. 4.5 – График показательной (экспоненциальной) зависимости вида $f(x;\Theta) = \Theta_0 e^{\Theta_1/x}$: *а*) случай $\Theta_1 > 0$; *б*) случай $\Theta_1 < 0$

Гибкую форму параметризации искомой регрессионной зависимости представляет один из частных случаев так называемой логистической кривой

$$y = \frac{1}{\Theta_0 + \Theta_1 e^{-x} + \varepsilon} \quad (0 \leqslant x \leqslant \infty).$$

Кривая $f(x;\Theta)$ имеет две горизонтальные асимптомы $y = 0$ и $y = 1/\Theta_0$ и «точку перегиба» ($x_0 = \ln(\Theta_1/\Theta_0), y_0 = 1/2\Theta_0$). Линеаризация этой зависимости производится с помощью перехода к переменным $\widetilde{y} = 1/y$ и $\widetilde{x} = e^{-x}$. Соответственно, вектор-столбец $\widetilde{Y}$ и матрица $\widetilde{X}$, участвующие в формулах МНК, определяются по исходным наблюдениям $\{(x_i, y_i)\}$, $i = 1, 2, \ldots, n$ следующим образом:

$$\widetilde{Y} = \begin{pmatrix} 1/y_1 & 1/y_2 & \ldots & 1/y_n \end{pmatrix}^T; \quad \widetilde{X} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ e^{-x1} & e^{-x2} & \ldots & e^{-xn} \end{pmatrix}.$$

График логистической кривой представлен на рисунке 4.6.

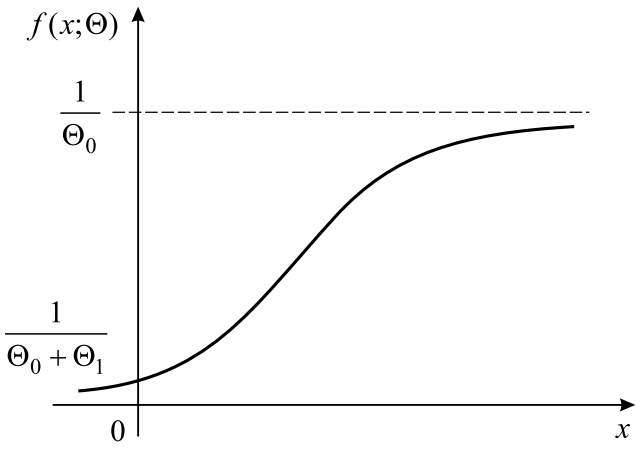


Рис. 4.6 – График логистической кривой вида $f(x;\Theta) = 1/(\Theta_0 + \Theta_1 e^{-x})$ (случай $\Theta_1 > 0$)

Логистические кривые используются для описания поведения показателей, имеющих определенные «уровни насыщения», например для описания зависимости спроса на товар (y) от дохода (x).

4.5 Зависимости степенного типа

Широко распространены в практике социально-экономических исследований так называемые степенные зависимости. Степенная модель множественной регрессии имеет вид

$$y = \Theta_0 \left(x^{(1)}\right)^{\Theta_1} \left(x^{(2)}\right)^{\Theta_2} \ldots \left(x^{(p)}\right)^{\Theta_p} e^{\varepsilon}.$$

При переходе к переменным $\widetilde{y} = \ln y$, $\widetilde{x}^{(j)} = \ln x^{(j)}$ $(j = 1, 2, \ldots p)$ можно представить эту зависимость в виде КЛММР:

$$\widetilde{y} = \widetilde{\Theta}_0 + \Theta_1 \widetilde{x}^{(1)} + \ldots + \Theta_p \widetilde{x}^{(p)} + \varepsilon,$$

где $\widetilde{\Theta}_0 = \ln \Theta_0$. При оценке параметров $\widetilde{\Theta}_0, \Theta_1, \ldots, \Theta_p$ участвующие в формулах МНК вектор-столбец $\widetilde{Y}$ и матрица $\widetilde{X}$ будут определяться по исходным наблюдениям $\left\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\right\}$ $i = 1, 2, \ldots, n$ следующим образом:

$$\widetilde{Y} = \left(\ln y_1 \quad \ln y_2 \quad \ldots \quad \ln y_n\right)^T,$$

а $(j+1)$-й столбец матрицы $\widetilde{X}$ есть $\left(\ln x_1^{(j)} \quad \ln x_2^{(j)} \quad \ldots \quad \ln x_n^{(j)}\right)^T$, $j = 1, 2, \ldots, p$ (первый столбец матрицы $\widetilde{X}$, как обычно, составлен из одних единиц). Графики зависимостей данного типа для случая $p = 1$ представлены на рисунке 4.7.

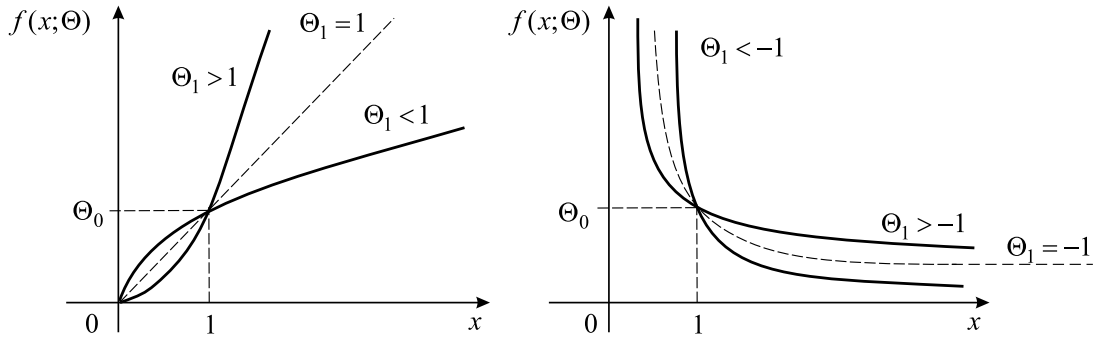


Рис. 4.7 – График степенной зависимости вида $f(x;\Theta) = \Theta_0 x^{\Theta_1}$: *а*) случай $\Theta_1 > 0$; *б*) случай $\Theta_1 < 0$

Важную роль играют зависимости степенного типа в задачах построения и анализа производственных функций (y — объем произведенной продукции, $x^{(1)}, x^{(2)}, \ldots$ — основные факторы производства: труд, капитал и т. д.). Достаточно часто используются степенные зависимости и при построении и анализе функций спроса (y — спрос на определенный вид товаров или услуг, $x^{(1)}$ — доход потребителя, $x^{(2)}, x^{(3)}, \ldots$ — цены на данный и другие виды товаров).

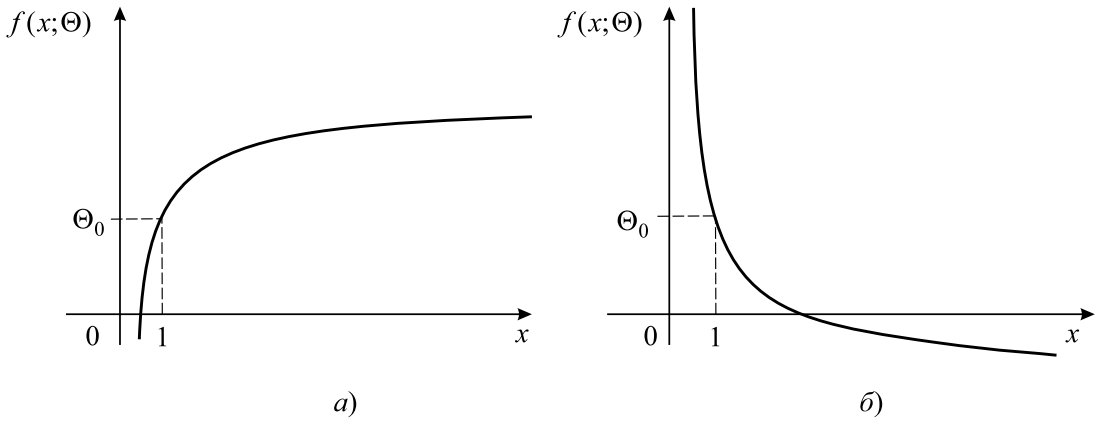
При анализе степенных регрессионных зависимостей прозрачную содержательную интерпретацию получают коэффициенты $\Theta_1, \Theta_2, \ldots, \Theta_p$, а именно: в соответствии с определением коэффициента эластичности признака y по объясняющей переменной $x^{(j)}$ величина $\Theta_j = \partial \ln f(x;\Theta) / \partial \ln x^{(j)}$ есть не что иное, как коэффициент эластичности анализируемого результирующего показателя по j-й объясняющей

переменной. Можно показать, что если эластичность y по каждой из объясняющих переменных $x^{(j)}$ постоянна (т. е. не зависит от того, при каких именно значениях объясняющих переменных она вычисляется), то y и X могут быть связаны только зависимостью степенного типа.

4.6 Зависимости логарифмического типа

На рисунке 4.8 представлены графики зависимостей логарифмического типа

$$y = \Theta_0 + \Theta_1 \ln x + \varepsilon \ (0 < x < \infty).$$



а) *б*)

Рис. 4.8 – График логарифмической зависимости вида $f(x;\Theta) = \Theta_0 + \Theta_1 \ln x$: *а*) случай $\Theta_1 > 0$; *б*) случай $\Theta_1 < 0$

Кривые на рисунке 4.8 проходят через точку $(1;\Theta_0)$ и имеют в качестве вертикальной асимптоты ось y (т. е. прямую $x = 0$). Переход к линейному виду зависимости осуществляется с помощью логарифмического преобразования объясняющей переменной $\widetilde{x} = \ln x$. Второй столбец матрицы $\widetilde{X}$, участвующей в формулах МНК, будет иметь вид $\begin{pmatrix}\ln x_1 & \ln x_2 & \dots & \ln x_n\end{pmatrix}^T$.

4.7 Оценка значимости уравнения регрессии. Коэффициент детерминации

Проверить значимость уравнения регрессии — значит, установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Проверка значимости уравнения регрессии производится на основе дисперсионного анализа.

В математической статистике дисперсионный анализ рассмотрен как самостоятельный инструмент статистического анализа.

Здесь же он применяется как вспомогательное средство для изучения качества регрессионной модели.

Связь между наблюденными значениями y_i и модельными данными $\hat{y}_i$ описывается формулой:

$$y_i = \hat{y}_i + \varepsilon_i,$$

где ε_i — регрессионные остатки, $i = 1, \ldots, n$.

Выполним расчет дисперсии (var):

$$\operatorname{var}(y) = \operatorname{var}(\hat{y} + \varepsilon) = \operatorname{var}(\hat{y}) + \operatorname{var}(\varepsilon) + 2\operatorname{cov}(\hat{y}, \varepsilon) = \operatorname{var}(\hat{y}) + \operatorname{var}(\varepsilon),$$

где $2\operatorname{cov}(\hat{y}, \varepsilon)$ — ковариация $\hat{y}$ и ε, равная нулю.

Используя формулу дисперсии, запишем полученное выражение:

$$\frac{1}{n}\sum (y - \overline{y})^2 = \frac{1}{n}\left(\hat{y} - \overline{\hat{y}}\right)^2 + \frac{1}{n}\sum (\varepsilon - \overline{\varepsilon})^2,$$

где $\overline{y}$, $\overline{\hat{y}}$, $\overline{\varepsilon}$ — средние значения величин y, $\hat{y}$, ε.

Среднее значение модельных данных равно среднему значению наблюденных значений, а согласно требованиям КЛММР $\overline{\varepsilon} = 0$, следовательно:

$$\sum (y - \overline{y})^2 = (\hat{y} - \overline{y})^2 + \sum \varepsilon^2$$

или

$$S_{\text{общ}} = S_{\text{откл}} + S_{\text{ост}},$$

где $S_{\text{общ}}$ — общая сумма квадратов; $S_{\text{откл}}$ — сумма квадратов отклонений; $S_{\text{ост}}$ — остаточная сумма квадратов отклонений.

Одной из наиболее эффективных оценок адекватности регрессионной модели, мерой качества уравнения регрессии, характеристикой прогностической силы анализируемой регрессионной модели является коэффициент детерминации, определяемый по формуле:

$$R^2 = 1 - \frac{S_{\text{ост}}}{S_{\text{общ}}} = \frac{S_{\text{откл}}}{S_{\text{общ}}}.$$

Величина R^2 показывает, какая часть (доля) вариации зависимой переменной обусловлена вариацией объясняющей переменной.

Так как $0 \leqslant S_{\text{откл}} \leqslant S_{\text{общ}}$, то $0 \leqslant R^2 \leqslant 1$.

Чем ближе R^2 к единице, тем лучше регрессия аппроксимирует эмпирические данные, тем теснее наблюдения примыкают к линии регрессии. Если $R^2 = 1$, то эмпирические точки (x_i, y_i) лежат на линии регрессии и между переменными Y и X существует линейная функциональная зависимость, так что $\hat{y}_i = y_i$ для всех i и все остатки равны нулю. Если $R^2 = 0$, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных.

Однако не следует абсолютизировать высокое значение R^2, т. к. коэффициент детерминации может быть близким к единице просто в силу того, что обе исследуемые величины X и Y имеют выраженный временной тренд, не связанный с их причинно-следственной зависимостью. В экономике обычно такой тренд имеет объемные показатели (ВНП, ВВП, доход, потребление). А темповые и относительные показатели (темпы роста, производительность, ставка процента) не всегда имеют тренд. Поэтому при оценивании регрессий по временным рядам объемных показателей (например, зависимость потребления от дохода или спроса от

цены) величина R^2 может быть весьма близкой к единице. Но это не обязательно свидетельствует о наличии значимой линейной связи между исследуемыми показателями, а может означать лишь то, что поведение зависимой переменной нельзя описать уравнением $Y = \overline{y}$.

Если уравнение регрессии строится по перекрестным данным, а не по временным рядам, то коэффициент детерминации R^2 для него обычно не превышает 0.6–0.7. Аналогичные значения R^2 обычно получаются и для регрессий по временным рядам, если они не имеют выраженного тренда (темп инфляции от уровня безработицы, темпы прироста выпуска от темпов прироста затрат ресурсов и т. п.).

Естественно, возникает вопрос, какое значение R^2 можно считать удовлетворительным. Точную границу приемлемости (статистической значимости) R^2 для всех случаев сразу указать невозможно. Нужно обращать внимание на объем выборки, число объясняющих переменных, наличие трендов и содержательную интерпретацию. R^2 может оказаться даже отрицательным. Обычно это случается для линейных уравнений регрессии, в которых отсутствует свободный член $Y = \sum_{j=1}^{p} \Theta_j x^{(j)}$. Оценивая такое уравнение по МНК, мы вынуждены рассматривать лишь те прямые (гиперплоскости), которые проходят через начало координат (рис. 4.9). Значение R^2 получается отрицательным тогда, когда разброс значений зависимой переменной вокруг линии $Y = \overline{y}$ меньше, чем вокруг любой из прямых (гиперплоскостей), проходящих через начало координат.

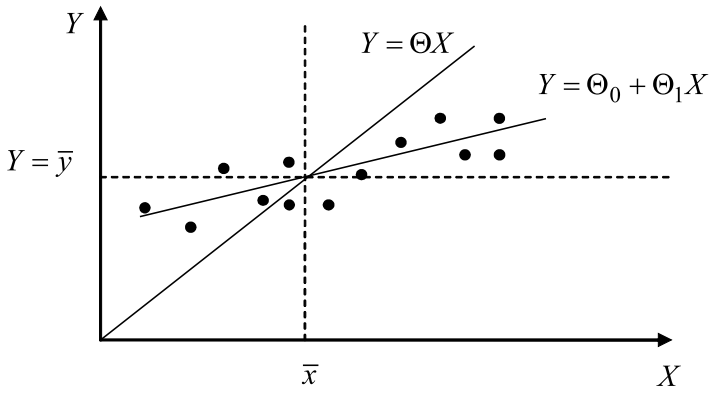


Рис. 4.9 – Графики регрессий с постоянным членом и без него

Из рисунка 4.9 видно, что разброс наблюдаемых значений переменной Y относительно прямой $Y = \overline{y}$ существенно меньше разброса относительно прямой $Y = \Theta X$. Отрицательное значение R^2 свидетельствует о целесообразности добавления в уравнение $Y = \sum_{j=1}^{p} \Theta_j x^{(j)}$ свободного члена ($Y = \Theta_0 + \Theta_1 X$, см. рис. 4.9).

Если индекс детерминации рассчитывается для уравнений регрессии, в которых меняется число коэффициентов, то для сравнения целесообразно использовать приведенный коэффициент детерминации $\widehat{\rho}^2$.

$$\widehat{R}^2 = 1 - \frac{(n-1)\cdot S_{\text{ост}}}{(n-m)\cdot S_{\text{общ}}} = 1 - \frac{n-1}{n-m}\cdot(1-R^2),$$

где m — количество вычисляемых коэффициентов регрессии. При неизменных $S_{\text{ост}}$, $S_{\text{общ}}$ увеличение m уменьшает значение $\widehat{R}^2$. Если количество коэффициентов у сравниваемых уравнений регрессии одинаково (например, $m = 2$), то отбор наилучшей регрессии можно осуществлять по величине R^2.

В случае парной линейной регрессионной модели коэффициент детерминации равен квадрату коэффициента корреляции, т. е. $R^2 = r^2$.

Коэффициент корреляции r вычисляется по формуле:

$$r = \Theta_1 \frac{\sigma_x}{\sigma_y} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}, \qquad (4.1)$$

где σ_x — среднее квадратическое отклонение x; σ_y — среднее квадратическое отклонение y.

Коэффициент корреляции принимает значения на отрезке $[-1; 1]$, т. е. — $0 \leqslant r \leqslant 1$.

Чем ближе r к единице, тем теснее связь (рис. 4.10). При $r = \pm 1$ корреляционная связь представляет линейную функциональную зависимость. При $r = 0$ линейная корреляционная связь отсутствует.

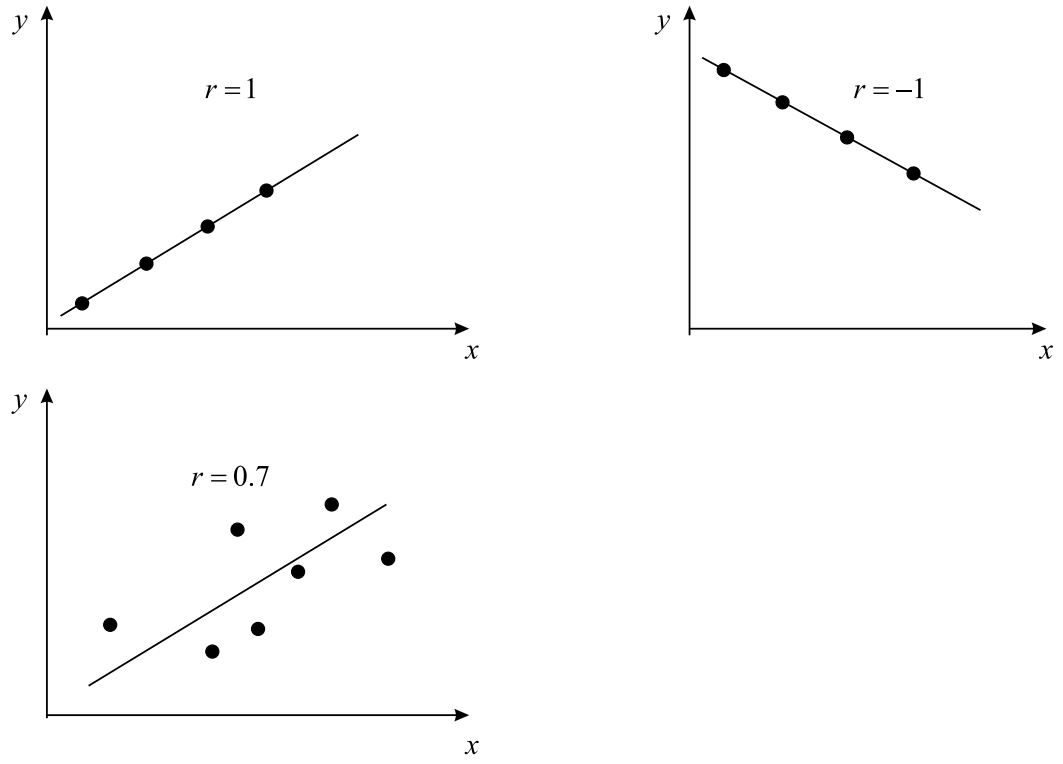


Рис. 4.10 – Виды корреляционной связи

Для оценки значимости уравнения регрессии также анализируется совокупная значимость коэффициентов. Такой анализ осуществляется на основе проверки *гипотезы об общей значимости* — гипотезы об одновременном равенстве нулю всех коэффициентов регрессии при объясняющих переменных:

$$H_0\colon\ \Theta_1 = \Theta_1 = \ldots = \Theta_{m-1} = 0.$$

Если данная гипотеза не отклоняется, то делается вывод о том, что совокупное влияние всех переменных модели на зависимую переменную Y можно считать статистически несущественным, а общее качество уравнения регрессии — невысоким.

Проверка данной гипотезы осуществляется на основе дисперсионного анализа — сравнения объясненной и остаточной дисперсий.

H_0: (объясненная дисперсия) = (остаточная дисперсия),

H_1: (объясненная дисперсия) > (остаточная дисперсия).

При отсутствии линейной зависимости между зависимой и объясняющими (ей) переменными случайные величины $s^2_{\text{откл}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2/(m-1)$ и $s^2_{\text{ост}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-m)$ имеют λ^2-распределение соответственно с $(m-1)$ и $(n-m)$ степенями свободы, а их отношение — F-распределение с теми же степенями свободы. Поэтому уравнение регрессии значимо на уровне α, если фактически наблюдаемое значение статистики

$$F = \frac{S_{\text{откл}}(n-m)}{S_{\text{ост}}(m-1)} = \frac{s^2_{\text{откл}}}{s^2_{\text{ост}}} > F_{\alpha;k_1;k_2},$$

где $F_{\alpha;k_1;k_2}$ — табличное значение F-критерия Фишера—Снедекора, определенное на уровне значимости α при $k_1 = m - 1$ и $k_2 = n - m$ степенях свободы.

Учитывая смысл величин $s^2_{\text{откл}}$ и $s^2_{\text{ост}}$, можно сказать, что значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению с ее средней.

В случае линейной парной регрессии $m = 2$ уравнение регрессии значимо на уровне α, если

$$F = \frac{S_{\text{откл}}(n-2)}{S_{\text{ост}}} > F_{\alpha;1;n-2}.$$

Следует отметить, что значимость уравнения парной линейной регрессии может быть проведена и другим способом, если оценить значимость коэффициента регрессии Θ_1, который имеет t-распределение Стьюдента с $k = n - 2$ степенями свободы.

Уравнение парной линейной регрессии или коэффициент регрессии Θ_1 значимы на уровне α (иначе — гипотеза H_0 о равенстве параметра Θ_1 нулю, т. е. H_0: $\Theta_1 = 0$, отвергается), если фактически наблюдаемое значение статистики

$$t = \frac{\Theta_1 - 0}{s}\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

больше критического (по абсолютной величине), т. е. $|t| > t_{1-\alpha;n-2}$.

Для парной линейной модели оба способа проверки значимости с использованием F- и t-критериев равносильны, т. к. эти критерии связаны соотношением $F = t^2$.

Вместо указанной гипотезы также проверяют тесно связанную с ней гипотезу о статистической значимости коэффициента детерминации R^2:

$$H_0\colon\ R^2 = 0,$$
$$H_0\colon\ R^2 > 0.$$

Для проверки данной гипотезы используется следующая F-статистика:

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

Показатели F и R^2 равны или не равны нулю одновременно. Если $F = 0$, то $R^2 = 0$, и линия регрессии $Y = y$ является наилучшей по МНК, и, следовательно, величина Y линейно не зависит от X. Для проверки нулевой гипотезы H_0: $F = 0$ при заданном уровне значимости α по таблицам критических точек распределения Фишера находится критическое значение $F_{\alpha; m-1; n-m}$. Нулевая гипотеза отклоняется, если $F > F_{кр}$. Это равносильно тому, что $R^2 > 0$, т. е. R^2 статистически значим.

Пример 4.1

Используя пространственную выборку таблицы 4.1, найти наилучшее уравнение нелинейной регрессии. Рассмотреть два вида регрессии:

- *степенную* $\hat{y} = \Theta_0 \cdot x^{\Theta_1}$;
- *экспоненциальную* $\hat{y} = \Theta_0 e^{\Theta_1 x}$.

Таблица 4.1 – Исходные данные

x_i	1	2	3	4
y_i	10	13	15	16

Решение:

1. Выполним расчет индекса детерминации для степенной функции регрессии.

