ВЫСШЕЕ ОБРАЗОВАНИЕ

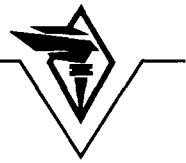
ЭКОНОМЕТРИКА

УЧЕБНОЕ ПОСОБИЕ

А.И. Новиков

www.infra-m.ru

ВЫСШЕЕ ОБРАЗОВАНИЕ
серия основана в 1996 г.



А.И. Новиков

ЭКОНОМЕТРИКА

УЧЕБНОЕ ПОСОБИЕ

Второе издание, исправленное и дополненное

Рекомендовано
Учебно-методическим объединением
по образованию в области экономики
и экономической теории в качестве учебного
пособия для студентов, обучающихся по направлению
521600 «Экономика» и экономическим специальностям

Москва
ИНФРА-М
2007

УДК 330.115(075.8)
ББК 65в6я73
Н73

Рецензенты:
Б.И. Клячин, канд. физ.-мат. наук, доцент, заведующий кафедрой математики Московской высшей школы бизнеса;
М.Ю. Крылков, канд. техн. наук, доцент кафедры математики Московского горного института

Новиков А.И.
Н73 Эконометрика: Учеб. пособие. — 2-е изд., испр. и доп. — М.: ИНФРА-М, 2007. — 144 с. – (Высшее образование).

ISBN 978-5-16-002974-0

Содержит систематическое изложение основ эконометрики, подготовлено в соответствии с требованиями государственного стандарта. Рассмотрены линейная модель парной и множественной регрессии, проверка гипотез, гетероскедастичность и автокорреляция ошибок. Отдельные главы посвящены динамическим моделям и системам одновременных уравнений.

Для студентов экономических вузов, ориентированных на прикладные задачи моделирования и прогнозирования в экономике.

ББК 65в6я73

ISBN 978-5-16-002974-0

Оригинал-макет изготовлен в Издательском Доме «ИНФРА-М»

ЛР № 070824 от 21.01.93 г.

Сдано в набор 10.08.2006. Подписано в печать 29.01.2007.
Формат 60×88/16. Бумага офсетная. Гарнитура Newton.
Усл. печ. л. 8,82. Уч.-изд. л. 8,97.
Тираж 3000 экз. Заказ № 2337

Издательский Дом «ИНФРА-М»
127282, Москва, ул. Полярная, д. 31в
Тел.: (495) 380-05-40, 380-05-43. Факс: (495) 363-92-12.
E-mail: books@infra-m.ru
http://www.infra-m.ru

Отдел «Книга — почтой»:
(495) 363-42-60 (доб. 246, 247)

Отпечатано в ОАО «Домодедовская типография»,
г. Домодедово, Каширское ш., д. 4, корп. 1.

Предисловие

В современных программах подготовки экономистов курс эконометрики наряду с микро- и макроэкономикой занял одно из ключевых мест.

Экономисты используют количественные данные для наблюдения за развитием экономики, для ее анализа и прогнозов. Набор статистических методов, используемых для этих целей, и составляет в совокупности эконометрику.

При изложении курса эконометрики используется минимальный математический аппарат, основанный на понятиях и свойствах ковариации и дисперсии. В начале курса приведены необходимые элементы математической статистики.

Все излагаемые методы и подходы в эконометрике иллюстрируются примерами и упражнениями с использованием пакета анализа данных Excel.

Эта книга предназначена студентам, впервые приступающим к изучению эконометрики.

Введение

Закономерности в экономике выражаются в виде зависимостей экономических показателей и математических моделей их поведения. Такие зависимости и модели могут быть получены только путем обработки реальных статистических данных, с учетом внутренних связей и случайных факторов.

Эконометрика — наука, изучающая количественные закономерности и взаимозависимости в экономике методами математической статистики.

Цель эконометрики — эмпирический вывод экономических законов.

Задачи эконометрики — построение экономических моделей и оценивание их параметров, проверка гипотез о свойствах экономических показателей и формах их связи.

Эконометрический анализ служит основой для экономического анализа и прогнозирования, создавая возможность для принятия обоснованных экономических решений.

ТИПЫ ДАННЫХ

При моделировании экономических процессов оперируют двумя типами данных: пространственными и временными.

Пространственные данные — это данные по какому-либо экономическому показателю, полученные от разных однотипных объектов (фирм, регионов и т.п.), но относящиеся к одному и тому же моменту времени (пространственный срез). Например, данные об объеме производства, количестве работников, доходе разных фирм в один и тот же момент времени.

Временные ряды — это данные, характеризующие один и тот же объект в различные моменты времени (временной срез). Например, ежеквартальные данные об инфляции, средней заработной плате, данные о национальном доходе за последние годы.

Отличительная черта временных данных — упорядоченность во времени. Кроме того, наблюдения в близкие моменты времени часто бывают зависимы.

Любые экономические данные представляют собой характеристики какого-либо экономического объекта. Они формируются под воздействием множества факторов, не все из которых доступны внешнему контролю. Неконтролируемые (неучтенные) факторы обусловливают случайность данных, которые они определяют.

Поскольку экономические данные имеют статистическую природу, для их анализа и обработки необходимо применять специальные методы.

КЛАССЫ МОДЕЛЕЙ

Можно выделить три основных класса моделей: модели временных рядов, регрессионные модели с одним уравнением и системы одновременных уравнений.

К **моделям временных рядов** относятся *модели тренда* и *модели сезонности*. Тренд представляет собой устойчивое изменение уровня показателя в течение длительного времени. Сезонность характеризует устойчивые внутригодовые колебания уровня показателя.

Кроме того, к этому классу относится множество более сложных моделей, таких, например, как модель адаптивного прогноза, модель авторегрессии.

Их общей чертой является то, что они объясняют поведение временного ряда исходя только из его предыдущих значений.

В **регрессионных моделях с одним уравнением** объясняемая переменная представляется в виде функции от объясняющих переменных. Примером служит модель спроса на некоторый товар в зависимости от его цены и дохода.

По виду функции регрессионные модели делятся на *линейные* и *нелинейные*. Существуют эффективные методы оценки и анализа линейных регрессионных моделей. Анализ линейных регрессионных моделей является базовым в прикладной эконометрике.

Область применения регрессионных моделей, даже линейных, значительно шире, чем моделей временных рядов.

Системы одновременных уравнений описываются системами уравнений, состоящими из тождеств и регрессионных уравнений, в каждом из которых аргументы содержат не только объясняющие переменные, но и объясняемые переменные из других уравнений системы. Примером служит модель формирования доходов.

Все три класса моделей могут использоваться при моделировании экономических процессов.

Обычно предполагают, что все факторы, не учтенные явно в экономической модели, оказывают на объект некое результирующее воздействие, величина которого задается случайной компонентой.

Введение случайной компоненты в экономическую модель делает ее доступной для эмпирической проверки на основе статистических данных.

ОСНОВНЫЕ ЭТАПЫ ЭКОНОМЕТРИЧЕСКОГО МОДЕЛИРОВАНИЯ

Укажем основные этапы эконометрического исследования.

1. **Постановочный.** Формулируется цель исследования, определяется набор участвующих в модели экономических переменных. Целью эконометрического моделирования могут быть анализ изучаемого экономического процесса (объекта), прогноз его экономических показателей, анализ возможного развития явления при различных значениях экзогенных (независимых) переменных, выработка управленческих решений. При выборе экономических переменных необходимо теоретическое обоснование каждой переменной. Объясняющие переменные не должны быть связаны функциональной или тесной корреляционной зависимостью, так как это может привести к невозможности оценки параметров модели (явление мультиколлинеарности). Для отбора переменных можно использовать процедуру пошагового отбора переменных, а для оценки влияния качественных признаков – фиктивные переменные.

2. **Априорный.** Проводится анализ сущности изучаемого объекта, формирование и формализация априорной (известной до начала моделирования) информации.

3. **Информационный.** Осуществляется сбор необходимой статистической информации, значений экономических переменных. Здесь используются данные наблюдения, полученные в условиях активного (с участием исследователя) и пассивного (без участия эконометриста) эксперимента.

4. **Спецификация модели.** В математической форме выражаются обнаруженные связи и соотношения, устанавливается состав экзогенных и эндогенных переменных; формируются исходные предпосылки и ограничения модели. От того, насколько точно выполнена задача спецификации, зависит успех эконометрического моделирования.

5. **Параметризация.** Оцениваются параметры (коэффициенты) выбранной зависимости. Эта оценка осуществляется на основе имеющихся статистических данных.

6. **Идентификация.** Осуществляются статистический анализ модели и оценка ее параметров.

7. **Верификация.** Проводится проверка адекватности модели, выясняется, насколько удачно решены проблемы спецификации, идентификации, какова точность расчетов по данной модели, насколько соответствует построенная модель реальному экономическому явлению.

ТИПЫ ЗАВИСИМОСТЕЙ

В экономических исследованиях одной из основных задач является анализ зависимостей между переменными. Зависимость может быть строгой (функциональной) либо статистической.

Функциональная зависимость задается в виде точной формулы, в которой каждому значению одной переменной соответствует строго определенное значение другой, воздействием случайных факторов при этом пренебрегают.

В экономике функциональная зависимость между переменными проявляется редко.

Статистической зависимостью называется связь переменных, на которую накладывается воздействие случайных факторов. При этом изменение одной переменной приводит к изменению математического ожидания другой переменной.

Уравнение регрессии — это формула статистической связи между переменными. Если эта формула линейна, то имеем линейную регрессию.

Формула статистической связи *двух* переменных называется **парной регрессией**, зависимость от *нескольких* переменных — **множественной регрессией.**

Глава 1
Элементы математической статистики

1.1. ОПЕРАЦИЯ СУММИРОВАНИЯ

Пусть величина X задается последовательностью данных x_1, x_2, ..., x_n, каждое из которых можно записать как x_i, $i = \overline{1, n}$.

Сумма этих чисел обозначается следующим образом:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n, \text{ причем } \sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j.$$

Если из контекста ясно, каковы начальный и конечный суммируемые члены, то часто используют сокращенные обозначения:

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

Сумма квадратов этих чисел обозначается следующим образом:

$$\sum x_i^2 = x_1^2 + x_2^2 + ... + x_n^2.$$

Обозначим средние значения величин X, X^2 и XY соответственно как:

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \overline{x^2} = \frac{1}{n}\sum x_i^2, \quad \overline{xy} = \frac{1}{n}\sum x_i y_i.$$

Имеет место неравенство

$$(\bar{x}) \le \overline{x^2}.$$

Правила суммирования (a, b — константы):

1. $\sum a = na.$
2. $\sum bx_i = b\sum x_i = bn\bar{x}.$
3. $\sum (a + bx_i) = na + bn\bar{x}.$
4. $\sum (x_i + y_i) = \sum x_i + \sum y_i = n(\bar{x} + \bar{y}).$
5. $\sum (x_i - \bar{x}) = 0.$

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

7. $\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}.$

1.2. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Случайной величиной (переменной) называется величина, которая под воздействием случайных факторов может с определенными вероятностями принимать те или иные значения из некоторого множества чисел.

Случайные величины обозначаются большими буквами, а их возможные значения — малыми.

Для полной характеристики случайной величины должны быть указаны не только все ее значения, но и их вероятности.

Универсальным способом задания случайной величины X является задание ее функции распределения.

Функцией распределения $F(x)$ случайной величины X называется вероятность того, что величина X принимает значение меньшее x, т.е.

$$F(x) = P(X < x), \qquad x \in \mathbf{R}.$$

Свойства функции распределения:

1. $0 \le F(x) \le 1$ при любых $x \in \mathbf{R}$.

2. $P(x_1 \le X \le x_2) = F(x_2) - F(x_1)$.

3. $F(x)$ — неубывающая функция.

4. $\lim\limits_{x\to-\infty} F(x) = 0, \quad \lim\limits_{x\to\infty} F(x) = 1.$

Различают *дискретные* и *непрерывные* случайные величины.

1. **Дискретной** называется случайная величина, которая принимает отдельные, изолированные друг от друга значения. Число возможных значений дискретной случайной величины *конечно* или *счетно*.

Дискретную случайную величину удобнее задавать не в виде функции распределения, а в виде ряда распределения.

При табличном задании **ряда распределения** первая строка таблицы содержит возможные значения случайной величины, а вторая — соответствующие им вероятности, т.е.

$\begin{pmatrix} x_1 & x_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$, где $p_i = P(X = x_i)$, $\sum p_i = 1$.

Графическое изображение ряда распределения называется **полигоном распределения**.

2. **Непрерывной** называется случайная величина, множество значений которой непрерывно заполняет некоторый числовой промежуток. Число возможных значений непрерывной случайной величины *бесконечно*.

Задать непрерывную случайную величину рядом распределения невозможно, поэтому ее задают функцией распределения $F(x)$.

Вместо функции распределения $F(x)$ для непрерывной случайной величины обычно используется плотность распределения вероятностей $f(x)$.

Плотностью распределения $f(x)$ непрерывной случайной величины называется производная от функции распределения, т.е. $f(x) = F'(x)$.

Из определения производной вытекает вероятностный смысл плотности распределения:

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x},$$

т.е. предел отношения вероятности попадания случайной величины X в интервал $(x, x + \Delta x)$ к длине этого интервала при $\Delta x \to 0$ равен значению плотности распределения вероятностей $f(x)$.

Из определения плотности распределения следует, что функция распределения $F(x)$ является первообразной для плотности распределения $f(x)$.

Свойства плотности распределения:

1. $f(x) \geq 0$ при любых $x \in \mathbf{R}$.

2. $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$.

3. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

В основе математической статистики лежат понятия *генеральной* и *выборочной совокупностей*.

Генеральная совокупность — это множество всех значений (исходов) случайной величины, которые она может принять в про-

цессе наблюдения. Например, данные о доходах всех жителей страны.

Выборочная совокупность (выборка) — это множество наблюдений, составляющих лишь часть генеральной совокупности.

Для любой случайной величины важную роль помимо функции распределения играют числовые характеристики ее распределения.

1.3. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЯ

I. Генеральная совокупность

Математическим ожиданием дискретной случайной величины называется сумма произведений всех ее значений на соответствующие им вероятности, т.е.

$$M(X) = \sum x_i p_i,$$

где суммирование осуществляется по всем возможным значениям случайной величины.

Математическое ожидание непрерывной случайной величины определяется выражением

$$M(X) = \int x f(x) \mathrm{d}x,$$

где интегрирование осуществляется на всем интервале, в котором определена $f(x)$.

Математическое ожидание случайной величины — это *среднее* ее значение по генеральной совокупности, обозначается $M(X) = \mu_x$.

Геометрически математическое ожидание случайной величины — это *центр* ее распределения.

Свойства математического ожидания (a, b — константы; X, Y — случайные величины):

1. $M(a) = a$.
2. $M(bX) = bM(X)$.
3. $M(a + bX) = a + bM(X)$.
4. $M(X + Y) = M(X) + M(Y)$.
5. $M(X - \mu_x) = 0$.

Математическое ожидание функции $g(X)$ определяется выражением

$$M[g(X)] = \sum g(x_i)p_i,$$

где суммирование осуществляется по всем возможным значениям x_i. В частности, если $g(X) = X^2$, то $M(X^2) = \sum x_i^2 p_i$.

Случайные величины X, Y называются **независимыми**, если $P(X=x;\ Y=y) = P(X=x)P(Y=y)$ для любых значений x, y.

Следствие. Если случайные величины X, Y независимы, то

$$M(XY) = M(X)M(Y),$$

$$M[(X-\mu_x)(Y-\mu_y)] = 0.$$

Теоретическая (генеральная) дисперсия случайной величины определяется как математическое ожидание квадрата отклонения случайной величины X относительно ее средней, т.е.

$$\sigma_x^2 = D(X) = M(X-\mu_x)^2.$$

Замечание. *Если ясно, о какой переменной идет речь, нижний индекс в* μ_x *или* σ_x^2 *можно не указывать.*

Для вычисления дисперсии часто используется другое выражение, получаемое из определения дисперсии:

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

Дисперсия является *мерой рассеяния* случайной величины относительно средней (центра). Размерность дисперсии не совпадает с размерностью случайной величины.

Стандартным отклонением случайной величины X называется корень квадратный из ее дисперсии, т.е.

$$\sigma_x = \sqrt{D(X)}.$$

Стандартное отклонение показывает, насколько в среднем отклоняется случайная величина в совокупности относительно средней (центра).

Свойства дисперсии:

1. $D(a) = 0$.
2. $D(bX) = b^2 D(X)$.
3. $D(a + bX) = b^2 D(X)$.

Следствие. Если случайные величины X, Y независимы, то

$D(X+Y) = D(X) + D(Y)$.

Заметим, что $M(X)$ и $D(X)$ — это числовые характеристики генеральной совокупности (числа), а не функции.

Нормальное распределение случайной величины X характеризуется лишь двумя параметрами: средним значением μ и дисперсией σ^2. Это обозначается как $X \sim N(\mu; \sigma^2)$.

График плотности нормального распределения $f(x)$ имеет колоколообразный симметричный вид (рис. 1). Максимум этой функции находится в точке $x = \mu$, а разброс относительно этой точки определяется параметром σ. Чем меньше значение σ, тем более острый и высокий максимум $f(x)$.

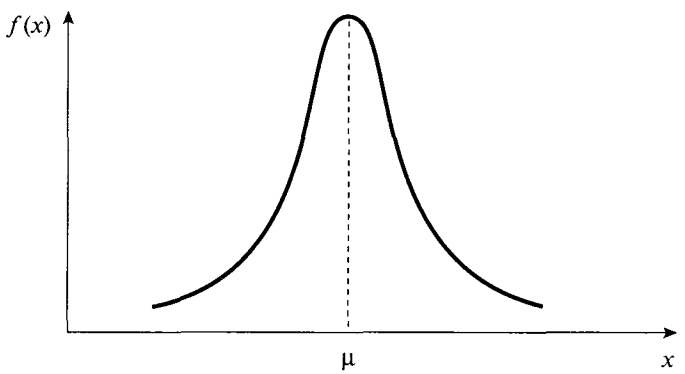


Рис. 1

II. Выборочная совокупность

Пусть из генеральной совокупности с распределением $F(x)$ извлекается выборка объема n. Считаем, что выборочные наблюдения $X_1, X_2, \ldots, X_n$ независимы и имеют одинаковые распределения.