Прежде всего, нужно найти оценки неизвестных параметров. Для этого запишем исходные данные в виде двух матриц:

$$X_1 = \begin{pmatrix} 1 & \ln(x_1) \\ 1 & \ln(x_2) \\ 1 & \ln(x_3) \\ 1 & \ln(x_4) \end{pmatrix} = \begin{pmatrix} 1 & \ln(1) \\ 1 & \ln(2) \\ 1 & \ln(3) \\ 1 & \ln(4) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0.693 \\ 1 & 1.099 \\ 1 & 1.386 \end{pmatrix},$$

$$Y_1 = \begin{pmatrix} \ln(y_1) & \ln(y_2) & \ln(y_3) & \ln(y_4) \end{pmatrix}^T = \begin{pmatrix} \ln(10) & \ln(13) & \ln(15) & \ln(16) \end{pmatrix}^T =$$
$$= \begin{pmatrix} 2.303 & 2.565 & 2.708 & 2.773 \end{pmatrix}^T.$$

Далее воспользуемся методом наименьших квадратов для случая парной регрессии ($n = 4$):

$$X_1^T X_1 = \begin{pmatrix} n & \sum_{i=1}^{n} \ln(x_i) \\ \sum_{i=1}^{n} \ln(x_i) & \sum_{i=1}^{n} \ln(x_i)^2 \end{pmatrix} = \begin{pmatrix} 4 & 3.178 \\ 3.178 & 3.608 \end{pmatrix},$$

$$X_1^T Y_1 = \begin{pmatrix} \sum_{i=1}^{n} \ln(y_i) \\ \sum_{i=1}^{n} \ln(x_i)\ln(y_i) \end{pmatrix} = \begin{pmatrix} 10.348 \\ 8.597 \end{pmatrix}.$$

Для нахождения обратной матрицы воспользуемся формулой:

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix},$$

$$\left(X_1^T X_1\right)^{-1} = \begin{pmatrix} 0.832 & -0.733 \\ -0.733 & 0.922 \end{pmatrix}.$$

Для нахождения оценок перемножим полученные матрицы:

$$\left(X_1^T X_1\right)^{-1} X_1^T Y_1 = \begin{pmatrix} 0.832 & -0.733 \\ -0.733 & 0.922 \end{pmatrix}\begin{pmatrix} 10.348 \\ 8.597 \end{pmatrix} = \begin{pmatrix} 2.312 \\ 0.346 \end{pmatrix}.$$

Умножение выполняется по правилу:

$$\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} a_1c_1 + b_1c_2 \\ a_2c_1 + b_2c_2 \end{pmatrix}.$$

Определим оценки $\widehat{\Theta}$:

$$\widehat{\Theta}_0 = e^{2.312} = 10.098,$$

$$\widehat{\Theta}_1 = 0.346.$$

Уравнение степенной регрессии имеет вид:

$$\widehat{y} = 10.098 \cdot x^{0.346}.$$

Подставив значения x из таблицы в уравнение регрессии, рассчитаем оценки $\widehat{y}$:

$$\widehat{Y} = \begin{pmatrix} 10.098 \\ 12.833 \\ 14.764 \\ 16.308 \end{pmatrix}.$$

Рассчитаем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = \sum_{i=1}^{4} (y_i - \widehat{y}_i)^2 = 0.18812.$$

Вычислим общую сумму квадратов:

$$MY = \frac{\sum_{i=1}^{4} y_i}{4} = 13.5,$$

$$S_{\text{общ}} = \sum_{i=1}^{4} (y_i - MY)^2 = 21.$$

Индекс детерминации равен:

$$R^2 = 1 - \frac{S_{\text{ост}}}{S_{\text{общ}}} = 1 - \frac{0.18812}{21} = 0.991.$$

2. Выполним аналогичные расчеты для экспоненциальной функции регрессии. Исходные данные:

$$X_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix},$$

$$Y_2 = \begin{pmatrix} \ln(y_1) & \ln(y_2) & \ln(y_3) & \ln(y_4) \end{pmatrix}^T = \begin{pmatrix} \ln(10) & \ln(13) & \ln(15) & \ln(16) \end{pmatrix}^T =$$
$$= \begin{pmatrix} 2.303 & 2.565 & 2.708 & 2.773 \end{pmatrix}^T.$$

Найдем произведения матриц:

$$X_2^T X_2 = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix},$$

$$X_2^T Y_2 = \begin{pmatrix} \sum_{i=1}^{n} \ln(y_i) \\ \sum_{i=1}^{n} x_i \ln(y_i) \end{pmatrix} = \begin{pmatrix} 10.348 \\ 26.647 \end{pmatrix},$$

$$\left(X_2^T X_2\right)^{-1} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix},$$

$$\left(X_2^T X_2\right)^{-1} X_2^T Y_2 = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix} \begin{pmatrix} 10.348 \\ 26.647 \end{pmatrix} = \begin{pmatrix} 2.199 \\ 0.155 \end{pmatrix}.$$

Определим оценки $\widehat{\Theta}$:

$$\widehat{\Theta}_0 = e^{2.199} = 9.014,$$

$$\widehat{\Theta}_1 = 0.155.$$

Уравнение экспоненциальной регрессии имеет вид:

$$\widehat{y} = 9.014 \cdot e^{0.155x}.$$

Подставив значения x из таблицы в уравнение регрессии, рассчитаем оценки $\widehat{y}$:

$$\widehat{Y} = \begin{pmatrix} 10.528 \\ 12.297 \\ 14.364 \\ 16.777 \end{pmatrix}.$$

Рассчитаем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = \sum_{i=1}^{4} \left(y_i - \widehat{y}_i\right)^2 = 1.782.$$

Общая сумма квадратов равна вычисленному для степенной функции значению:

$$S_{\text{общ}} = \sum_{i=1}^{4} \left(y_i - MY\right)^2 = 21.$$

Индекс детерминации равен:

$$R^2 = 1 - \frac{S_{\text{ост}}}{S_{\text{общ}}} = 1 - \frac{1.782}{21} = 0.915.$$

В ходе выполненных вычислений получили, что степенная функция регрессии описывает данные наилучшим образом, т. к. индекс детерминации степенной функции равен 0.991, а экспоненциальной — 0.915.

4.8 Подбор линеаризующего преобразования (подход Бокса—Кокса)

В предыдущем пункте описан набор зависимостей, поддающихся линеаризации с помощью подходящих преобразований анализируемых переменных. Но решение вопроса о том, к какому именно из перечисленных линеаризуемых типов зависимостей следует отнести наш конкретный случай, является задачей не простой. Можно, конечно, действовать методом «проб и ошибок»: последовательно построить по имеющимся у нас исходным статистическим данным каждую из альтернативного набора линеаризуемых моделей, а затем выбрать из них наилучшую в смысле какого-то «критерия качества» (например, по максимальному значению подправленной на несмещенность оценки коэффициента детерминации R).

Английские статистики Г. Бокс и Л. Кокс предложили более формализованную процедуру подбора линеаризующего преобразования. Их метод основан на предположении, что искомое преобразование принадлежит определенному однопараметрическому семейству преобразований вида

$$\widetilde{y}_i(\lambda) = \frac{y_i^{\lambda} - 1}{\lambda}, \quad \widetilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)}\right)^{\lambda} - 1}{\lambda}, \quad i = 1, 2, \ldots n. \tag{4.2}$$

Точнее их гипотезу можно сформулировать следующим образом: существует такое вещественное (положительное или отрицательное) число λ^*, что один из двух нижеследующих вариантов представления искомой регрессионной зависимости между наблюдаемыми переменными y и $X = \left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right)$:

$$\widetilde{y}_i\left(\lambda^*\right) = \Theta_0 + \Theta_1 \widetilde{x}_i^{(1)}\left(\lambda^*\right) + \ldots + \Theta_p \widetilde{x}_i^{(p)}(\lambda^*) + \varepsilon_i \tag{4.3}$$

или

$$\widetilde{y}_i\left(\lambda^*\right) = \Theta_0 + \Theta_1 x_i^{(1)} + \ldots + \Theta_p x_i^{(p)} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

будет удовлетворять всем требованиям нормальной классической линейной модели множественной регрессии.

Замечание 1. Преобразования вида (4.2) применяются обычно к переменным, принимающим только положительные значения. Поэтому если это не так, то вначале подбирают «сдвиговые» константы $c^{(0)}, c^{(1)}, \ldots, c^{(p)}$, которые обеспечивают положительность значений $y_i + c^{(0)}$ и $x_i^{(j)} + c^{(j)}$ $(j = 1, 2, \ldots, p)$, а затем к сдвинутым значениям переменных применяют данное преобразование, т. е.:

$$\widetilde{y}_i(\lambda) = \frac{\left(y_i + c^{(0)}\right)^{\lambda} - 1}{\lambda}, \quad \widetilde{x}_i(\lambda) = \frac{\left(x_i^{(j)} + c^{(j)}\right)^{\lambda} - 1}{\lambda} \ (i = 1, 2, \ldots, n).$$

Замечание 2. Семейство степенных преобразований вида (4.2) весьма широко и гибко. При $\lambda = 1$ модели (4.3) являются линейными относительно y_i и $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$. При $\lambda = 0$ мы имеем степенную зависимость между y и X, поскольку $\widetilde{y}_i(0) = \lim\limits_{\lambda \to 0} \left(y_i^{\lambda} - 1\right)/\lambda = \ln y_i$ и $\widetilde{x}_i(0) = \lim\limits_{\lambda \to 0} \left(\left(x_i^{(j)}\right)^{\lambda} - 1\right)/\lambda = \ln x_i^{(j)}$. При других значениях λ уравнения (4.3) будут связывать между собой какие-то степени исходных переменных.

Оценка неизвестного значения параметра λ. Таким образом, если исходить из справедливости сформулированной выше гипотезы, подбор линеаризующего преобразования анализируемых переменных сводится к оценке параметра λ по имеющимся в нашем распоряжении исходным статистическим данным. Один из способов решения этой проблемы — с помощью метода максимального правдоподобия.

С этой целью определяется априорный диапазон $(\lambda_{\min}, \lambda_{\max})$ возможных значений λ (обычно достаточно рассмотреть в качестве области возможных значений λ отрезок от $\lambda_{\min} = -1$ до $\lambda_{\max} = 2$), на этом диапазоне выбирается сетка («решето») значений $\lambda_i = \lambda_{\min} + i(\lambda_{\max} - \lambda_{\min})/N$, $i = 0, 1, \ldots, N$ и для каждого такого значения λ_i, последовательно вычисляются $\widehat{\Theta}(\lambda_i)$, $\widehat{\sigma}^2(\lambda_i)$, $l_{\max}(\lambda_i)$, то значение λ^*, при котором

$$l_{\max}\left(\lambda^*\right) = \max_{\lambda = \lambda_0, \ldots, \lambda_N} l_{\max}(\lambda_i)$$

и будет определять искомое линеаризующее преобразование. Оценки λ^*, $\widehat{\Theta}(\lambda^*)$, $\widehat{\sigma}^2(\lambda^*)$ являются оценками метода максимального правдоподобия, а процедуру их поиска часто называют «решетчатой».

Также может быть использована следующая процедура выбора формы зависимости.

Шаг 1. Вычисляем среднее геометрическое значений зависимой переменной, и все её значения делим на это среднее:

$$y_i^* = \frac{y_i}{\sqrt[n]{y_1 y_2 \ldots y_n}}.$$

Шаг 2. Рассчитываем новые переменные (преобразование Бокса—Кокса) при значениях λ от 0 до 1:

$$y_{i(B-C)} = \frac{\left(y_i^{*\lambda} - 1\right)}{\lambda}, \quad x_{i(B-C)} = \frac{\left(x_i^{\lambda} - 1\right)}{\lambda}.$$

Шаг 3. Рассчитываем регрессии для новых переменных при значениях λ от 0 до 1:

$$y_{i(B-C)} = \Theta_0 + \Theta_1 x_{i(B-C)} + \varepsilon_i.$$

Шаг 4. Выбираем минимальное значение суммы квадратов остатков, выбираем одну из крайних регрессий, к которой ближе точка минимума.

4.9 Тест Зарембки

Для выбора функции регрессии могут быть использованы тесты. Рассмотрим тест Зарембки, который позволяет сравнить линейную и логарифмическую регрессии и оценить значимость наблюдаемых различий [5].

Шаг 1. Вычисляем среднее геометрическое значений зависимой переменной, и все её значения делим на это среднее:

$$y_i^* = \frac{y_i}{\sqrt[n]{y_1 y_2 \cdots y_n}}.$$

Шаг 2. Рассчитываем линейную ($y_i^* = \Theta_0 + \Theta_1 x_i + \varepsilon$) и логарифмическую регрессии ($\ln y_i^* = \Theta_0 + \Theta_1 \ln x_i + \varepsilon$) и сравниваем значения их суммы квадратов остатков.

Шаг 3. Вычисляем хи-квадрат статистику для оценки значимости различий:

$$\chi^2 = \frac{n}{2}\left|\ln \frac{S1_{\text{ост}}}{S2_{\text{ост}}}\right|.$$

Шаг 4. Сравниваем с критическим значением хи-квадрат распределения с одной степенью свободы, различия различимы, если $\chi^2 > \chi^2_{\text{кр}}$ (например, для уровня значимости 0.95 значение равно 0.00393).

Контрольные вопросы по главе 4

1. Какие примеры нелинейных регрессионных моделей Вам известны?
2. Какие должны быть дальнейшие действия, если получена нелинейная модель регрессия?
3. Какую часть исследования называют процедурой линеаризацией модели?
4. Какие Вам известны регрессионные зависимости, поддающиеся линеаризации?
5. Каким рекомендациям желательно следовать при выборе вида функции?
6. С помощью какого преобразования объясняющей переменной зависимость гиперболического типа $f(x,\Theta) = \Theta_0 + \Theta_1/x$ можно привести к линейному виду, если $0 < x < \infty$?
7. Какую переменную следует рассматривать в качестве результирующего признака при гиперболической зависимости вида $f(x,\Theta) = 1/(\Theta_0 + \Theta_1 x)$, $x > -\Theta_0/\Theta_1$?

8. Какой переход к новым переменным обеспечит линеаризацию гиперболической зависимости вида $f(x,\Theta) = x/(\Theta_0 x + \Theta_1)$, $x > -\Theta_1/\Theta_0$?
9. В каких ситуациях используются функции гиперболической зависимости?
10. Как должны формироваться матрицы $\widetilde{X}$ и $\widetilde{Y}$, используемые в формулах МНК, при переходе к новым переменным гиперболической зависимости?
11. Что представляют собой графики показательной зависимости для случаев $\Theta_1 > 1$ и $\Theta_1 < 1$?
12. Какой переход к новой переменной позволяет свести показательную зависимость $f(x,\Theta) = \Theta_0 e^{\Theta_1 x}$ к линейному виду (случаи $\Theta_1 > 0$ и $\Theta_1 < 0$)?
13. С помощью каких преобразований переменных осуществляется линеаризация зависимости $f(x,\Theta) = \Theta_0 e^{\Theta_1 x}$ (случаи $\Theta_1 > 0$ и $\Theta_1 < 0$)?
14. Как определяются матрицы $\widetilde{X}$ и $\widetilde{Y}$, используемые в формулах МНК, для случая зависимости $f(x,\Theta) = \Theta_0 e^{\Theta_1 x}$?
15. Какой вид имеет график логистической кривой вида $f(x,\Theta) = 1/(\Theta_0 + \Theta_1 e^{-x})$, $\Theta_1 > 0$?
16. С помощью каких преобразований переменных осуществляется линеаризация зависимости $f(x,\Theta) = 1/(\Theta_0 + \Theta_1 e^{-x})$, $\Theta_1 > 0$ ($\Theta_1 > 0$)?
17. Как определяются матрицы $\widetilde{X}$ и $\widetilde{Y}$, используемые в формулах МНК, для случая зависимости $f(x,\Theta) = 1/(\Theta_0 + \Theta_1 e^{-x})$, $\Theta_1 > 0$?
18. Где используются логистические кривые?
19. Какой вид имеет график степенной зависимости $f(x,\Theta) = \Theta_0 x^{\Theta_1}$, $\Theta_1 > 0$, $\Theta_1 < 0$?
20. С помощью каких преобразований переменных осуществляется линеаризация зависимости $f(x,\Theta) = \Theta_0 x^{\Theta_1}$?
21. Где используются степенные зависимости?
22. Какой вид имеет график логарифмической зависимости?
23. С помощью какого преобразования осуществляется переход логарифмической зависимости к линейному виду?
24. Как определяется общая сумма квадратов?
25. Как рассчитывается коэффициент детерминации?
26. Какие значения может принимать коэффициент детерминации?
27. В каком случае значение коэффициента детерминации может оказаться отрицательным?
28. Как рассчитывается приведенный коэффициент детерминации?
29. Когда целесообразно использовать приведенный коэффициент детерминации?
30. Каким образом можно вычислить коэффициент корреляции?
31. Какие значения может принимать коэффициент корреляции?
32. Как можно оценить значимость уравнения регрессии?
33. Как осуществляется подбор линеаризующего преобразования с помощью подхода Бокса—Кокса?
34. Как выполняется тест Зарембки?
35. Какие формы зависимости можно выбрать с помощью теста Зарембки?

Глава 5

ГЕТЕРОСКЕДАСТИЧНОСТЬ

5.1 Понятие гетероскедастичности

Равенство дисперсий возмущений (ошибок) регрессии (гомоскедастичность) является существенным условием линейной классической регрессионной модели множественной регрессии, записываемым в виде $M\varepsilon_i^2 = \sigma^2$.

Однако на практике это условие нередко нарушается, и мы имеем дело с гетероскедастичностью модели (в том случае условие будет записано $M\varepsilon_i^2 = \sigma_i^2$).

Пример 5.1

Предположим, что необходимо изучить зависимость размера оплаты труда Y (в усл. ден. ед.) сотрудников фирмы от разряда X, принимающего значения от 1 до 10. Получены $n = 100$ пар наблюдений (x_i, y_i). График зависимости переменной Y от номеров наблюдений, упорядоченных по возрастанию уровня значений объясняющей переменной X, показан на рисунке 5.1.

Из графика видно, что вариация размера оплаты труда сотрудников высоких уровней значительно превосходит его вариацию для сотрудников низких уровней. Следовательно, можно предположить, что регрессионная модель получится гетероскедастичной и условие $M\varepsilon_i^2 = \sigma^2$ не выполняется.

Гетероскедастичность чаще всего встречается в пространственных выборках, а также во временных рядах, когда зависимая переменная имеет большой интервал качественно неоднородных значений или высокий темп изменения.

Гетероскедастичность увеличивает дисперсию распределения оценок коэффициентов.

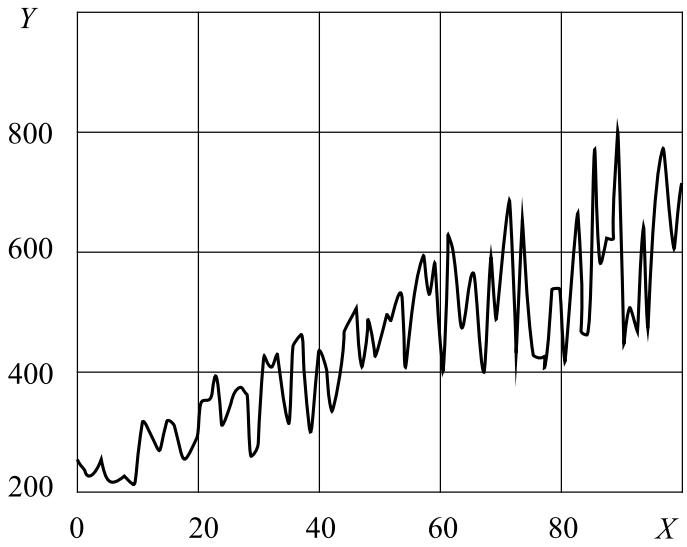


Рис. 5.1 – График зависимости Y от номеров наблюдений

5.2 Графический анализ остатков

Использование графического представления отклонений позволяет определиться с наличием гетероскедастичности. В этом случае по оси абсцисс откладывается объясняющая переменная X (либо линейная комбинация объясняющих переменных $Y = \Theta_0 + \Theta_1 X^{(1)} + \ldots + \Theta_p X^{(p)}$), а по оси ординат — либо отклонения ε_i, либо их квадраты ε_i^2. Примеры таких графиков приведены на рисунке 5.2.

На рисунке 5.2, *а* все отклонения ε_i^2 находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс. Это говорит о независимости дисперсий ε_i^2 от значений переменной X и их постоянстве, т. е. в этом случае мы находимся в условиях гомоскедастичности.

На рисунке 5.2, *б–г* наблюдаются некие систематические изменения в соотношениях между значениями x_i переменной X и квадратами отклонений ε_i^2. На рисунке 5.2, *в* отражена линейная; 5.2, *г* — квадратичная; 5.2, *д* — гиперболическая зависимости между квадратами отклонений и значениями объясняющей переменной X. Другими словами, ситуации, представленные на рисунке 5.2, *б–д*, отражают большую вероятность наличия гетероскедастичности для рассматриваемых статистических данных.

Отметим, что графический анализ отклонений является удобным и достаточно надежным в случае парной регрессии. При множественной регрессии графический анализ возможен для каждой из объясняющих переменных $X^{(j)}$, $j = 1,\ldots,p$ отдельно. Чаще же вместо объясняющих переменных $X^{(j)}$ по оси абсцисс откладывают значения $\hat{y}_i$, получаемые из эмпирического уравнения регрессии. Поскольку по уравнению множественной линейной регрессии $\hat{y}_i$ является линейной комбинацией $x_i^{(j)}$, $j = 1,\ldots,p$, то график, отражающий зависимость ε_i^2 от $\hat{y}_i$, может указать на

наличие гетероскедастичности аналогично ситуациям на рисунке 5.2, *б–д*. Такой анализ наиболее целесообразен при большом количестве объясняющих переменных.

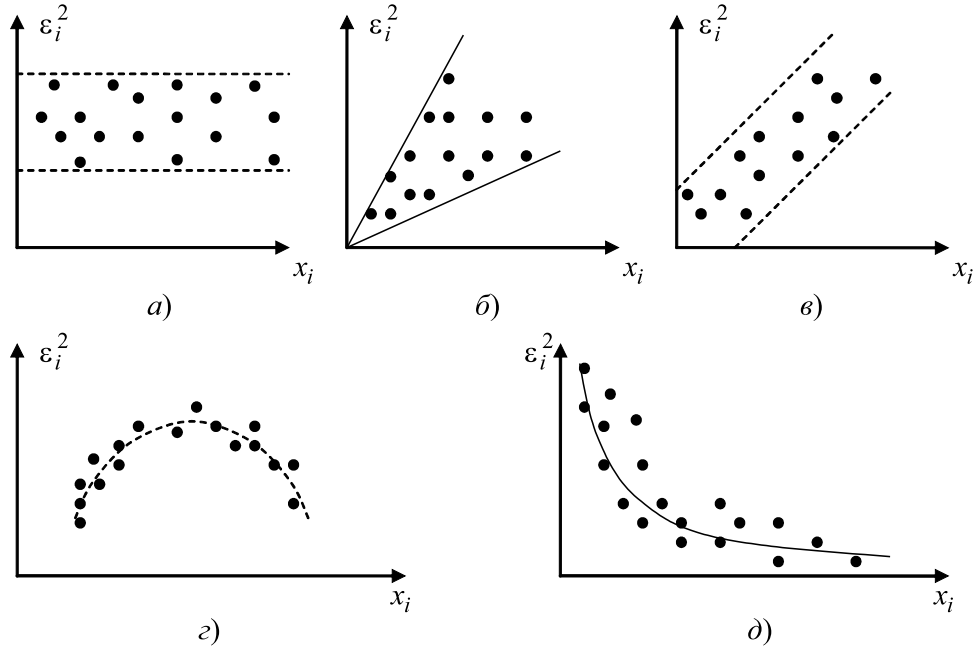


Рис. 5.2 – Графики остатков

5.3 Тесты на гетероскедастичность

В примере 5.1 наличие гетероскедастичности не вызывает сомнения, — чтобы убедиться в этом, достаточно взглянуть на рисунке 5.1. Однако в некоторых случаях гетероскедастичность визуально не столь очевидна.

Рассмотрим еще один пример, в котором исследуется зависимость дохода индивидуума (Y) от уровня его образования X_1, принимающего значения от 1 до 5, по данным n = 150 наблюдений. В число объясняющих переменных (регрессоров) включен также и возраст X_2.

На рисунке 5.3 приведен график зависимости переменной Y от номеров наблюдений, упорядоченных по возрастанию уровня значений объясняющей переменной X_1.

Хотя диаграмма имеет локально расположенные пики, в целом подобный рисунок может соответствовать как гомо-, так и гетероскедастичной выборке.

Чтобы определить, какая же именно ситуация имеет место, используются тесты на гетероскедастичностъ. Все они используют в качестве нулевой гипотезы H_0 гипотезу об отсутствии гетероскедастичности.