Выборочной средней называется среднее арифметическое наблюдаемых значений случайной величины в выборке, т.е.

$$\bar{x} = \frac{1}{n}\sum x_i.$$

Выборочной дисперсией (вариацией) называется среднее арифметическое квадратов отклонения наблюдаемых значений случайной величины от среднего значения, т.е.

$$\mathrm{var}(X) = \frac{1}{n}\sum (x_i - \bar{x})^2, \quad \text{или} \quad \mathrm{var}(X) = \overline{x^2} - (\bar{x})^2.$$

Свойства выборочной дисперсии:

1. $\mathrm{var}(a) = 0$.

2. $\mathrm{var}(bX) = b^2\mathrm{var}(X)$.

3. $\mathrm{var}(a + bX) = b^2\mathrm{var}(X)$.

Значения $\bar{x}$, $\mathrm{var}(X)$ являются числовыми характеристиками выборочной совокупности.

Для разных выборок, взятых из одной и той же генеральной совокупности, выборочные средние и выборочные дисперсии будут различны, т.е. выборочные характеристики являются случайными величинами.

Из условия, что выборочные наблюдения $X_1, X_2, \ldots, X_n$ независимы и имеют одинаковые распределения, вытекают следующие соотношения:

$$M(\bar{x}) = \mu_x, \quad D(\bar{x}) = \frac{\sigma_x^2}{n}, \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}.$$

Центральная предельная теорема закона больших чисел устанавливает, что распределение средней выборочной $\bar{x}$ при достаточно большом n является нормальным, т.е.

$$\bar{x} \sim N\left(\mu_x; \frac{\sigma_x^2}{n}\right).$$

Пример 1.1. Вычислить выборочные характеристики по исходным данным

№ п/п	1	2	3	4	5
x	2	6	10	14	18

Исходные данные (x) и расчетные показатели (x^2) представим в виде расчетной таблицы:

№ п/п	x	x^2
1	2	4
2	6	36
3	10	100

Окончание таблицы

№ п/п	x	x^2
4	14	196
5	18	324
Итого	50	660
Среднее	10	132
	$\bar{x}$	$\overline{x^2}$

Выборочные характеристики:

$$\bar{x} = 24, \quad \mathrm{var}(x) = \overline{x^2} - (\bar{x})^2 = 132 - 100 = 32.$$

Для вычисления выборочной средней и выборочной дисперсии в Excel можно использовать функции

$\bar{x}$ = СРЗНАЧ(массив x), var(x) = ДИСПР(массив x).

1.4. ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

Характеристики генеральной совокупности обычно неизвестны. Задача заключается в их оценке по характеристикам выборочной совокупности.

Характеристики генеральной совокупности принято называть **параметрами**, а выборочной совокупности — **оценками**.

Пусть искомый параметр генеральной совокупности есть ϑ_0, а на основе выборки объема n определяется оценка ϑ.

Различают *точечные* и *интервальные* оценки параметров генеральной совокупности.

Точечной оценкой ϑ параметра ϑ_0 называется числовое значение этого параметра, полученное по выборке, т.е. $\vartheta_0 \approx \vartheta$.

Для того чтобы выборочная оценка давала хорошее приближение оцениваемого параметра, она должна удовлетворять определенным требованиям (*несмещенности*, *эффективности* и *состоятельности*).

1. Несмещенность оценок. Оценка ϑ называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру ϑ_0 при любом объеме выборки, т.е.

$$M(\vartheta) = \vartheta_0.$$

Если это не так, то оценка называется **смещенной**, а разность $M(\vartheta) - \vartheta_0$ — **смещением**.

Таким образом, требование несмещенности гарантирует отсутствие систематических ошибок при оценивании.

Выборочная средняя $\bar{x}$ является *несмещенной оценкой* генеральной средней μ, так как $M(\bar{x}) = \mu$. Тем не менее оценка $\bar{x}$ не единственная возможная несмещенная оценка μ.

Выборочная дисперсия $\mathrm{var}(X)$ является *смещенной оценкой* генеральной дисперсии σ_x^2, при этом

$$M[\mathrm{var}(X)] = \frac{n-1}{n}\sigma_x^2.$$

В качестве *несмещенной оценки* генеральной дисперсии используется величина (**исправленная**, или **остаточная, дисперсия**)

$$S_x^2 = \frac{n}{n-1}\mathrm{var}(X) = \frac{\sum(x_i - \bar{x})^2}{n-1},$$

для которой $M(S_x^2) = \sigma_x^2$.

Отметим, что в знаменателе остаточной дисперсии стоит число степеней свободы $(n-1)$, а не n, так как одна степень свободы теряется при определении выборочной средней.

З а м е ч а н и е. *Для получения несмещенной оценки дисперсии случайной величины соответствующую сумму квадратов отклонений от средней делят на число степеней свободы независимого варьирования случайной величины. Число степеней свободы равно разности между числом единиц совокупности n и числом определяемых по ней констант.*

Стандартным отклонением S_x случайной величины в выборке называется корень квадратный из ее исправленной дисперсии:

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

По данным примера 1.1 определим S_x^2 и S_x:

$$S_x^2 = \frac{n}{n-1}\mathrm{var}(X) = \frac{5}{5-1}\cdot 32 = 40, \quad S_x = \sqrt{40} = 6{,}324.$$

Для вычисления исправленной дисперсии и стандартного отклонения в Excel можно использовать функции

S_x^2 = `ДИСП(массив x)`, S_x = `СТАНДОТКЛОН(массив x)`.

2. Эффективность оценок. Несмещенная оценка ϑ называется **эффективной**, если она имеет минимальную дисперсию по сравнению с другими выборочными оценками, т.е. $\min D(\vartheta)$.

Предположим, что имеются две оценки параметра ϑ_0, рассчитанные на основе одной и той же информации (рис. 2). Оценка A является более эффективной, чем оценка B.

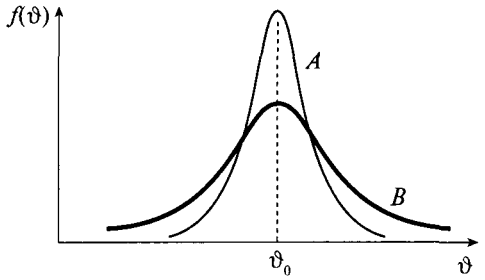


Рис. 2

Выборочная средняя $\bar{x}$ является *эффективной оценкой* генеральной средней μ, т.е. имеет наименьшую дисперсию в классе несмещенных оценок.

Покажем это для выборки из двух наблюдений X_1, X_2. Обобщенная оценка

$$Z = \lambda_1 X_1 + \lambda_2 X_2, \quad \text{где} \quad \lambda_1 + \lambda_2 = 1,$$

является *несмещенной*, так как

$$M(Z) = M(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 M(X_1) + \lambda_2 M(X_2) = (\lambda_1 + \lambda_2)\mu = \mu.$$

Теоретическая дисперсия обобщенной оценки

$$D(Z) = D(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1^2 D(X_1) + \lambda_2^2 D(X_2) = (\lambda_1^2 + \lambda_2^2)\sigma_x^2.$$

Минимизация дисперсии $D(Z)$ эквивалентна минимизации выражения

$$\lambda_1^2 + \lambda_2^2 = \lambda_1^2 + (1 - \lambda_1)^2 = 2\lambda_1^2 - 2\lambda_1 + 1 \to \min.$$

Минимум этого выражения достигается при $\lambda_1 = \lambda_2 = {}^1/_2$, следовательно, выборочная средняя $(X_1 + X_2)/2$ имеет наименьшую дисперсию.

Этот вывод можно обобщить для выборок любого размера, если наблюдения независимы друг от друга.

3. Состоятельность оценок. Оценка ϑ называется **состоятельной**, если при $n \to \infty$ она стремится по вероятности к оцениваемому параметру ϑ_0, т.е.

$$\lim_{n \to \infty} P(|\vartheta - \vartheta_0| < \varepsilon) = 1.$$

Иначе говоря, состоятельной называется такая оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений.

На рис. 3 показано, как при различном объеме выборки может выглядеть распределение вероятностей (состоятельная оценка, смещенная на малой выборке).

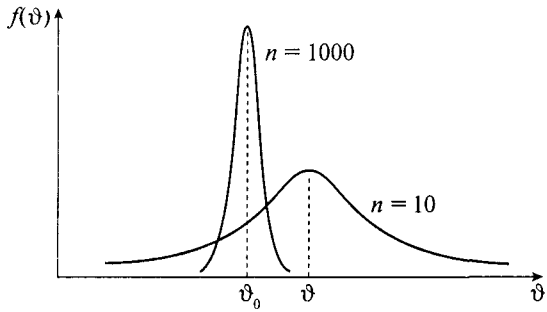


Рис. 3

Теорема Чебышёва закона больших чисел утверждает, что

$$\lim_{n\to\infty} P(|\bar{x} - \mu| < \varepsilon) = 1,$$

т.е. выборочная средняя $\bar{x}$ является *состоятельной оценкой* генеральной средней μ.

Пусть выборочная характеристика ϑ служит оценкой неизвестного параметра ϑ_0. Наряду с точечными оценками параметров (в виде одного числа) рассматривают интервальные оценки (в виде двух чисел — концов интервала).

Интервальной называют оценку, определяющую числовой интервал $(\vartheta - \varepsilon; \vartheta + \varepsilon)$, $\varepsilon > 0$, содержащий оцениваемый параметр ϑ_0, т.е.

$$\vartheta - \varepsilon < \vartheta_0 < \vartheta + \varepsilon, \quad \text{или} \quad |\vartheta - \vartheta_0| < \varepsilon.$$

Доверительным интервалом называется интервал $|\vartheta - \vartheta_0| < \varepsilon$, в котором с заданной вероятностью γ заключен неизвестный параметр ϑ_0, а сама вероятность называется *доверительной вероятностью*, т.е.

$$P(|\vartheta - \vartheta_0| < \varepsilon) = \gamma.$$

Уровнем значимости α называется вероятность $P(|\vartheta - \vartheta_0| > \varepsilon) = \alpha$, причем $\alpha = 1 - \gamma$.

1.5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистической гипотезой H называется предположение относительно параметров или вида распределения случайной величины.

Нулевой гипотезой H_0 называют выдвигаемую гипотезу. Обычно считают, что H_0 — гипотеза об отсутствии различий.

Конкурирующей гипотезой H_1 называют гипотезу, которая противоречит нулевой. Таким образом, H_1 — гипотеза о значимости различий.

Проверку статистической гипотезы выполняют на основе результатов выборки. Поскольку выборка имеет ограниченный объем, то появляется возможность принятия ошибочного решения.

Статистическим критерием называется случайная величина, которая служит для проверки нулевой гипотезы. В качестве статистического критерия выбирается такая случайная величина, например t, точное или приближенное распределение которой известно.

Наблюдаемым значением t называется значение критерия, вычисленное по данным выборки.

Уровнем значимости α называется вероятность того, что будет отвергнута правильная нулевая гипотеза, т.е. $P_{H_0}(\overline{H}_0) = \alpha$.

Уровень значимости α устанавливается заранее. Выбор, например, 5%-го уровня значимости означает, что в пяти случаях применения критерия из ста верная гипотеза будет отвергнута. Стремление к уменьшению α ведет к одновременному уменьшению вероятности отвергнуть гипотезу, когда она является ложной.

В экономических исследованиях проверку гипотез осуществляют при 5%- и 1%-ном уровнях значимости, которые называются **стандартными уровнями**.

Между переменными t и α существует взаимно однозначное соответствие.

Проверку статистических гипотез можно выполнить двумя способами.

Способ 1. Стандартным уровням значимости α соответствуют определенные значения $t_{кр} = t(\alpha)$, называемые **критическими точками**.

Практически значения критических точек $t_{кр.1}$ для $\alpha = 0{,}05$ и $t_{кр.2}$ для $\alpha = 0{,}01$ определяются по таблицам известного распределения выбранного критерия.

Для наглядности процесса принятия решения на координатной оси t указывают эти критические точки (рис. 4).

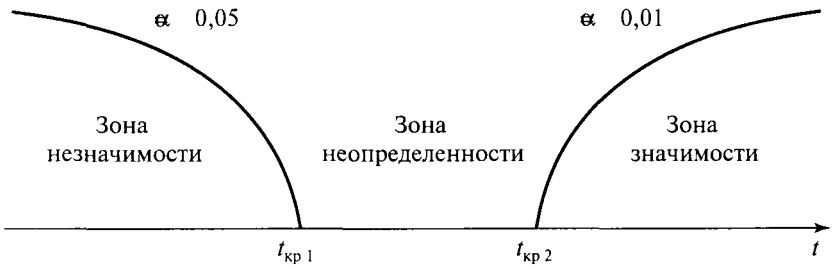


Рис. 4

Критические точки разбивают множество значений критерия t на три непересекающиеся области.

Область левее критической точки $t_{кр\ 1}$ называется **зоной незначимости**. Если $t < t_{кр\ 1}$, то H_0 *принимается* на уровне значимости $\alpha = 0{,}05$, и тем более на уровне $\alpha = 0{,}01$.

Область правее критической точки $t_{кр\ 2}$ называется **зоной значимости**. Если $t > t_{кр\ 2}$, то H_0 *отвергается* на уровне значимости $\alpha = 0{,}01$, и тем более на уровне $\alpha = 0{,}05$.

Область между двумя критическими точками называется **зоной неопределенности**. Если $t_{кр\ 1} < t < t_{кр\ 2}$, то H_0 *отвергается* на уровне $\alpha = 0{,}05$, но *принимается* на уровне $\alpha = 0{,}01$.

Таким образом, если *наблюдаемое* значение критерия *t больше критического* значения $t_{кр}$, то гипотеза H_0 отвергается на заданном уровне значимости и исследуемый показатель является статистически *значимым*.

Способ 2. Наблюдаемому значению критерия t соответствует определенный уровень значимости $\alpha(t)$, который в дальнейшем будем обозначать как *значимость* $t = \alpha$ (наблюдаемое значение t). Практически *значимость* t можно определить с помощью функции Excel.

Для наглядности процесса принятия решения на координатной оси α указывают его стандартные значения 0,01 и 0,05 (рис. 5).

Стандартные значения 0,01 и 0,05 разбивают множество значений α на три непересекающиеся области.

Область левее стандартной точки 0,01 является **зоной значимости**. Если *значимость* $t < 0{,}01$, то H_0 *отвергается* на уровне 0,01, и тем более на уровне 0,05.

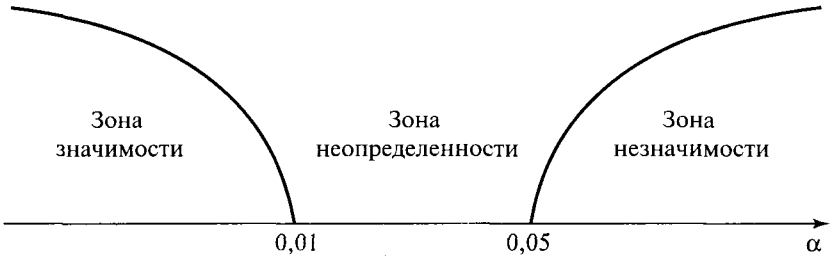


Рис. 5

Область правее стандартной точки 0,05 является **зоной незначимости**. Если *значимость* $t > 0{,}05$, то H_0 *принимается* на уровне 0,05, и тем более на уровне 0,01.

Область между двумя стандартными точками является **зоной неопределенности.** Если $0{,}01 <$ *значимость* $t < 0{,}05$, то H_0 *принимается* на уровне 0,01, но *отвергается* на уровне 0,05.

Таким образом, если *значимость t меньше* заданного *стандартного уровня*, то гипотеза H_0 отвергается и исследуемый показатель является статистически *значимым*.

Такая проверка осуществляется в современных статистических пакетах на компьютере, в которых значимость критерия подсчитывается непосредственно в процессе работы.

Замечание. *Если в качестве критерия проверки нулевой гипотезы используется случайная величина, подчиненная распределению Стьюдента, то ее обозначают через t* (***t*-статистика**), *а если используется случайная величина, подчиненная распределению Фишера, — через F* (***F*-статистика**).

***t*-статистика** часто используется для проверки *гипотезы о значимости выборочной оценки исследуемого параметра* и для нахождения интервальных оценок параметра. В качестве критерия *t* принимают отношение выборочной оценки параметра к ее стандартной ошибке: $t = \dfrac{|\theta|}{S(\theta)}$.

***F*-статистика** используется для проверки *гипотезы о равенстве дисперсий*. В качестве критерия *F* принимают отношение исправленных выборочных дисперсий: $F = \dfrac{S_1^2}{S_2^2}$.

В дальнейшем для проверки статистических гипотез будем использовать в основном второй способ.

1.6. КОВАРИАЦИЯ И КОРРЕЛЯЦИЯ

Различают *выборочную* и *теоретическую* ковариацию.

Выборочной ковариацией двух переменных x, y называется средняя величина произведения отклонений этих переменных от своих средних, т.е.

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}), \quad \text{или} \quad \operatorname{cov}(x, y) = \overline{xy} - \bar{x}\,\bar{y},$$

где $\bar{x}$, $\bar{y}$ — выборочные средние переменных x, y.

Выборочная ковариация является *мерой взаимосвязи* между двумя переменными.

Пусть данные наблюдений переменных x, y представлены в виде точечного графика — *диаграммы рассеяния наблюдений* (рис. 6).

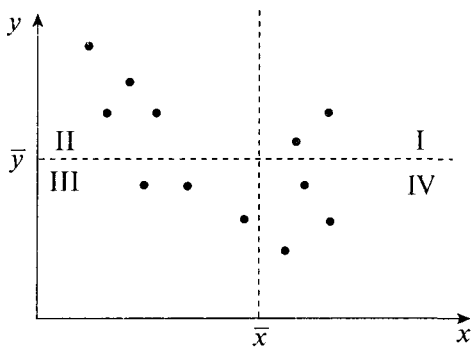


Рис. 6

Точка $(\bar{x}, \bar{y})$ на диаграмме является центром рассеяния переменных x, y.

Вертикальная и горизонтальная прямые, проведенные через точку $(\bar{x}, \bar{y})$, разделяют диаграмму рассеяния на четыре области.

Наблюдения в областях I, III, в которых $(x_i - \bar{x})(y_i - \bar{y}) > 0$, дают положительный вклад в ковариацию, а в областях II, IV, в которых $(x_i - \bar{x})(y_i - \bar{y}) < 0$, — отрицательный.

Если положительные вклады преобладают над отрицательными, то ковариация будет *положительной*, в противном случае она будет *отрицательной*. Положительной ковариации отвечает положительная связь, а отрицательной — отрицательная.

При положительной (прямой) связи с увеличением одной переменной другая переменная в среднем также увеличивается, и наоборот при отрицательной (обратной) связи.

Заметим, что $\operatorname{cov}(x, x) = \frac{1}{n}\sum (x_i - \bar{x})^2 = \operatorname{var}(X)$.

Правила расчета ковариации:

1. $\operatorname{cov}(x, u + v) = \operatorname{cov}(x, u) + \operatorname{cov}(x, v)$.
2. $\operatorname{cov}(x, a) = 0$, если $a = \text{const}$.
3. $\operatorname{cov}(x, bu) = b\operatorname{cov}(x, u)$, если $b = \text{const}$.
4. $\operatorname{cov}(u, v) = \operatorname{cov}(v, u)$.
5. $\operatorname{var}(u + v) = \operatorname{var}(u) + \operatorname{var}(v) + 2\operatorname{cov}(u, v)$.

Доказательство вытекает из определения ковариации. Например:

$$\operatorname{cov}(x, a) = \frac{1}{n}\sum (x_i - \bar{x})(a - a) = 0;$$

$$\operatorname{var}(u + v) = \operatorname{cov}(u + v, u + v) = \operatorname{cov}(u, u) + \operatorname{cov}(v, v) + 2\operatorname{cov}(u, v) =$$
$$= \operatorname{var}(u) + \operatorname{var}(v) + 2\operatorname{cov}(u, v).$$

Теоретической ковариацией случайных величин X, Y называется математическое ожидание произведения отклонений этих величин от своих средних значений, т.е.

$$\operatorname{Cov}(X, Y) = M[(X - \mu_x)(Y - \mu_y)],$$

где $\mu_x = M(X)$, $\mu_y = M(Y)$.

Запись $\operatorname{Cov}(X, Y)$ указывает на то, что ковариация рассматривается по генеральной совокупности.

Заметим, что $\operatorname{Cov}(X, X) = M(X - \mu_x)^2 = \sigma_x^2$.

Свойство. Если случайные величины X, Y независимы, то теоретическая ковариация равна нулю, т.е. $\operatorname{Cov}(X, Y) = 0$.

Более точной мерой зависимости между величинами является коэффициент корреляции. Различают *теоретический* и *выборочный* коэффициенты корреляции.

Теоретический коэффициент корреляции определяется выражением

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

где σ_X, σ_Y — средние квадратичные отклонения случайных величин X, Y.

Коэффициент корреляции является безразмерной величиной, изменяющейся в пределах $-1 \le \rho \le 1$.

Теоретический коэффициент корреляции показывает *тесноту линейной связи* двух случайных величин:

- $\rho > 0$ при положительной связи и $\rho = 1$ при строгой положительной линейной связи;
- $\rho < 0$ при отрицательной связи и $\rho = -1$ при строгой отрицательной линейной связи;
- $\rho = 0$ при отсутствии линейной связи.

Случайные величины X, Y называются **некоррелированными**, если $\rho = 0$, и **коррелированными**, если $\rho \neq 0$.

Независимость случайных величин X, Y означает отсутствие любой связи (линейной и нелинейной), а *некоррелированность* — отсутствие только линейной связи.

Свойства. Если случайные величины X, Y независимы, то они некоррелированы ($\rho = 0$), но из некоррелированности не следует их независимость, т.е. равенство $\rho = 0$ указывает на отсутствие линейной связи между переменными, но не на отсутствие связи между ними вообще.

Выборочный коэффициент корреляции определяется выражением

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}.$$

При каждом конкретном значении ρ_{XY} выборочный коэффициент корреляции является случайной величиной, изменяющейся в пределах $-1 \le r \le 1$. Он является безразмерной величиной и показывает *степень линейной связи* двух переменных.

На рис. 7 отражен геометрический смысл коэффициента корреляции.

Если для генеральной совокупности $\rho = 0$, это не всегда означает, что и для выборочной совокупности $r = 0$.

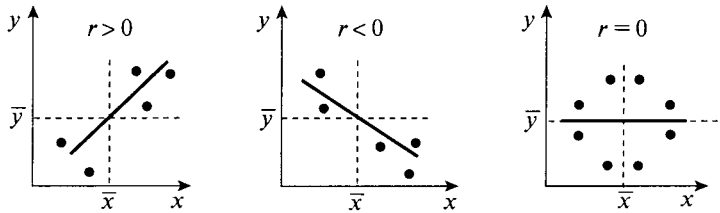


Рис. 7

Проверка гипотезы о корреляции случайных величин. Пусть по данным выборки объема n получен выборочный коэффициент корреляции $r \neq 0$. Требуется проверить гипотезу о равенстве нулю истинного значения коэффициента корреляции, т.е.

$$\begin{cases} H_0: \ \rho = 0, \\ H_1: \ \rho \neq 0. \end{cases}$$

В качестве критерия проверки гипотезы H_0 принимается случайная величина

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Величина t при справедливости гипотезы H_0 имеет распределение Стьюдента (t-статистика) с $\nu = n - 2$ степенями свободы.

Проверку значимости r выполним двумя способами.

1. Критическое значение $t_{кр}$ при заданных α и ν определяется по таблице t-распределения Стьюдента или в Excel с помощью функции

$t_{кр}$ = СТЬЮДРАСПОБР(α; ν).

Из сравнения *наблюдаемого* значения t с *критическим* получаем:

- **если $|t| < t_{кр}$, то H_0 принимается, т.е. r незначим;**
- **если $|t| > t_{кр}$, то H_0 отвергается, т.е. r значим.**

2. Наблюдаемому (расчетному) значению критерия t соответствует *значимость* t, которая может быть определена в Excel с помощью функции

Значимость t = СТЬЮДРАСП(t; ν; 2),

где $\nu = n - 2$ — число степеней свободы.

Из сравнения *значимости* t с заданным *стандартным уровнем значимости* получаем:

- **если значимость** t **больше стандартного уровня, то** r **незначим;**
- **если значимость** t **меньше стандартного уровня, то** r **значим.**

Пример 1.2. По приведенным ниже исходным данным вычислить ковариацию и коэффициент корреляции между переменными x, y, установить его значимость:

№ п/п	1	2	3	4	5
x	2	6	10	14	18
y	1	2	4	11	12

Представим исходные данные и расчетные показатели в виде следующей расчетной таблицы:

№ п/п	x	y	x^2	xy	y^2
1	2	1	4	2	1
2	6	2	36	12	4
3	10	4	100	40	16
4	14	11	196	154	121
5	18	12	324	216	144
Итого	50	30	660	424	286
Среднее	10	6	132	84,8	57,2
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{xy}$	$\overline{y^2}$

Окончательно имеем

$$\mathrm{var}(x) = \overline{x^2} - (\bar{x})^2 = 132 - 100 = 32,$$

$$\mathrm{var}(y) = \overline{y^2} - (\bar{y})^2 = 57,2 - 36 = 21,2,$$

$$\mathrm{cov}(x, y) = \overline{xy} - \bar{x}\,\bar{y} = 84,8 - 60 = 24,8,$$

$$r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} = \frac{24,8}{\sqrt{32 \cdot 21,2}} = 0,952.$$

Замечание. *В Excel можно по исходным данным получить коэффициент корреляции и его квадрат с помощью функций*

`r = ПИРСОН(массив x; массив y),`

r^2 `= КВПИРСОН(массив x; массив y).`

Проверим значимость выборочного коэффициента корреляции. Наблюдаемое значение критерия

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}952\sqrt{3}}{\sqrt{0{,}0937}} = 5{,}5.$$

Выполним проверку значимости r двумя способами.

1. При $\alpha = 0{,}05$ и $\nu = 3$ по таблице или с помощью функции `СТЬЮДРАСПОБР(α; ν)` находим $t_{кр} = 3{,}18$. Поскольку $|t| = 5{,}5 > t_{кр} = 3{,}18$, то $r = 0{,}952$ значим при 5%-ном уровне.

2. Наблюдаемому (расчетному) значению критерия $t = 5{,}5$ соответствует *значимость* $t = 0{,}0124$, которая может быть определена в Excel с помощью функции

`Значимость t = СТЬЮДРАСП(t; ν; 2),`

где $\nu = n - 2$ — число степеней свободы.

Поскольку *значимость* $t = 0{,}0124 < 0{,}05$, то коэффициент $r = 0{,}952$ значим при 5%-ном уровне, следовательно, имеется линейная зависимость между переменными.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем сходство и отличие генеральной и выборочной совокупностей?
2. В чем отличие генеральных характеристик генеральной совокупности от выборочных?
3. Что является численной характеристикой центра распределения случайной величины?
4. Что является численной характеристикой меры рассеяния случайной величины?
5. Сколькими параметрами характеризуется нормальное распределение случайной величины?
6. Что называется статистическим критерием?
7. Какие случайные величины используются в качестве статистического критерия?
8. Что характеризуют ковариация и корреляция случайных величин?

Глава 2
Модель парной регрессии

В **модели парной линейной регрессии** зависимость между переменными в генеральной совокупности представляется в виде

$$Y = \alpha + \beta X + \varepsilon, \tag{2.1}$$

где X — неслучайная величина, а Y и ε — случайные величины.

Величина Y называется **объясняемой** (зависимой) переменной, а X — **объясняющей** (независимой) переменной. Постоянные α, β — параметры уравнения.

Наличие случайного члена ε (ошибки регрессии) связано с воздействием на зависимую переменную других неучтенных в уравнении факторов.

На основе выборочного наблюдения оценивается выборочное уравнение регрессии (*линия регрессии*):

$$\hat{y} = a + bx, \tag{2.2}$$

где (a, b) — оценки параметров (α, β).

2.1. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений (x_i, y_i), $i = \overline{1, n}$ линейным уравнением (2.2).

На рис. 8 приведены диаграмма рассеяния наблюдений и линия регрессии.

Величина $\hat{y}_i$ описывается как расчетное значение переменной y, соответствующее x_i. Наблюдаемые значения y_i не лежат в точности на линии регрессии, т.е. не совпадают с $\hat{y}_i$.

Определим остаток e_i в i-м наблюдении как разность между фактическим и расчетным значениями зависимой переменной, т.е.

$$e_i = y_i - \hat{y}_i.$$

Неизвестные значения (a, b) определяются **методом наименьших квадратов** (МНК).

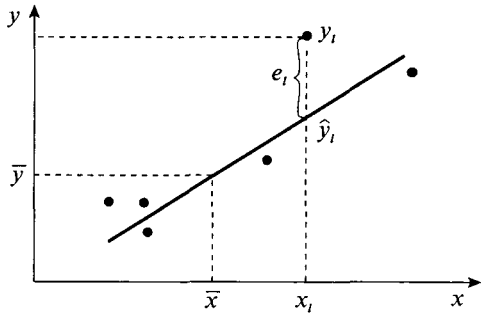


Рис. 8

Суть МНК заключается в *минимизации суммы квадратов остатков*:

$$Q = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2 \to \min.$$

Здесь (x_i, y_i) — известные значения наблюдения (числа), (a, b) — неизвестные.

Запишем необходимые условия экстремума:

$$\begin{cases} Q'_a = -2\sum (y_i - a - bx_i) = 0, \\ Q'_b = -2\sum (y_i - a - bx_i)x_i = 0. \end{cases}$$

После преобразования получим следующую систему нормальных уравнений:

$$\begin{cases} a + b\bar{x} = \bar{y}, \\ a\bar{x} + b\overline{x^2} = \overline{xy}. \end{cases}$$

Решение системы:

$$\begin{cases} b = \dfrac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}, \\ a = \bar{y} - b\bar{x}. \end{cases} \qquad (2.3)$$

Линия регрессии (расчетное значение зависимой переменной):

$$\hat{y} = a + bx, \quad \text{или} \quad \hat{y} - \bar{y} = b(x - \bar{x}).$$

Линия регрессии проходит через точку $(\bar{x}, \bar{y})$, и выполняются равенства

$$\bar{e} = 0, \quad y = \bar{\hat{y}}.$$

Коэффициент b есть **угловой коэффициент регрессии**, он показывает, на сколько единиц в среднем изменяется переменная y при увеличении независимой переменной x на единицу.

Постоянная a дает *прогнозируемое* значение зависимой переменной при $x = 0$. Это может иметь смысл в зависимости от того, как далеко находится $x = 0$ от выборочных значений x.

После построения уравнения регрессии наблюдаемые значения y можно представить как

$$y_i = \hat{y}_i + e_i. \tag{2.4}$$

Остатки e_i, как и ошибки ε_i, являются случайными величинами, однако они, в отличие от ошибок ε_i, наблюдаемы.

Докажем, что $\operatorname{cov}(\hat{y}, e) = 0$.

Действительно, используя равенства

$$\hat{y} = a + bx, \quad e = y - a - bx, \quad \operatorname{cov}(x, y) - b\operatorname{var}(x) = 0,$$

получим

$$\operatorname{cov}(\hat{y}, e) = \operatorname{cov}(a + bx, y - a - bx) = b\operatorname{cov}(x, y) - b^2 \operatorname{var}(x) =$$

$$= b[\operatorname{cov}(x, y) - b\operatorname{var}(x)] = 0.$$

Можно показать, что

$$b = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = r\sqrt{\frac{\operatorname{var}(y)}{\operatorname{var}(x)}} = r\frac{S_y}{S_x},$$

где r – коэффициент корреляции между x и y, а S_x, S_y – их стандартные отклонения.

Если коэффициент r уже рассчитан, то можно получить коэффициенты (a, b) парной регрессии.

Определим **выборочные дисперсии** величин y, $\hat{y}$, e:

$\operatorname{var}(y) = \frac{1}{n}\sum(y_i - \bar{y})^2$ — дисперсия *наблюдаемых* значений y;

$\operatorname{var}(\hat{y}) = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$ — дисперсия *расчетных* значений y;

$\operatorname{var}(e) = \frac{1}{n}\sum(e_i - \bar{e})^2 = \frac{1}{n}\sum e_i^2$ — дисперсия *остатков*.

2.2. АНАЛИЗ ВАРИАЦИИ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Цель регрессионного анализа состоит в объяснении поведения зависимой переменной y.

Пусть на основе выборочных наблюдений построено уравнение регрессии $\hat{y}$, тогда значение зависимой переменной y в каждом наблюдении можно разложить на две составляющие:

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

где остаток e_i есть та часть зависимой переменной y, которую невозможно объяснить с помощью уравнения регрессии.

Разброс значений зависимой переменной характеризуется выборочной дисперсией $\mathrm{var}(y)$. Разложим дисперсию $\mathrm{var}(y)$:

$$\mathrm{var}(y) = \mathrm{var}(\hat{y} + e) = \mathrm{var}(\hat{y}) + \mathrm{var}(e) + 2\mathrm{cov}(\hat{y}, e).$$

Поскольку $\mathrm{cov}(\hat{y}, e) = 0$, то

$$\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(e). \qquad (2.5)$$

Таким образом, дисперсия $\mathrm{var}(y)$ разложена на две части:

- $\mathrm{var}(\hat{y})$ — часть, объясненная регрессионным уравнением;
- $\mathrm{var}(e)$ — необъясненная часть.

Коэффициентом детерминации R^2 называется отношение

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{\mathrm{var}(e)}{\mathrm{var}(y)}, \quad 0 \le R^2 \le 1,$$

характеризующее долю вариации (разброса) зависимой переменной, *объясненную* с помощью уравнения регрессии.

Отношение $\dfrac{\mathrm{var}(e)}{\mathrm{var}(y)}$ представляет собой долю *необъясненной* дисперсии.

Если $R^2 = 1$, то подгонка точная:

$$\mathrm{var}(y) = \mathrm{var}(\hat{y}), \quad \mathrm{var}(e) = 0, \quad y_i = \hat{y}_i, \quad i = \overline{1, n},$$

т.е. все точки наблюдения лежат на регрессионной прямой.

Если $R^2 = 0$, то регрессия ничего не дает:

$$\mathrm{var}(y) = \mathrm{var}(e), \quad \mathrm{var}(\hat{y}) = 0, \quad \hat{y}_i = \bar{y}, \quad i = \overline{1, n},$$

т.е. переменная x не улучшает качества предсказания y по сравнению с горизонтальной прямой $\hat{y} = \bar{y}$.

Чем ближе к единице R^2, тем лучше качество подгонки, т.е. $\hat{y}$ более точно аппроксимирует y.

Замечание. *Вычисление R^2 корректно, если константа a включена в уравнение регрессии.*

Пример 2.1. Покажем, что $\sqrt{R^2} = r_{\hat{y},y}$, где $r_{\hat{y},y}$ — коэффициент корреляции между $\hat{y}$ и y.

Действительно, учитывая соотношение

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

получим

$$r_{\hat{y},y} = \frac{\operatorname{cov}(\hat{y}, y)}{\sqrt{\operatorname{var}(\hat{y})\operatorname{var}(y)}} = \sqrt{\frac{\operatorname{var}(\hat{y})}{\operatorname{var}(y)}} = \sqrt{R^2}.$$

Пример 2.2. Покажем, что $r_{\hat{y},y} = r_{x,y}$ в случае парной регрессии $\hat{y} = a + bx$.

Действительно, из соотношений

$$\operatorname{cov}(\hat{y}, y) = \operatorname{cov}(a + bx, y) = b\operatorname{cov}(x, y),$$

$$\operatorname{var}(\hat{y}) = \operatorname{var}(a + bx) = b^2 \operatorname{var}(x)$$

имеем

$$r_{\hat{y},y} = \frac{\operatorname{cov}(\hat{y}, y)}{\sqrt{\operatorname{var}(\hat{y})\operatorname{var}(y)}} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = r_{x,y}.$$

Вывод. В случае парной регрессии коэффициент детерминации есть квадрат коэффициента корреляции переменных x и y, т.е. $R^2 = r^2_{x,y}$.