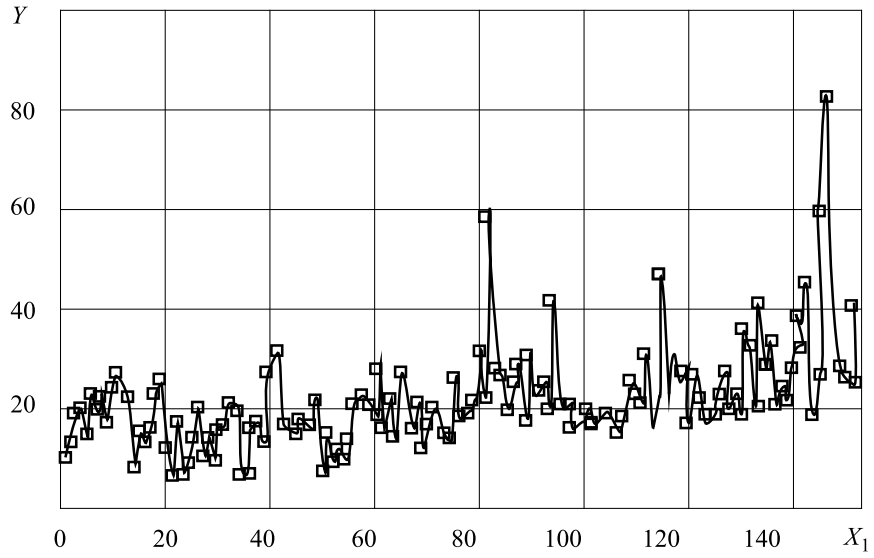


Рис. 5.3 – График зависимости переменной Y от номеров наблюдений, упорядоченных по возрастанию уровня значений объясняющей переменной X_1

5.3.1 Тест ранговой корреляции Спирмена

Этот тест использует наиболее общие предположения о зависимости дисперсий ошибок регрессии от значений регрессоров: $\sigma_i^2 = f_i(x_i)$, $i = 1, \ldots, n$.

При этом никаких дополнительных предположений относительно вида функций f_i не делается. Не накладываются также ограничения на закон распределения возмущений (ошибок) регрессии.

Идея теста заключается в том, что абсолютные величины остатков регрессии ε_i являются оценками σ_i, поэтому в случае гетероскедастичности абсолютные величины остатков ε_i и значения регрессоров x_i будут коррелированны.

Тест состоит из следующих шагов.

Шаг 1. Выборка упорядочивается по фактору x. Рассчитываются ранги x (*порядковый номер*).

Шаг 2. Рассчитывается уравнение регрессии $y_i = \Theta_0 + \Theta_1 x + \varepsilon_i$.

Шаг 3. Вычисляются остатки $\varepsilon_i = y_i - \widehat{y}_i$.

Шаг 4. Выборка снова упорядочивается по величине остатков (по модулю). Рассчитываются ранги (*порядковый номер*) остатков.

Шаг 5. Рассчитывается коэффициент ранговой корреляции Спирмена:

$$r_{x,\varepsilon} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

где d_i — разность рангов x и ε.

Шаг 6. Рассчитывается статистика

$$t_r = \frac{r_{x,\varepsilon} \cdot \sqrt{n-1}}{\sqrt{1 - r_{x,\varepsilon}^2}}.$$

Шаг 7. Проверяется гипотеза.

Если в модели регрессии имеется более одной объясняющей переменной, то проверка гипотезы может выполняться с использованием каждой из них.

Условие принятия гипотез: $t_r > t_{\alpha, n-2}$.

Если данное условие выполняется, то нулевая гипотеза об отсутствии гетероскедастичности отклоняется при уровне значимости α.

Пример 5.2

В таблице 5.1 представлена зависимость спроса Y от цены X. Выполнить тест на гетероскедастичность с помощью теста Спирмена.

Таблица 5.1 – Исходные данные

Цена, *X*	15.91	15.54	16.76	15.21	15.28	15.92	15.95	16.69
Спрос, *Y*	117.088	119.864	110.023	123.809	121.175	116.17	118.344	110.106
Цена, *X*	15.09	15.62	16.31	16.33	16.60	15.49	15.70	
Спрос, *Y*	125.178	118.068	116.201	111.457	115.103	116.914	123.589	

Решение:

Такой тест удобнее всего выполнять с помощью электронных таблиц Excel. Для проведения теста ранговой корреляции Спирмена необходимо выполнить следующие действия.

1. Отсортировать данные в таблице по возрастанию значений x.

2. Придать каждому наблюдению ранг, для чего необходимо добавить новый столбец, в котором задать числа от 1 до n.

3. Вычислить регрессионные остатки. В Excel это можно выполнить различными способами. Один из них — с помощью функции ЛИНЕЙН (или ОТРЕЗОК, НАКЛОН) найти оценку функции $\hat{y}$, а затем вычислить остатки ε_i как разность $y_i - \hat{y}_i$. Другой способ — с использованием надстройки «Анализ данных. . .». Нужно вызвать опцию «Регрессия», указать диапазон Y и X и выбрать в диалоговом окне опцию «Остатки». После выполнения данной надстройки появится дополнительная таблица, в которой содержатся номера наблюдений, прогнозы и остатки. Тот столбец таблицы, в котором находятся остатки, необходимо перенести к исходным данным. После выполнения этих действий наша таблица будет содержать четыре столбца: ранг наблюдения, упорядоченные значения регрессора x, значения y и значения остатков (табл. 5.2).

Таблица 5.2 – Расчет характеристик

Ранг по *X*	**Цена *X* (р.)**	**Спрос *Y* (тыс. шт.)**	**Остатки** ε
1	15.09	125.178	1.450697114
2	15.21	123.809	1.006051072
3	15.28	121.175	−1.088742452
4	15.49	116.914	−3.732123026

продолжение на следующей странице

Таблица 5.2 – Продолжение

Ранг по *X*	**Цена *X* (р.)**	**Спрос *Y* (тыс. шт.)**	**Остатки ε**
5	15.54	119.864	−0.396975543
6	15.62	118.068	−1.576739571
7	15.70	123.589	4.560496401
8	15.91	117.088	−0.322884172
9	15.92	116.17	−1.163854676
10	15.95	118.344	1.241233814
11	16.31	116.201	1.871295688
12	16.33	111.457	−2.718645319
13	16.60	115.103	3.007151087
14	16.69	110.106	−1.296583445
15	16.76	110.023	−0.840376969

4. Отсортировать данные по возрастанию модулей остатков и добавить новый столбец рангов остатков, аналогичным образом задав значения от 1 до n.

5. В дополнительном столбце вычислить значения разности между двумя полученными рангами (это и будет значение d_i) (табл. 5.3).

Таблица 5.3 – Расчет характеристик

Ранг по *X*	**Цена *X* (р.)**	**Спрос *Y* (тыс. шт.)**	**Остатки ε (модуль)**	**Ранг по остаткам**	**Разность рангов d_i**
8	15.91	117.088	0.322884172	1	7
5	15.54	119.864	0.396975543	2	3
15	16.76	110.023	0.840376969	3	12
2	15.21	123.809	1.006051072	4	−2
3	15.28	121.175	1.088742452	5	−2
9	15.92	116.17	1.163854676	6	3
10	15.95	118.344	1.241233814	7	3
14	16.69	110.106	1.296583445	8	6
1	15.09	125.178	1.450697114	9	−8
6	15.62	118.068	1.576739571	10	−4
11	16.31	116.201	1.871295688	11	0
12	16.33	111.457	2.718645319	12	0
13	16.60	115.103	3.007151087	13	0
4	15.49	116.914	3.732123026	14	−10
7	15.70	123.589	4.560496401	15	−8

6. Подсчитать коэффициент ранговой корреляции и статистику:

$$r_{x,e} = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot \left(n^2 - 1\right)} = 1 - \frac{6 \cdot 508}{15 \cdot \left(225 - 1\right)} = 0.0928.$$

$$t_r = 0.0928 \cdot \sqrt{15 - 1} = 0.35.$$

7. Проверить гипотезу.

Рассчитаем t-статистику: $t_{0.95,13} = 2.16$.

Так как условие $t_r > t_{\alpha,n-2}$ не выполняется, то нулевая гипотеза об отсутствии гетероскедастичности принимается при уровне значимости $\alpha = 0.95$.

. . . . . .

5.3.2 Тест Парка

Тест Парка осуществляется с помощью следующих шагов.

Шаг 1. Рассчитывается уравнение регрессии $y_i = \Theta_0 + \Theta_1 x + \varepsilon_i$.

Шаг 2. Вычисляются остатки $\varepsilon_i = y_i - \hat{y}_i$.

Шаг 3. Оценивается вспомогательное уравнение регрессии

$$\ln \varepsilon_i^2 = a + b \cdot \ln x_{ij} + \nu_i,$$

где x_{ij} — i-е значение фактора; v_i — случайный остаток.

Шаг 4. Проверяется значимость коэффициента b. Условие принятия гипотезы: $t_b > t_{\alpha,n-2}$.

Если данное условие выполняется, то нулевая гипотеза о наличии гетероскедастичности будет принята при уровне значимости α.

5.3.3 Тест Гольдфельда—Квандта

Этот тест применяется в том случае, если ошибки регрессии можно считать нормально распределенными случайными величинами. Предположим, что средние квадратические отклонения возмущений σ_i пропорциональны значениям объясняющей переменной X (это означает постоянство часто встречающегося на практике относительного (а не абсолютного, как в классической модели) разброса возмущений ε_i регрессионной модели).

В этом случае все наблюдения необходимо упорядочить по мере возрастания значений x. Разделить исходную модель на три равные части. Если количество наблюдение не делится нацело на 3, то уменьшается количество наблюдений в средней части, а первая и вторая части остаются одинаковыми по количеству наблюдений. Затем построить регрессионную модель для первых k и последних k наблюдений. Соответственно обозначим через $S^{(1)}$ и $S^{(3)}$ сумму квадратов отклонений в каждой регрессии. Тогда статистика имеет вид:

$$F = \frac{S^{(3)}}{S^{(1)}}.$$

Если выполняется условие $F > F_\gamma(k - m - 1, k - m - 1)$, то гипотеза об отсутствии гетероскедастичности отвергается. Здесь m — количество объясняющих переменных в уравнении, k — объем подвыборки.

5.4 Устранение гетероскедастичности

Если дисперсии σ_i^2 $(i = 1, \ldots, n)$ случайной величины известны, то гетероскедастичность устраняется путем нормирования переменных. Для модели парной регрессии такое нормирование будет иметь вид:

$$\frac{Y_i}{\sigma_i} = \frac{\Theta_0}{\sigma_i} + \frac{\Theta_1 x}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}.$$

Далее может быть использован метод наименьших квадратов.

Для парной регрессии минимизируемая функция имеет вид:

$$\sum_{i=1}^{n} \left(\frac{y_i - \Theta_0 - \Theta_1 x_i}{\sigma_i} \right)^2 \to \min_{\Theta_0, \Theta_1}.$$

Оценки неизвестных параметров могут быть вычислены по формулам:

$$\widehat{\Theta}_1 = \frac{\left(\sum\limits_{i=1}^{n} \lambda_i \right) \left(\sum\limits_{i=1}^{n} \lambda_i x_i y_i \right) - \left(\sum\limits_{i=1}^{n} \lambda_i x_i \right) \left(\sum\limits_{i=1}^{n} \lambda_i y_i \right)}{\left(\sum\limits_{i=1}^{n} \lambda_i \right) \left(\sum\limits_{i=1}^{n} \lambda_i x_i^2 \right) - \left(\sum\limits_{i=1}^{n} \lambda_i x_i \right)^2}, \tag{5.1}$$

$$\widehat{\Theta}_0 = \frac{\sum\limits_{i=1}^{n} \lambda_i y_i}{\sum\limits_{i=1}^{n} \lambda_i} - \widehat{\Theta}_1 \frac{\sum\limits_{i=1}^{n} \lambda_i x_i}{\sum\limits_{i=1}^{n} \lambda_i},$$

где $\lambda_i = 1/\sigma_i^2$.

Также для расчета оценок может быть использован обобщенный МНК. В этом случае оценка определяется по формуле:

$$\widehat{\Theta} = \left(X^T V_\varepsilon^{-1} X \right)^{-1} X^T V_\varepsilon^{-1} Y,$$

где V — единичная матрица, диагональные элементы которой равны σ^2.

Однако в большинстве случаев информация о дисперсиях σ_i^2 отсутствует. В этом случае используется двухшаговый метод взвешенных наименьших квадратов. Кратко опишем этот метод для гетероскедастичной линейной регрессионной модели с двумя объясняющими переменными ($m = 3$), в которой дисперсии σ_i^2 возмущений ε_i зависят от значений x_{ij}, как

$$\sigma_i^2 = \alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)},$$

где параметры $\alpha_0, \alpha_1, \alpha_2$ — неизвестны.

На первом шаге находим вектор b с помощью МНК:

$$b = \left(X^T X \right)^{-1} X^T Y,$$

и вычисляется вектор остатков (невязок):

$$\varepsilon = Y - \widehat{Y} = Y - X^T b.$$

Предполагая, что $D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma_i^2$, формируем линейную регрессионную модель, в которой объясненной частью является σ_i^2, т. е.

$$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)} + \gamma_i, \quad i = 1, 2, \ldots, n.$$

Применяя к этой модели обычный МНК, находим оценки $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ для параметров $\alpha_0, \alpha_1, \alpha_2$.

На втором шаге подставляем оценки $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ в уравнение $\sigma_i^2 = \alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)}$ и вычисляем значения σ_i^2, из которых формируем ковариационную матрицу $V_\varepsilon = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}$ и вычисляем оценку Θ:

$$\hat{\Theta} = \left(X^T V_\varepsilon^{-1} X\right)^{-1} X^T V_\varepsilon^{-1} Y.$$

Пример 5.3

Используя исходные данные таблицы 5.4, найти оценки неизвестных параметров функции регрессии: а) используя обобщенный МНК, б) используя двухфазный МНК, считая, что значения σ_i^2 неизвестны.

Таблица 5.4 – Исходные данные

$X^{(1)}$	8	11	12	9	8	8	9	9	8	12
$X^{(2)}$	5	8	8	5	7	8	6	4	5	7
Y	5.11	8.81	9.71	5.82	4.79	5.74	5.36	5.46	4.57	8.01
σ_2	0.22	0.66	0.79	0.3	0.29	0.32	0.34	0.26	0.22	0.73

Решение:

а) Запишем исходные данные:

$$X = \begin{pmatrix} 1 & 8 & 5 \\ 1 & 11 & 8 \\ 1 & 12 & 8 \\ 1 & 9 & 5 \\ 1 & 8 & 7 \\ 1 & 8 & 8 \\ 1 & 9 & 6 \\ 1 & 9 & 4 \\ 1 & 8 & 5 \\ 1 & 12 & 7 \end{pmatrix},$$

$$Y = \begin{pmatrix} 5.11 & 8.81 & 9.71 & 5.82 & 4.79 & 5.74 & 5.36 & 5.46 & 4.57 & 8.01 \end{pmatrix}^T,$$

$$V = \begin{pmatrix} 0.22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.66 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.79 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.29 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.32 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.34 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.26 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.22 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.73 \end{pmatrix}.$$

Используя МНК, находим оценки неизвестных параметров:

$$\hat{\Theta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y = \begin{pmatrix} -3.846 \\ 0.915 \\ 0.25 \end{pmatrix}.$$

б) Используя исходные матрицы X и Y из а), находим вектор b с помощью МНК:

$$b = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} -4.037 \\ 0.897 \\ 0.308 \end{pmatrix}.$$

Вычисляем вектор остатков (невязок):

$$\varepsilon = Y - X^T b;$$

$$\varepsilon = \left(0.429 \quad 0.513 \quad 0.515 \quad 0.242 \quad -0.508 \quad 0.134 \quad -0.527 \quad 0.19 \quad -0.111 \quad -0.876\right)^T.$$

Формируем линейную регрессионную модель, в которой объясненной частью является σ_i^2, т. е.

$$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)}.$$

Применяя к этой модели обычный МНК, находим оценки $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ для параметров $\alpha_0, \alpha_1, \alpha_2$:

$$\hat{\alpha} = \left(X^T X\right)^{-1} X^T \varepsilon^2 = \begin{pmatrix} -0.69 \\ 0.09 \\ 8.653 \cdot 10^{-3} \end{pmatrix}.$$

Переходим ко второму шагу этого метода: подставляем оценки $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ в уравнение $\sigma_i^2 = \alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)}$ и вычисляем значения σ_i^2, из которых формируем ковариационную матрицу $V_\varepsilon = diag\left\{\sigma_1^2, \sigma_2^2 \ldots, \sigma_n^2\right\}$:

$$V_\varepsilon = \begin{pmatrix} 0.076 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.373 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.464 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.167 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.094 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.102 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.175 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.158 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.076 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.455 \end{pmatrix}$$

и вычисляем оценку $\widehat{\Theta}$:

$$\widehat{\Theta} = \left(X^T V_\varepsilon^{-1} X\right)^{-1} X^T V_\varepsilon^{-1} Y = \begin{pmatrix} -3.845 \\ 0.92 \\ 0.244 \end{pmatrix}.$$

Контрольные вопросы по главе 5

1. Что понимается под гетероскедастичностью?
2. Какие можно привести примеры гетероскедастичности?
3. Как с помощью графического анализа остатком можно сделать вывод о гетероскедастичности?
4. Как с помощью графического анализа остатком можно сделать вывод о гомоскедастичности?
5. Как выполняется тест ранговой корреляции Спирмена?
6. Как выполняется тест Парка?
7. Как выполняется тест Гольдфельда—Квандта?
8. На сколько частей разделяется выборка в тесте Гольдфельда—Квандта?
9. Какими способами может быть устранена гетероскедастичность?
10. Как выглядит функция минимизации при нормировании для парной регрессии в случае гетероскедастичности?
11. Как могут быть записаны формулы для расчета неизвестных оценок параметров в случае гетероскедастичности (в случае парной регрессии)?
12. Какую форму имеет ковариационная матрица в обобщенном МНК в случае гетероскедастичности?
13. Какими действиями реализуется двухшаговый метод наименьших квадратов?

Глава 6

АВТОКОРРЕЛЯЦИЯ

6.1 Основные понятия

Встречаются ситуации, когда прослеживается механизм влияния результатов предыдущих наблюдений на результаты последующих. Математически это выражается в том, что случайные величины ε_i в регрессионной модели не оказываются независимыми, в частности условие $M(\varepsilon_i \varepsilon_j) = 0$ (i не равно j) не выполняется.

Такие модели называются моделями с наличием автокорреляции, на практике ими чаще всего оказываются временные ряды.

Рассмотрим в качестве примера временной ряд y_i — ряд последовательных значений курса ценной бумаги A, наблюдаемых в моменты времени $1, \ldots, 100$. Результаты наблюдений графически изображены на рисунке 6.1.

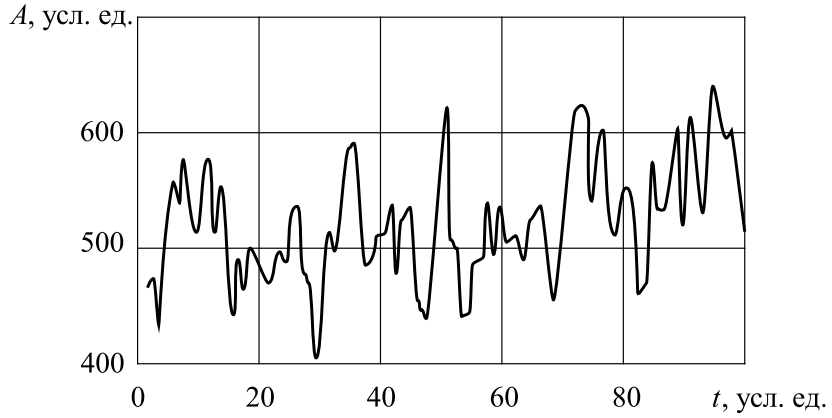


Рис. 6.1 – Результаты наблюдений

Очевидно, курс ценной бумаги A имеет тенденцию к росту, что можно проследить на графике. Естественно предположить, что результаты предыдущих торгов оказывают влияние на результаты последующих: если в какой-то момент курс ока-

жется завышенным по сравнению с реальным, то, скорее всего, он будет завышен на следующих торгах, т. е. имеет место положительная автокорреляция.

Графически положительная автокорреляция выражается в чередовании зон, где наблюдаемые значения оказываются выше объясненных (предсказанных), и зон, где наблюдаемые значения ниже.

Так, на рисунке 6.2 представлены графики наблюдаемых значений y_i и объясненных, сглаженных $\hat{y}_i$.

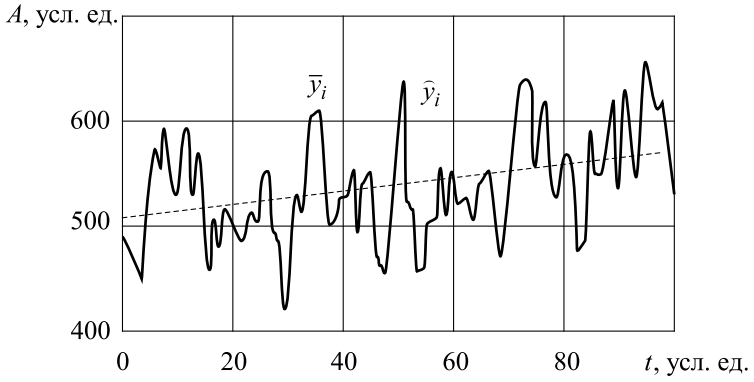


Рис. 6.2 – Положительная автокорреляция

Отрицательная автокорреляция встречается в тех случаях, когда наблюдения действуют друг на друга по принципу «маятника» — завышенные значения в предыдущих наблюдениях приводят к занижению их в наблюдениях последующих. Графически это выражается в том, что результаты наблюдений y_i «слишком часто» «перескакивают» через график объясненной части $\hat{y}_i$. Примерное поведение графика наблюдаемых значений временного ряда изображено на рисунке 6.3.

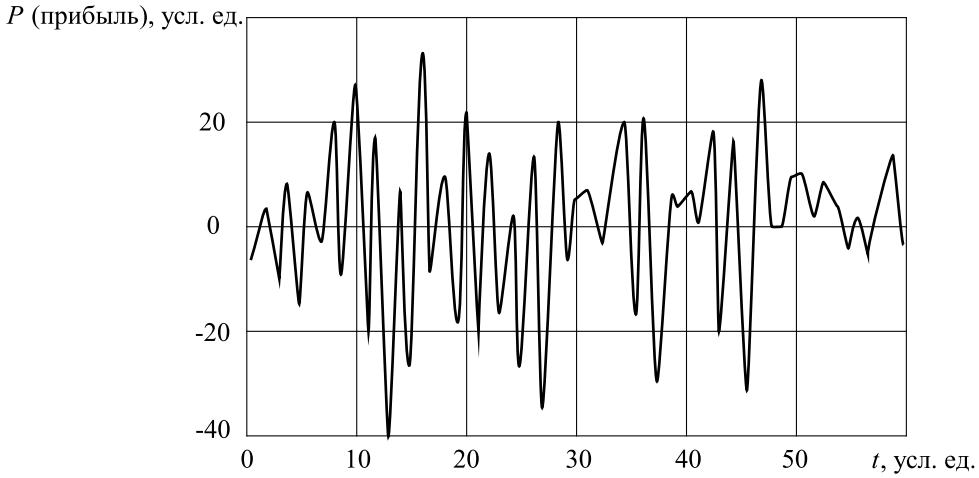


Рис. 6.3 – Отрицательная автокорреляция

Рассмотрим ещё один пример. Пусть исследуется спрос Y на прохладительные напитки от дохода X по ежемесячным данным. Трендовая зависимость, отражающая увеличение спроса с ростом дохода, может быть представлена линейной функцией $Y = \Theta_0 + \Theta_1 X$, изображенной на рисунке 6.4, *а*. Однако фактические

точки наблюдений обычно будут превышать трендовую линию в летние периоды и будут ниже ее в зимние.

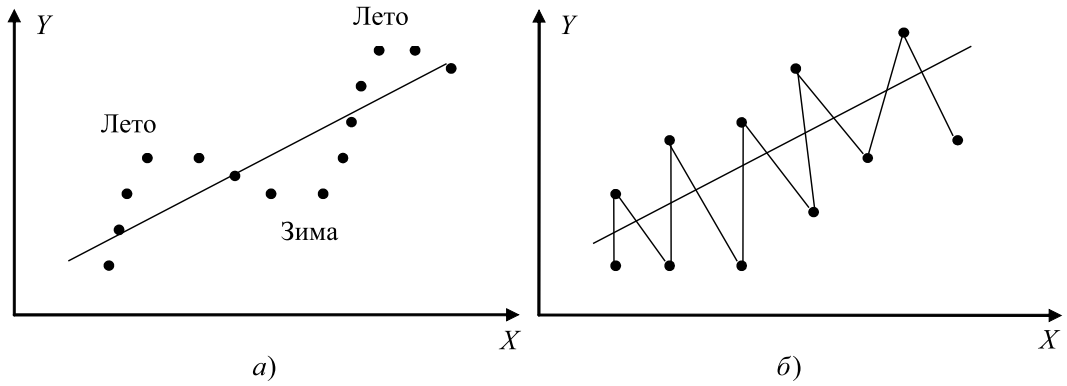


Рис. 6.4 – Графики зависимости спроса на прохладительные напитки

Если ту же зависимость между спросом на прохладительные напитки и доходами рассматривать по сезонным данным (зима — лето), то будет иметь место отрицательная автокорреляция (рис. 6.4, *б*).