Пример 2.3. Зависимость переменной в регрессии $y = \alpha + \beta x + \varepsilon$ разбивается на две компоненты: $y = y_1 + y_2$. Рассмотрим две регрессии для компонент:

$$y_1 = \alpha_1 + \beta_1 x + \varepsilon_1, \quad y_2 = \alpha_2 + \beta_2 x + \varepsilon_2.$$

Докажем следующие соотношения для МНК-оценок параметров двух регрессий: $a = a_1 + a_2$, $b = b_1 + b_2$.

Действительно,

$$b = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \frac{\operatorname{cov}(x, y_1 + y_2)}{\operatorname{var}(x)} = \frac{\operatorname{cov}(x, y_1) + \operatorname{cov}(x, y_2)}{\operatorname{var}(x)} = b_1 + b_2,$$

$$a = \bar{y} - b\bar{x} = \overline{(y_1 + y_2)} - \bar{x}(b_1 + b_2) = a_1 + a_2.$$

Пример 2.4. Покажем, что если все значения переменных изменить на одно и то же число или в одно и то же число раз, то величина коэффициента b в парной регрессии не изменится.

Пусть $x' = x + c, \quad y' = y + c$, тогда

$$b' = \frac{\operatorname{cov}(x', y')}{\operatorname{var}(x')} = \frac{\operatorname{cov}(x + c, y + c)}{\operatorname{var}(x + c)} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = b.$$

Пусть $x' = kx, \quad y' = ky$, тогда

$$b' = \frac{\operatorname{cov}(x', y')}{\operatorname{var}(x')} = \frac{\operatorname{cov}(kx, ky)}{\operatorname{var}(kx)} = \frac{k^2 \operatorname{cov}(x, y)}{k^2 \operatorname{var}(x)} = b.$$

F-ТЕСТ НА КАЧЕСТВО ОЦЕНИВАНИЯ

Для определения статистической значимости коэффициента детерминации R^2 проверяется гипотеза H_0: $F = 0$ для F-статистики:

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

Величина F имеет распределение Фишера с $\nu_1 = 1$, $\nu_2 = n - 2$.

Проверку значимости R^2 можно выполнить двумя способами.

1. Критическое значение $F_{кр}$ при заданных α, ν_1, ν_2 определяется по таблице F-распределения Фишера или в Excel с помощью функции

$F_{кр}$ = FРАСПОБР(α; ν_1; ν_2).

Из сравнения *наблюдаемого* значения F с *критическим* получаем:

- **если $F < F_{кр}$, то H_0 принимается, т.е. R^2 незначим;**
- **если $F > F_{кр}$, то H_0 отвергается, т.е. R^2 значим.**

2. Наблюдаемому (расчетному) значению критерия F соответствует определенная *значимость* F, которую можно вычислить в Excel с помощью функции

Значимость F = FРАСП(F; ν_1; ν_2).

Из сравнения *значимости* F с заданным *стандартным уровнем значимости* получаем:

- **если значимость F больше стандартного уровня, то R^2 незначим;**
- **если значимость F меньше стандартного уровня, то R^2 значим.**

3-2337

Чаще всего F-тест используется для оценки того, значимо ли объяснение, даваемое уравнением в целом.

СРЕДНЯЯ ОШИБКА АППРОКСИМАЦИИ

Оценку качества построенной модели дает коэффициент детерминации, а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации — среднее отклонение расчетных значений зависимой переменной от фактических:

$$A = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right|\cdot 100\%.$$

Допустимый предел значений A — не более 8–10%.

Пример 2.5. Построим регрессионные зависимости: а) расходов на питание y и личного дохода x; б) расходов на питание y и времени t — по следующим данным (усл. ед.):

Год	1990	1991	1992	1993	1994
x	2	6	10	14	18
y	1	2	4	11	12

и оценим качество подгонки.

а) Пусть истинная модель описывается выражением $y = \alpha + \beta x + \varepsilon$.

По выборочным наблюдениям определяем оценки (a, b).

Исходные данные и расчетные показатели удобно представить в виде следующей таблицы:

Год	x	y	x^2	xy	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1990	2	1	4	2	−0,2	25	38,44	1,44
1991	6	2	36	12	2,9	16	9,61	0,81
1992	10	4	100	40	6	4	0	4
1993	14	11	196	154	9,1	25	9,61	3,61
1994	18	12	324	216	12,2	36	38,44	0,04
Итого	50	30	660	424	30	106	96,1	9,9
Среднее	10	6	132	84,8	6	21,2	19,22	1,98
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{xy}$	$\bar{\hat{y}}$	var(y)	var($\hat{y}$)	var(e)

Окончательно имеем

$$\operatorname{cov}(x, y) = \overline{xy} - \bar{x}\,\bar{y} = 84{,}8 - 60 = 24{,}8,$$

$$\operatorname{var}(x) = \overline{x^2} - (\bar{x})^2 = 132 - 100 = 32,$$

$$b = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \frac{24{,}8}{32} = 0{,}775, \quad a = \bar{y} - b\bar{x} = 6 - 0{,}775 \cdot 10 = -1{,}75.$$

Следовательно, $\hat{y} = -1{,}75 + 0{,}775x$.

Коэффициент $b = 0{,}775$ показывает, что при увеличении дохода на 1 усл. ед. расходы на питание увеличиваются в среднем на 0,775 усл. ед.

Замечание. *В Excel оценки (a, b) можно также определить с помощью функций*:

a = ОТРЕЗОК(массив y; массив x),

b = НАКЛОН(массив y; массив x).

Условие $\operatorname{var}(y) = \operatorname{var}(\hat{y}) + \operatorname{var}(e)$ выполняется.

Качество подгонки оцениваем коэффициентом детерминации:

$$R^2 = \frac{\operatorname{var}(\hat{y})}{\operatorname{var}(y)} = \frac{19{,}22}{21{,}2} = 0{,}907,$$

т.е. 90,7% вариации зависимой переменной (расходы на питание) объясняется регрессией.

Значимость коэффициента R^2 проверяем по F-тесту:

$$F = \frac{R^2(n-2)}{1-R^2} = \frac{0{,}907 \cdot 3}{0{,}093} = 29{,}2.$$

Выполним проверку значимости R^2 двумя способами.

1. При $\alpha = 0{,}05$, $\nu_1 = 1$ и $\nu_2 = 3$ по таблице или с помощью функции **FРАСПОБР(α; ν_1; ν_2)** находим $F_{кр} = 10{,}13$. Поскольку $F = 29{,}2 > F_{кр} = 10{,}13$, то $R^2 = 0{,}907$ значим при 5%-ном уровне.

2. Наблюдаемому (расчетному) значению критерия $F = 29{,}2$ соответствует *значимость* $F = 0{,}0124$, которую можно определить в Excel с помощью функции

Значимость F = FРАСП(F; ν_1; ν_2),

где $\nu_1 = 1$, $\nu_2 = 3$.

Поскольку *значимость* $F = 0{,}0124 < 0{,}05$, то R^2 значим при уровне 5%.

3*-2337

б) Пусть истинная модель $y = \alpha + \beta t + \varepsilon$ (модель временного ряда). Выборочная регрессия $\hat{y} = a + bt$, где t — время, определяемое как $t = 1$ для 1990 г., $t = 2$ для 1991 г. и т.д.

Представим исходные и расчетные показатели в виде таблицы:

Год	t	y	t^2	ty	$\hat{y}$
1990	1	1	1	1	−0,2
1991	2	2	4	4	2,9
1992	3	4	9	12	6
1993	4	11	16	44	9,1
1994	5	12	25	60	12,2
Итого	15	30	55	121	30
Среднее	3	6	11	24,2	6
	$\overline{t}$	$\overline{y}$	$\overline{t^2}$	$\overline{ty}$	$\overline{\hat{y}}$

Окончательно имеем

$$b = \frac{\overline{ty} - \overline{t}\,\overline{y}}{\overline{t^2} - (\overline{t})^2} = \frac{24{,}2 - 18}{11 - 9} = 3{,}1, \quad a = \overline{y} - b\overline{t} = 6 - 3{,}1 \cdot 3 = -3{,}3.$$

Следовательно, $\hat{y} = -3{,}3 + 3{,}1t$.

Коэффициент $b = 3{,}1$ показывает, что за год расходы на питание в среднем возрастают на 3,1 усл. ед.

Пример 2.6. Покажем, что в модели регрессии *без свободного члена* $Y = \beta X + \varepsilon$ оценка МНК для β есть

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\overline{xy}}{\overline{x^2}}.$$

Выборочная регрессия для этой модели $\hat{y} = bx$. Наблюдаемые значения зависимой переменной связаны с расчетными значениями уравнением $y_i = \hat{y}_i + e_i$. Оценку b найдем из минимизации величины

$$Q = \sum e_i^2 = \sum (y_i - bx_i)^2 = \sum y_i^2 - 2b\sum x_i y_i + b^2 \sum x_i^2.$$

Получаем

$$Q'_b = -2\sum x_i y_i + 2b\sum x_i^2 = 0,$$

откуда

$$b = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Вычисление R^2 при отсутствии свободного члена некорректно.

Пример 2.7. Покажем, что в модели регрессии $Y = \alpha + \varepsilon$ оценка МНК для α есть $a = \bar{y}$.

Выборочная регрессия для заданной модели есть $\hat{y}_t = a$. Наблюдаемые значения зависимой переменной связаны с расчетными значениями уравнением $y_t = \hat{y}_t + e_t = a + e_t$. Оценку a найдем из минимизации величины

$$Q = \sum e_t^2 = \sum (y_t - a)^2 = \sum y_t^2 - 2a\sum y_t + na^2.$$

Получаем

$$Q'_a = -2\sum y_t + 2an = 0,$$

откуда

$$a = \frac{1}{n}\sum y_t = \bar{y}.$$

Выборочная регрессия $\hat{y} = \bar{y}$.

Упражнение 2.1. По данным примера 2.5 покажите, что зависимость расходов на питание y от личного дохода x для модели регрессии *без свободного члена* есть $\hat{y} = 0{,}642x$, при этом $\hat{y} \neq \bar{y}$ и $\mathrm{var}(y) \neq \mathrm{var}(\hat{y}) + \mathrm{var}(e)$.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. С чем связана ошибка регрессии?
2. В чем заключается метод наименьших квадратов?
3. Каков смысл коэффициента регрессии и каким способом его оценивают?
4. Что характеризует коэффициент детерминации?
5. Для чего используется F-критерий Фишера?
6. В чем смысл средней ошибки аппроксимации и как она определяется?

Глава 3
Свойства коэффициентов регрессии и проверка гипотез

3.1. СЛУЧАЙНЫЕ СОСТАВЛЯЮЩИЕ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Величина Y в модели регрессии $Y = \alpha + \beta X + \varepsilon$ имеет две составляющие:

- *неслучайную* $(\alpha + \beta X)$;
- *случайную* (ε).

Оценки коэффициентов регрессии (a, b) являются линейными функциями Y, и теоретически их также можно представить в виде двух составляющих.

Воспользовавшись разложением показателей

$$\operatorname{cov}(x, y) = \operatorname{cov}(x, \alpha + \beta x + \varepsilon) = \beta \operatorname{var}(x) + \operatorname{cov}(x, \varepsilon),$$

$$\bar{y} = \frac{1}{n}\sum y_t = \frac{1}{n}\sum(\alpha + \beta x_t + \varepsilon_t) = \alpha + \beta\bar{x} + \frac{1}{n}\sum \varepsilon_t,$$

получим преобразованные соотношения для (a, b):

$$\begin{cases} b = \beta + \dfrac{\operatorname{cov}(x, \varepsilon)}{\operatorname{var}(x)}, \\ a = \alpha + \left[\dfrac{1}{n}\sum \varepsilon_t - \dfrac{\bar{x}\operatorname{cov}(x, \varepsilon)}{\operatorname{var}(x)}\right]. \end{cases} \tag{3.1}$$

Таким образом, коэффициенты (a, b) разложены на две составляющие: *неслучайную*, равную истинным значениям (α, β), и *случайную*, зависящую от ε.

На практике нельзя разложить коэффициенты регрессии на составляющие, так как значения (α, β) или фактические значения ε в выборке неизвестны.

3.2. ПРЕДПОСЫЛКИ РЕГРЕССИОННОГО АНАЛИЗА

УСЛОВИЯ ГАУССА — МАРКОВА

Линейная регрессионная модель с двумя переменными имеет вид

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = \overline{1, n}),$$

где Y — объясняемая переменная, X — объясняющая переменная, ε — случайный член.

Для того чтобы регрессионный анализ, основанный на МНК, давал наилучшие из всех возможных результаты, должны выполняться определенные условия (**условия Гаусса — Маркова**).

1. *Математическое ожидание случайного члена в любом наблюдении должно быть равно нулю*, т.е.

$$M(\varepsilon_i) = 0 \quad (i = \overline{1, n}).$$

2. *Дисперсия случайного члена должна быть постоянной для всех наблюдений*, т.е.

$$D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma^2 \quad (i = \overline{1, n}).$$

3. *Случайные члены должны быть статистически независимы (некоррелированы) между собой*, т.е.

$$M(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j).$$

4. *Объясняющая переменная x_i есть величина неслучайная.*

При выполнении условий Гаусса — Маркова модель называется **классической нормальной линейной регрессионной моделью**.

Наряду с условиями Гаусса — Маркова обычно предполагается, что *случайный член распределен нормально*, т.е. $\varepsilon_i \sim N(0; \sigma^2)$.

Замечание. *Если случайный член имеет нормальное распределение, то требование некоррелированности случайных членов эквивалентно их независимости.*

Рассмотрим подробнее условия и предположения, лежащие в основе регрессионного анализа.

Первое условие означает, что случайный член не должен иметь систематического смещения. Если постоянный член включен в уравнение регрессии, то это условие выполняется автоматически.

Второе условие означает, что дисперсия случайного члена в каждом наблюдении имеет только одно значение.

Под дисперсией σ^2 имеется в виду возможное поведение случайного члена до того, как сделана выборка. Величина σ^2 неизвестна, и одна из задач регрессионного анализа состоит в ее оценке.

Условие *независимости* дисперсии случайного члена от номера наблюдения называется **гомоскедастичностью** (что означает одинаковый разброс). *Зависимость* дисперсии случайного члена от номера наблюдения называется **гетероскедастичностью**.

Таким образом:

- $D(\varepsilon_i) = \sigma^2 \quad (i = \overline{1, n})$ — гомоскедастичность;
- $D(\varepsilon_i) = \sigma_i^2 \quad (i = \overline{1, n})$ — гетероскедастичность.

Характерные диаграммы рассеяния для случаев гомоскедастичности и гетероскедастичности показаны на рис. 9, *а* и *б* соответственно.

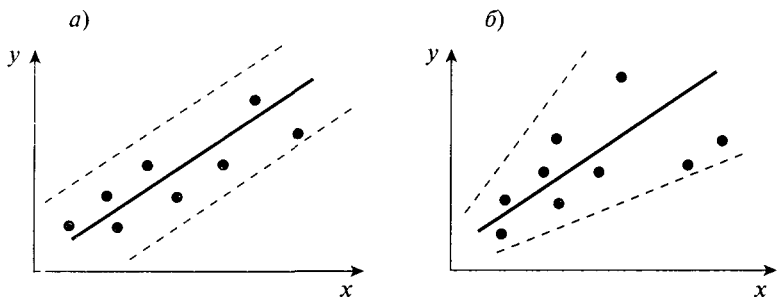


Рис. 9

Если условие гомоскедастичности не выполняется, то оценки коэффициентов регрессии будут *неэффективными*, хотя и *несмещенными*.

Существуют специальные методы диагностирования и устранения гетероскедастичности.

Третье условие указывает на некоррелированность случайных членов для разных наблюдений. Это условие часто нарушается, когда данные являются временными рядами. В случае когда третье условие не выполняется, говорят об **автокорреляции остатков**.

Типичный вид данных при наличии автокорреляции показан на рис. 10.

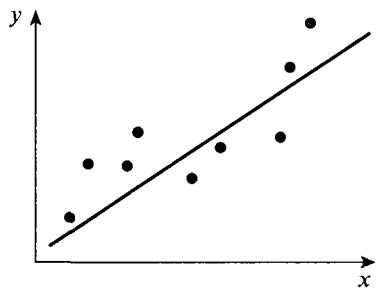


Рис. 10

Если условие независимости случайных членов не выполняется, то оценки коэффициентов регрессии, полученные по МНК, оказываются *неэффективными*, хотя и *несмещенными*.

Существуют методы диагностирования и устранения автокорреляции.

Четвертое условие является особенно важным. Если условие о неслучайности объясняющей переменной не выполняется, то оценки коэффициентов регрессии оказываются *смещенными* и *несостоятельными*.

Нарушение этого условия может быть связано с ошибками измерения объясняющих переменных или с использованием лаговых переменных.

В регрессионном анализе часто вместо *условия о неслучайности объясняющей переменной* используется более слабое *условие о независимости (некоррелированности) распределений объясняющей переменной и случайного члена.* Получаемые при этом оценки коэффициентов регрессии обладают теми же основными свойствами, что и оценки, полученные при использовании условия о неслучайности объясняющей переменной.

Предположение о нормальности распределения случайного члена необходимо для проверки значимости параметров регрессии и для их интервального оценивания.

ТЕОРЕМА ГАУССА — МАРКОВА

Теорема Гаусса — Маркова. Если условия 1—4 регрессионного анализа выполняются, то оценки (a, b), сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками, т.е. обладают следующими свойствами:

1) *несмещенность*: $M(a) = \alpha, \quad M(b) = \beta$.

(Это означает отсутствие систематической ошибки в положении линии регрессии);

2) *эффективность*: имеют наименьшую дисперсию в классе всех линейных несмещенных оценок, равную

$$D(a) = \frac{\overline{x^2}\sigma^2}{n\,\mathrm{var}(x)}, \quad D(b) = \frac{\sigma^2}{n\,\mathrm{var}(x)};$$

3) *состоятельность*: $\lim\limits_{n\to\infty} D(a) = 0, \quad \lim\limits_{n\to\infty} D(b) = 0$.

(Это означает, что при достаточно большом n оценки (a, b) близки к (α, β).)

Для проверки выводов теоремы воспользуемся оценками (a, b) в виде разложения (3.1) и соотношением

$$\mathrm{cov}(x, \varepsilon) = \frac{1}{n}\sum (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) = \frac{1}{n}\sum (x_i - \overline{x})\varepsilon_i - \frac{\overline{\varepsilon}}{n}\sum (x_i - \overline{x}) =$$

$$= \frac{1}{n}\sum (x_i - \overline{x})\varepsilon_i.$$

Пусть x — неслучайная величина, тогда

$$M[\mathrm{cov}(x, \varepsilon)] = M\left[\frac{1}{n}\sum (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n}\sum (x_i - \overline{x})M(\varepsilon_i) = 0,$$

$$D[\mathrm{cov}(x, \varepsilon)] = D\left[\frac{1}{n}\sum (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n^2}\sum (x_i - \overline{x})^2 D(\varepsilon_i) = \frac{\mathrm{var}(x)}{n}\sigma^2,$$

$$M\left[\frac{\sum \varepsilon_i}{n}\right] = \frac{\sum M(\varepsilon_i)}{n} = 0,$$

$$D\left[\frac{\sum \varepsilon_i}{n}\right] = \frac{\sum D(\varepsilon_i)}{n^2} = \frac{\sigma^2}{n},$$

$$M\left[\frac{\overline{x}\,\mathrm{cov}(x, \varepsilon)}{\mathrm{var}(x)}\right] = \frac{\overline{x}}{\mathrm{var}(x)} M[\mathrm{cov}(x, \varepsilon)] = 0,$$

$$D\left[\frac{\overline{x}\,\mathrm{cov}(x, \varepsilon)}{\mathrm{var}(x)}\right] = \frac{(\overline{x})^2}{\mathrm{var}^2(x)} D[\mathrm{cov}(x, \varepsilon)] = \frac{(\overline{x})^2\sigma^2}{n\,\mathrm{var}(x)}.$$

Вычислим математическое ожидание и дисперсию оценок b и a:

$$M(b) = \beta + M\left[\frac{\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \beta + \frac{M[\operatorname{cov}(x,\varepsilon)]}{\operatorname{var}(x)} = \beta,$$

$$D(b) = D\left[\frac{\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \frac{D[\operatorname{cov}(x,\varepsilon)]}{\operatorname{var}^2(x)} = \frac{\sigma^2}{n\operatorname{var}(x)};$$

$$M(a) = \alpha + M\left[\frac{\sum \varepsilon_i}{n}\right] - M\left[\frac{\bar{x}\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \alpha,$$

$$D(a) = D\left[\frac{\sum \varepsilon_i}{n}\right] + D\left[\frac{\bar{x}\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \frac{\sigma^2}{n}\left[1 + \frac{(\bar{x})^2}{\operatorname{var}(x)}\right] = \frac{\sigma^2\overline{x^2}}{n\operatorname{var}(x)}.$$

РАСЧЕТ СТАНДАРТНЫХ ОШИБОК КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Полученные теоретические дисперсии $D(a)$, $D(b)$ зависят от дисперсии σ^2 случайного члена.

По данным выборки отклонения ε_i, а следовательно, и их дисперсии σ^2 неизвестны, поэтому они заменяются наблюдаемыми остатками e_i и их выборочной дисперсией.

Однако оценка $\operatorname{var}(e)$ является *смещенной*, т.е.

$$M[\operatorname{var}(e)] = \frac{n-2}{n}\sigma^2.$$

Несмещенной оценкой дисперсии σ^2 является величина (остаточная дисперсия)

$$S^2 = \frac{n}{n-2}\operatorname{var}(e) = \frac{1}{n-2}\sum e_i^2,$$

которая служит *мерой разброса* зависимой переменной вокруг линии регрессии.

Величина S называется **стандартной ошибкой регрессии**.

Отметим, что в знаменателе остаточной дисперсии стоит число степеней свободы $(n-2)$, а не n, так как две степени свободы теряются при определении двух параметров (a, b).

Заменив в теоретических дисперсиях неизвестную σ^2 на оценку S^2, получим оценки дисперсии:

$$S_a^2 = \frac{\overline{x^2}S^2}{n\operatorname{var}(x)}, \quad S_b^2 = \frac{S^2}{n\operatorname{var}(x)}.$$

Величины S_a, S_b называются **стандартными ошибками коэффициентов регрессии**.

Пример 3.1. Для полученной в примере 2.5 зависимости расходов на питание от личного дохода $\hat{y} = -1{,}75 + 0{,}775x$ рассчитаем стандартные ошибки коэффициентов регрессии.

Исходные данные: $n = 5$, $\text{var}(x) = 32$, $\overline{x^2} = 132$, $\text{var}(e) = 1{,}98$.

Остаточная дисперсия S^2 и стандартная ошибка регрессии S равны соответственно

$$S^2 = \frac{n}{n-2}\text{var}(e) = \frac{5}{3}\cdot 1{,}98 = 3{,}3, \quad S = \sqrt{3{,}3} = 1{,}816.$$

Для расчета стандартной ошибки можно также воспользоваться функцией Excel:

`S = СТОШYX(массив y; массив x).`

Стандартные ошибки коэффициентов регрессии

$$S_a = S\sqrt{\frac{\overline{x^2}}{n\,\text{var}(x)}} = 1{,}816\sqrt{\frac{132}{5\cdot 32}} = 1{,}65,$$

$$S_b = \frac{S}{\sqrt{n\,\text{var}(x)}} = \frac{1{,}816}{\sqrt{5\cdot 32}} = 0{,}143.$$

Пример 3.2. Покажем, что в выборочной регрессии *без свободного члена* $\hat{y} = bx$ стандартная ошибка оценки b

$$S_b = \frac{S}{\sqrt{n\overline{x^2}}},$$

где $S^2 = \dfrac{1}{n-1}\sum(y_i - bx_i)^2$.

Подставим в оценку для b выражение $y_i = \beta x_i + \varepsilon_i$:

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2} = \beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}.$$

Оценка b является *несмещенной*, так как $M(b) = \beta$.

Дисперсия оценки b

$$D(b) = \frac{\sum x_i^2 D(\varepsilon_i)}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2}{n\overline{x^2}}.$$

В исходной модели оценивается один параметр, поэтому оценкой σ^2 является

$$S^2 = \frac{1}{n-1}\sum e_i^2 = \frac{1}{n-1}\sum (y_i - bx_i)^2.$$

Следовательно, $S_b = \frac{S}{\sqrt{nx^2}}$.

Пример 3.3. Покажем, что в выборочной регрессии $\hat{y} = a$ стандартная ошибка оценки a

$$S_a = \frac{S}{\sqrt{n}},$$

где $S^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$.

Подставим в оценку для a выражение $y_i = \alpha + \varepsilon_i$:

$$a = \frac{\sum y_i}{n} = \frac{\sum(\alpha + \varepsilon_i)}{n} = \alpha + \frac{\sum \varepsilon_i}{n}.$$

Оценка a является *несмещенной*, так как $M(a) = \alpha$.

Дисперсия оценки a

$$D(a) = \frac{\sum D(\varepsilon_i)}{n^2} = \frac{\sigma^2}{n}.$$

В исходной модели оценивается один параметр, поэтому оценкой σ^2 является

$$S^2 = \frac{1}{n-1}\sum e_i^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2.$$

Следовательно, $S_a = \frac{S}{\sqrt{n}}$.

Пример 3.4. По данным примера 2.5 построим зависимость расходов на питание y от личного дохода x для модели регрессии *без свободного члена* и рассчитаем стандартную ошибку коэффициента регрессии.

Исходные данные и расчетные показатели представим в виде таблицы:

Год	x	y	x^2	xy	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1990	2	1	4	2	1,28	25	26,378	0,0806
1991	6	2	36	12	3,85	16	6,594	3,429

Окончание таблицы

Год	x	y	x^2	xy	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1992	10	4	100	40	6,42	4	0	5,856
1993	14	11	196	154	8,99	25	6,594	4,048
1994	18	12	324	216	11,56	36	26,378	0,193
Итого	50	30	660	424	32,1	106	65,946	13,608
Среднее	10	6	132	84,8	6,42	21,2	13,189	2,721
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{xy}$	$\bar{\hat{y}}$	var(y)	var($\hat{y}$)	var(e)

Коэффициент b определяется выражением

$$b = \frac{\overline{xy}}{\overline{x^2}} = \frac{84,8}{132} = 0,642.$$

Следовательно, $\hat{y} = 0,642x$.

Заметим, что в отсутствие свободного члена $\bar{\hat{y}} \neq \bar{y}$, $\text{var}(y) \neq \text{var}(\hat{y}) + \text{var}(e)$.

Остаточная дисперсия S^2 и стандартная ошибка регрессии S равны соответственно

$$S^2 = \frac{1}{n-1}\sum e_i^2 = \frac{13,608}{4} = 3,397, \quad S = \sqrt{3,397} = 1,843.$$

Стандартная ошибка коэффициента регрессии

$$S_b = \frac{S}{\sqrt{n\overline{x^2}}} = \frac{1,843}{\sqrt{5 \cdot 132}} = 0,071.$$

СТАТИСТИЧЕСКИЕ СВОЙСТВА МНК-ОЦЕНОК (a, b)

Пусть выполняется условие нормальности распределения случайного члена: $\varepsilon_i \sim N(0; \sigma^2)$. Тогда МНК-оценки коэффициентов регрессии также имеют нормальное распределение, поскольку являются линейными функциями от ε_i, т.е.

$$a \sim N\left(\alpha; \frac{\sigma^2\overline{x^2}}{n\,\text{var}(x)}\right); \quad b \sim N\left(\beta; \frac{\sigma^2}{n\,\text{var}(x)}\right).$$

Если условие нормальности распределения случайного члена не выполняется, то оценки (a, b) имеют асимптотически нормальное распределение.

3.3. ПРОВЕРКА ГИПОТЕЗ, ОТНОСЯЩИХСЯ К КОЭФФИЦИЕНТАМ РЕГРЕССИИ (a, b)

ПРОВЕРКА ГИПОТЕЗЫ H_0: $\beta = \beta_0$

Пусть в теоретической зависимости

$$Y = \alpha + \beta X + \varepsilon$$

случайный член ε распределен нормально с неизвестной дисперсией σ^2.

Хотя величина β и неизвестна, имеется основание предполагать, что она равна заданной величине β_0.

Выдвигаются гипотезы

$$\begin{cases} H_0: \ \beta = \beta_0, \\ H_1: \ \beta \neq \beta_0. \end{cases}$$

Задача заключается в проверке гипотезы H_0 на основании выборочных данных.

Пусть по выборочным данным получена оценка b.

В качестве критерия проверки гипотезы H_0 принимают случайную величину

$$t = \frac{b - \beta_0}{S_b},$$

которая имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы.

Вычисляется наблюдаемое значение критерия t. По таблице критических точек распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $\nu = n - 2$ находят критическую точку $t_{кр}$.

Сравнивая наблюдаемое значение критерия с критическим, можно принять или отвергнуть нулевую гипотезу.

Результаты оценивания регрессии совместимы не только с конкретной гипотезой H_0: $\beta = \beta_0$, но и с некоторым их множеством.

Любое значение β, совместимое с оценкой b, удовлетворяет условию

$$\left| \frac{b - \beta}{S_b} \right| < t_{кр}, \quad \text{или} \quad -t_{кр} < \frac{\beta - b}{S_b} < t_{кр}.$$

Разрешив это неравенство относительно β, получим

$$b - t_{кр}S_b < \beta < b + t_{кр}S_b,$$

т.е. **доверительный интервал** для величины β.

Посредине интервала лежит величина b. Границы интервала одинаково отстоят от b, зависят от выбора уровня значимости и являются случайными числами.

Доверительный интервал покрывает значение параметра β с заданной вероятностью $(1 - \alpha)$, т.е.

$$P(b - t_{кр}S_b < \beta < b + t_{кр}S_b) = 1 - \alpha.$$

ПРОВЕРКА ГИПОТЕЗЫ H_0: β = 0

Пусть по выборке получена оценка коэффициента регрессии b.

Гипотеза H_0: β = 0 используется для установления значимости коэффициента регрессии b.

Соответствующая t-статистика есть оценка коэффициента регрессии, деленная на ее стандартную ошибку, т.е.

$$t = \frac{b - 0}{S_b} = \frac{b}{S_b}.$$

Величина t имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы.

Наблюдаемому (расчетному) значению критерия t соответствует определенная *значимость* t, которую можно вычислить в Excel с помощью функции

`Значимость t = СТЬЮДРАСП(t; ν; 2).`

Из сравнения *значимости* t с заданным *стандартным уровнем значимости* получаем:

- **если значимость t больше стандартного уровня, то b незначим;**
- **если значимость t меньше стандартного уровня, то b значим.**

Пример 3.5. Зависимость расходов на питание y от личного дохода x по данным примера 2.5 имеет вид

$$\hat{y} = \underset{(1,65)}{-1,75} + \underset{(0,143)}{0,775x}$$

(в скобках указаны стандартные ошибки).

Оценим значимость коэффициента регрессии $b = 0{,}775$ и построим доверительный интервал для β при 5%-ном уровне значимости.

Наблюдаемое значение критерия $t = \dfrac{b}{S_b} = \dfrac{0{,}775}{0{,}143} = 5{,}4$.

Значимость $t = 0{,}0124$, соответствующую расчетному значению критерия $t = 5{,}4$, определяем с помощью функции **Значимость t = СТЬЮДРАСП(t; ν; 2)**, где $\nu = 3$. Поскольку *значимость* $t = 0{,}0124 < 0{,}05$, то коэффициент регрессии $b = 0{,}775$ значим.

При $\alpha = 0{,}05$ критическое значение критерия $t_{кр} = 3{,}18$ определяем с помощью функции $t_{кр}$ **= СТЬЮДРАСПОБР(α; ν)**.

Доверительный интервал для β

$0{,}775 - 3{,}18 \cdot 0{,}143 < \beta < 0{,}775 + 3{,}18 \cdot 0{,}143$, или $0{,}32 < \beta < 1{,}23$.

ПАКЕТ АНАЛИЗА EXCEL (ПРОГРАММА «РЕГРЕССИЯ»)

Построение линейной регрессии, оценивание ее параметров и их значимости можно выполнить значительно быстрее при использовании пакета анализа Excel (программа «Регрессия»).

Пример 3.6. Имеются данные о расходах на питание y и душевом доходе x для девяти групп семей (усл. ед.):

x	63	158	260	370	480	593	728	935	1880
y	43	62	90	111	130	149	165	191	241

Рассмотрим интерпретацию полученных результатов в общем случае (k объясняющих переменных) по данным примера 3.6.

I

Регрессионная статистика	
Множественный *R*	0,940
R-квадрат	0,884
Нормированный *R*-квадрат	0,868
Стандартная ошибка	22,87
Наблюдения	9

В таблице **Регрессионная статистика** приводятся значения:

1. **Множественный** R — коэффициент множественной корреляции $R = \sqrt{R^2}$.
2. ***R*-квадрат** — коэффициент детерминации R^2.

3. **Нормированный R-квадрат** — скорректированный R^2 с поправкой на число степеней свободы.

4. **Стандартная ошибка** — стандартная ошибка регрессии S.

5. **Наблюдения** — число наблюдений n.

II

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	Значимость *F*
Регрессия	1	28102,2	28102,2	53,69	0,00016
Остаток	7	3663,7	523,3		
Итого	8	31766			

В таблице **Дисперсионный анализ** приведены:

1. Столбец ***df*** — число степеней свободы, равное:
   $df = k$ для строки **Регрессия**;
   $df = n - k - 1$ для строки **Остаток**;
   $df = n - 1$ для строки **Итого**.
2. Столбец ***SS*** — сумма квадратов отклонений, равная:
   $\sum(\hat{y} - \bar{y})^2$ для строки **Регрессия**;
   $\sum(y - \hat{y})^2$ для строки **Остаток**;
   $\sum(y - \bar{y})^2$ для строки **Итого**.
3. Столбец ***MS*** — дисперсии, определяемые по формуле ***MS = SS/df***:
   - *факторная* для строки **Регрессия**;
   - *остаточная* для строки **Остаток**.
4. Столбец ***F*** — расчетное значение F-критерия, вычисляемое по формуле

   ***F = MS*(регрессия)/*MS*(остаток).**

5. Столбец **Значимость *F*** — значение уровня значимости, соответствующее вычисленной F-статистике:

   Значимость *F* = FРАСП(*F*-статистика; *df*(регрессия); *df*(остаток)).

Если значимость *F* меньше стандартного уровня значимости, то R^2 статистически значим.

III

	Коэффициенты	Стандартная ошибка	*t*-статистика	*P*-значение	Нижние 95%	Верхние 95%
	65,92	11,74	5,61	0,00080	38,16	93,68
x	0,107	0,014	7,32	0,00016	0,0728	0,142

В этой таблице указаны:

1. **Коэффициенты** — значения коэффициентов a, b.

2. **Стандартная ошибка** — стандартные ошибки коэффициентов регрессии S_a, S_b.

3. ***t*-статистика** — расчетные значения t-критерия, вычисляемые по формуле

***t*-статистика = Коэффициенты/Стандартная ошибка.**

4. ***P*-значение (значимость *t*)** — это значение уровня значимости, соответствующее вычисленной t-статистике:

***P*-значение = СТЬЮДРАСП(*t*-статистика; *df*(остаток)).**

Если *P*-значение меньше стандартного уровня значимости, то соответствующий коэффициент статистически значим.

5. **Нижние 95%** и **Верхние 95%** — нижние и верхние границы 95%-ных доверительных интервалов для коэффициентов теоретического уравнения линейной регрессии.

IV

Вывод остатка		
Наблюдение	Предсказанное *y*	Остатки *e*
1	72,70	-29,70
2	82,91	-20,91
3	94,53	-4,53
4	105,72	5,27
5	117,56	12,44
6	129,70	19,29
7	144,22	20,77
8	166,49	24,50
9	268,13	-27,13

В таблице **Вывод остатка** указаны:

1. **Наблюдение** — номер наблюдения.