Как правило, если автокорреляция присутствует, то наибольшее влияние на последующее наблюдение оказывает результат предыдущего наблюдения. Так, например, если рассматривается ряд значений курса какой-либо ценной бумаги, то, очевидно, именно результат последних торгов служит отправной точкой для формирования курса на следующих торгах. Такую автокорреляцию называют автокорреляцией первого порядка:

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i,$$

где ρ — коэффициент автокорреляции, принадлежащий интервалу $(-1;1)$; u_i — случайный элемент, не подверженный автокорреляции.

Ситуация, когда на значение наблюдения y_i оказывает основное влияние не результат y_{i-1}, а более ранние значения, является достаточно редкой. Чаще всего при этом влияние носит сезонный (циклический) характер, например на значение y_i оказывает наибольшее влияние y_{i-7}, если наблюдения осуществляются ежедневно и имеют недельный цикл (например, сбор кинотеатра). В этом случае можно составить ряды наблюдений отдельно по субботам, воскресеньям и так далее, после чего наиболее сильная корреляция будет наблюдаться между соседними членами. В этом случае связь остатков будет иметь вид:

$$\varepsilon_i = \rho\varepsilon_{i-7} + u_i.$$

Таким образом, отсутствие корреляции между соседними членами служит хорошим основанием считать, что корреляция отсутствует в целом и обычный метод наименьших квадратов дает адекватные и эффективные результаты.

В том случае, если на значение наблюдения y_i оказывает влияние не только y_{i-1}, но y_{i-2}, то такую автокорреляцию называют автокорреляцией второго порядка:

$$\varepsilon_i = \rho_1\varepsilon_{i-1} + \rho_2\varepsilon_{i-2} + u_i.$$

Если число, оказывающее влияние на значение наблюдения, больше двух, то такую автокорреляцию называю автокорреляцией высших порядков.

Кроме того, на практике встречаются динамические модели: модели с распределенными лагами и авторегресионные модели.

Модели с распределенными лагами. Лаговые переменные — переменные, влияние которых характеризуется определенным запаздыванием. Это модели, содержащие в качестве лаговых переменных лишь независимые переменные:

$$y_i = \Theta_0 + \Theta_1 x_i + \Theta_2 x_{i-1} + \ldots + \Theta_i x_{i-l} + \varepsilon_i.$$

Авторегрессионные модели. Это модели, уравнения которых в качестве лаговых объясняющих переменных включают значения зависимых переменных:

$$y_i = \Theta_0 + \Theta_1 x_i + \Theta_2 y_{i-1} + \varepsilon_i.$$

Причин наличия лагов достаточно много, и среди них можно выделить следующие:

- психологические причины — обычно выражаются через инерцию в поведении людей. Например, люди тратят свой доход постепенно, а не мгновенно. Привычка к определенному образу жизни приводит к тому, что люди приобретают те же блага в течение некоторого времени даже после падания реального дохода;
- технологические причины — например, изобретение персональных компьютеров не привело к мгновенному вытеснению ими больших ЭВМ в силу необходимости замены соответствующего программного обеспечения, которое потребовало продолжительного времени;
- институциональные причины — например, контракты между фирмами, трудовые договоры требуют определенного постоянства в течение времени контракта;
- механизм формирования экономических показателей, например инфляция, во многом является инерционным процессом, денежный мультипликатор также проявляет себя на определенном временном интервале;
- сглаживание данных. Зачастую данные по некоторому продолжительному временному периоду получают усреднением данных по составляющим его подынтервалам. Это может привести к определенному сглаживанию колебаний, которые имелись внутри рассматриваемого периода, что, в свою очередь, может послужить причиной автокорреляции;
- ошибки спецификации. Неучет в модели какой-либо важной объясняющей переменной либо неправильный выбор формы зависимости обычно приводит к системным отклонениям точек наблюдений от линии регрессии, что может привести к автокорреляции.

6.2 Графический метод обнаружения автокорреляции

Существует несколько вариантов графического определения автокорреляции. Один из них, увязывающий отклонения ε_t с моментами t их получения (их порядковыми номерами i), приведен на рисунке 6.5. Это так называемые последовательно-временные графики. В этом случае по оси абсцисс обычно откладываются либо момент получения статистических данных, либо порядковый номер наблюдения, а по оси ординат – отклонения ε_t (либо оценки отклонений $\hat{\varepsilon}_t$).

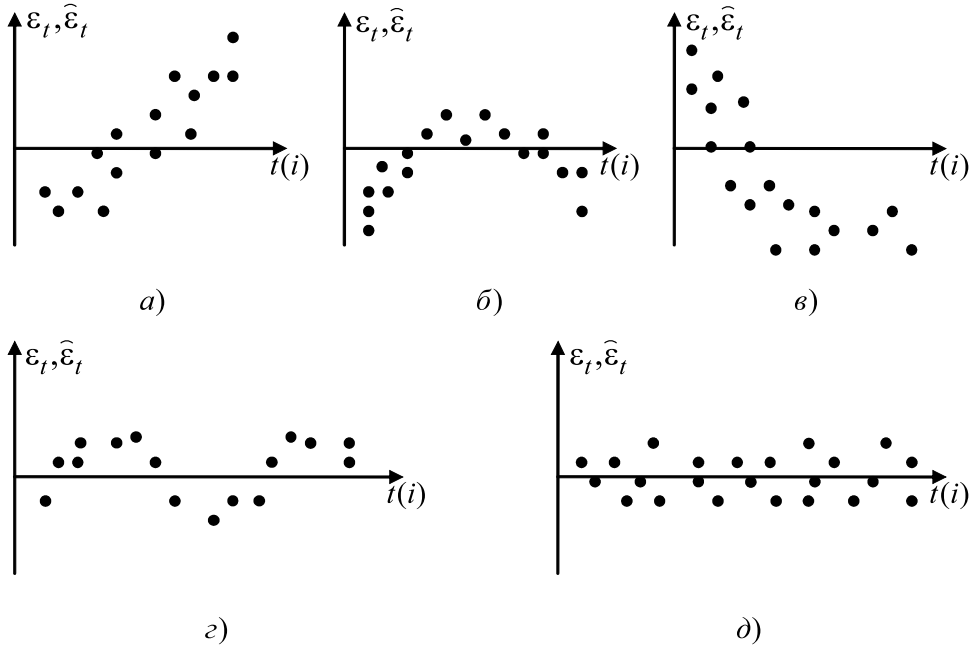


Рис. 6.5 – Последовательно-временные графики

Естественно предположить, что на рисунке 6.5, *а–г* имеются определенные связи между отклонениями, т. е. автокорреляция имеет место. Отсутствие зависимости на рисунке 6.5, *д*, скорее всего, свидетельствует об отсутствии автокорреляции.

Например, на рисунке 6.5, *б* отклонения вначале в основном отрицательные, затем положительные, потом снова отрицательные. Это свидетельствует о наличии между отклонениями определенной зависимости. Более того, можно утверждать, что в этом случае имеет место положительная автокорреляция остатков. Она становится весьма наглядной, если график 6.5, *б* дополнить графиком зависимости ε_t от ε_{t-1}, который в этом случае ориентировочно будет выглядеть, как на рисунке 6.6.

Подавляющее большинство точек на этом графике расположено в I и III четвертях декартовой системы координат, подтверждая положительную зависимость между соседними отклонениями.

Следует сказать, что в современных эконометрических пакетах аналитическое выражение регрессии дополняется графическим представлением результатов. На график реальных колебаний зависимой переменной накладывается график колебаний переменной по уравнению регрессии. Сопоставив эти два графика, можно

выдвинуть гипотезу о наличии автокорреляции остатков. Если эти графики пересекаются редко, то можно предположить наличие положительной автокорреляции остатков.

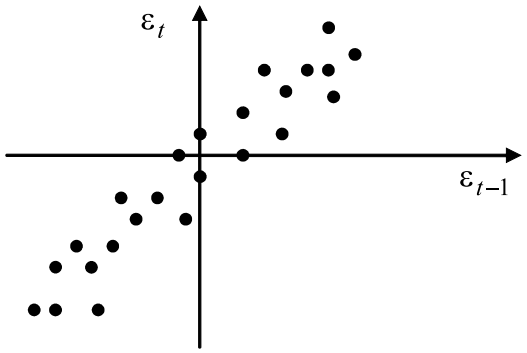


Рис. 6.6 – График зависимости ε_t от ε_{t-1}

6.3 Метод рядов

Этот метод достаточно прост: последовательно определяются знаки отклонений ε_t. Например,

$$(-----)(++++++++)(---)(++++)(-)$$

т. е. 5 «–», 7 «+», 3 «–», 4 «+», 1 «–» при 20 наблюдениях.

Ряд определяется как непрерывная последовательность одинаковых знаков. Количество знаков в ряду называется *длиной ряда*.

Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями. Если рядов слишком мало по сравнению с количеством наблюдений n, то вполне вероятна положительная автокорреляция. Если же рядов слишком много, то вероятна отрицательная автокорреляция. Для более детального анализа предлагается следующая процедура. Пусть

- n — объем выборки;
- n_1 — общее количество знаков «+» при n наблюдениях (количество положительных отклонений ε_t);
- n_2 — общее количество знаков «–» при n наблюдениях (количество отрицательных отклонений ε_t);
- k — количество рядов.

При достаточно большом количестве наблюдений ($n_1 > 10, n_2 > 10$) и отсутствии автокорреляции СВ k имеет асимптотически нормальное распределение с

$$M(k) = \frac{2n_1n_2}{n_1+n_2} + 1; \quad D(k) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}.$$

Тогда, если $M(k) - u_{\alpha/2}D(k) < k < M(k) + u_{\alpha/2}D(k)$, то гипотеза об отсутствии автокорреляции не отклоняется.

При небольшом числе наблюдений ($n_1 < 20$, $n_2 < 20$) Свед и Эйзенхарт разработали таблицы критических значений количества рядов при n наблюдениях (Приложение Е). Суть таблиц в следующем.

На пересечении строки n_1 и столбца n_2 определяются нижнее k_1 и верхнее k_2 значения при уровне значимости $\alpha = 0.05$.

Если $k_1 < k < k_2$, то говорят об отсутствии автокорреляции.

Если $k \leqslant k_1$, то говорят о положительной автокорреляции остатков.

Если $k \geqslant k_2$, то говорят об отрицательной автокорреляция остатков.

6.4 Тест Дарбина—Уотсона

Этот простой критерий (тест) определяет наличие автокорреляции между соседними членами.

Тест Дарбина—Уотсона основан на следующей идее: если корреляция ошибок регрессии не равна нулю, то она присутствует и в остатках регрессии ε_i, получающихся в результате применения обычного метода наименьших квадратов. В тесте Дарбина—Уотсона для оценки корреляции используется статистика вида:

$$d = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2}.$$

Статистика Дарбина—Уотсона следующим образом связана с выборочным коэффициентом корреляции между соседними наблюдениями:

$$d \approx 2(1 - r).$$

В случае отсутствия автокорреляции выборочный коэффициент r окажется не сильно отличающимся от нуля, а значение статистики d будет близко к двум. Близость наблюдаемого значения к нулю должна означать наличие положительной автокорреляции, к четырем — отрицательной.

Существуют два пороговых значения $d_в$ и $d_н$, зависящие только от числа наблюдений, числа регрессоров и уровня значимости, такие, что выполняются следующие условия.

Если фактически наблюдаемое значение d:

1) $d_в < d < 4 - d_в$, то гипотеза об отсутствии автокорреляции не отвергается (принимается);
2) $d_н < d < d_в$ или $4 - d_в < d < 4 - d_н$, то вопрос об отвержении или принятии гипотезы остается открытым (область неопределенности критерия);
3) $0 < d < d_н$, то принимается альтернативная гипотеза о положительной автокорреляции;
4) $4 - d_н < d < 4$, то принимается альтернативная гипотеза об отрицательной автокорреляции.

Изобразим результат Дарбина—Уотсона графически (рис. 6.7).

Рис. 6.7 – Результат теста Дарбина—Уотсона

Значения $d_н$ и $d_в$ приводятся в специальных таблицах и определены для $n \geqslant 15$ (Приложение Д). В таблице Приложения Д с помощью интерполяции определены значения и для $n < 15$.

Недостатками критерия Дарбина—Уотсона является наличие области неопределенности критерия, а также то, что критические значения d-статистики определены для объемов выборки не менее 15. Тем не менее тест Дарбина—Уотсона является наиболее употребляемым.

Пример 6.1

Выявить на уровне значимости 0.05 наличие автокорреляции возмущений для временного ряда y_i (по данным табл. 6.1).

Таблица 6.1 – Исходные данные

Год, x_i	1	2	3	4	5	6	7	8
Спрос, y_i	213	171	291	309	317	362	351	361

Решение:

Построенное по этим данным уравнение регрессии имеет вид:

$$\hat{y}_i = 181.32 + 25.679x_i.$$

В таблице 6.2 приведен расчет сумм, необходимых для вычисления d-статистики.

Таблица 6.2 – Вычисление показателей

x_i	y_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	ε_{i-1}	$\varepsilon_i\varepsilon_{i-1}$	ε_i^2
1	213	207.0	6.0	—	—	36.0
2	171	232.7	−61.7	6.0	−370.2	3806.9
3	291	258.4	32.6	−61.7	−2011.4	1062.8
4	309	284.0	25.0	32.6	815.0	625.0
5	317	309.7	7.3	25.0	182.5	53.3
6	362	335.4	26.6	7.3	194.2	707.6
7	351	361.1	−10.1	26.6	−268.7	102.0

продолжение на следующей странице

Таблица 6.2 – Продолжение

x_i	y_i	$\widehat{y}_i$	$\varepsilon_i = y_i - \widehat{y}_i$	ε_{i-1}	$\varepsilon_i\varepsilon_{i-1}$	ε_i^2
8	361	386.8	25.8	−10.1	260.6	665.6
$\sum_{i=1}^{8}$	—	—	—	—	−1198.0	7059.2

Статистика:

$$d \approx 2\left(1 + \frac{1198}{7059.2}\right) = 2.34.$$

По таблице Приложения Д при $n = 15$ критические значения $d_н = 1.08$ и $d_в = 1.36$, т. е. фактически найденное $d = 2.34$ находится в пределах от $d_в$ до $4 - d_в$ ($1.36 < d < 2.64$). Как уже отмечено, при $n < 15$ критических значений d-статистики в таблице нет, но судя по тенденции их изменений с уменьшением n, можно предполагать, что найденное значение останется в интервале $(d_в; 4 - d_в)$, т. е. для рассматриваемого временного ряда спроса на уровне значимости 0.05 гипотеза об отсутствии автокорреляции возмущений не отвергается (принимается).

. . . . . .

В случае присутствия лаговой переменной применяется h-статистика Дарбина—Уотсона:

$$h = (1 - 0.5d)\sqrt{\frac{n}{\left(1 - n\sigma^2_{y_{i-1}}\right)}},$$

где d — статистика Дарбина—Уотсона; $\sigma^2_{y_{i-1}}$ — дисперсия оценки коэффициента при лаговой переменной y_{i-1}; n — число наблюдений.

При увеличении объёма выборки распределение h-статистики стремится к нормальному с нулевым математическим ожиданием и дисперсией, равной 1. Поэтому гипотеза об отсутствии автокорреляции остатков отвергается, если фактическое значение h-статистики оказывается больше, чем критическое значение нормального распределения.

Также для проверки гипотезы по h-критерию используется следующее простое правило:

1) если $h > 1.96$, то нулевая гипотеза об отсутствии положительной автокорреляции в остатках отклоняется;
2) если $h < -1.96$, то нулевая гипотеза об отсутствии отрицательной автокорреляции в остатках отклоняется;
3) если $-1.96 < h < 1.96$, то нет основания отклонять нулевую гипотезу об отсутствии автокорреляции в остатках.

Ограничение данной статистики следует из её формулировки: в формуле присутствует квадратный корень, следовательно, если дисперсия коэффициента при y_{i-1} велика, то процедура невыполнима.

Кроме теста Дарбина—Уотсона также могут быть использованы и другие тесты. Рассмотрим тест серий (Бреуша—Годфри).

Тест основан на следующей идее: если имеется корреляция между соседними наблюдениями, то естественно ожидать, что в уравнении

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i, \quad (i = 1, \ldots, n) \tag{6.1}$$

(где ε_i — остатки регрессии, полученные обычным методом наименьших квадратов) коэффициент ρ окажется значимо отличающимся от нуля.

Это уравнение является авторегрессионным уравнением первого порядка.

Практическое применение теста заключается в оценивании методом наименьших квадратов регрессии (6.1) (временной ряд ε_{i-1} представляет ряд ε_i со сдвигом по времени на единицу).

Преимущество теста Бреуша—Годфри по сравнению с тестом Дарбина—Уотсона заключается в первую очередь в том, что он проверяется с помощью статистического критерия, между тем как тест Дарбина—Уотсона содержит зону неопределенности для значений статистики d. Другим преимуществом теста является возможность обобщения: в число регрессоров могут быть включены не только остатки с лагом 1, но и с лагом 2, 3 и т. д., что позволяет выявить корреляцию не только между соседними, но и между более отдаленными наблюдениями.

Например, при рассмотрении модели зависимости курса ценной бумаги от времени с помощью метода наименьших квадратов было получено уравнение:

$$\varepsilon_i = \underset{(0.10)}{0.56}\varepsilon_{i-1} - \underset{(0.12)}{0.12}\varepsilon_{i-2} - \underset{(0.10)}{0.01}\varepsilon_{i-3}.$$

Как видно, значимым оказывается только регрессор ε_{i-1}, т. е. существенное влияние на результат наблюдения ε_i оказывает только одно предыдущее значение ε_{i-1}. Положительность оценки соответствующего коэффициента регрессии указывает на положительную корреляцию между ошибками регрессии ε_i и ε_{i-1}.

6.5 Устранение автокорреляции

Одной из причин автокорреляции ошибок регрессии является наличие «скрытых» регрессоров, влияние которых в результате проявляется через случайный член. Выявление этих «скрытых» регрессоров часто позволяет получить регрессионную модель без автокорреляции.

Наиболее часто «скрытыми» регрессорами оказываются лаговые объясняемые переменные. В случае временного ряда вполне естественно предположить, что значения объясняемых переменных зависят не только от включенных уже регрессоров, но и от предыдущих значений объясняемой переменной. Рассмотренные тесты показывают, что это почти всегда имеет место в случае автокорреляции.

Другой механизм образования автокорреляции следующий. Случайные возмущения представляют собой белый шум ξ_i, но на результат наблюдения y_i влияет не только величина ξ_i, но (хотя обычно и в меньшей степени) несколько предыдущих величин $\xi_{i-1}, \ldots, \xi_{i-p}$.

Например, рассматривая модель формирования курса ценной бумаги A, мы можем считать, что кроме временной тенденции на курс еще влияет конъюнктура рынка, которую в момент времени x_i можно считать случайной величиной ξ_i с нулевым средним и некоторой дисперсией. Будем предполагать, что величины ξ_i независимы. Естественно ожидать, что на формирование курса в момент времени x_i будет оказывать влияние в первую очередь конъюнктура ξ_i и (в меньшей степени) конъюнктуры в дни предыдущих торгов ξ_{i-1} и т. д.

Рассмотрим регрессионную модель вида:

$$y_i = \Theta_0 + \sum_{j=1}^{p} \Theta_j x_i^{(j)} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{6.2}$$

Будем полагать, что случайные возмущения коррелированны и образуют наиболее простой процесс — авторегрессионный процесс первого порядка, т. е.:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i, \tag{6.3}$$

где $\nu_i \ (i = 1, \ldots, n)$ представляет белый шум, т. е. последовательность независимых нормально распределенных случайных величин с нулевой средней и дисперсией σ_0^2; ρ — коэффициент авторегрессии.

Исключая $\varepsilon_i \ (i = 1, \ldots, n)$ из равенств (6.2), (6.3), получим

$$y_i - \rho y_{i-1} = \Theta_0(1 - \rho) + \sum_{j=1}^{p} \Theta_j \left(x_i^{(j)} - \rho x_{i-1}^{j} \right) + \nu_i, \quad i = 1, \ldots, n. \tag{6.4}$$

Полученная модель является классической, так как теперь случайные возмущения $\nu_i \ (i = 1, \ldots, n)$ независимы и имеют постоянную дисперсию σ_0^2.

Равенство (6.4) имеет смысл только при $i \geqslant 2$, так как при $i = 1$ не определены значения лаговых переменных. Параметры модели сохранятся, если при $i = 1$ умножить обе части уравнения (6.2) на $\sqrt{1 - \rho^2}$:

$$\sqrt{1 - \rho^2} y_1 = \sqrt{1 - \rho^2}\Theta_0 + \sqrt{1 - \rho^2} \sum_{j=1}^{p} \Theta_j x_1^{(j)} + \sqrt{1 - \rho^2}\varepsilon_1. \tag{6.5}$$

Преобразование (6.5) называется поправкой Прайса—Уинстона для первого наблюдения. При большом количестве наблюдений поправка Прайса—Уинстона практически не изменяет результат, поэтому ее часто не учитывают, оставляя значение первого наблюдения неопределенным.

Таким образом, при известном значении ρ автокорреляция легко устраняется путем замены переменных. Например, для парной регрессии:

$$y_i^* = y_i - \rho y_{i-1}, \quad x_i^* = x_i - \rho x_{i-1}, \quad \Theta_0^* = \Theta_0(1 - \rho),$$

$$y_1^* = \sqrt{1 - \rho^2} y_1, \quad x_1^* = \sqrt{1 - \rho^2} x_1.$$

Полученное уравнение: $y_i^* = \Theta_0^* + \Theta_1 x_i^* + \nu_i$.

На практике, однако, значение ρ не бывает известно, поэтому в равенстве (6.4) присутствует не точное значение ρ, а наблюдаемое значение его оценки $\hat{\rho}$.

Коэффициент может быть определен на основе d-статистики:

$$\hat{\rho} \approx 1 - \frac{d}{2}.$$

Этот метод дает удовлетворительные результаты при большом числе наблюдений.

Другой способ оценить ρ — применить обычный метод наименьших квадратов к регрессионному уравнению (6.3).

Двухшаговая процедура Дарбина

Как правило, более точную оценку параметра Θ_1 дает двухшаговая процедура Дарбина, которая заключается в следующем. Исключая ε_i из уравнений (6.2)–(6.3), запишем регрессионную модель в виде

$$y_i = \Theta_0(1-\rho) + \Theta_1 x_i + \rho y_{i-1} - \Theta_1 \rho x_{i-1} + \nu_i, \quad i = 1,\ldots,n. \tag{6.6}$$

Применим к уравнению (6.6) обычный метод наименьших квадратов, включая ρ в число оцениваемых параметров. Получим оценки r и θ величин ρ и $-\Theta_1\rho$. Тогда оценкой Дарбина является величина:

$$\widehat{\Theta}_1 = -\frac{\theta}{r}.$$

Пример 6.2

В таблице 6.3 представлены данные о ВВП (x_i) и расходах на потребительские товары (y_i). Необходимо выполнить тест на автокорреляцию и при необходимости устранить её.

Решение:

По данным таблицы 6.3, используя метод наименьших квадратов, находим оценки неизвестных параметров $\widehat{\Theta}_0, \widehat{\Theta}_1$:

$$\widehat{\Theta} = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} -2.32 \\ 0.07 \end{pmatrix}.$$

Подставляя значения x_i в уравнение $\widehat{y} = -2.32 + 0.07x$, определим величины оценок $\widehat{y}$ (табл. 6.3, 3 столбец), регрессионные остатки $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ (табл. 6.3, 4 столбец), квадрат остатков $\widehat{\varepsilon}_i^2$ (табл. 6.3, 5 столбец), квадрат разности остатков $(\varepsilon_i - \varepsilon_{i-1})^2$ (табл. 6.3, 6 столбец). Определяем значение d-статистики:

$$d = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2} = \frac{832.17}{710.17} = 1.172.$$

Граничные значения для $n = 34$, $\alpha = 0.05$: $d_н = 1.39$, $d_в = 1.51$.

Выполняется условие $0 < d < d_н$, значит, между данными значениями существует положительная автокорреляция.

Для устранения автокорреляции определим $\widehat{\rho}$. Для этого добавим ещё один столбец ε_{i-1} и определим с помощью МНК $\widehat{\rho}$ из уравнения $\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i$. Получим: $\widehat{\rho} = 0.3936$.

Добавим в таблицу 6.3 столбцы y_i^*, x_i^*, вычисленные с помощью правил:

$$y_i^* = y_i - \rho y_{i-1}, \quad x_i^* = x_i - \rho x_{i-1},$$

$$y_1^* = \sqrt{1-\rho^2}y_1, \quad x_1^* = \sqrt{1-\rho^2}x_1.$$

Например,

$$y_1^* = \sqrt{1-\rho^2}y_1 = \sqrt{1-0.3936^2}\cdot 0.34 = 0.31,$$
$$y_2^* = y_2 - \rho y_1 = 0.22 - 0.3936\cdot 0.31 = 0.09$$

и т. д.