2. **Предсказанное *y*** — расчетные значения зависимой переменной.

3. **Остатки *e*** — разница между наблюдаемыми и расчетными значениями зависимой переменной.

Используя результаты работы пакета анализа Excel (программа «Регрессия»), проанализируем зависимость расходов на питание от величины душевого дохода.

Результаты регрессионного анализа принято записывать в виде

$$\underset{(11.74)}{\hat{y} = 65{,}92} + \underset{(0{,}014)}{0{,}107x}, \quad R^2 = 0{,}884$$

(в скобках указаны стандартные ошибки коэффициентов регрессии).

Коэффициенты регрессии $a = 65{,}92$ и $b = 0{,}107$. Направление связи между y и x определяет знак коэффициента регрессии $b = 0{,}107$, т.е. связь является прямой и положительной. Коэффициент $b = 0{,}107$ показывает, что при увеличении душевого дохода на 1 усл. ед. расходы на питание увеличиваются на 0,107 усл. ед.

Оценим значимость коэффициентов полученной модели. Значимость коэффициентов (a, b) проверяется по t-тесту:

P-значение $(a) = 0{,}00080 < 0{,}01 < 0{,}05$;

P-значение $(b) = 0{,}00016 < 0{,}01 < 0{,}05$.

Следовательно, коэффициенты (a, b) значимы при 1%-ном уровне, а тем более при 5%-ном уровне значимости. Таким образом, коэффициенты регрессии значимы и модель адекватна исходным данным.

Результаты оценивания регрессии совместимы не только с полученными значениями коэффициентов регрессии, но и с некоторым их множеством (доверительным интервалом). С вероятностью 95% доверительные интервалы для коэффициентов есть (38,16–93,68) для a и (0,0728–0,142) для b.

Качество модели оценивается коэффициентом детерминации R^2. Величина $R^2 = 0{,}884$ означает, что фактором душевого дохода можно объяснить 88,4% вариации (разброса) расходов на питание.

Значимость R^2 проверяется по F-тесту:

Значимость $F = 0{,}00016 < 0{,}01 < 0{,}05$.

Следовательно, R^2 значим при 1%-ном уровне, а тем более при 5%-ном уровне значимости.

В случае парной линейной регрессии коэффициент корреляции можно определить как $r = \sqrt{R^2} = 0{,}94$. Полученное значение коэффициента корреляции свидетельствует, что связь между расходами на питание и душевым доходом очень тесная.

В парном регрессионном анализе эквивалентны t-критерий для H_0: $\beta = 0$; t-критерий для H_0: $\rho = 0$; F-критерий для R^2:

$$t_b = \frac{b}{S_b}, \quad t_r = r\sqrt{\frac{n-2}{1-r^2}}, \quad F = \frac{R^2(n-2)}{1-R^2}.$$

Связь между критериями выражается равенством

$$t_b = t_r = \sqrt{F},$$

причем для критических значений критериев при любом уровне значимости

$$(t_b)_{\text{кр}} = (t_r)_{\text{кр}} = \sqrt{F_{\text{кр}}},$$

и эти критерии дают один и тот же результат.

Вывод. Проверки значимости коэффициента b в парной линейной регрессии, коэффициента корреляции r и коэффициента детерминации R^2 эквивалентны.

3.4. ПРОГНОЗИРОВАНИЕ В РЕГРЕССИОННЫХ МОДЕЛЯХ

Под **прогнозированием** в эконометрике понимается построение оценки зависимой переменной для некоторого набора независимых переменных, которых нет в исходных наблюдениях.

Различают *точечное* и *интервальное* прогнозирование. В первом случае оценка — некоторое число, во втором — интервал, в котором находится истинное значение зависимой переменной с заданным уровнем значимости.

Рассмотрим регрессионную модель

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

Действительное значение зависимой переменной при $x = x_p$

$$y_p = \alpha + \beta x_p + \varepsilon_p,$$

где $M(\varepsilon_p) = 0$, $D(\varepsilon_p) = \sigma^2$. Значения α, β, ε_p неизвестны.

Предсказанным значением является оценка y_p (точечный прогноз):

$$\hat{y}_p = a + bx_p.$$

Ошибка предсказания равна разности между предсказанным и действительным значениями:

$$\Delta_p = \hat{y}_p - y_p.$$

Ошибка предсказания имеет нулевое математическое ожидание:

$$M(\Delta_p) = 0.$$

Действительно,

$$M(\Delta_p) = M(\hat{y}_p) - M(y_p) = M(a + bx_p) - M(\alpha + \beta x_p + \varepsilon_p) = 0.$$

Вычислим дисперсию прогноза. Учитывая, что в случае парной регрессии

$$\hat{y}_p = \bar{y} + b(x_p - \bar{x}), \quad D(\bar{y}) = \frac{\sigma^2}{n}, \quad D(y_p) = D(\varepsilon_p) = \sigma^2,$$

для дисперсии прогноза получим

$$D(\Delta_p) = \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n\,\mathrm{var}(x)}\right]\sigma^2.$$

Из формулы следует, что чем больше x_p отклоняется от выборочного среднего $\bar{x}$, тем больше дисперсия ошибки предсказания, и чем больше объем выборки n, тем меньше дисперсия.

Заменяя в дисперсии прогноза σ^2 на ее оценку S^2 и извлекая квадратный корень, получим **стандартную ошибку предсказания**

$$S_p = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n\,\mathrm{var}(x)}}.$$

Доверительный интервал для действительного значения y_p определяется выражением

$$\hat{y}_p - t_{\text{кр}} S_p < y_p < \hat{y}_p + t_{\text{кр}} S_p,$$

где $t_{\text{кр}}$ — критическое значение t-статистики при заданном уровне значимости и числе степеней свободы.

На рис. 11 в общем виде показано соотношение между доверительным интервалом предсказания и значением объясняющей переменной. Отрезок, отмеченный на рисунке стрелками, определяет доверительный интервал предсказания в точке x_p.

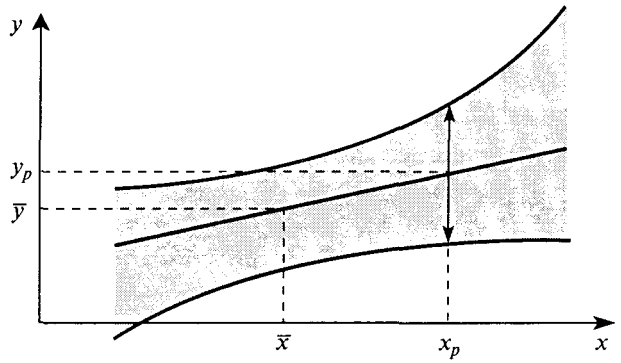


Рис. 11

Пример 3.7. По данным о зависимости объема продаж y фирмы от затрат на рекламу x оценить объем продаж при затратах на рекламу, равных 5,5 усл. ед. Найти стандартную ошибку предсказания и 99%-ный доверительный интервал для полученной оценки.

Исходные данные (усл. ед.):

x	5	8	6	5	3	9	12	4	3	10
y	72	76	78	70	68	80	82	65	62	90

Объем продаж фирмы $y_p = 71{,}87$ при затратах на рекламу $x_p = 5{,}5$ можно определить с помощью статистической функции Excel

y_p = ПРЕДСКАЗ(x_p; массив x; массив y).

Значения $\bar{x} = 6{,}5$, $\mathrm{var}(x) = 9{,}61$, $S = 4{,}24$, $t_{кр} = 3{,}35$ можно получить с помощью функций:

$\bar{x}$ = СРЗНАЧ(массив x);

var(x) = ДИСПР(массив x);

S = СТОШУХ(массив y; массив x);

$t_{кр}$ = СТЬЮДРАСПОВР(1 − α; ν),

где $n = 10$, $\nu = n - 2 = 8$, $\alpha = 1 - 0{,}99 = 0{,}01$.

Вычисляем стандартную ошибку предсказания и доверительный интервал для полученной оценки:

$$S_p = 4{,}24\sqrt{1+\frac{1}{10}+\frac{(5{,}5-6{,}5)^2}{10\cdot 9{,}61}} = 4{,}47,$$

$$71{,}87 - 3{,}35\cdot 4{,}47 < y_p < 71{,}87 + 3{,}35\cdot 4{,}47,$$

или

$$56.9 < y_p < 86{,}84.$$

3.5. НЕЛИНЕЙНЫЕ РЕГРЕССИИ

Нелинейность регрессии проявляется как по переменным, так и по параметрам.

Нелинейность по переменным устраняется путем *замены переменной*. Например, нелинейное уравнение $y = \alpha + \beta\sqrt{x} + \varepsilon$ после замены переменной $z = \sqrt{x}$ становится линейным: $y = \alpha + \beta z + \varepsilon$, и для оценки его параметров используется МНК.

Пример 3.8. Имеются данные о зависимости между ежегодным потреблением бананов y и годовым доходом x 10 американских семей (усл. ед.):

x	1	2	3	4	5	6	7	8	9	10
y	2	7	9	12	10	12	11	12	13	12

Рассмотрим различные варианты уравнения регрессии.

1. Оценка линейного уравнения $y = \alpha + \beta x + \varepsilon$ по выборочным наблюдениям (x, y) приводит к уравнению

$$\hat{y} = 5{,}13 + 0{,}88x, \quad R^2 = 0{,}64, \quad S = 2{,}10.$$

2. Если рассмотрим нелинейное уравнение $y = \alpha + \beta\sqrt{x} + \varepsilon$ и определим $z = \sqrt{x}$, то уравнение примет линейный вид $y = \alpha + \beta z + \varepsilon$.

Оценив регрессию между y и z, получим

$$\hat{y} = 0{,}774 + 4{,}106z, \quad R^2 = 0{,}762, \quad S = 1{,}72.$$

Подставив $z = \sqrt{x}$, имеем $\hat{y} = 0{,}774 + 4{,}106\sqrt{x}$.

3. Если рассмотрим нелинейное уравнение $y = \alpha + \beta/x + \varepsilon$ и определим $z = 1/x$, то уравнение примет линейный вид $y = \alpha + \beta z + \varepsilon$.

Оценив регрессию между y и z, получим

$$\hat{y} = 13{,}42 - 11{,}67z, \quad R^2 = 0{,}942, \quad S = 0{,}85.$$

Подставив $z = 1/x$, имеем $\hat{y} = 13{,}42 - \dfrac{11{,}67}{x}$.

Качество оценивания последнего варианта уравнения выше, чем у других.

В таблице представлены исходные данные для построения рассмотренных уравнений регрессии с помощью пакета анализа Excel (программа «Регрессия»):

y	x	$z = \sqrt{x}$	$z = 1/x$
2	1	1	1
7	2	1,414	0,5
9	3	1,732	0,333
12	4	2	0,25
10	5	2,236	0,2
12	6	2,449	0,166
11	7	2,645	0,142
12	8	2,828	0,125
13	9	3	0,111
12	10	3,162	0,1

Нелинейность по параметру часто устраняется путем *логарифмического преобразования уравнения*. Например, следующие нелинейные уравнения после логарифмирования сводятся к линейным:

- степенная функция $y = \alpha x^{\beta}\varepsilon \sim \ln y = \ln\alpha + \beta \ln x + \ln\varepsilon$;
- экспоненциальная функция $y = \alpha e^{\beta x}\varepsilon \sim \ln y = \ln\alpha + \beta x + \ln\varepsilon$.

Использование МНК для нахождения оценок параметров этих уравнений требует, чтобы $\ln\varepsilon$ имел нормальное распределение.

Однако уравнение $y = \alpha x^{\beta} + \varepsilon$, в котором случайный член ε является аддитивным, уже никакими преобразованиями не приводится к линейному. В этом случае используют специальные *итерационные методы оценивания нелинейной регрессии*.

В экономике применяются функции вида:

- $y = \alpha x^{\beta}\varepsilon$ при моделировании кривых спроса;
- $y = \alpha e^{\beta x}\varepsilon$ при моделировании временных трендов, при этом вместо x используется время t, а вместо β — постоянный темп прироста r, т.е. $y = \alpha e^{rt}\varepsilon$.

Пример 3.9. По данным примера 2.5 построим зависимость расходов на питание от доходов в виде степенной функции и экспоненциальный временной тренд.

В таблице представлены исходные данные (усл. ед.) для построения указанных уравнений с помощью пакета анализа Excel (программа «Регрессия»).

t	x	y	$\ln x$	$\ln y$
1	2	1	0,693147	0
2	6	2	1,791759	0,693147
3	10	4	2,302585	1,386294
4	14	11	2,639057	2,397895
5	18	12	2,890372	2,484907

1. Уравнение $y = \alpha x^{\beta}\varepsilon$ после логарифмирования приводится к линейному виду $\ln y = \ln\alpha + \beta \ln x + \ln\varepsilon$.

Оценив регрессию между $\ln y$ и $\ln x$, получим преобразованное выражение

$$\ln \hat{y} = -1{,}049 + 1{,}183 \ln x.$$

Выполнив обратные преобразования, получим

$$\hat{y} = e^{-1{,}049} x^{1{,}183} = 0{,}350 x^{1{,}183}.$$

2. Уравнение $y = \alpha e^{rt}\varepsilon$ после логарифмирования приводится к линейному виду $\ln y = \ln\alpha + rt + \ln\varepsilon$.

Оценив регрессию между $\ln y$ и t, получим преобразованное выражение

$$\ln \hat{y} = -0{,}61 + 0{,}667t.$$

Выполнив обратные преобразования, получим

$$\hat{y} = e^{-0{,}61} e^{0{,}667t} = 0{,}543 e^{0{,}667t}.$$

В экономическом анализе часто используется эластичность функции. **Эластичность функции** $y = f(x)$ рассчитывается как относительное изменение y к относительному изменению x, т.е.

$$Э = \left(\frac{dy}{y}\right) \Big/ \left(\frac{dx}{x}\right) = \frac{x}{y} f'(x).$$

Эластичность показывает, на сколько процентов изменяется функция $y = f(x)$ при изменении независимой переменной на 1%.

Для степенной функции $y = ax^b$ эластичность представляет собой постоянную величину, равную b.

Например, для зависимости расходов на питание от дохода $\hat{y} = 0{,}350 x^{1{,}183}$ эластичность спроса на продукты питания по доходу

составляет 1,183. Это означает, что увеличение личного дохода на 1% приведет к увеличению расходов на питание на 1,183%.

Коэффициент 0,350 не имеет экономического смысла. Он помогает прогнозировать значение y при заданных значениях x, приводя их к единому масштабу.

Для экспоненциального временного тренда $\hat{y} = 0{,}543e^{0{,}667t}$ постоянный темп роста $r = 0{,}667$. Это означает, что расходы на продукты питания в течение выборочного периода росли с темпом 66,7% в год.

Постоянный множитель 0,543 показывает, что в момент $t = 0$ общие расходы на питание составили 0,543 усл. ед.

В силу того что эластичность линейной функции $y = a + bx$ не является постоянной величиной, а зависит от x, т.е.

$$Э = b\frac{x}{y},$$

обычно вычисляется средний показатель эластичности по формуле

$$\overline{Э} = b\frac{\overline{x}}{\overline{y}},$$

где $\overline{x}, \overline{y}$ — средние значения переменных x, y в выборке.

Например, для зависимости расходов на питание от доходов $\hat{y} = -1{,}75 + 0{,}775x$ ($\overline{x} = 10, \overline{y} = 6$) средний показатель эластичности равен 1,29 и показывает, что с увеличением дохода на 1% расходы на питание возрастут в среднем на 1,29%.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Каковы предпосылки регрессионного анализа?
2. Что является мерой разброса зависимой переменной вокруг линии регрессии?
3. Какие существуют критерии проверки гипотез, относящихся к коэффициентам регрессии?
4. В чем отличие стандартной ошибки положения линии регрессии от средней ошибки прогнозирования индивидуального результативного признака при заданном значении фактора?
5. Какие существуют виды моделей, нелинейных относительно включаемых переменных и оцениваемых параметров?
6. Как определяются коэффициенты эластичности по разным видам регрессионных моделей?
7. Какова взаимозависимость различных критериев в парном регрессионном анализе?

Глава 4
Модель множественной регрессии

Обобщением линейной регрессионной модели с одной объясняющей переменной является линейная регрессионная модель с k объясняющими переменными (**модель множественной регрессии**):

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon,$$

где β_0, β_1, ... β_k — параметры модели, а ε — случайный член.

Как и в модели с парной регрессией, случайный член ε удовлетворяет **условиям Гаусса — Маркова**:

1. *Математическое ожидание случайного члена в любом наблюдении должно быть равно нулю*, т.е.

$$M(\varepsilon_i) = 0 \quad (i = \overline{1, n}).$$

2. *Дисперсия случайного члена должна быть постоянной для всех наблюдений*, т.е.

$$D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma^2 \quad (i = \overline{1, n}).$$

3. *Случайные члены должны быть статистически независимы (некоррелированы) между собой*, т.е.

$$M(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j).$$

4. *Случайные члены в любом наблюдении должны быть статистически независимы от объясняющих переменных.*

При выполнении условий Гаусса — Маркова модель называется **классической нормальной линейной регрессионной моделью**.

Предполагается, что объясняющие переменные *некоррелированы* друг с другом.

На основе n наблюдений оценивается выборочное уравнение регрессии

$$\hat{y} = b_0 + b_1 x_1 + ... + b_k x_k,$$

где b_0, b_1, ..., b_k — оценки параметров β_0, β_1, ..., β_k.

Для оценки параметров регрессии используется метод наименьших квадратов. В соответствии с МНК минимизируется сумма квадратов остатков:

$$Q = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \to \min.$$

Необходимым условием ее минимума является равенство нулю всех ее частных производных по $b_0, b_1, ..., b_k$.

В результате приходим к системе из $(k+1)$ линейного уравнения с $(k+1)$ неизвестным, называемой *системой нормальных уравнений*. Ее решение в явном виде обычно записывается в матричной форме, иначе оно становится слишком громоздким.