По рассчитанным значениям y_i^*, x_i^* с помощью метода наименьших квадратов находим оценки неизвестных параметров $\widehat{\Theta}_0^*, \widehat{\Theta}_1^*$:

$$\widehat{\Theta}^* = \left(X^{*T}X^*\right)^{-1}X^{*T}Y^* = \begin{pmatrix}-1.87\\ 0.07\end{pmatrix}.$$

Аналогично, подставляя значения x_i в уравнение $\widehat{y}^* = -1.87 + 0.07x^*$, определим величины оценок $\widehat{y}^*$, регрессионные остатки $\widehat{\varepsilon}_i^* = y_i^* - \widehat{y}_i^*$, квадрат остатков $\widehat{\varepsilon}_i^{*2}$, квадрат разности остатков $\left(\varepsilon_i^* - \varepsilon_{i-1}^*\right)^2$. Определяем значение d^*-статистики:

$$d^* = \frac{\sum\limits_{i=2}^{n}\left(\varepsilon_i^* - \varepsilon_{i-1}^*\right)^2}{\sum\limits_{i=1}^{n}\varepsilon_i^{*2}} = \frac{938.37}{565.50} = 1.66.$$

Теперь выполняется условие $d_в < d^* < 4 - d_в$, следовательно, автокорреляция отсутствует.

. . . . . .

Таблица 6.3 – Исходные и рассчитанные характеристики

Расходы, y_i	**ВВП, x_i**	$\hat{y}_i$	ε_i	ε_i^2	ε_{i-1}	$(\varepsilon_i - \varepsilon_{i-1})^2$	y_i^*	x_i^*	$\hat{y}_i^*$	ε_i^*	ε_i^{*2}	ε_{i-1}^*	$(\varepsilon_i^* - \varepsilon_{i-1}^*)^2$
0.34	5.67	−1.94	2.28	5.20			0.31	5.21	−1.51	1.82	3.32		
0.22	10.13	−1.64	1.86	3.47	2.28	0.18	0.09	7.90	−1.32	1.41	1.98	1.82	0.17
0.32	11.34	−1.56	1.88	3.54	1.86	0.00	0.23	7.35	−1.36	1.59	2.53	1.41	0.03
1.23	18.88	−1.06	2.29	5.23	1.88	0.16	1.10	14.42	−0.86	1.97	3.87	1.59	0.14
1.81	20.94	−0.92	2.73	7.45	2.29	0.20	1.33	13.51	−0.93	2.25	5.08	1.97	0.08
1.02	22.16	−0.84	1.86	3.45	2.73	0.76	0.31	13.92	−0.90	1.21	1.46	2.25	1.10
1.27	23.83	−0.73	2.00	3.98	1.86	0.02	0.87	15.11	−0.82	1.68	2.84	1.21	0.23
1.07	24.67	−0.67	1.74	3.03	2.00	0.07	0.57	15.29	−0.80	1.37	1.88	1.68	0.10
0.67	27.56	−0.48	1.15	1.31	1.74	0.35	0.25	17.85	−0.62	0.87	0.76	1.37	0.25
1.25	27.57	−0.48	1.73	2.98	1.15	0.34	0.99	16.72	−0.70	1.69	2.85	0.87	0.67
0.75	40.15	0.37	0.38	0.15	1.73	1.80	0.26	29.30	0.18	0.08	0.01	1.69	2.59
2.8	51.62	1.13	1.67	2.78	0.38	1.65	2.50	35.82	0.63	1.87	3.50	0.08	3.21
4.9	57.71	1.54	3.36	11.29	1.67	2.87	3.80	37.39	0.74	3.05	9.32	1.87	1.40
3.5	63.03	1.90	1.60	2.57	3.36	3.08	1.57	40.32	0.95	0.62	0.39	3.05	5.91
4.45	66.32	2.12	2.33	5.45	1.60	0.53	3.07	41.51	1.03	2.04	4.16	0.62	2.01
1.6	66.97	2.16	−0.56	0.31	2.33	8.37	−0.15	40.87	0.99	−1.14	1.30	2.04	10.10
4.26	76.88	2.82	1.44	2.07	−0.56	3.99	3.63	50.52	1.66	1.97	3.87	−1.14	9.65
5.31	101.65	4.48	0.83	0.69	1.44	0.37	3.63	71.39	3.12	0.51	0.26	1.97	2.12
6.4	115.97	5.44	0.96	0.93	0.83	0.02	4.31	75.96	3.44	0.87	0.75	0.51	0.13
7.15	119.49	5.67	1.48	2.18	0.96	0.26	4.63	73.85	3.30	1.34	1.78	0.87	0.22
11.22	124.15	5.98	5.24	27.41	1.48	14.12	8.41	77.12	3.53	4.88	23.82	1.34	12.57
8.66	140.98	7.11	1.55	2.40	5.24	13.58	4.24	92.12	4.57	−0.33	0.11	4.88	27.16
5.56	153.85	7.97	−2.41	5.81	1.55	15.69	2.15	98.37	5.01	−2.86	8.18	−0.33	6.40
13.41	169.38	9.01	4.40	19.36	−2.41	46.39	11.22	108.83	5.74	5.48	30.00	−2.86	69.51

продолжение на следующей странице

Таблица 6.3 – Продолжение

Расходы, y_i	**ВВП, x_i**	$\widehat{y}_i$	ε_i	ε_i^2	ε_{i-1}	$(\varepsilon_i - \varepsilon_{i-1})^2$	y_i^*	x_i^*	$\widehat{y}_i^*$	ε_i^*	ε_i^{*2}	ε_{i-1}^*	$(\varepsilon_i^* - \varepsilon_{i-1}^*)^2$
5.46	186.33	10.14	−4.68	21.94	4.40	82.52	0.18	119.67	6.50	−6.32	39.95	5.48	139.19
4.79	211.78	11.85	−7.06	49.79	−4.68	5.63	2.64	138.45	7.82	−5.18	26.80	−6.32	1.31
8.92	249.72	14.38	−5.46	29.86	−7.06	2.54	7.03	166.37	9.77	−2.74	7.49	−5.18	5.95
18.9	261.41	15.17	3.73	13.94	−5.46	84.60	15.39	163.13	9.55	5.84	34.16	−2.74	73.64
15.95	395.52	24.14	−8.19	67.02	3.73	142.10	8.51	292.64	18.61	−10.10	101.98	5.84	254.17
29.9	534.97	33.46	−3.56	12.71	−8.19	21.36	23.62	379.31	24.68	−1.05	1.11	−10.10	81.80
33.59	655.29	41.51	−7.92	62.77	−3.56	18.99	21.82	444.75	29.26	−7.43	55.27	−1.05	40.71
38.62	815	52.20	−13.58	184.31	−7.92	31.96	25.40	557.10	37.12	−11.72	137.38	−7.43	18.38
61.61	1040.5	67.28	−5.67	32.15	−13.58	62.51	46.41	719.75	48.51	−2.09	4.39	−11.72	92.66
181.3	2586.4	170.69	10.61	112.66	−5.67	265.16	157.05	2176.90	150.50	6.55	42.96	−2.09	74.81
$\sum$				710.17		832.17					565.50		938.37

Контрольные вопросы по главе 6

1. Что понимается под автокорреляцией?
2. Какие можно привести примеры автокорреляции?
3. Что представляет собой положительная автокорреляция?
4. Что представляет собой отрицательная автокорреляция?
5. Как записывается формула вычисления случайных остатков в случае автокорреляции первого порядка?
6. Как записывается формула вычисления случайных остатков в случае автокорреляции второго порядка?
7. Что представляют собой модели с распределенными лагами?
8. Какие модели называются авторегресионными?
9. Каковы причины появления лагов?
10. Как с помощью графического метода можно обнаружить автокорреляцию?
11. В чём заключается метод рядов?
12. Что понимается под рядом?
13. В каком случае в методе рядов можно сделать вывод об отрицательной, положительной автокорреляции?
14. Как выполняется тест Дарбина—Уотсона?
15. Как вычисляется статистика d в тесте Дарбина—Уотсона?
16. Какие интервалы попадания статистики d рассматриваются в тесте Дарбина—Уотсона?
17. Как вычисляется статистика в случае присутствия лаговой переменной?
18. Как выполняется тест серий?
19. В чем преимущество теста серий по сравнению с тестом Дарбина—Уотсона?
20. Как можно устранить автокорреляцию?
21. Какие переменные нужно заменить для устранения автокорреляции?
22. Как оценивается коэффициент ρ при устранении автокорреляции?
23. Что можно определить с помощью двухшаговой процедуры Дарбина?
24. Как реализуется двухшаговая процедура Дарбина?

Глава 7

НЕКОТОРЫЕ ВОПРОСЫ ПРАКТИЧЕСКОГО ИСПОЛЬЗОВАНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ

7.1 Расчет эластичностей

Эластичность *— показатель, который говорит о том, на сколько процентов изменится значение результирующей переменной при изменении объясняющей переменной на 1%:*

$$e = \frac{\partial(\ln y)}{\partial(\ln x)} = \frac{\Delta Y}{\Delta X} \cdot \frac{X}{Y}.$$

Рассмотрим расчет эластичности для различных функций регрессии:
1. Линейная функция:

$$Y = \Theta_0 + \Theta_1 X, \quad e = \frac{\Delta Y}{\Delta X}\frac{X}{Y} = (\Theta_0 + \Theta_1 X)' \frac{X}{Y} = \Theta_1 \frac{X}{Y}.$$

2. Двойная логарифмическая функция:

$$\ln Y = \Theta_0 + \Theta_1 \ln X, \quad e = \frac{\Delta Y}{\Delta X}\frac{X}{Y} = Y\Theta_1 \frac{1}{X}\frac{X}{Y} = \Theta_1.$$

3. Линейно-логарифмическая функция:

$$Y = \Theta_0 + \Theta_1 \ln X, \quad e = \frac{\Delta Y}{\Delta X}\frac{X}{Y} = (\Theta_0 + \Theta_1 \ln X)' = \Theta_1 \frac{1}{X}\frac{X}{Y} = \Theta_1 \frac{1}{Y}.$$

4. Обратная функция:

$$Y = \Theta_0 + \Theta_1 \frac{1}{X}, \quad e = \frac{\Delta Y}{\Delta X}\frac{X}{Y} = \left(\Theta_0 + \Theta_1 \frac{1}{X}\right) = \Theta_1 \frac{1}{X^2}\frac{X}{Y} = -\Theta_1 \frac{1}{XY}.$$

7.2 Мультиколлинеарность

*Под **мультиколлинеарностью** понимается высокая взаимная коррелированность объясняющих переменных. Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах.*

При функциональной форме мультиколлинеарности (совершенная мультиколлинеарность) по крайней мере одна из парных связей между объясняющими переменными является линейной функциональной зависимостью. В этом случае матрица особенная, так как содержит линейнозависимые векторы-столбцы и ее определитель равен нулю, т. е. нарушается предпосылка регрессионного анализа. Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

Мультиколлинеарность может быть проблемой лишь в случае множественной регрессии. Ее суть можно представить на следующем примере совершенной мультиколлинеарности.

Пусть уравнение регрессии имеет вид

$$Y = \Theta_0 + \Theta_1 X_1 + \Theta_2 X_2 + \varepsilon.$$

Пусть также между объясняющими переменными существует строгая линейная зависимость:

$$X_2 = \gamma_0 + \gamma_1 X_1.$$

Подставив X_2 в общее уравнение, получим:

$$Y = \Theta_0 + \Theta_1 X_1 + \Theta_2 (\gamma_0 + \gamma_1 X_1) + \varepsilon$$

или

$$Y = (\Theta_0 - \Theta_2 \gamma_0) + (\Theta_1 + \Theta_2 \gamma_1) X_1 + \varepsilon.$$

Обозначив $\Theta_0 + \Theta_2 \gamma_0 = a$, $\Theta_1 + \Theta_2 \gamma_1 = b$, получаем уравнение парной линейной регрессии:

$$Y = a + bX_1 + \varepsilon.$$

По МНК нетрудно определить коэффициенты a и b. Тогда получим систему двух уравнений:

$$\begin{cases} \Theta_0 + \Theta_2 \gamma_0 = a, \\ \Theta_1 + \Theta_2 \gamma_1 = b. \end{cases}$$

В систему входят три неизвестные $\Theta_0, \Theta_1, \Theta_2$. Такая система в подавляющем числе случаев имеет бесконечно много решений. Таким образом, совершенная мультиколлинеарность не позволяет однозначно определить коэффициенты регрессии уравнения.

Однако в экономических исследованиях мультиколлинеарность чаще проявляется в стохастической форме (несовершенная мультиколлинеарность), когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Матрица $X^T X$ в этом случае является неособенной, но ее определитель очень мал.

Уравнение регрессии в случае несовершенной мультиколлинеарности имеет вид:

$$Y = \Theta_0 + \Theta_1 X_1 + \Theta_2 X_2 + \varepsilon, \text{ где } X_2 = \gamma_0 + \gamma_1 X_1 + u.$$

На рисунке 7.1 представлена совершенная и несовершенная мультиколлинеарность.

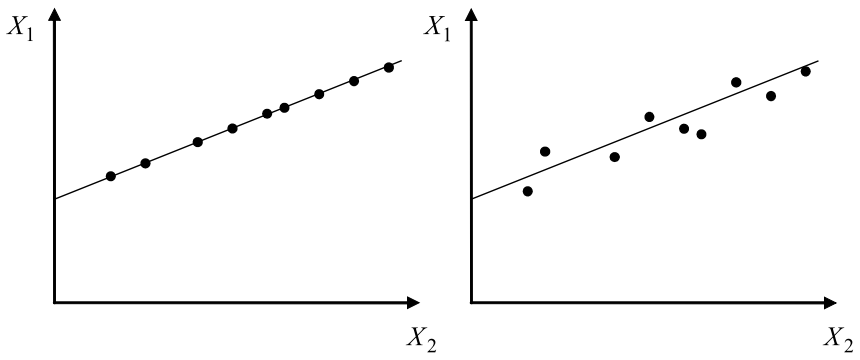


Рис. 7.1 – Мультиколлинеарность: *а*) совершенная; *б*) несовершенная

На рисунке 7.2, *а* коррелированность между объясняющими переменными X_1 и X_2 отсутствует и влияние каждой из них на Y находит отражение в наложении кругов X_1 и X_2 на круг Y. По мере усиления линейной зависимости между X_1 и X_2 соответствующие круги все больше накладываются друг на друга. Заштрихованная область отражает совпадающие части влияния X_1 и X_2 на Y. На рисунке 7.2, *г* при совершенной мультиколлинеарности невозможно разграничить степени индивидуального влияния объясняющих переменных X_1 и X_2 на зависимую переменную Y.

В результате мультиколлинеарности получаются значительные средние квадратические отклонения (стандартные ошибки) коэффициентов регрессии $\Theta_0, \ldots, \Theta_p$, и оценка их значимости по t-критерию не имеет смысла.

Оценки становятся очень чувствительными к незначительному изменению результатов наблюдений и объема выборки.

Уравнения регрессии в этом случае, как правило, не имеют реального смысла, так как некоторые из его коэффициентов могут иметь неправильные с точки зрения экономической теории знаки и неоправданно большие значения.

Точных количественных критериев для определения наличия или отсутствия мультиколлинеарности не существует. Тем не менее имеются некоторые эвристические подходы по ее выявлению.

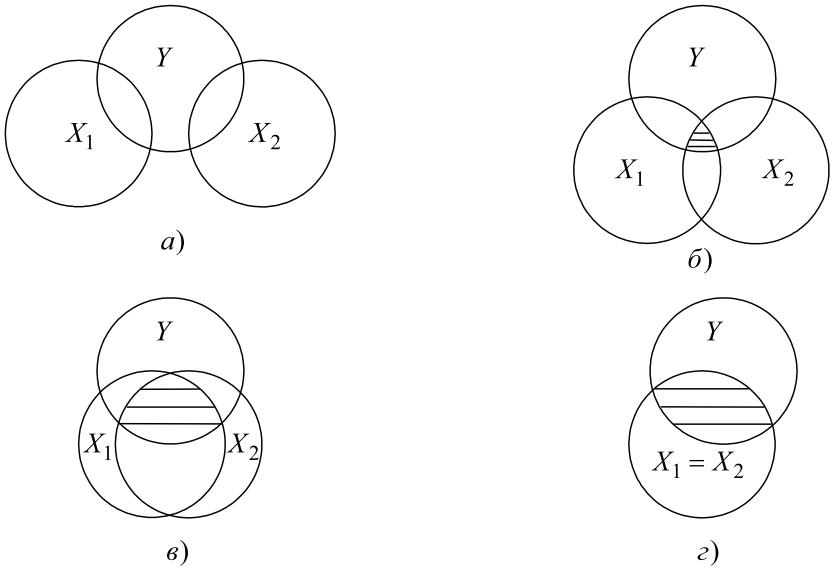


Рис. 7.2 – Графическое представление мультиколлинеарности

1. Один из таких подходов заключается в анализе корреляционной матрицы между объясняющими переменными $x_1, x_2, \ldots, x_p$ и выявлении пар переменных, имеющих высокие коэффициенты корреляции (обычно больше 0.8). Если такие переменные существуют, то говорят о мультиколлинеарности между ними.

2. Также признаком мультиколлинеарности является то, что один или более параметров являются статистически незначимыми при высоком значении индекса детерминации.

3. Ещё одним признаком мультиколлинеарности является чувствительность оценок коэффициентов и оценок дисперсий к добавлению и исключению наблюдений из выборок.

4. Полезно также находить множественные коэффициенты детерминации между одной из объясняющих переменных и некоторой группой из них. Наличие высокого множественного коэффициента детерминации (обычно больше 0.6) свидетельствует о мультиколлинеарности.

5. Рекомендуется также вычислить коэффициент *vif* (variance inflation factor). Для этого строятся регрессионные зависимости для каждой объясняющей переменной. Например, для первой переменной:

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \Theta_2 x_i^{(2)} + \ldots + \Theta_p x_i^{(p)} + \varepsilon_i,$$
$$x_i^{(1)} = \alpha_0 + \alpha_2 x_i^{(2)} + \alpha_3 x_i^{(3)} + \ldots + \alpha_p x_i^{(p)} + \varepsilon_i.$$

Далее по полученным данным определяется значение индекса детерминации R_1^2 и коэффициент *vif*:

$$vif\left(\widehat{\Theta}_1\right) = \frac{1}{\left(1 - R_1^2\right)}.$$

Среди эконометристов существует убеждение, что если для одной из переменных $vif > 10$ [6], то в регрессии есть мультиколлинеарность. При отсутствии мультиколлинеарности $vif = 1$.

6. Другой подход состоит в исследовании матрицы $X^T X$. Если определитель матрицы $X^T X$ либо ее минимальное собственное значение $\lambda_{\min}$ близки к нулю (например, одного порядка с накапливающимися ошибками вычислений), то это говорит о наличии мультиколлинеарности. О том же может свидетельствовать и значительное отклонение максимального собственного значения $\lambda_{\max}$ матрицы $X^T X$ от ее минимального собственного значения $\lambda_{\min}$. Для целей сравнения также рассчитывается индекс обусловленности:

$$k = \sqrt{\frac{\max(\lambda_i)}{\min(\lambda_i)}}.$$

Идея состоит в том, что число ненулевых собственных чисел — это и есть число независимых объясняющих факторов, которые можно выделить из иксов. Если есть собственные числа, близкие к нулю, значит, число этих факторов меньше, чем число переменных в регрессии, поэтому можно утверждать, что есть мультиколлинеарность. Если $k < 10$, то мультиколлинеарности нет. Если $10 < k < 30$, то есть допустимая мультиколлинеарность. Если $k > 30$, то мультиколлинеарность сильная.

Для устранения или уменьшения мультиколлинеарности используется ряд методов. Самый простой из них (но далеко не всегда возможный) состоит в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0.8), одну переменную исключают из рассмотрения. При этом, какую переменную оставить, а какую удалить из анализа, решают в первую очередь на основании экономических соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший коэффициент корреляции с зависимой переменной. Также изменение выборки (добавление данных, обновление) в ряде случаев позволяет устранить мультиколлинеарность.

Для устранения мультиколлинеарности может быть использован переход от исходных объясняющих переменных $x_1, x_2, \ldots, x_n$, связанных между собой достаточно тесной корреляционной зависимостью, к новым переменным, представляющим линейные комбинации исходных. При этом новые переменные должны быть слабо коррелированными либо вообще не коррелированными.

Например, пусть уравнение регрессии имеет вид

$$Y = \Theta_0 + \Theta_1 X^{(1)} + \Theta_2 X^{(2)} + \varepsilon,$$

причем $X^{(1)}$ и $X^{(2)}$ — коррелированные переменные. В этой ситуации можно попытаться определять регрессионные зависимости относительных величин

$$\frac{\hat{Y}}{X^{(1)}} = \frac{\Theta_0}{X^{(1)}} + \Theta_1 + \Theta_2 \frac{X^{(2)}}{X^{(1)}} + \frac{\varepsilon}{X^{(1)}},$$

$$\frac{\hat{Y}}{X^{(2)}} = \frac{\Theta_0}{X^{(2)}} + \Theta_1 \frac{X^{(1)}}{X^{(2)}} + \Theta_2 + \frac{\varepsilon}{X^{(2)}}.$$

Вполне вероятно, что в таких моделях проблема мультиколлинеарности будет отсутствовать.

Преобразования могут быть и другими, например логарифмические преобразования разности. Последний вид преобразования можно записать в форме:

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \Theta_2 x_i^{(2)} + \varepsilon_i,$$
$$y_{i-1} = \Theta_0 + \Theta_1 x_{i-1}^{(1)} + \Theta_2 x_{i-1}^{(2)} + \varepsilon_{i-1},$$
$$y_i - y_{i-1} = \Theta_1 \left(x_i^{(1)} - x_{i-1}^{(1)} \right) + \Theta_2 \left(x_i^{(2)} - x_{i-1}^{(2)} \right) + \nu_i.$$

Также в качестве таких переменных берут, например, так называемые главные компоненты вектора исходных объясняющих переменных, изучаемые в компонентном анализе, и рассматривают регрессию на главных компонентах, в которой последние выступают в качестве обобщенных объясняющих переменных, подлежащих в дальнейшем содержательной (экономической) интерпретации.

7.3 Отбор наиболее существенных объясняющих переменных в регрессионной модели

В исследовательской и практической статистической работе приходится сталкиваться с ситуациями, когда общее число p признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, регистрируемых на каждом из множества обследуемых объектов (стран, городов, предприятий, семей, пациентов, технических или экологических систем), очень велико — порядка ста и более. Тем не менее имеющиеся многомерные наблюдения следует подвергнуть статистической обработке, осмыслить либо ввести в базу данных для того, чтобы иметь возможность их использовать в нужный момент.

Желание специалиста представить каждое из наблюдений в виде вектора Z некоторых вспомогательных показателей $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ существенно меньшим (чем p) числом компонент p' бывает обусловлено в первую очередь следующими причинами:

- необходимостью наглядного представления (визуализации) исходных данных, что достигается их проецированием на специально подобранное трехмерное пространство ($p = 3$), плоскость ($p = 2$) или числовую прямую;
- стремлением к лаконизму исследуемых моделей, обусловленному необходимостью упрощения счета и интерпретации полученных статистических выводов;
- необходимостью существенного сжатия объемов хранимой статистической информации (без видимых потерь в ее информативности), если речь идет о записи и хранении массивов в специальной базе данных.

При этом новые (вспомогательные) признаки $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ могут выбираться из числа исходных или определяться по какому-либо правилу по совокупности исходных признаков, например как их линейные комбинации. При формировании новой системы признаков к последним предъявляются разного рода требования, такие, как наибольшая информативность (в определенном смысле), взаимная некоррелированность, наименьшее искажение геометрической структуры множества исходных данных и т. п. В зависимости от варианта формальной конкретизации этих требований может быть выбран тот или иной алгоритм снижения размерности.

Имеется, по крайней мере, три основных типа принципиальных предпосылок, обусловливающих возможность перехода от большого числа p исходных показателей состояния (поведения, эффективности функционирования) анализируемой системы к существенно меньшему числу p' наиболее информативных переменных. Это, во-первых, дублирование информации, доставляемой сильно взаимосвязанными признаками (мультиколлинеарность); во-вторых, неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); в-третьих, возможность агрегирования, т. е. простого или «взвешенного» суммирования, по некоторым признакам.

Одним из способов снижения размерности исходных данных является использование пошаговых процедур отбора наиболее информативных переменных. Например, на первом шаге рассматривается лишь одна объясняющая переменная, имеющая с зависимой переменной Y наибольший коэффициент детерминации. На втором шаге в регрессию включается новая объясняющая переменная, которая вместе с первоначально отобранной образует пару объясняющих переменных, имеющую с Y наиболее высокий (скорректированный) коэффициент детерминации. На третьем шаге в регрессию вводится еще одна объясняющая переменная, которая вместе с двумя первоначально отобранными образует тройку объясняющих переменных, имеющую с Y наибольший (скорректированный) коэффициент детерминации, и т. д.

Процедура введения новых переменных продолжается до тех пор, пока будет увеличиваться соответствующий (скорректированный) коэффициент детерминации $\widehat{R}^2$.