Оценки параметров модели и их теоретические дисперсии в матричной форме определяются выражениями

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, \quad D(b_j) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}_{jj}\sigma^2,$$

где $\boldsymbol{b}$ — вектор с компонентами $b_0, b_1, ..., b_k$; $\boldsymbol{X}$ — матрица значений объясняющих переменных; $\boldsymbol{Y}$ — вектор значений зависимой переменной; σ^2 — дисперсия случайного члена.

Несмещенной оценкой σ^2 является величина S^2 (остаточная дисперсия):

$$S^2 = \frac{1}{n-k-1}\sum e_i^2.$$

Величина S называется **стандартной ошибкой регрессии**.

Заменяя в теоретических дисперсиях неизвестную дисперсию σ^2 на ее оценку S^2 и извлекая квадратный корень, получим стандартные ошибки коэффициентов регрессии

$$S_{b_j} = S\sqrt{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}_{jj}}.$$

Если предпосылки относительно случайного члена ε выполняются, оценки параметров множественной регрессии являются *несмещенными*, *состоятельными* и *эффективными*.

При использовании компьютерных программ коэффициенты регрессии $b_0, b_1, ..., b_k$ и их стандартные отклонения вычисляются одновременно.

Пример 4.1. По данным бюджетного обследования семи случайно выбранных семей изучалась зависимость накоплений y от дохода x_1 и стоимости имущества x_2.

Исходные данные (усл. ед.):

x_1	40	55	45	30	30	60	50
x_2	60	40	40	15	90	30	30
y	2	7	5	4	2	7	6

Используя компьютерную программу, получим оцененное уравнение регрессии

$$\hat{y} = \underset{(2,06)}{0,45} + \underset{(0,036)}{0,129}x_1 - \underset{(0,017)}{0,034}x_2, \quad S = 0,97$$

(в скобках указаны стандартные ошибки).

Из этого уравнения можно сделать следующие выводы.

1. Прогнозируемые накопления семьи с доходом 40 усл. ед. и имуществом стоимостью 25 усл. ед. составляют

$$\hat{y} = 0,45 + 0,129 \cdot 40 - 0,034 \cdot 25 = 4,76.$$

2. Если доход семьи возрастет на 10 усл. ед., а стоимость имущества не изменится, то накопления возрастут на величину

$$\Delta y = 0,129\Delta x_1 = 0,129 \cdot 10 = 1,29.$$

3. Если доход семьи увеличится на 5 усл. ед., а стоимость имущества — на 15 усл. ед., то накопления возрастут на величину

$$\Delta y = 0,129\Delta x_1 - 0,034\Delta x_2 = 0,129 \cdot 5 - 0,034 \cdot 15 = 0,135.$$

Замечание. *Для прогноза значения переменной y при заданных значениях x объясняющих переменных можно воспользоваться статистической функцией Excel*: **ТЕНДЕНЦИЯ**.

4.1. АНАЛИЗ ВАРИАЦИИ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Пусть в уравнении регрессии содержится k объясняющих переменных. Допустим, что можно разложить дисперсию зависимой переменной на *объясненную* и *необъясненную* составляющие:

$$\text{var}(y) = \text{var}(\hat{y}) + \text{var}(e).$$

Используя определение выборочной дисперсии, это уравнение можно представить в виде

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2.$$

Обозначим:

$TSS = \sum (y_i - \bar{y})^2$ — общий разброс зависимой переменной;

$ESS = \sum (\hat{y}_i - \bar{y})^2$ — разброс, объясненный регрессией;

$USS = \sum e_i^2$ — разброс, не объясненный регрессией.

Тогда

$$\underset{(n-1)}{TSS} = \underset{(k)}{ESS} + \underset{(n-k-1)}{USS}$$

(в скобках указано число степеней свободы, соответствующее каждому члену уравнения).

Замечание. *Любая сумма квадратов связана с числом степеней свободы, т.е. с числом независимого варьирования переменной. Существует равенство между числами степеней свободы для этого уравнения. Отнесение каждой суммы квадратов этого уравнения на одну степень свободы приводит их к сравнимому виду.*

Коэффициент детерминации есть доля объясненной части разброса зависимой переменной, т.е.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{USS}{TSS}.$$

Величина R^2 является мерой объясняющего качества уравнения регрессии по сравнению с горизонтальной линией $\hat{y} = \bar{y}$.

Поскольку коэффициент R^2 измеряет долю дисперсии, совместно объясненной независимыми переменными, то, казалось бы, можно определить отдельный вклад каждой независимой переменной и таким образом получить меру ее относительной важности. Однако такое разложение невозможно, если независимые переменные коррелированы, поскольку в этом случае их объясняющие способности будут перекрываться.

На рис. 12, *а* и *б* отражен геометрический смысл коэффициента детерминации при использовании одной и двух объясняющих переменных соответственно.

С увеличением объясненной части разброса *ESS* коэффициент R^2 приближается к единице. Кроме того, с добавлением еще одной переменной R^2 обычно увеличивается.

Для компенсации такого увеличения R^2 вводится **скорректированный коэффициент детерминации** с поправкой на число степеней свободы:

$$\overline{R^2} = 1 - \frac{n-1}{n-k-1}\frac{USS}{TSS}.$$

Если увеличение доли объясненной регрессии при добавлении новой переменной мало, то скорректированный коэффициент детерминации может уменьшиться, следовательно, добавлять переменную нецелесообразно.

Кроме того, если объясняющие переменные x_1 и x_2 сильно коррелируют между собой, то они объясняют одну и ту же часть разброса переменной y, и в этом случае трудно оценить вклад каждой из переменных в объяснение поведения y.

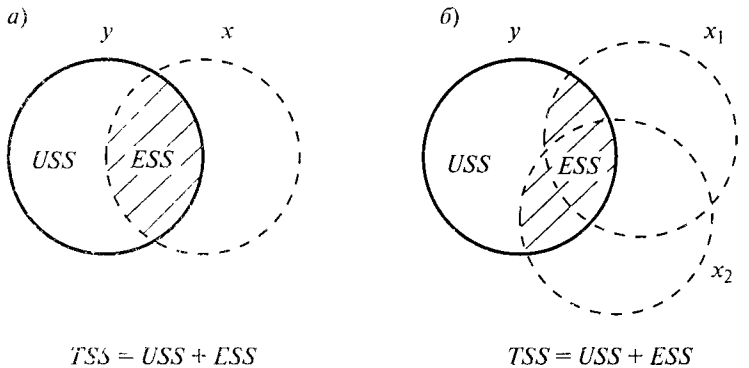


Рис. 12

4.2. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

ПРОВЕРКА ГИПОТЕЗЫ H_0: $\beta_i = 0$

Статистическая значимость коэффициентов множественной линейной регрессии с k объясняющими переменными проверяется на основе t-статистики:

$$t_j = \frac{b_j}{S_{b_j}},$$

имеющей распределение Стьюдента с $\nu = n - k - 1$ степенями свободы.

***t*-тесты** для коэффициентов множественной регрессии выполняются так же, как и в парном регрессионном анализе. Отметим,

что критический уровень t при любом уровне значимости зависит от *числа степеней свободы*, которое равно $(n - k - 1)$: *число наблюдений* минус *число оцененных параметров* (один коэффициент для каждой независимой переменной и постоянный член). Доверительные интервалы определяются точно так же, как и в парном регрессионном анализе, в соответствии с указанием относительно числа степеней свободы.

t-статистика обеспечивает эффективную проверку значимости переменной при допущении, что все другие переменные уже включены в уравнение.

Последовательный отсев несущественных факторов составляет основу многошагового регрессионного анализа. Однако по коэффициентам регрессии нельзя определить, какой из факторов оказывает наибольшее влияние на зависимую переменную, так как коэффициенты регрессии между собой несопоставимы (они измерены разными единицами).

Различия в единицах измерения факторов устраняют с помощью *частных коэффициентов эластичности*, рассчитываемых по формуле

$$Э_j = b_j \frac{\bar{x}_j}{\bar{y}},$$

где $\bar{x}_j$ — среднее значение изучаемого фактора.

Частные коэффициенты эластичности показывают, на сколько процентов в среднем изменяется зависимая переменная с изменением на 1% каждого фактора при фиксированном значении других факторов.

Упражнение 4.1. Регрессия зависимой переменной y на три независимые переменные на основе $n = 30$ наблюдений дала следующие результаты:

$\hat{y}$	$= (\ ) + 1{,}2x_1 +$	$1{,}0x_2 -$	$0{,}5x_3$	
Стандартные ошибки:	(2,1)	( )	(0,6)	( )
t-значения:	(11,9)	(2,4)	( )	(2,5)

Заполните пропуски и постройте 95%-ный доверительный интервал для значимых коэффициентов регрессии.

5-2337

ПРОВЕРКА ГИПОТЕЗЫ H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$

Предположим, что в модель множественной регрессии включен *свободный член*, тогда $TSS = ESS + USS$, где ESS — объясненная сумма квадратов отклонения с $\nu_1 = k$ степенями свободы, использованными на ее объяснение, а USS — остаточная (необъясненная) сумма квадратов с $\nu_2 = n - k - 1$ степенями свободы.

Для определения того, действительно ли объясненный разброс ESS больше случайного USS, используется ***F*-тест**.

Построим F-статистику:

$$F = \frac{ESS/k}{USS/(n-k-1)}$$

(для сопоставимости ESS и USS их значения привели на одну степень свободы).

После деления числителя и знаменателя этого выражения на TSS можно вычислить F-статистику на основе R^2:

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

Показатели F и R^2 равны или не равны нулю одновременно, поэтому принятие гипотезы H_0: $F = 0$ равнозначно статистической незначимости R^2.

Величина F имеет распределение Фишера с $\nu_1 = k$, $\nu_2 = n - k - 1$ степенями свободы.

Наблюдаемому (расчетному) значению критерия F соответствует определенная *значимость F*, которую можно вычислить в Excel с помощью функции

Значимость *F* = FРАСП(*F*; ν_1; ν_2).

Из сравнения *значимости F* с заданным *стандартным уровнем значимости* получаем:

- **если значимость *F* больше стандартного уровня, то R^2 незначим;**
- **если значимость *F* меньше стандартного уровня, то R^2 значим.**

Чаще всего F-тест используется для оценки того, значимо ли объяснение, даваемое уравнением в целом.

Проверка гипотезы H_0: $F = 0$ равнозначна проверке гипотезы H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$ об одновременном равенстве нулю всех коэффициентов линейной регрессии, за исключением свободного члена.

Замечание. *Если объясняющие способности независимых переменных перекрываются (сильная корреляция между ними), то t-тест для каждой переменной окажется незначимым, в то время как F-тест для уравнения в целом может быть значимым.*

ПРОВЕРКА ГИПОТЕЗЫ H_0: $\beta_{k+1} = \beta_{k+2} = \ldots = \beta_{k+m} = 0$

Распределение Фишера можно использовать для проверки гипотезы об одновременном равенстве нулю части коэффициентов регрессии.

Пусть сначала была оценена регрессия с k объясняющими переменными

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$$

и объясненная сумма квадратов составляет ESS_k.

Затем добавлено еще m переменных и по тем же данным оценено уравнение

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_{k+m} x_{k+m},$$

при этом объясненная сумма квадратов возрастает до ESS_{k+m}.

Таким образом, объяснили дополнительно величину $(ESS_{k+m} - ESS_k)$, использовав для этого m степеней свободы.

Требуется выяснить, превышает ли данное увеличение объясненной части то увеличение, которое может быть получено случайно (USS_{k+m}). Используя F-тест, соответствующую F-статистику можно записать в виде

$$F = \frac{(ESS_{k+m} - ESS_k)/m}{USS_{k+m}/(n-k-m-1)} = \frac{R^2_{k+m} - R^2_k}{1 - R^2_{k+m}} \frac{n-k-m-1}{m},$$

и в соответствии с нулевой гипотезой H_0: $F = 0$ она распределена с $\nu_1 = m$ и $\nu_2 = n-k-m-1$ степенями свободы.

Значимость F, соответствующая расчетному значению F, сравнивается со *стандартным уровнем значимости*. Если *значимость F меньше стандартного уровня*, то дополнительное включение в модель m переменных *оправданно*.

Гипотеза H_0: $F = 0$ равнозначна гипотезе H_0: $\beta_{k+1} = \ldots = \beta_{k+m} = 0$.

В частности, если добавить только одну переменную, то

$$F = \frac{R^2_{k+1} - R^2_k}{1 - R^2_{k+1}} \frac{n-k-2}{1} \qquad (\nu_1 = 1, \quad \nu_2 = n-k-2).$$

5*-2337

Пример 4.2. Пусть по данным примера 4.1 изучалась зависимость накоплений y от дохода x_1 и стоимости имущества x_2. При включении в модель только переменной x_1 уравнение регрессии y на x_1 имело коэффициент детерминации $R^2 = 0{,}733$. После дополнительного включения в модель переменной x_2 уравнение регрессии y на x_1 и x_2 имело $R^2 = 0{,}86$.

Определим: а) значимость в целом регрессии y на x_1; б) значимость регрессии y на x_1 и x_2; в) верна ли гипотеза H_0: $\beta_2 = 0$.

а) При $\nu_1 = 1$, $\nu_2 = 5$ определяем расчетное значение критерия

$$F = \frac{R^2}{1 - R^2}\frac{\nu_2}{\nu_1} = \frac{0{,}733}{0{,}267} \cdot 5 = 13{,}72.$$

Поскольку *значимость* $F = 0{,}0139 < 0{,}05$, то уравнение регрессии в целом значимо.

б) При $\nu_1 = 2$, $\nu_2 = 4$ определяем расчетное значение критерия

$$F = \frac{R^2}{1 - R^2}\frac{\nu_2}{\nu_1} = \frac{0{,}86}{0{,}14}\frac{4}{2} = 12{,}28.$$

Поскольку *значимость* $F = 0{,}0196 < 0{,}05$, то уравнение регрессии в целом значимо.

в) При $\nu_1 = 1$, $\nu_2 = 4$ определяем расчетное значение критерия

$$F = \frac{R_2^2 - R_1^2}{1 - R_2^2}\frac{\nu_2}{\nu_1} = \frac{0{,}86 - 0{,}733}{1 - 0{,}86} \cdot 4 = 3{,}62.$$

Поскольку *значимость* $F = 0{,}129 > 0{,}05$, то уравнение с включением переменной x_2 улучшения в объяснении дисперсии y не дало, т.е. коэффициент $\beta_2 = 0$.

ПРОВЕРКА ГИПОТЕЗЫ H_0: $\beta' = \beta''$ (ТЕСТ ЧОУ)

Пусть имеются две выборки объема n_1 и n_2. Для каждой из этих выборок оценено уравнение регрессии с k объясняющими переменными:

$$(\hat{y})' = b_0' + b_1'x_1 + \ldots + b_k'x_k$$

с необъясненной суммой квадратов USS_1 ($\nu = n_1 - k - 1$);

$$(\hat{y})'' = b_0'' + b_1''x_1 + \ldots + b_k''x_k$$

с необъясненной суммой квадратов USS_2 ($\nu = n_2 - k - 1$).

Проверяется нулевая гипотеза H_0: $\beta' = \beta''$, т.е. все соответствующие коэффициенты этих уравнений равны друг другу.

Пусть оценено уравнение регрессии того же вида сразу для всех $(n_1 + n_2)$ наблюдений с необъясненной суммой квадратов USS_0 $(\nu = n_1 + n_2 - k - 1)$.

Тогда рассматривается F-статистика:

$$F = \frac{USS_0 - (USS_1 + USS_2)}{USS_1 + USS_2} \frac{n_1 + n_2 - 2k - 2}{k + 1},$$

которая имеет распределение Фишера с $\nu_1 = k + 1$ и $\nu_2 = n_1 + n_2 - 2k - 2$ степенями свободы.

F-статистика будет близка к нулю, если $USS_0 = USS_1 + USS_2$, т.е. если уравнения регрессии для обеих выборок одинаковы.

Если *значимость F меньше стандартного значения*, то H_0 *отклоняется*, т.е. нельзя построить единое уравнение регрессии для обеих выборок.

4.3. МУЛЬТИКОЛЛИНЕАРНОСТЬ

Мультиколлинеарность — это коррелированность двух или нескольких объясняющих переменных в уравнении регрессии. При наличии мультиколлинеарности МНК-оценки формально существуют, но обладают рядом недостатков:

1) небольшое изменение исходных данных приводит к существенному изменению оценок регрессии;

2) оценки имеют большие стандартные ошибки, малую значимость, в то время как модель в целом является значимой (высокое значение R^2).

Если при оценке уравнения регрессии несколько факторов оказались незначимыми, то нужно выяснить, нет ли среди них сильно коррелированных между собой.

При наличии корреляции один из пары связанных между собой факторов исключается либо в качестве объясняющего фактора берется какая-то их функция. Если статистически незначим лишь один фактор, то он должен быть исключен либо заменен другим показателем.

Для отбора факторов в модель регрессии и оценки их мультиколлинеарности можно использовать *матрицу парных коэффициентов корреляции* (расчет корреляционной матрицы предусмотрен в стандартном программном обеспечении).

В модель регрессии включаются те факторы, которые более сильно связаны с зависимой переменной, но слабо связаны с другими факторами.

Упражнение 4.2. Пусть по данным бюджетного обследования семи случайно выбранных семей изучалась зависимость накоплений y от дохода x_1, расходов на питание x_2 и стоимости имущества x_3. Исходные данные (усл. ед.):

x_1	40	55	45	30	30	60	50
x_2	10	15	12	8	10	20	15
x_3	60	40	40	15	90	30	30
y	2	7	5	4	2	7	6

Используя компьютерную программу «Корреляция», получите следующую матрицу парных коэффициентов корреляции:

	y	x_1	x_2	x_3
y	1			
x_1	0,85	1		
x_2	0,81	0,93	1	
x_3	−0,65	−0,38	−0,28	1

Проанализируйте целесообразность включения в модель каждого фактора.

4.4. СПЕЦИФИКАЦИЯ И КЛАССИФИКАЦИЯ ПЕРЕМЕННЫХ В УРАВНЕНИЯХ РЕГРЕССИИ

Построение экономической модели включает выбор объясняющих переменных. Свойства оценок коэффициентов регрессии в значительной мере зависят от правильной спецификации модели. Рассмотрим два случая:

- отсутствие в модели переменной, которая должна быть включена;
- наличие в модели переменной, которая не должна быть включена.