Пример 7.1

По данным $n = 20$ сельскохозяйственных районов области исследуется зависимость переменной Y — урожайности зерновых культур (в ц/га) от ряда переменных — факторов сельскохозяйственного производства:

- $x^{(1)}$ — *число тракторов (приведенной мощности на 100 га);*
- $x^{(2)}$ — *число зерноуборочных комбайнов на 100 га;*
- $x^{(3)}$ — *число орудий поверхностной обработки почвы на 100 га;*
- $x^{(4)}$ — *количество удобрений, расходуемых на 1 га (т/га);*
- $x^{(5)}$ — *количество химических средств защиты растений, расходуемых на 1 га (ц/га).*

Исходные данные приведены в таблице 3.1.

В случае обнаружения мультиколлинеарности принять меры по ее устранению (уменьшению), используя пошаговую процедуру отбора наиболее информативных переменных.

Решение:

Найдем вектор оценок параметров регрессионной модели:

$$\widehat{\Theta} = \left(X^T X\right)^{-1} X^T Y = \left(3.515 \quad -0.006 \quad 15.542 \quad 60.110 \quad 4.475 \quad -2.932\right)^T.$$

Уравнение выборочной регрессии будет иметь вид:

$$\underset{}{\hat{y}} = \underset{(5.41)}{3.515} - \underset{(0.60)}{0.006}x^{(1)} + \underset{(21.59)}{15.542}x^{(2)} + \underset{(0.85)}{0.110}x^{(3)} + \underset{(1.54)}{4.475}x^{(4)} - \underset{(3.09)}{2.932}x^{(5)}.$$

В скобках указаны средние квадратические отклонения (стандартные ошибки) коэффициентов регрессии $\widehat{\Theta}$.

Сравнивая значения t-статистики (по абсолютной величине) каждого коэффициента регрессии $\widehat{\Theta}$ по формуле $t_{\Theta_j} = \widehat{\Theta}_j / s_{\Theta_j}$ ($j = 0, 1, 2, 3, 4, 5$), т. е. $t_{\Theta_0} = 0.65$; $t_{\Theta_1} = -0.01$; $t_{\Theta_2} = 0.72$; $t_{\Theta_3} = 0.13$; $t_{\Theta_4} = 2.91$; $t_{\Theta_5} = -0.95$ с критическим значением $t_{0.95;14} = 2.14$, определенным по таблице приложений на уровне значимости $\alpha = 0.05$ при числе степеней свободы $k = n - p - 1 = 20 - 5 - 1 = 14$, мы видим, что значимым оказался только коэффициент регрессии $\widehat{\Theta}_4$ при переменной $x^{(4)}$ — количество удобрений, расходуемых на гектар земли.

По формуле (4.1) была рассчитана матрица парных коэффициентов корреляции (табл. 7.1).

Таблица 7.1 – Корреляционная матрица

Переменные	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
Y	1.00	0.43	0.37	0.40	0.58 (*)	0.33
$x^{(1)}$	0.43	1.00	0.85 (*)	0.98 (*)	0.11	0.34
$x^{(2)}$	0.37	0.85 (*)	1.00	0.88 (*)	0.03	0.46 (*)
$x^{(3)}$	0.40	0.98 (*)	0.88 (*)	1.00	0.03	0.28
$x^{(4)}$	0.58 (*)	0.11	0.03	0.03	1.00	0.57 (*)
$x^{(5)}$	0.33	0.34	0.46 (*)	0.28	0.57 (*)	1.00

Знаком * отмечены коэффициенты корреляции, значимые по t-критерию на 5%-ном уровне.

Анализируя матрицу парных коэффициентов корреляции, можно отметить тесную корреляционную связь между переменными $x^{(1)}$ и $x^{(2)}$ ($r_{12} = 0.85$), $x^{(1)}$ и $x^{(3)}$ ($r_{13} = 0.98$), $x^{(2)}$ и $x^{(3)}$ ($r_{23} = 0.88$), что, очевидно, свидетельствует о мультиколлинеарности объясняющих переменных.

Для устранения мультиколлинеарности применим процедуру пошагового отбора наиболее информативных переменных.

Шаг 1. Из объясняющих переменных $x^{(1)} - x^{(5)}$ выделяется переменная $x^{(4)}$, имеющая с зависимой переменной Y наибольший коэффициент детерминации R^2 (равный для парной модели квадрату коэффициента корреляции r). Очевидно, это переменная $x^{(4)}$, так как коэффициент детерминации $R^2_{y4} = r^2_{y4} = 0.58^2 = 0.336$ — максимальный. С учетом поправки на несмещенность скорректированный коэффициент детерминации $\widehat{R}^2_{y4} = 1 - 19(1 - 0.336)/18 = 0.299$.

Шаг 2. Среди возможных пар объясняющих переменных $x^{(4)}$, $x^{(j)}$, $j = 1, 2, 3, 5$ выбирается пара $(x^{(4)}, x^{(3)})$, имеющая с зависимой переменной Y наиболее высокий коэффициент детерминации $R^2_{y4j} = R^2_{y43} = 0.483$ и с учетом поправки $\widehat{R}^2_{y43} = 1 - 19(1 - 0.483)/17 = 0.422$.

Шаг 3. Среди всевозможных троек объясняющих переменных $(x^{(4)}, x^{(3)}, x^{(j)})$, $j = 1, 2, 5$ наиболее информативной оказалась тройка $(x^{(4)}, x^{(3)}, x^{(5)})$, имеющая мак-

симальный коэффициент детерминации $R^2_{y43j} = R^2_{y435} = 0.513$ и соответственно скорректированный коэффициент $R^2_{y435} = 0.422$.

Так как скорректированный коэффициент детерминации на 3-м шаге не увеличился, то в регрессионной модели достаточно ограничиться лишь двумя отобранными ранее объясняющими переменными $x^{(4)}$ и $x^{(3)}$.

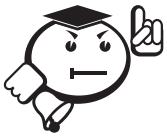
Рассчитанное уравнение регрессии по этим переменным примет вид:

$$\hat{y} = \underset{(0.66)}{7.29} + \underset{(0.13)}{3.48}x^{(3)} + \underset{(1.07)}{3.48}x^{(4)}.$$

Теперь все коэффициенты регрессии значимы, так как каждое из значений t-статистики:

$$t_{\Theta_0} = \frac{7.29}{0.66} = 11.0; \quad t_{\Theta_3} = \frac{3.48}{0.13} = 26.8; \quad t_{\Theta_4} = \frac{3.48}{1.07} = 3.25$$

больше соответствующего табличного значения $t_{0.95;17} = 2.11$.



Замечание. Так как значения коэффициентов корреляции весьма высокие (больше 0.8): $r_{12} = 0.85$, $r_{13} = 0.98$, $r_{23} = 0.88$, то, очевидно, из соответствующих трех переменных $x^{(1)}, x^{(2)}, x^{(3)}$ две переменные можно было сразу исключить из регрессии без проведения пошагового отбора, но какие именно переменные исключить — следовало решать, исходя из качественных соображений, основанных на знании предметной области (в данном случае влияния на урожайность факторов сельскохозяйственного производства).

Кроме рассмотренной выше пошаговой процедуры присоединения объясняющих переменных используются также пошаговые процедуры присоединения — удаления и процедура удаления объясняющих переменных. Следует отметить, что какая бы пошаговая процедура ни использовалась, она не гарантирует определения оптимального (в смысле получения максимального коэффициента детерминации $\widehat{R}^2$) набора объясняющих переменных. Однако в большинстве случаев получаемые с помощью пошаговых процедур наборы переменных оказываются оптимальными или близкими к оптимальным.

Метод главных компонент

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем признаков. Так,

например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер-рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показатели исследования, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам.

Первой главной компонентой $\widetilde{z}^{(1)}(X),\ldots,\widetilde{z}^{(p)}(X)$ исследуемой системы показателей $X = \left(x^{(1)} \quad \ldots \quad x^{(p)}\right)^T$ называется такая нормированно-центрированиая линейная комбинация этих показателей, которая среди всех прочих нормировочно-центрированных линейных комбинаций переменных $x^{(1)},\ldots,x^{(p)}$ обладает наибольшей дисперсией.

k-й главной компонентой $\widetilde{z}^{(k)}(X)$ $(k = 2,3,\ldots,p)$ исследуемой системы показателей $X = \left(x^{(1)} \quad \ldots \quad x^{(p)}\right)^T$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k-1$ предыдущими главными компонентами и среди всех прочих нормированно-центрированных и некоррелированных с предыдущими $k-1$ главными компонентами линейных комбинаций переменных $x^{(1)},\ldots,x^{(p)}$ обладает наибольшей дисперсией.

Замечание 1 (переход к центрированным переменным). Поскольку, как увидим ниже, решение задачи (а именно вид матрицы линейного преобразования L) зависит только от элементов ковариационной матрицы Σ, которые в свою очередь не изменяются при замене исходных переменных $x^{(j)}$ переменными $x^{(j)} - c^{(j)}$ ($c^{(j)}$ — произвольные постоянные числа), то в дальнейшем будем считать, что исходная система показателей уже центрирована, т. е. что $Ex^{(j)} = 0$, $j = 1,2,\ldots,p$. В статистической практике этого добиваются, переходя к наблюдениям $\widetilde{x}_i^{(j)} = x_i^{(j)} - \overline{x}^{(j)}$, где $\overline{x}^{(j)} = \sum_{i=1}^{n} x_i^{(j)}/n$ (для упрощения обозначений волнистую черту над центрированной переменной и над главной компонентой в дальнейшем ставить не будем).

Замечание 2. Использование главных компонент оказывается наиболее естественным и плодотворным в ситуациях, в которых все компоненты $x^{(1)},x^{(2)},\ldots,x^{(p)}$ исследуемого вектора X имеют общую физическую природу и соответственно измерены в одних и тех же единицах. К таким примерам можно отнести исследование структуры бюджета времени индивидуумов (все $x^{(i)}$ измеряются в единицах времени), исследование структуры потребления семей (все $x^{(i)}$ измеряются в денежных единицах), исследование общего развития и умственных способностей индивидуумов с помощью специальных тестов (все $x^{(i)}$ измеряются в баллах), разного рода антропологические исследования (все $x^{(i)}$ измеряются в единицах меры длины) и т. д. Если же признаки $x^{(1)},x^{(2)},\ldots,x^{(p)}$ измеряются в различных единицах, то результаты исследования с помощью главных компонент будут существенно зависеть от выбора масштаба и природы единиц измерения. Поэтому в подобных ситуациях

исследователь предварительно переходит к вспомогательным безразмерным признакам $x^{*(i)}$, например с помощью нормирующего преобразования

$$x_{\nu}^{*(i)} = \frac{x_{\nu}^{(i)}}{\sqrt{\widehat{\sigma}_{ii}}} \quad \begin{pmatrix} i = 1, 2, \ldots, p \\ \nu = 1, 2, \ldots, n \end{pmatrix},$$

(где $\widehat{\sigma}_{ii}$ — ковариация), а затем строит главные компоненты относительно этих вспомогательных признаков X^* и их ковариационной матрицы, которая является одновременно выборочной корреляционной матрицей исходных наблюдений.

Для расчета главных компонент в первую очередь определяется ковариационная матрица Σ связи между переменными $x^{(1)}, \ldots, x^{(p)}$.

Далее необходимо вычислить собственные числа этой матрицы:

$$|\Sigma - \lambda I| = 0. \tag{7.1}$$

Это уравнение (относительно λ) называется характеристическим для матрицы Σ. При симметричности и неотрицательной определенности матрицы Σ (каковой она и является как всякая ковариационная матрица) это уравнение имеет p вещественных неотрицательных корней $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$, называемых характеристическими (или собственными) значениями матрицы Σ.

Подставляем λ_1 в систему уравнений

$$(\Sigma - \lambda I) l_1^T = 0 \tag{7.2}$$

и, решая ее относительно $l_{11}, \ldots, l_{1p}$, определяем компоненты вектора l_1.

Таким образом, первая главная компонента получается как линейная комбинация $\widetilde{z}^{(1)}(X) = l_1 X$, где l_1 — собственный вектор матрицы Σ, соответствующий наибольшему собственному числу этой матрицы.

Аналогично $\widetilde{z}^{(k)}(X) = l_k X$, где l_k — собственный вектор матрицы Σ, соответствующий k-му по величине собственному значению λ_k этой матрицы.

Таким образом, соотношения для определения всех p главных компонент вектора X могут быть представлены в виде

$$\widetilde{Z} = LX,$$

где $\widetilde{Z} = \left(\widetilde{z}^{(1)} \quad \ldots \quad z^{(p)}\right)^T$, $X = \left(x^{(1)} \quad \ldots \quad x^{(p)}\right)^T$, а матрица L состоит из строк $l_j = (l_{j1}, \ldots, l_{jp})$, $j = \overline{1,p}$, являющихся собственными векторами матрицы Σ, соответствующими собственным числам λ_j. При этом сама матрица L в является ортогональной, т. е.

$$LL^T = L^T L = I.$$

Критерий информативности метода главных компонент может быть представлен в виде

$$I_{p'}\left(Z(X)\right) = \frac{\lambda_1 + \ldots + \lambda_{p'}}{\lambda_1 + \ldots + \lambda_p}, \tag{7.3}$$

где $\lambda_1, \lambda_2, \ldots, \lambda_p$ — собственные числа ковариационной матрицы Σ вектора X, расположенные в порядке убывания.

Этот критерий дает исследователю некоторую основу, опорную точку зрения, при вынесении решения о том, сколько последних главных компонент можно без

особого ущерба изъять из рассмотрения, сократив тем самым размерность исследуемого пространства.

Действительно, анализируя с помощью (7.3) изменение относительно доли дисперсии, вносимой первыми p' главными компонентами, в зависимости от числа этих компонент, можно разумно определить число компонент, которое целесообразно оставить в рассмотрении. Так, при изменении $I_{p'}$, изображенном на рисунке 7.3, очевидно, целесообразно было бы сократить размерность пространства с $p = 10$ до $p' = 3$, так как добавление всех остальных семи главных компонент может повысить суммарную характеристику рассеяния не более чем на 10%.

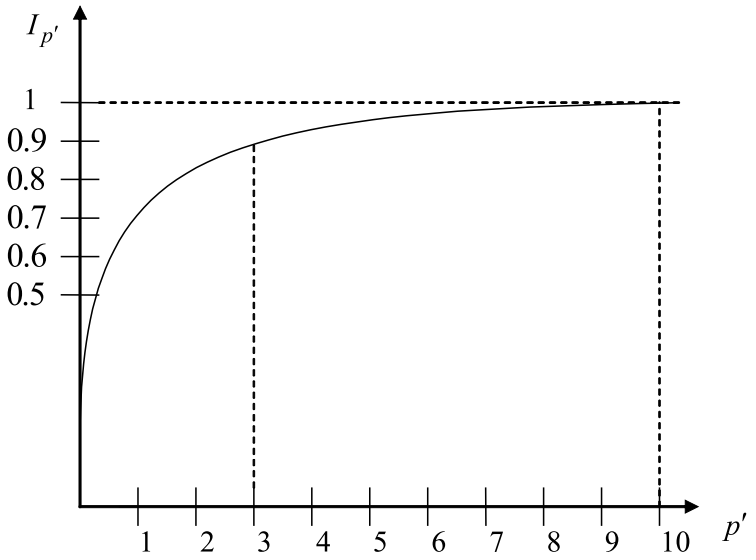


Рис. 7.3 – Изменение относительной доли суммарной дисперсии используемых признаков, обусловленной первыми p' главными компонентами, в зависимости от p' (случай $p = 10$)

Пример 7.2

При формировании типообразующих признаков предприятий отрасли были обследованы 24 предприятия ($n = 24$) по трем технико-экономическим показателям: объему выпускаемой продукции $x^{(1)}$, основным фондам $x^{(2)}$ и себестоимости $x^{(3)}$ (все переменные измерялись в денежных единицах). По полученным в результате обследования исходным статистическим данным $\left(x_i^{(1)}, x_i^{(2)}, x_i^{(3)}\right)$, $i = 1, 2, \ldots, 24$ была определена выборочная ковариационная матрица

$$\Sigma = \begin{pmatrix} 451.39 & 271.17 & 168.70 \\ 271.17 & 171.73 & 103.29 \\ 168.70 & 103.29 & 66.65 \end{pmatrix}.$$

Решая кубическое уравнение (относительно λ) вида

$$\begin{vmatrix} 451.39-\lambda & 271.17 & 168.70 \\ 271.17 & 171.73-\lambda & 103.29 \\ 168.70 & 101.29 & 66.65-\lambda \end{vmatrix} = 0,$$

находим $\lambda_1 = 680.40$, $\lambda_2 = 6.50$, $\lambda_3 = 2.86$.

Подставляя последовательно численные значения λ_1, λ_2 и λ_3 в систему (6.3) и решая эти системы относительно неизвестных $l_i = (l_{i1}, l_{i2}, l_{i3})$ $(i = 1,2,3)$, получаем

$$l_1 = \begin{pmatrix} 0.8126 & 0.4955 & 0.3068 \end{pmatrix},$$
$$l_2 = \begin{pmatrix} -0.5454 & 0.8321 & 0.1006 \end{pmatrix},$$
$$l_3 = \begin{pmatrix} -0.2054 & -0.2491 & 0.9465 \end{pmatrix}.$$

В качестве главных компонент получаем

$$z^{(1)} = 0.81x^{(1)} + 0.50x^{(2)} + 0.31x^{(3)},$$
$$z^{(2)} = -0.55x^{(1)} + 0.83x^{(2)} + 0.10x^{(3)},$$
$$z^{(3)} = -0.21x^{(1)} - 0.25x^{(2)} + 0.95x^{(3)}.$$

Здесь под $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ подразумеваются отклонения объема выпускаемой продукции $(x^{(1)})$, основных фондов $(x^{(2)})$ и себестоимости $(x^{(3)})$ предприятия от соответствующих средних значений.

Вычисление относительной доли суммарной дисперсии, обусловленной одной, двумя и тремя главными компонентами, в соответствии с формулой (6.4) дает

$$q(1) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0.9864;$$

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0.9958;$$

$$q(3) = 1.$$

Отсюда можно сделать вывод, что почти вся информация о специфике предприятия данного типа содержится в одной лишь первой главной компоненте, которую и естественно использовать при соответствующей типологизации предприятий.

. . . . . .

Матрица «нагрузок» $A = (a_{ij})$, $i,j = 1,2,\ldots,p$ главных компонент на исходные признаки также является важной характеристикой главных компонент. Если анализируемые переменные $X = \begin{pmatrix} x^{(1)} & x^{(2)} & \ldots & x^{(p)} \end{pmatrix}^T$ предварительно процентрированы и пронормированы (см. замечания 1 и 2), т. е. если главные компоненты строятся для признаков $X^* = \begin{pmatrix} x^{*(1)} & x^{*(2)} & \ldots & x^{*(p)} \end{pmatrix}^T$, $Ex^{*(i)} = 0$, $Dx^{*(i)} = 1$, $i = 1,2,\ldots,p$, то элементы матрицы нагрузок a_{ij} определяют одновременно степень тесноты парной линейной связи (т. е. парный коэффициент корреляции) между $x^{*(i)}$ и $z^{(j)}$ и удельный вес влияния пронормированной j-й главной компоненты на признак $x^{*(i)}$.

Матрица нагрузок A определяется соотношением

$$A = L^T \Lambda^{\frac{1}{2}}, \text{ где } \Lambda^{\frac{1}{2}} = \Sigma_Z^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \cdots & \\ 0 & & & \sqrt{\lambda_p} \end{pmatrix}.$$

Пример 7.3

Компонентный анализ проведен по данным двадцати сельскохозяйственных районов (n = 20) области, которые содержат результаты измерений следующих показателей: $x^{(1)}$ — число колесных тракторов на 100 га; $x^{(2)}$ — число зерноуборочных комбайнов на 100 га; $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га; $x^{(4)}$ — количество удобрений, расходуемых на гектар; $x^{(5)}$ — количество средств защиты растений, расходуемых на гектар.

Требовалось выделить $p' < 5$ первых главных компонент для анализа и дать им содержательную интерпретацию.

Расчеты проводились по нормированным данным и представлены в таблице 7.2.

Таблица 7.2 – Данные расчета

Главные компоненты $z^{(i)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$
Собственные значения λ_i	3.04	1.41	0.43	0.10	0.02
Вклад i-ой главной компоненты в суммарную дисперсию, %	60.8	28.2	8.6	2.0	0.4
Суммарный вклад первых главных компонент (%)	60.8	89.0	97.6	99.6	100.0

При расчете относительного вклада главных компонент учитывалось, что $\sum_{i=1}^{p} \lambda_i =$ $= p = 5$. Для анализа были оставлены две первые главные компоненты ($p = 2$), на которые приходится 89% суммарной вариации.

Для интерпретации главных компонент построена матрица факторных нагрузок

$$A = \begin{pmatrix} 0.95^* & 0.97^* & 0.94^* & 0.24 & 0.56 \\ -0.19 & -0.17 & -0.28 & 0.88^* & 0.67^* \end{pmatrix}^T.$$

Звездочкой (*) отмечены элементы $|a_{ij}| > 0.6$, которые следует учитывать при интерпретации главных компонент $z^{(1)}$ и $z^{(2)}$.

Из вида матрицы нагрузок A следует, что первая главная компонента наиболее тесно связана с показателями: $x^{(1)}$ — число колесных тракторов ($a_{11} = r(x^{(1)}, z^{(1)}) =$ $= 0.95$); $x^{(2)}$ — число зерноуборочных комбайнов ($a_{21} = r(x^{(2)}, z^{(1)}) = 0.97$); $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га ($a_{31} = r(x^{(3)}, z^{(1)}) = 0.94$). Поэтому первая главная компонента $z^{(1)}$ интерпретирована как уровень механизации работ.

Вторая главная компонента $z^{(2)}$ тесно связана с количествами удобрения $(x^{(4)})$ и средств защиты растений $(x^{(5)})$, расходуемых на гектар. Соответственно $z^{(2)}$ интерпретируется как уровень химизации растениеводства.

. . . . . .

7.4 Линейные регрессионные модели с переменной структурой. Фиктивные переменные

До сих пор мы рассматривали регрессионную модель, в которой в качестве объясняющих переменных (регрессоров) выступали количественные переменные (производительность труда, себестоимость продукции, доход и т. п.). Однако на практике достаточно часто возникает необходимость исследования влияния качественных признаков, имеющих два или несколько уровней (градаций). К числу таких признаков можно отнести: пол (мужской, женский), образование (начальное, среднее, высшее), фактор сезонности (зима, весна, лето, осень) и т. п.

Качественные признаки могут существенно влиять на структуру линейных связей между переменными и приводить к скачкообразному изменению параметров регрессионной модели. В этом случае говорят об исследовании регрессионных моделей с переменной структурой или построении регрессионных моделей по неоднородным данным.

Например, нам надо изучить зависимость размера заработной платы Y работников не только от количественных факторов $x_1, \ldots, x_n$, но и от качественного признака $z^{(1)}$ (например, фактора «пол работника»).

В принципе, можно было получить оценки регрессионной модели

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \ldots + \Theta_p x_i^{(p)} + \varepsilon_i, \quad i = 1, \ldots, n$$

для каждого уровня качественного признака (т. е. выборочное уравнение регрессии отдельно для работников-мужчин и отдельно — для женщин), а затем изучать различия между ними.

Но есть и другой подход, позволяющий оценивать влияние значений количественных переменных и уровней качественных признаков с помощью одного уравнения регрессии. Этот подход связан с введением так называемых фиктивных (манекенных) переменных, или манекенов.

В качестве фиктивных переменных обычно используются дихотомические (бинарные, булевы) переменные, которые принимают всего два значения: «0» или «1» (например, значение такой переменной z_1 по фактору «пол»: $z^{(1)} = 0$ для работников-женщин и $z^{(1)} = 1$ — для мужчин).

В этом случае первоначальная регрессионная модель заработной платы изменится и примет вид:

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \ldots + \Theta_p x_i^{(p)} + \alpha_1 z_i^{(1)} + \varepsilon_i, \quad i = 1, \ldots, n,$$

где $z_i^{(1)} = \begin{cases} 1, & \text{если } i\text{-й работник мужского пола;} \\ 0, & \text{если } i\text{-й работник женского пола.} \end{cases}$

Таким образом, принимая эту модель, мы считаем, что средняя заработная плата у мужчин на α_1 выше, чем у женщин, при неизменных значениях других параметров модели. А проверяя гипотезу H_0: $\alpha_1 = 0$, мы можем установить существенность влияния фактора «пол» на размер заработной платы работника.

Следует отметить, что в принципе качественное различие можно формализовать с помощью любой переменной, принимающей два разных значения, не обязательно «0» или «1». Однако в эконометрической практике почти всегда используются фиктивные переменные типа «0–1», так как при этом интерпретация полученных результатов выглядит наиболее просто. Так, если бы в рассмотренной модели в качестве фиктивной выбрали переменную $z^{(1)}$, принимающую значения $z_i^{(1)} = 4$ (для работников-мужчин) и $z_i^{(1)} = 1$ (для женщин), то коэффициент регрессии α_1 при этой переменной равнялся бы $1/(4-1)$, т. е. одной трети среднего изменения заработной платы у мужчин.

Если рассматриваемый качественный признак имеет несколько (k) уровней (градаций), то, в принципе, можно было ввести в регрессионную модель дискретную переменную, принимающую такое же количество значений (например, при исследовании зависимости заработной платы Y от уровня образования Z можно рассматривать $k = 3$ значения: $z_i^{(1)} = l$ при наличии начального образования, $z_i^{(1)} = 2$ — среднего и $z_i^{(1)} = 3$ при наличии высшего образования). Однако обычно так не поступают из-за трудности содержательной интерпретации соответствующих коэффициентов регрессии, а вводят $(k-1)$ бинарных переменных.

В рассматриваемом примере для учета фактора образования можно было в рассмотренную выше регрессионную модель ввести $k - 1 = 3 - 1 = 2$ бинарные переменные $z_i^{(21)}$ и $z_i^{(22)}$:

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \ldots + \Theta_p x_i^{(p)} + \alpha_1 z_i^{(1)} + \alpha_{21} z_i^{(21)} + \alpha_{22} z_i^{(22)} + \varepsilon_i,$$

где $z_i^{(21)} = \begin{cases} 1, & \text{если } i\text{-й работник имеет высшее образование;} \\ 0, & \text{во всех остальных случаях;} \end{cases}$

$z_i^{(22)} = \begin{cases} 1, & \text{если } i\text{-й работник имеет среднее образование;} \\ 0, & \text{во всех остальных случаях.} \end{cases}$

Третьей бинарной переменной не требуется: если i-й работник имеет начальное образование, это будет отражено парой значений $z_i^{(21)} = 0$, $z_i^{(22)} = 0$.

Более того, вводить третью бинарную переменную $z_i^{(23)}$ (со значениям $z_i^{(23)} = 1$, если i-й работник имеет начальное образование; $z_i^{(23)} = 0$ — в остальных случаях) нельзя, так как при этом для любого i-го работника $z_i^{(21)} + z_i^{(22)} + z_i^{(23)} = 1$, т. е. при суммировании элементов столбцов общей матрицы плана, соответствующих фиктивным переменным $z_i^{(21)}, z_i^{(22)}, z_i^{(23)}$, мы получили бы столбец, состоящий из одних единиц. А так как в матрице плана такой столбец из единиц уже есть (это первый столбец, соответствующий свободному члену уравнения регрессии), то это означало бы линейную зависимость значений (столбцов) общей матрицы плана X, т. е. нарушило бы предпосылку регрессионного анализа. Таким образом, мы оказались бы в условиях мультиколлинеарности в функциональной форме и как следствие — невозможности получения оценок методом наименьших квадратов.

Такая ситуация, когда сумма значений нескольких переменных, включенных в регрессию, равна постоянному числу (единице), получила название «ловушки».

Чтобы избежать такой ловушки, число вводимых бинарных переменных должно быть на единицу меньше числа уровней (градаций) качественного признака.

Рассматриваемые выше регрессионные модели отражали влияние качественного признака (фиктивных переменных) только на значения переменной Y, т. е. на свободный член уравнения регрессии. В более сложных моделях может быть отражена также зависимость фиктивных переменных на сами параметры при переменных регрессионной модели. Например, при наличии в модели объясняющих переменных — количественной $x^{(1)}$ и фиктивных $z_i^{(11)}$, $z_i^{(12)}$, $z_i^{(21)}$, $z_i^{(22)}$, из которых $z_i^{(11)}$, $z_i^{(12)}$ влияют только на значение коэффициента при $x^{(1)}$, а $z_i^{(21)}$, $z_i^{(22)}$ — только на величину свободного члена уравнения, такая регрессионная модель примет вид:

$$y_i = \Theta_0 + \Theta_1 x_i^{(1)} + \Theta_{11}\left(z_i^{(11)} x_i^{(1)}\right) + \Theta_{12}\left(z_i^{(12)} x_i^{(1)}\right) + \alpha_{21} z_i^{(21)} + \alpha_{22} z_i^{(22)} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Модели такого типа используются, например, при исследовании зависимости объема потребления Y некоторого продукта от дохода потребителя X, когда одни качественные признаки (например, фактор сезонности) влияют лишь на количество потребляемого продукта (свободный член уравнения регрессии), а другие (например, уровень доходности домашнего хозяйства) — на параметр Θ_1 при X, интерпретируемый как «склонность к потреблению».

7.5 Критерий Г. Чоу

В практике эконометриста нередки случаи, когда имеются две выборки пар значений зависимой и объясняющих переменных (x_i, y_i). Например, одна выборка пар значений переменных объемом n_1 получена при одних условиях, а другая, объемом n_2, — при несколько измененных условиях. Необходимо выяснить, действительно ли две выборки однородны в регрессионном смысле? Другими словами, можно ли объединить две выборки в одну и рассматривать единую модель регрессии Y по X?

При достаточных объемах выборок можно было, например, построить интервальные оценки параметров регрессии по каждой из выборок и в случае пересечения соответствующих доверительных интервалов сделать вывод о единой модели регрессии. Возможны и другие подходы.

В случае если объем хотя бы одной из выборок незначителен, то возможности такого и аналогичных подходов резко сужаются из-за невозможности построения сколько-нибудь надежных оценок.

В критерии (тесте) Г. Чоу эти трудности в существенной степени преодолеваются. По каждой выборке строятся две линейные регрессионные модели:

$$y_i = \Theta_0' + \sum_{j=1}^{p} \Theta_p' x_i^{(j)} + \varepsilon_i', \quad i = 1, \ldots, n_1;$$

$$y_i = \Theta_0'' + \sum_{j=1}^{p} \Theta_p'' x_i^{(j)} + \varepsilon_i'', \quad i = n_1 + 1, \ldots, n_1 + n_2.$$

Проверяемая нулевая гипотеза имеет вид — H_0: $\Theta' = \Theta''$; $D(\varepsilon') = D(\varepsilon'') = \sigma^2$, где $\Theta' = \Theta''$ — векторы параметров двух моделей; $\varepsilon', \varepsilon''$ — их случайные возмущения.

Если нулевая гипотеза H_0 верна, то две регрессионные модели можно объединить в одну объема $n = n_1 + n_2$:

$$y_i = \Theta_0 + \sum_{j=1}^{p} \Theta_p x_i^{(j)} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Согласно критерию Г. Чоу нулевая гипотеза H_0 отвергается на уровне значимости α, если статистика:

$$F = \frac{\left(\sum_{i=1}^{n} \varepsilon_i^2 - \sum_{i=1}^{n_1} \varepsilon_i^2 - \sum_{i=n_1}^{n} \varepsilon_i^2 \right) (n - 2p - 2)}{\left(\sum_{i=1}^{n_1} \varepsilon_i^2 + \sum_{i=n_1}^{n} \varepsilon_i^2 \right) (p + 1)} > F_{\alpha; p+1; n-2; p-2}, \tag{7.4}$$

где $\sum_{i=1}^{n} \varepsilon_i^2$, $\sum_{i=1}^{n_1} \varepsilon_i^2$, $\sum_{i=n_1}^{n} \varepsilon_i^2$ — остаточные суммы квадратов соответственно для объединенной, первой и второй выборок; $n = n_1 + n_2$.

Критерий Г. Чоу может быть использован при построении регрессионных моделей при воздействии качественных признаков, когда имеется возможность разделения совокупности наблюдений по степени воздействия этого фактора на отдельные группы и требуется установить возможность использования единой модели регрессии.

Оценивание регрессии с использованием фиктивных переменных более информативно в том отношении, что позволяет использовать t-критерий для оценки существенности влияния каждой фиктивной переменной на зависимую переменную.

7.6 Частная корреляция

Если переменные коррелируют друг с другом, то на значении коэффициента корреляции частично сказывается влияние других переменных. В связи с этим часто возникает необходимость исследовать частную корреляцию между переменными при исключении (элиминировании) влияния одной или нескольких переменных.

Выборочным частным коэффициентом корреляции (или просто частным коэффициентом корреляции) между переменными X_i и X_j при фиксированных значениях остальных $(p - 2)$ переменных называется выражение

$$r_{ij.1,2\ldots p} = \frac{-q_{ij}}{\sqrt{q_{ii} q_{jj}}},$$

где q_{ii} и q_{jj} — алгебраические дополнения элементов r_{ii} и r_{jj} матрицы выборочных коэффициентов корреляции

$$q_p = \begin{pmatrix} 1 & r_{12} & \ldots. & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}.$$

В частности, в случае трех переменных ($n = 3$) следует, что

$$r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{\left(1 - r_{ik}^2\right)\left(1 - r_{jk}^2\right)}}. \qquad (7.5)$$

Поясним полученную формулу (7.5). Предположим, что имеется обычная регрессионная модель $x_i = \Theta_0 + \Theta_1 x_j + \Theta_2 x_k + \varepsilon_i$ и необходимо оценить корреляцию между зависимой переменной X_i и объясняющей переменной X_j при исключении (элиминировании) влияния другой объясняющей переменной X_k. С этой целью найдем уравнения парной регрессии X_i по X_k $\left(\hat{x}_i = \Theta_0 + \Theta_1 x_k\right)$ и X_j по X_k $\left(\hat{x}_j = \Theta_0' + \Theta_1' x_k\right)$, а затем удалим влияние переменной X_k, взяв остатки $\varepsilon_{x_i} = x_i - \hat{x}_i$ и $\varepsilon_{x_j} = x_j - \hat{x}_j$. Очевидно, что коэффициент корреляции между остатками ε_{x_i} и ε_{x_j} будет отражать тесноту частной корреляции между переменными X_i и X_j при исключении влияния переменной X_k. Обычный коэффициент корреляции между остатками ε_{x_i} и ε_{x_j} равен частному коэффициенту корреляции $r_{ij.k}$, определенному по формуле (6.2).

Частный коэффициент корреляции $r_{ij.12\ldots p}$, как и парный коэффициент r_{ij}, может принимать значения от –1 до +1. Кроме того, $r_{ij.12\ldots p}$, вычисленный на основе выборки объема n, имеет такое же распределение, как и r_{ij}, вычисленный по $n' = n - p + 2$ наблюдениям. Поэтому значимость частного коэффициента корреляции $r_{ij.12\ldots p}$ оценивают так же, как и обычного коэффициента корреляции r, но при этом полагают $n' = n - p + 2$.

7.7 Построение КЛММР по неоднородным данным в условиях, когда значения сопутствующих переменных неизвестны

Если воздействия сопутствующих качественных факторов на структуру КЛММР скрыты, т. е. их «значения» не поддаются учету и контролю или не были своевременно (в процессе сбора исходных статистических данных) зарегистрированы, то мы можем оказаться в ситуации, когда собранные исходные статистические данные в действительности представляют собой смесь нескольких регрессионно однородных подвыборок, однако выявить эти подвыборки по значениям сопутствующих переменных мы не можем. Игнорирование этого обстоятельства, выражающееся в построении единой регрессионной зависимости на основании всех имеющихся в распоряжении эконометрика исходных данных, является причиной многих недоразумений и неудач в прикладных экономических исследованиях. Для подтверждения этого тезиса рассмотрим следующий пример.

Пример 7.4

В целях исследования зависимости интенсивности региональных эмиграционных процессов (y%) от уровня (общей продолжительности) полученного образования (x лет) были собраны исходные статистические данные вида за определенный промежуток времени в анализируемом регионе. Так что элемент (x_i, y_i) выборки

интерпретируется в данном случае следующим образом: x_i (лет) — общая продолжительность процесса обучения взрослого (т. е. в возрасте не менее 25 лет) жителя региона, y_i — процент уехавших из региона за рассматриваемый промежуток времени взрослых жителей среди всех взрослых жителей с уровнем образования x_i (лет). Несмотря на гипотетическую уверенность специалистов в существовании достаточно тесной статистической связи между x и y, регрессионный анализ данных свидетельствовал об отсутствии какой бы то ни было зависимости между этими переменными. Обратившись к визуальному анализу исходных статистических данных, естественно предположить, что наша выборка регрессионно не однородна и состоит из двух пересекающихся «крестом» подвыборок, каждая из которых имеет вид линейно вытянутого эллиптического облака точек-наблюдений. Более внимательный содержательный анализ каждой из этих подвыборок позволил обнаружить и скрытый (не учтенный в ходе сбора статистических данных) сопутствующий признак z. Им оказался тип полученного образования с двумя градациями: 1 — гуманитарное, 2 — естественно-научно-техническое. Разделение всех имеющихся данных на подвыборки по этой сопутствующей переменной и построение искомой регрессионной зависимости отдельно для каждой из этих подвыборок дают две различные модели достаточно высокой прогностической силы и противоположной направленности: если для жителей с гуманитарным образованием (им соответствуют «штрихи» на рисунке 7.4) интенсивность эмиграции падает (по линейному закону) с ростом уровня образования, то для жителей с естественно-научно-техническим образованием (им соответствуют точки на рисунке 7.4) мы наблюдаем обратную картину.

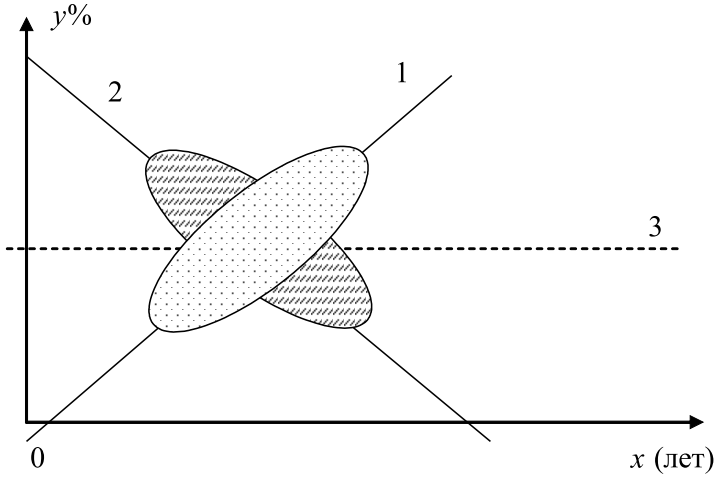


Рис. 7.4 – Графики аппроксимирующих функций регрессии, построенных соответственно по наблюдениям подвыборок: 1 (точки), 2 (штрихи) и по объединенной выборке, состоящей из тех и других наблюдений

Возвращаясь к общей проблеме построения линейной регрессионной модели по неоднородным данным в условиях, когда значения сопутствующих переменных неизвестны, отметим, что уже при трех объясняющих переменных (т. е. при $p = 3$) визуальный анализ геометрической структуры исходной выборки принци-

пиально невозможен. Поэтому с целью предварительного анализа геометрической структуры данных и выделения в них регрессионно однородных подвыборок необходимо использовать методы кластер-анализа, в том числе методы расщепления смеси многомерных распределений вероятностей. Подчеркнем, что разбиение исходных данных на регрессионно однородные подвыборки (кластеры) проводится в объединенном $(p+1)$-мерном пространстве (X, y) (а не в пространстве только объясняющих переменных X). Причем процесс этот либо предшествует процессу построения регрессионных моделей, либо проводится в режиме итерационного взаимодействия с последним.

Контрольные вопросы по главе 7

1. Что понимается под эластичностью?
2. Как рассчитывается эластичность для линейной, двойной логарифмической, линейно-логарифмической, обратной функции регрессии?
3. Что понимается под мультиколлинеарностью?
4. Как можно выявить мультиколлинеарность?
5. Какие значения коэффициента *vif* говорят о наличии мультиколлинеарности?
6. Какие можно привести примеры преобразования переменных для устранения мультиколинеарности?
7. Какие существуют способы устранения мультиколлинеарности?
8. Для чего нужно выполнять отбор наиболее существенных объясняющих переменных?
9. Из каких шагов состоит процедура пошагового отбора параметров?
10. Как происходит отбор параметров с помощью метода главных компонент?
11. Как выполняется переход к центрированным переменным?
12. Как привести параметры исследуемого вектора к одним единицам измерения?
13. Каким образом можно определить ковариационную матрицу Σ связи между переменными $x^{(1)}, \ldots, x^{(p)}$?
14. Как можно вычислить первую главную компоненту?
15. Как рассчитывается критерий информативности метода главных компонент?
16. Что понимают под матрицей нагрузок?
17. Как определяют матрицу нагрузок?
18. Как можно учитывать в регрессионной модели качественные переменные?
19. Какие можно привести примеры фиктивных переменных?
20. Какая ситуация называется «ловушкой»?
21. Для каких целей используется критерий Чоу?

22. Какая статистика вычисляется для проверки критерия Чоу?
23. Для чего необходимо вычислять частную корреляцию?
24. Что понимают под выборочным коэффициентом корреляции?
25. Какие значения может принимать частный коэффициент корреляции?
26. Как можно оценить результирующую переменную, если значения некоторых объясняющих переменных не поддаются учету или не были своевременно зарегистрированы?

ЗАКЛЮЧЕНИЕ

В данном курсе Вы познакомились с эконометрическими моделями, с помощью которых можно исследовать различные экономические явления.

Модели реальных объектов часто оказываются намного сложнее, так что без использования персонального компьютера и специализированных эконометрических пакетов получить необходимую информацию довольно сложно. Компьютерные эконометрические пакеты сделали эконометрические методы более доступными и наглядными, так как наиболее трудоемкую (рутинную) работу по расчету различных статистик, параметров, характеристик, построению таблиц и графиков в основном стал выполнять компьютер, а эконометристу осталась главным образом творческая работа: постановка задачи, выбор соответствующей модели и метода ее решения, интерпретация результатов.

Для исследования экономических процессов применяют и другие виды моделей: оптимизационные, имитационные и т. д. В последующих курсах (математическое и имитационное моделирование экономических процессов, исследование операций в экономике) Вы продолжите изучение метода моделирования.

ЛИТЕРАТУРА

[1] Кремер Н. Ш. Эконометрика : учебник для вузов / Н. Ш. Кремер, Б. А. Путко. — М. : ЮНИТИ-ДАНА, 2002. — 311 с.

[2] Айвазян С. А. Прикладная статистика и основы эконометрики : учебник для вузов / С. А. Айвазян, В. С. Мхитарян. — М. : ЮНИТИ, 1998. — 1005 с.

[3] Бородич С. А. Вводный курс эконометрики / С. А. Бородич. — Мн. : БГУ, 2000. — 354 с.

[4] Доугерти К. Введение в эконометрику / К. Доугерти. — М. : ИНФРА-М., 1997. — 335 с.

[5] http://crow.academy.ru/econometrics/ (дата обращения 19.11.2014).

[6] Gujarati D. N. Basics Econometrics / D. N. Gujarati. — New York : McGraw-Hill, 1995. — 838 p.

Приложение А

ФУНКЦИЯ СТАНДАРТНОГО НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	05596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.6948	0.5987	0.6026	0.6064	0.6301	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7167	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7464	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.9707	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	09370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9723	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

продолжение на следующей странице

z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
2.0	0.9772	0.9778	0.9783	0.9783	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9830	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9868	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9898	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9922	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9941	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9956	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9967	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9976	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9982	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9987	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Приложение Б

КВАНТИЛИ РАСПРЕДЕЛЕНИЯ $\chi^2(\nu)$

ν / α	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**
1	0.00004	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	0.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	63.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.6	112.3	116.3
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.6	135.8	140.2
120	83.85	86.92	91.58	95.70	100.62	140.2	146.57	152.2	159.0	163.6

ν — число степеней свободы

Приложение В

ДВУСТОРОННИЕ КВАНТИЛИ РАСПРЕДЕЛЕНИЯ СТЬЮДЕНТА

Число степеней свободы	α		
	0.1	**0.05**	**0.01**
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.449
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
25	1.708	2.060	2.878
30	1.697	2.042	2.750
40	1.684	2.021	2.704
50	1.676	2.009	2.678
100	1.660	1.984	2.626
200	1.652	1.972	2.601
∞	1.645	1.960	2.576

Приложение Г

ТАБЛИЦА КРИТЕРИЯ ФИШЕРА ДЛЯ $\alpha = 0.05$

ν_2 \ ν_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	162	200	216	225	230	234	237	239	241	242
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.26	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.69	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.38	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

продолжение на следующей странице

ν_2 \ ν_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
200	3.92	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

ν_1 — число степеней свободы числителя,

ν_2 — число степеней свободы знаменателя

Приложение Д

ТАБЛИЦА КРИТЕРИЯ ДАРБИНА—УОТСОНА ДЛЯ $\alpha = 0.05$

	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
n	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
6	0.610	1.400	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—
9	0.824	1.320	0.624	1.699	0.455	2.128	0.296	2.588	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.814	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.824	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786

продолжение на следующей странице

	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
n	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
50	1.503	1.585	1.452	1.628	1.421	1.674	1.378	1.721	1.335	1.771
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
90	1.635	1.679	1.612	1.703	1.589	1.726	1.568	1.751	1.542	1.776
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.789	1.665	1.802
200	1.758	1.778	1.748	1.789	1.736	1.799	1.728	1.810	1.718	1.820

n — число наблюдений, m — число объясняющих переменных

	$m = 6$		$m = 7$		$m = 8$		$m = 9$		$m = 10$	
n	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
6	—	—	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—
11	0.203	3.005	—	—	—	—	—	—	—	—
12	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
30	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
40	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.945	2.149
50	1.219	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
60	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
70	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
80	1.480	1.801	1.428	1.831	1.425	1.861	1.397	1.893	1.369	1.925
90	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
100	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

Приложение Е

ТАБЛИЦА КРИТИЧЕСКИХ ЗНАЧЕНИЙ КОЛИЧЕСТВА РЯДОВ

Нижняя граница k_1

n_1	n_2																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	6	7	7	8	8	9	9	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

Верхняя граница k_2

n_1	n_2																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
4				9	9														
5			9	10	10	11	11												
6			9	10	11	12	12	13	13	13	13								
7				11	12	13	13	14	14	14	14	15	15	15					
8				11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9					13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10					13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11					13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12					13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13						15	16	17	18	19	19	20	20	21	21	22	22	23	23
14						15	16	17	18	19	20	20	21	22	22	23	23	23	24
15						15	16	18	18	19	20	21	22	22	23	23	24	24	25
16							17	18	19	20	21	21	22	23	23	24	25	25	25
17							17	18	19	20	21	22	23	23	24	25	25	26	26
18							17	18	19	20	21	22	23	24	25	25	26	26	27
19							17	18	20	21	22	23	23	24	25	26	26	27	27
20							17	18	20	21	22	23	24	25	25	26	27	27	28

Пример: пусть при $n = 20$ будет 11 знаков «+» ($= n_1$) и 9 знаков «–» ($= n_2$). Тогда при $\alpha = 0.05$ нижняя граница $k_1 = 6$, верхняя граница $k_2 = 16$. Если $k < 6$ или $k > 16$, то гипотеза об отсутствии автокорреляции должна быть отклонена.

Приложение Ж

НЕОБХОДИМЫЕ СВЕДЕНИЯ ИЗ МАТРИЧНОЙ АЛГЕБРЫ

Квадратной матрицей *называется матрица, у которой число строк равно числу столбцов, т. е.* $n \times n$ *— матрица (при любом целом* $n > 1$*). Элементы* $a_{11}, a_{22}, \ldots, a_{nn}$ *квадратной матрицы образуют ее главную диагональ.*

Среди квадратных матриц можно выделить:

- *диагональную матрицу D*, у которой все элементы, кроме элементов, стоящих на главной диагонали, равны нулю, т. е.

$$D = \begin{pmatrix} d_{11} & 0 & 0 & \ldots & 0 & 0 \\ 0 & d_{22} & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & d_{nn} \end{pmatrix} = diag\left(d_{11}, d_{22}, \ldots, d_{nn}\right);$$

- *единичную матрицу* I_n, которая является частным случаем $n \times n$ диагональной матрицы, поскольку все ее диагональные элементы равны единице:

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

Нижний индекс n определяет размерность матрицы, и в тех случаях, когда эта размерность очевидна из контекста, он может опускаться.

Равенство матриц. Две матрицы $A = (a_{ij})$ и $B = (b_{ij})$ называются равными, если они имеют одну и ту же размерность и если $a_{ij} = b_{ij}$ для всех i и j. Это означает, что равные матрицы совпадают поэлементно.

Вектор-строка *— это матрица размерности* $1 \times m$ *(*$m > 1$*), т. е. матрица, состоящая из единственной строки длины m. Например, результаты наблюдения значений p анализируемых переменных* $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ *на одном объекте, зарегистрированные в определенный момент времени t, образуют вектор-столбец*

$$X_t = \left(x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(p)}\right).$$

Вектор-столбец *— это матрица размерности* $n \times 1$ *(*$n \geqslant 1$*), т. е. матрица, состоящая из единственного столбца длины n.*

Например, результаты наблюдения одной какой-либо переменной $x^{(j)}$ на n статистически обследованных объектах (или на одном объекте, но зарегистрированные в n последовательных моментов времени) можно представить в виде вектора-столбца

$$X^{(j)} = \begin{pmatrix} x_1^{(j)} \\ x_2^{(j)} \\ \ldots \\ x_n^{(j)} \end{pmatrix}.$$

Транспонирование матрицы A определяется как действие, в результате которого из A получается новая матрица A^T, строками которой служат столбцы матрицы A, а столбцами — строки матрицы A при сохранении их порядка. Таким образом, первая строка матрицы A становится первым столбцом матрицы A^T, вторая строка матрицы A — вторым столбцом матрицы A^T и т. д., так что (i,j)-й элемент a_{ij} матрицы A становится (j,i)-м элементом матрицы A. Так, при построении регрессионных моделей мы оперировали как с $n \times (p+1)$-матрицей наблюдений объясняющих переменных

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(p)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p)} \end{pmatrix} \qquad \text{(Ж.1)}$$

так и с транспонированной $(p+1) \times n$-матрицей

$$X^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1^{(1)} & x_2^{(1)} & \ldots & x_n^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{(p)} & x_2^{(p)} & \ldots & x_n^{(p)} \end{pmatrix}.$$

Сложение двух матриц. Если A и B — матрицы одной размерности, то можно определить новую матрицу $C = A + B$, которая будет иметь ту же размерность, что и A и B, а ее элементы c_{ij} определяются для всех i и j соотношениями $c_{ij} = a_{ij} + b_{ij}$.

Умножение матрицы на число. Произведение матрицы A на число λ определяется как

$$\lambda A = (\lambda a_{ij}),$$

т. е. каждый элемент матрицы A умножается на это число.

Произведение матриц. Если число столбцов $n \times m$-матрицы A равно числу строк $m \times k$-матрицы B, то может быть определена операция умножения AB матрицы A на матрицу B (при этом говорят, что «матрица B умножается на матрицу A *слева*»). Элементы c_{ij} такого произведения $C = AB$ определяются для всех $i = 1, 2, \ldots, n$ и $j = 1, 2, \ldots, k$ соотношениями

$$c_{ij} = \sum_{l=1}^{m} a_{il} b_{lj}.$$

Таким образом, (i,j)-й элемент произведения C вычисляется как *скалярное произведение* i-й строки матрицы A и j-го столбца матрицы B, т. е. как сумма попарных произведений элементов i-й строки первой матрицы на соответствующие (т. е. стоящие на тех же по порядку местах) элементы j-го столбца второй матрицы. При этом автоматически определяется размерность матрицы-произведения C: она будет иметь столько же строк (n), сколько имел первый сомножитель, и столько же столбцов (k), сколько их имел второй сомножитель.

Заметим, что при перестановке сомножителей произведение BA может просто не существовать, но даже если оно существует (как это и бывает в случае, если матрицы A и B — квадратные и одной размерности или если сомножители имеют размерности $n \times m$ и $m \times n$), то, *вообще говоря, коммутативный закон для умножения матриц не имеет места, т. е.*

$$AB \neq BA.$$

Отметим еще несколько полезных для эконометрических приложений свойств произведения матриц.

Ассоциативный закон: $ABC = (AB)C = A(BC)$.

Дистрибутивный закон: $A(B + C) = AB + AC$, $(B + C)A = BA + CA$.

Умножение на единичную матрицу: для любой $n \times n$-матрицы A имеют место тождества: $AI_n = I_n A = A$.

Транспонирование произведения матриц: $(A_1 A_2 \ldots A_k)^T = A_k^T A_{k-1}^T \ldots A_2^T A_1^T$.

Произведения вида AA^T и $A^T A$ играют заметную роль в эконометрических построениях. Так, произведение $X^T X$, в котором матрица X определена соотношением (Ж.1), является непременным «участником» всех основных формул классического метода наименьших квадратов. Заметим, что если A — матрица размерности $n \times m$, то произведение AA^T будет иметь размерность $n \times n$, в то время как произведение $A^T A$ — это матрица размерности $m \times m$. Но в любом случае *произведения вида* AA^T *и* $A^T A$ *всегда являются квадратными симметрическими матрицами*. Обратим внимание на специальный случай, когда $X = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix}^T$ — это вектор-столбец, состоящий из n элементов (или *вектор-столбец длины* n). Тогда

$$X^T X = \sum_{l=1}^{n} x_l^2 — \qquad \text{(Ж.2)}$$

это 1×1-матрица, т. е. число, а

$$XX^T = \begin{pmatrix} x_1^2 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2^2 & \dots & x_2x_n \\ \dots & \dots & \dots & \dots \\ x_nx_1 & x_nx_2 & \dots & x_n^2 \end{pmatrix} — \qquad (Ж.3)$$

это матрица размерности $n \times n$.

Матрицы типа (Ж.2) и (Ж.3) играют заметную роль в регрессионном анализе и в системах одновременных уравнений. Действительно, если в качестве компонент x_i вектора X рассмотреть регрессионные «невязки» $y_i - \widehat{\Theta}_0 - \widehat{\Theta}_1 x_i^{(1)} - \dots - \widehat{\Theta}_p x_i^{(p)}$ $(i = 1, 2, \dots, n)$, то произведение (Ж.2) даст нам сумму квадратов «невязок», которая играет важную роль в анализе точности регрессионной модели. Если же в качестве компонент x_i вектора X рассмотреть отклонения i-й объясняющей переменной $x^{(i)}$ от своего среднего значения $a^{(i)}$ $(i = 1, 2, \dots, p)$, то произведение (Ж.3) после применения к нему операции усреднения (математического ожидания) E даст $p \times p$ ковариационную матрицу объясняющих переменных Σ_X.

Матричное дифференцирование. Пусть $\Theta = \begin{pmatrix} \Theta_1 & \Theta_2 & \dots & \Theta_m \end{pmatrix}^T$ — $m \times 1$-матрица (т. е. вектор-столбец длины m), компоненты которой играют роль неизвестных параметров эконометрической модели, а $A(\Theta) = \begin{pmatrix} a_1(\Theta) & a_2(\Theta) & \dots & a_n(\Theta) \end{pmatrix}^T$ — $n \times 1$-матрица (т. е. вектор-столбец длины n), компоненты который интерпретируются как некоторые характеристики этой модели, зависящие от Θ.

Производной $n \times 1$ векторной функции $A(\Theta)$ по $m \times 1$-векторному аргументу Θ называется $m \times n$-матрица

$$\frac{\partial A(\Theta)}{\partial \Theta} = \begin{pmatrix} \dfrac{\partial a_1(\Theta)}{\partial \Theta_1} & \dfrac{\partial a_2(\Theta)}{\partial \Theta_1} & \dots & \dfrac{\partial a_n(\Theta)}{\partial \Theta_1} \\ \dfrac{\partial a_1(\Theta)}{\partial \Theta_2} & \dfrac{\partial a_2(\Theta)}{\partial \Theta_2} & \dots & \dfrac{\partial a_n(\Theta)}{\partial \Theta_2} \\ \dots & \dots & \dots & \dots \\ \dfrac{\partial a_1(\Theta)}{\partial \Theta_m} & \dfrac{\partial a_2(\Theta)}{\partial \Theta_m} & \dots & \dfrac{\partial a_n(\Theta)}{\partial \Theta_m} \end{pmatrix}. \qquad (Ж.4)$$

Рассмотрим важные для эконометрических приложений частные случаи функции $A(\Theta)$:

1. $A(\Theta) = X\Theta$, где X — симметрическая матрица размерности $n \times m$.
   Тогда $\dfrac{\partial A(\Theta)}{\partial \Theta} = \dfrac{\partial (X\Theta)}{\partial \Theta} = X$.
2. $A(\Theta) = \Theta^T B\Theta$, где B — симметрическая квадратная матрица размерности $m \times m$.
   Тогда $\dfrac{\partial A(\Theta)}{\partial \Theta} = \dfrac{\partial \left(\Theta^T B\Theta\right)}{\partial \Theta} = 2\Theta^T B$.
3. $A(\Theta) = X^T\Theta$, где X — вектор-столбец длины m.
   Тогда $\dfrac{\partial A(\Theta)}{\partial \Theta} = \dfrac{\partial \left(X^T\Theta\right)}{\partial \Theta} = X^T$.

Любой квадратной матрице A можно сопоставить некоторый набор ее числовых характеристик. Одна из таких характеристик называется *определителем (детерминантом)* матрицы A и обозначается $\det A$ или $|A|$.

Определитель (детерминант) $n \times n$-матрицы A вычисляется по формуле:

$$\det A = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \dots \sum_{j_n=1}^{n} (-1)^{\upsilon(j_1,j_2,\dots,j_n)} a_{1j_1} a_{2j_2} \dots a_{nj_n}, \qquad (Ж.5)$$

где суммирование ведется по всем возможным комбинациям *различных* столбцов (т. е. по всем возможным перестановкам вторых индексов), а $\upsilon(j_1,j_2,\dots,j_n)$ — это минимальное число инверсий (т. е. парных обменов местами), которое надо совершить с элементами *исходной* перестановки $(1,2,\dots,n)$, чтобы получить перестановку $(j_1,j_2,\dots,j_n)$. Очевидно, общее число слагаемых в правой части (Ж.5) составит при таком определении n-факториал ($n!$).

Для малых размерностей матриц это определение приводит, в частности, к следующим результатам:

1) $n = 2$: $\det A = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}$;

2) $n = 3$: $\det A = \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$.

Приведем здесь важнейшие свойства определителя:

1) $\det(AB) = \det(BA) = \det A \cdot \det B$;
2) $\det(\lambda A) = \lambda^n \det A$ (λ — число, n — размерность матрицы A);
3) $\det\left[diag(a_{11}, a_{22}, \dots, a_{nn})\right] = a_{11}a_{22}\dots a_{nn}$;
4) $\det I_n = 1$;
5) $\det A^T = \det A$;
6) $\det A = 0$, если в матрице A есть две одинаковые строки (два одинаковых столбца);
7) инверсия (обмен местами) двух строк (столбцов) матрицы A приводит к изменению знака ее определителя;
8) значение определителя матрицы A не изменится, если к любой его строке (столбцу) добавить линейную комбинацию других строк (столбцов).

Отметим, что если квадратная матрица имеет отличный от нуля определитель, то она называется *невырожденной*.

В операциях с *числами* для любого *отличного от нуля* числа a существует число $a^{-1} = 1/a$, которое мы называем *обратным* и которое обладает тем характеристическим свойством, что $aa^{-1} = a^{-1}a = 1$. В матричной алгебре роль единицы, как мы видели, выполняет *единичная матрица* I_n, поскольку при умножении на I_n любой квадратной матрицы размерности $n \times n$ справа и слева эта матрица не меняется.

Поэтому по аналогии с алгеброй чисел определим: *пусть A — квадратная невырожденная (т. е. $\det A \neq 0$) матрица; тогда матрица A^{-1} называется обратной,*

если $AA^{-1} = A^{-1}A = I$. Можно показать, что такое определение обратной матрицы $A^{-1} = (a_{ij}^{\text{обр}})$ приводит к следующей формуле для вычисления ее элементов:

$$a_{ij}^{\text{обр}} = \frac{(-1)^{i+j} \det A_{ji}}{\det A},$$

где A_{ji} — как и прежде, матрица, получающаяся из матрицы A вычеркиванием из нее j-й строки и i-го столбца (т. е. числитель правой части является алгебраическим дополнением элемента a_{ji} в *исходной* матрице A, или, что то же, — *алгебраическим дополнением* (j,i)-го элемента в *транспонированной* матрице A^T).

Например, пусть дана матрица $A = \begin{pmatrix} 1-a_{11} & -a_{12} \\ -a_{21} & 1-a_{22} \end{pmatrix}$, тогда

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} 1-a_{22} & a_{12} \\ a_{21} & 1-a_{11} \end{pmatrix}.$$

Основные свойства обратной матрицы:

1) матрица A^{-1} для любой невырожденной матрицы A — *единственна*;
2) $\det A^{-1} = (\det A)^{-1}$;
3) $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$;
4) $\left(A^{-1}\right)^{-1} = 1$;
5) $(AB)^{-1} = B^{-1}A^{-1}$, $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ и т. д.

(напомним, что все матрицы, участвовавшие в формулировке свойств обратной матрицы, — квадратные и невырожденные).

Квадратная невырожденная матрица A называется ***ортогональной****, если* $A^T = A^{-1}$.

Из определения ортогональной матрицы непосредственно следует, в частности, что $A^T A = AA^T = I$. Нетрудно также вывести, что определитель ортогональной матрицы всегда равен по абсолютной величине единице, т. е. $|\det A| = 1$. Действительно, поскольку $\det A = \det A^T$, а $\det(AB) = \det A \cdot \det B$, то для ортогональной матрицы $\det\left(AA^T\right) = \det A \det A^T = (\det A)^2 = \det I = 1$. Отсюда получаем, что $|\det A| = 1$, если A ортогональна.

В общем случае, включающем и *прямоугольные* матрицы, существует еще одна очень важная числовая характеристика матрицы — *ее ранг*.

Перед тем, как сформулировать строгое определение этого понятия, рассмотрим понятие *линейной зависимости строк* (столбцов) анализируемой $n \times m$ матрицы A.

Строки $A_i = (a_{i1}, a_{i2}, \ldots, a_{im})$, $i = 1,2,\ldots,n$, *матрицы A называются линейно* ***зависимыми****, если существуют числа* $\lambda_1, \lambda_2, \ldots, \lambda_n$, *не все равные нулю, и такие, что*

$$\lambda_1 A_1 + \lambda_2 A_2 + \ldots + \lambda_n A_n = 0_{1.m}$$

(здесь $0_{1.m}$, в соответствии с принятыми выше обозначениями, — это строка длины m, состоящая из нулей).
Аналогично: столбцы $A_{.j} = (a_{1j}, a_{2j}, \ldots, a_{nj})^T$, $j = 1, 2, \ldots, m$, матрицы A называются линейно зависимыми, если существуют числа $\mu_1, \mu_2, \ldots, \mu_m$, не все равные нулю, и такие, что

$$\mu_1 A_{.1} + \mu_2 A_{.2} + \ldots + \mu_m A_{.m} = 0_{n.1}.$$

*В противном случае строки (столбцы) называются **линейно независимыми**.*

. . . . . .

Максимальное число линейно независимых строк $n \times m$-матрицы A совпадает с максимальным числом ее линейно независимых столбцов и, одновременно, — с максимальным порядком ее не равного нулю минора (напомним, что минором порядка k матрицы A называется определитель $k \times k$-матрицы, получающейся из матрицы A вычеркиванием из нее $n - k$ строк и $m - k$ столбцов).

. . . . . .

***Ранг** $n \times m$-матрицы A (будем обозначать его ранг A) определяется как максимальное число ее линейно независимых столбцов.*

. . . . . .

Ранг матрицы A может быть определен и как максимальное число ее линейно независимых строк, и как максимальный порядок ее отличного от нуля минора. Кстати, последнее определение часто бывает наиболее удобным с точки зрения возможности *практического вычисления* ранга конкретной матрицы. При этом, определяя ранг матрицы как максимальный порядок ее отличного от нуля минора, подразумеваем, что достаточно того, чтобы нашелся *хотя бы один* ненулевой минор порядка k, в то время как все миноры порядка $k + 1$ уже будут равны нулю. Так, например, в 4×3-матрице

$$A = \begin{pmatrix} 4 & 8 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & -2 \\ 3 & 6 & -3 \end{pmatrix}$$

все четыре минора 3-го порядка равны нулю. Следовательно, *ранг* $A < 3$. И хотя большинство миноров 2-го порядка тоже равно нулю, все-таки существуют миноры этого порядка, отличные от нуля. А это значит, что *ранг* $A = 2$.

Заметим, что применительно к матрице X наблюденных значений объясняющих переменных в регрессионном анализе линейная зависимость столбцов означает линейную зависимость объясняющих переменных.

Широкий круг задач многомерного статистического анализа и эконометрики сводится (в вычислительном плане) к необходимости анализа и решения параметрического семейства систем уравнений типа

$$(A - \lambda I_n)X = 0_{n.1}, \quad \det(A - \lambda I_n) = 0, \tag{Ж.6}$$

где A — некоторая $n \times n$-матрица; $X = (x_1, x_2, \ldots, x_n)^T$ — вектор-столбец неизвестных, а λ — некоторый числовой параметр. Для того, чтобы существовало нетривиальное (отличное от нулевого) решение, необходимо, чтобы матрица $A - \lambda I_n$ была вырожденной, т. е. необходимо потребовать, чтобы

$$\det(A - \lambda I_n) = 0. \tag{Ж.7}$$

Из правил вычисления определителя матрицы следует, что левая часть представляет собой алгебраический полином от λ степени n, так что соотношение (Ж.7) — это алгебраическое уравнение степени n относительно A. Само это уравнение, а следовательно, и его корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ по построению полностью определяются элементами матрицы A. Приходим к определению.

Характеристическими (собственными) *числами $n \times n$-матрицы A называются корни характеристического уравнения* (Ж.7).

Беря *любое* из собственных чисел λ_i и подставляя его в исходное соотношение (Ж.6), мы получаем уже *конкретную* систему уравнений (относительно X) вида

$$(A - \lambda_i I_n)X = 0_{n.1}.$$

Существование нетривиального решения $X(i)$ этого уравнения обеспечивается равенством нулю определителя $\det(A - \lambda_i I_n)$. Соответственно, приходим к еще одному определению.

Вектор-столбец $X(i) = \left(x_1^{(i)} \quad x_2^{(i)} \quad \ldots \quad x_n^{(i)}\right)^T$, являющийся решением уравнения (Ж.7), называется ***характеристическим (собственным) вектором*** матрицы A, соответствующим характеристическому (собственному) числу λ_i.

А поскольку алгебраическое уравнение степени n имеет n корней (среди которых, вообще говоря, могут быть совпадающие и комплексные), то всякая квадратная $n \times n$-матрица A имеет n собственных чисел (не обязательно различных) и n соответствующих им собственных векторов.

ГЛОССАРИЙ

Автокорреляция — корреляционная зависимость между наблюдениями временного ряда.

Бесповторная выборка — выборка, в которой отобранный в выборку объект не возвращается в генеральную совокупность.

Вариация — изменение значений признака внутри изучаемой совокупности.

Вероятность $P(A)$ события A — численная мера степени объективной возможности появления события A, отношение числа исходов, благоприятствующих наступлению этого события, к общему числу равновозможных исходов.

Вероятностно-статистическая модель — это вероятностная модель, значения отдельных характеристик (параметров) которой оцениваются по результатам наблюдений, характеризующих функционирование моделируемого конкретного явления.

Временной ряд — это данные, характеризующие один и тот же объект в различные моменты времени (временной срез). Например, еженедельные данные по объему продаж фирмы или ежеквартальные данные по инфляции.

Выборка — часть генеральной совокупности, отобранная для изучения.

Генеральная совокупность — множество всех возможных значений или реализаций исследуемой случайной величины X при данном реальном комплексе условий.

Гетероскедастичность — зависимость дисперсии случайных остатков от номера наблюдения.

Гомоскедастичность — зависимость дисперсии случайных остатков от номера наблюдения.

Дискретная случайная величина — СВ, которая принимает отдельные, изолированные значения с определенными вероятностями.

Дисперсия — характеристика рассеяния, разброса, вариации значений случайной величины относительно среднего значения. Дисперсией случайной величины называется математическое ожидание квадрата ее отклонения от математического ожидания.

Доверительная вероятность — достоверность (надежность) определения неизвестного значения параметра с помощью оценки параметра.

Доверительный интервал (при интервальной оценке неизвестного параметра генеральной совокупности) — числовой интервал, который с заданной доверительной вероятностью накрывает неизвестное значение параметра.

Достоверное событие — событие, которое происходит всегда в условиях данного эксперимента.

Зависимая переменная — в регрессионной модели некоторая переменная Y, являющаяся функцией регрессии с точностью до случайного возмущения.

Закон распределения дискретной случайной величины — соответствие между всеми возможными значениями СВ и их вероятностями.

Корреляция — статистическая взаимосвязь двух или более случайных величин.

Коэффициент авторегрессии — коэффициент корреляции между соседними возмущениями.

Коэффициент детерминации R^2 показывает качество подбора функции и характеризует долю дисперсии результативного признака, объясняемую регрессией, в общей дисперсии результативного признака.

Коэффициент эластичности — показатель, который говорит о том, на сколько процентов изменится значение результирующей переменной при изменении объясняющей переменной на 1%.

Лаг — смещение во времени изменения одного показателя по сравнению с изменением другого.

Лаговые переменные — переменные, взятые в предыдущий момент времени и выступающие в качестве эндогенных и экзогенных переменных.

Линеаризация модели — подбор таких преобразований к анализируемым переменным $y, x^{(1)}, \ldots, x^{(p)}$, которые позволили бы представить искомую зависимость в виде линейного соотношения между преобразованными переменными.

Линейная функция множественной регрессии — $y = a + b_1x_1 + b_2x_2 + \ldots + b_px_p + \varepsilon$.

Математическая модель — это абстракция реального мира, в которой интересующие исследователя отношения между реальными элементами заменены подходящими отношениями между математическими категориями.

Непрерывная случайная величина — СВ, которая может принимать любое значение из некоторого конечного или бесконечного промежутка (т. е. число возможных значений непрерывной СВ бесконечно).

Множественная регрессия широко используется в решении проблем спроса доходности акций, при изучении функций издержек производства, в макроэкономических расчетах и целого ряда других вопросов эконометрики. Основная цель множественной регрессии — построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное воздействие их на моделируемый показатель.

Мультиколлинеарность — высокая взаимная коррелированность объясняющих переменных; может проявляться в функциональной (явной) и стохастической (скрытой) формах.

Невозможное событие — событие, которое не происходит никогда в условиях данного эксперимента.

Несмещенная оценка — такая оценка $\hat{\Theta}$ параметра Θ, что $M\hat{\Theta} = \Theta$.

Несовместимые события — события, которые не могут происходить одновременно.

Оптимальная оценка — оценка, которая удовлетворяет условию $M\left(\hat{\Theta}_{\text{опт}} - \Theta\right)^2 = \min M\left(\hat{\Theta} - \Theta\right)^2$, $\hat{\Theta} \in \tilde{M}$.

Парная регрессия представляет собой модель, где среднее значение зависимой (объясняемой) переменной y рассматривается как функция одной независимой (объясняющей) переменной x, то есть это модель вида: $\hat{y} = f(X)$.

Повторная выборка — выборка, в которой объект перед отбором следующего возвращается в генеральную совокупность.

Предопределенные переменные — переменные, выступающие в системе в роли факторов-аргументов, или объясняющих переменных.

Простая гипотеза — гипотеза, которая содержит одно конкретное предположение $\left(H_1^{(1)}: \theta = \theta_0;\ H_1^{(4)}: \theta = \theta_1\right)$.

Противоположные события — одно из событий происходит тогда и только тогда, когда не происходит другое.

Сложная гипотеза — гипотеза, которая состоит из конечного или бесконечного числа простых гипотез $\left(H_1^{(1)}: \theta \neq \theta_0;\ H_1^{(2)}: \theta > \theta_0;\ H_1^{(3)}: \theta < \theta_0\right)$.

Случайная величина — величина, которая в результате наблюдения принимает то или иное значение, заранее неизвестное и зависящее от случайных обстоятельств.

Случайное событие — событие, которое может произойти или не произойти в условиях данного эксперимента.

Событие — это любой исход или совокупность исходов какого-либо вероятностного эксперимента.

Совместимые события — события, которые могут происходить одновременно.

Составное событие — событие, представимое в виде совокупности (суммы) нескольких элементарных событий.

Состоятельные оценки — оценки $\hat{\Theta}$ и $\hat{\sigma}^2$ являются состоятельными тогда и только тогда, когда наименьшее собственное значение матрицы $X^T X$ стремится к бесконечности при $n \to \infty$.

Среднее квадратическое отклонение — квадратный корень из дисперсии $D(X)$: $\sigma(X) = \sqrt{D}$.

Статистическая гипотеза — гипотеза о виде закона распределения или о параметрах известного распределения. В первом случае гипотеза называется непараметрической, а во втором — параметрической.

Степень свободы — число возможных направлений варьирования признака. Существует равенство между числом степеней свободы общей, факторной и остаточной суммами квадратов: $n - 1 = 1 + (n - 2)$.

Точечная оценка $\hat{\theta}$ *параметра* θ — числовое значение этого параметра, полученное по выборке объема n.

Фиктивные переменные — переменные, полученные путем перевода качественных признаков переменных в количественные, то есть при присвоении цифровых меток.

Функция регрессии y по X — функция, которая описывает изменение условного среднего значения результирующей переменной y (при условии, что значения объясняющих переменных X зафиксированы на уровнях X^*) в зависимости от изменения значений X^* объясняющих переменных.

Экзогенные переменные — переменные задаваемые как бы «извне», автономно, в определенной степени управляемые (планируемые).

Эконометрическая модель — вероятностно-статистическая модель, описывающая механизм функционирования экономической или социально-экономической системы.

Эластичность — показатель, который говорит о том, на сколько процентов изменится значение результирующей переменной при изменении объясняющей переменной на 1%.

Элементарное событие — событие, которое нельзя разбить на более простые.

Эндогенные переменные — переменные, значения которых формируются в процессе и внутри функционирования анализируемой социально-экономической системы в существенной мере под воздействием экзогенных переменных и, конечно, во взаимодействии друг с другом; в эконометрической модели они являются предметом объяснения.

Учебное издание

Грибанова Екатерина Борисовна

ЭКОНОМЕТРИКА

Учебное пособие

Корректор Осипова Е. А.
Компьютерная верстка Перминова М. Ю.

Издано в Томском государственном университете
систем управления и радиоэлектроники.
634050, г. Томск, пр. Ленина, 40
Тел. (3822) 533018.