1. Влияние отсутствия в модели переменной, которая должна быть включена. Предположим, что переменная y зависит от двух переменных x_1 и x_2 в соответствии с соотношением $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Однако считается, что модель выглядит как $y = \alpha + \beta_1 x_1 + \varepsilon$, и оценивается регрессия $\hat{y} = a + b_1 x_1$.

В этом случае оценка b_1 и ее дисперсия являются *смещенными*. Смещенность оценки b_1 связана с тем, что если не учесть x_2 в регрессии, то переменная x_1 будет играть двойную роль: отражать

свое прямое влияние и заменять переменную x_2 в описании ее влияния.

Коэффициент R^2 для данной регрессии отражает общую объясняющую способность переменной x_1 в обеих ролях и является завышенной оценкой.

2. Влияние включения в модель переменной, которая не должна быть включена. Допустим, что истинная модель представляется в виде $y = \alpha + \beta_1 x_1 + \varepsilon$. Однако считается, что ею является $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, и оценивается регрессия $\hat{y} = a + b_1 x_1 + b_2 x_2$.

Оценки коэффициентов регрессии и их дисперсии в этом случае являются *несмещенными*, но *неэффективными*. Практически обнаруживается, что коэффициент b_2 статистически незначим, и переменная x_2 исключается из модели.

ЗАМЕЩАЮЩИЕ ПЕРЕМЕННЫЕ

Предположим, что истинной моделью является

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

и допустим, что не имеется данных по существенной переменной x_1.

Если не включить в модель эту переменную, то регрессия может пострадать от смещения оценок и статистическая проверка будет некорректной.

Если вместо отсутствующей переменной x_1 использовать ее заменитель z, линейно связанный с x_1, и построить регрессию

$$\hat{y} = a + b_2 x_2 + \ldots + b_k x_k + cz,$$

то оценки b_2, ..., b_k, их стандартные ошибки и коэффициент R^2 будут такими же, как и с использованием x_1. Единственным недостатком является то, что отсутствует оценка коэффициента при самой величине x_1, а величина a не является оценкой α.

В качестве замещающей переменной, например, для показателя технического прогресса может использоваться время.

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

При исследовании влияния качественных признаков в модель можно вводить фиктивные переменные, принимающие, как правило, два значения: *единица*, если данный признак присутствует в наблюдении, и *нуль* при его отсутствии.

Если включаемый в рассмотрение качественный признак имеет не два, а несколько значений, то используют несколько фиктивных переменных, число которых должно быть на единицу меньше числа значений признака.

При назначении фиктивных переменных исследуемая совокупность по числу значений качественного признака разбивается на группы. Одну из групп выбирают как эталонную (группа 0) и определяют фиктивные переменные для остальных.

Например, если качественный признак имеет три значения, то две фиктивные переменные определяются следующим образом:

группа 0: $z_1 = z_2 = 0$,
группа 1: $z_1 = 1, \quad z_2 = 0$,
группа 2: $z_1 = 0, \quad z_2 = 1$,

или

$$z_1 = \begin{cases} 1 & \text{(группа 1)}, \\ 0 & \text{(остальные)}, \end{cases} \qquad z_2 = \begin{cases} 1 & \text{(группа 2)}, \\ 0 & \text{(остальные)}. \end{cases}$$

Введение в регрессию фиктивных переменных существенно улучшает качество ее оценивания.

Пример 4.3. Имеются данные о весе новорожденного y в граммах и количестве сигарет x, выкуриваемых в день будущей матерью во время беременности в случаях первых и непервых родов:

№ п/п	Первенец	y	x	z	№ п/п	Первенец	y	x	z
1	Нет	3450	8	1	11	Нет	3200	31	1
2	Нет	3300	21	1	12	Нет	3400	13	1
3	Нет	3400	18	1	13	Да	3450	5	0
4	Нет	3300	24	1	14	Да	3400	10	0
5	Нет	3450	6	1	15	Да	3200	19	0
6	Нет	3450	16	1	16	Да	3350	12	0
7	Нет	3100	19	1	17	Да	3000	20	0
8	Нет	3500	7	1	18	Да	3300	8	0
9	Нет	3400	20	1	19	Да	3300	16	0
10	Нет	3500	10	1	20	Да	3400	9	0

Оценив регрессию между y и x, получим выражение

$$\hat{y} = \underset{(56,8)}{3519} - \underset{(3,5)}{12,11x}, \quad R^2 = 0,394$$

(в скобках указаны стандартные ошибки).

Это означает, что ребенок, рожденный некурящей матерью, будет иметь при рождении средний вес около 3500 г, а уменьшение веса новорожденного по причине курения его матери составляет около 12 г на каждую сигарету, выкуриваемую в день.

Для учета качественного фактора (первый или не первый ребенок) введем в модель фиктивную переменную:

$$z = \begin{cases} 0 & \text{(первенец)}, \\ 1 & \text{(непервенец)}. \end{cases}$$

Оценив регрессию между y и x, z, получим выражение

$$\hat{y} = \underset{(49,2)}{3480} - \underset{(3,0)}{14,56}\,x + \underset{(42,1)}{124}\,z, \quad R^2 = 0,600$$

(в скобках указаны стандартные ошибки).

Коэффициент 124 при фиктивной переменной z статистически значим.

Это выражение можно переписать в виде двух уравнений:

$\hat{y} = 3480 - 14,56x$ (для первенца);

$\hat{y} = 3604 - 14,56x$ (для непервенца).

Параметр сдвига (эффект от фактора «первенец — непервенец») составляет $3604 - 3480 = 124$ г.

Как видим, добавление в регрессию фиктивной переменной существенно улучшило качество оценки.

ЛАГОВЫЕ ПЕРЕМЕННЫЕ

При использовании данных временного ряда на текущие значения зависимой переменной могут влиять не только текущие значения объясняющих переменных, но также их значения с некоторым запаздыванием.

В общем, если какая-то переменная появляется в модели с запаздыванием на τ периодов, она записывается с нижним индексом $(t - \tau)$, например:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-\tau} + \varepsilon_t.$$

Сдвиг τ, характеризующий запаздывание в воздействии фактора на результат, называется **лагом**. Переменная, влияние которой характеризуется некоторым запаздыванием, называется **лаговой**.

4.5. СТОХАСТИЧЕСКИЕ ОБЪЯСНЯЮЩИЕ ПЕРЕМЕННЫЕ И ОШИБКИ ИЗМЕРЕНИЯ

Обычно в предпосылках регрессионного анализа считается, что объясняющие переменные являются *неслучайными*. Пусть некоторые объясняющие переменные являются *стохастическими* (*случайными*).

Рассмотрим три случая.

1. Объясняющие переменные распределены независимо от случайного члена. В этом случае оценки, полученные обычным МНК, сохраняют все свои свойства: *несмещенность*, *эффективность* и *состоятельность*.

2. Объясняющие переменные и случайный член являются зависимыми, но одномоментно некоррелированы. Например, имеется модель с лаговой зависимой переменной в качестве одной из объясняющих переменных:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t.$$

В этом случае y_{t-1} находится непосредственно под воздействием ε_{t-1} и косвенно под воздействием всех предшествующих значений случайного члена. Следовательно, объясняющая переменная y_{t-1} и случайный член ε_t зависимы, и МНК не дает несмещенных оценок.

Тем не менее, если y_{t-1} и ε_t некоррелированы, МНК-оценки будут *состоятельными*, хотя и *смещенными*.

3. Объясняющие переменные одномоментно коррелированы со случайным членом. В этом случае оценки, полученные обычным МНК, являются *смещенными* и *несостоятельными*.

Рассмотрим влияние ошибок измерения.

1. Ошибки в измерениях зависимой переменной. Предположим, что истинной является модель $y = \alpha + \beta x + \varepsilon$, но измеряемое значение зависимой переменной есть $y^* = y + \nu$, где ν — ошибки, имеющие нулевое математическое ожидание и не зависящие от ε и x. Следовательно, зависимость между наблюдаемым значением зависимой переменной и x представляется выражением $y^* = \alpha + \beta x + (\varepsilon + \nu)$.

Оценки параметров (α, β) будут *несмещенными* и *состоятельными*, так как $M(\varepsilon + \nu) = 0$ и $\mathrm{Cov}(x, (\varepsilon + \nu)) = 0$.

Наличие же ошибки ν приводит лишь к увеличению дисперсии регрессии: $D(\varepsilon + \nu) = \sigma_\varepsilon^2 + \sigma_\nu^2$.

2. Ошибки в измерениях объясняющих переменных. Пусть истинной моделью является $y = \alpha + \beta x + \varepsilon$, но измеряемое значение объясняющей переменной есть $x^* = x + \nu$, где ν — ошибки, имеющие нулевое математическое ожидание и не зависящие от ε и x. Реально будет осуществляться регрессия: $y = \alpha + \beta x^* + \varepsilon^*$, где $\varepsilon^* = \varepsilon - \beta\nu$, в которой объясняющая переменная x^* и ошибка ε^* уже являются коррелированными, так как обе зависят от ν. Это означает, что МНК-оценки будут *смещенными* и *несостоятельными*.

Таким образом, при наличии ошибок в измерениях *зависимой* переменной МНК-оценки остаются *несмещенными* и *состоятельными*, а наличие ошибок в измерениях *независимых* переменных приводит к возникновению корреляции между объясняющими переменными и ошибками и, как следствие, к *смещенности* и *несостоятельности* МНК-оценок.

4.6. МЕТОД ИНСТРУМЕНТАЛЬНЫХ ПЕРЕМЕННЫХ

При наличии *корреляции* между объясняющими переменными и случайным членом МНК-оценки являются *смещенными* и *несостоятельными*. Для получения состоятельных оценок можно воспользоваться **методом инструментальных переменных** (ИП).

Суть метода ИП заключается в *замене непригодной объясняющей переменной такой переменной, которая некоррелирована со случайным членом и коррелирована с исходной переменной.*

Пусть в модели $y = \alpha + \beta x + \varepsilon$ переменная x коррелирована со случайным членом. Предположим, что можно найти другую переменную z, которая коррелирована с x, но некоррелирована с ε.

Основанная на использовании ИП оценка параметра β, определяемая как

$$b_{\text{ИП}} = \frac{\operatorname{cov}(z, y)}{\operatorname{cov}(z, x)},$$

является состоятельной.

Действительно, из соотношений

$$\operatorname{cov}(z, y) = \operatorname{cov}(z, \alpha + \beta x + \varepsilon) = \beta \operatorname{cov}(z, x) + \operatorname{cov}(z, \varepsilon)$$

следует, что

$$b_{\text{ИП}} = \beta + \frac{\operatorname{cov}(z, \varepsilon)}{\operatorname{cov}(z, x)},$$

т.е. оценка методом ИП равна истинному значению β плюс ошибка, равная $\mathrm{cov}(z, \varepsilon)/\mathrm{cov}(z, x)$.

В больших выборках ошибка исчезает при условии, что переменная z действительно распределена независимо от ε и $\mathrm{cov}(z, x) \neq 0$. Следовательно, на больших выборках $b_{ИП}$ будет стремиться к истинному значению β.

Таким образом, оценка $b_{ИП}$ является *состоятельной*, но в общем случае *смещенной* и *неэффективной*.

4.7. ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ КОББА — ДУГЛАСА

Макроэкономическая производственная функция — это статистически значимая связь между объемом выпуска Y, капитальными затратами K и затратами труда L.

Для моделирования и решения задач как на макро-, так и на микроэкономическом уровне часто используют **производственную функцию Кобба — Дугласа** (КД):

$$Y = AK^{\alpha}L^{\beta}\varepsilon, \tag{4.1}$$

где A, α, β — параметры функции, причем $A > 0$, $0 < \alpha < 1$, $0 < \beta < 1$.

Свойства производственной функции Кобба — Дугласа:

1) *эластичность выпуска продукции*. Эластичность выпуска продукции по капиталу и труду равна соответственно α и β:

$$\frac{\partial Y}{Y}\Big/\frac{\partial K}{K} = \alpha, \quad \frac{\partial Y}{Y}\Big/\frac{\partial L}{L} = \beta.$$

Это означает, что увеличение затрат капитала на 1% приведет к росту выпуска продукции на α%, а увеличение затрат труда на 1% — к росту выпуска на β%;

2) *эффект от масштаба производства*. При росте затрат каждого из факторов K, L в λ раз выпуск возрастает в $\lambda^{\alpha+\beta}$ раз. Это означает следующее:

- если $\alpha + \beta > 1$, то функция КД имеет *возрастающую* отдачу от масштаба производства;
- если $\alpha + \beta < 1$, то функция КД имеет *убывающую* отдачу от масштаба производства;
- если $\alpha + \beta = 1$, то функция КД имеет *постоянную* отдачу от масштаба производства;

3) *прогнозируемые доли производственных факторов*. В рыночной экономике оценки α и β интерпретируются как прогнозируемые доли дохода, полученные соответственно за счет капитала и труда.

Для оценки параметров производственной функции КД с помощью модели множественной линейной регрессии необходимо прологарифмировать уравнение (4.1):

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L + \ln \varepsilon.$$

При этом обычно предполагается, что ошибки $\ln\varepsilon$ обладают свойствами, необходимыми для оценивания линейной регрессионной модели.

По рядам данных Y, K, L рассчитываются ряды их логарифмов, и для них оценивается уравнение регрессии.

Пример 4.4. Известны данные об объеме производства Y, капитальных затратах K и затратах труда L некоторой страны за 12 лет (усл. ед.):

T	Y	K	L	T	Y	K	L
1	100	100	100	7	153	216	145
2	112	114	110	8	184	236	154
3	124	131	123	9	189	266	154
4	143	149	125	10	227	335	196
5	151	176	138	11	218	397	193
6	155	198	140	12	179	417	147

Используя компьютерную программу «Регрессия», получим следующее оцененное уравнение регрессии:

$$\ln \hat{Y} = -0{,}302 + 0{,}15 \ln K + 0{,}92 \ln L,$$

т.е. оценки $\alpha = 0{,}15$, $\beta = 0{,}92$. Это означает, что увеличение затрат капитала на 1% приведет к росту выпуска продукции на 0,15%, а увеличение затрат труда на 1% — к росту выпуска на 0,92%.

При построении производственной функции КД с использованием данных временного ряда следует иметь в виду, что на выпуск продукции оказывает также влияние технический прогресс.

Влияние технического прогресса можно учесть, включив экспоненциальный временной тренд, т.е. записав функцию КД, например, в виде $Y = AK^{\alpha}L^{\beta}e^{rt}\varepsilon$, где t — время, а r — темп прироста выпуска благодаря техническому прогрессу.

Оценив это соотношение по данным рассматриваемого примера, получим $\ln\hat{Y} = 2{,}7 - 0{,}61\ln K + 1{,}01\ln L + 0{,}095t$, где оценки $\alpha = -0{,}61$ и $r = 0{,}095$ неправдоподобны и статистически незначимы. Это связано с мультиколлинеарностью, так как коэффициент корреляции между $\ln K$ и t составляет 0,997.

4.8. ПОНЯТИЕ О ВРЕМЕННЫХ РЯДАХ

Временной ряд — это совокупность значений какого-либо показателя за несколько последовательных моментов времени.

Каждый уровень $y(t)$ временного ряда формируется под совместным влиянием длительных, кратковременных и случайных факторов.

Длительные, постоянно действующие факторы оказывают на изучаемое явление определяющее влияние и формируют основную тенденцию ряда — тренд $T(t)$. *Кратковременные*, периодические факторы формируют сезонные колебания ряда $S(t)$. *Случайные* факторы отражаются случайными изменениями уровней ряда $\varepsilon(t)$.

Модель, в которой временной ряд представлен как сумма перечисленных компонент, т.е. $y(t) = T(t) + S(t) + \varepsilon(t)$, называется **аддитивной**.

Модель, в которой временной ряд представлен как произведение перечисленных компонент, т.е. $y(t) = T(t)S(t)\varepsilon(t)$, называется **мультипликативной**.

Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний.

Если амплитуда сезонных колебаний приближенно *постоянная*, используют *аддитивную модель*. Если амплитуда *возрастает* или *уменьшается*, используют *мультипликативную модель*.

Основная задача эконометрического исследования временного ряда — выявить каждую из перечисленных компонент ряда.

ВЫЯВЛЕНИЕ ОСНОВНОЙ ТЕНДЕНЦИИ РАЗВИТИЯ

Основной тенденцией развития (трендом) называется плавное и устойчивое изменение уровня явления во времени, свободное от случайных колебаний.

Задача выявления основной тенденции развития в статистике называется **выравниванием временного ряда**.

Методами выявления тренда являются:

- метод укрупнения интервалов;
- метод скользящей средней;
- аналитическое выравнивание.

Метод укрупнения интервалов основан на укрупнении периодов времени, к которым относятся уровни ряда. При суммировании уровней по укрупненным интервалам колебания в уровнях, обусловленные случайными причинами, взаимопогашаются и более четко обнаруживается общая тенденция.

Метод скользящей средней заключается в том, что рассматривается средний уровень из определенного числа первых по счету уровней ряда, затем — из такого же числа уровней, но начиная со второго по счету и т.д. Таким образом, средняя как бы «скользит» по ряду динамики, продвигаясь на один срок.

Пример 4.5. Рассчитаем скользящую среднюю по данным об урожайности зерновых культур (ц/га) за 10 лет.

Исходные данные и расчетные показатели представим в следующей таблице:

Год	Фактический уровень	Скользящая средняя		Центрированная скользящая средняя
		трехлетняя	четырехлетняя	
1996	15	–	–	–
1997	13	14,33	14,75	–
1998	15	14,67	15,50	15,125
1999	16	16,33	16,50	16
2000	18	17,00	16,75	16,625
2001	17	17,00	17,50	17,125
2002	16	17,33	17,25	17,375
2003	19	17,33	18	17,625
2004	17	18,67	–	–
2005	20	–	–	–

Период скольжения может быть четным и нечетным. Для *нечетного* периода (трехлетнего) первое значение скользящей средней есть (15 + 13 + 15)/3 = 14,33, второе — (13 + 15 + 16)/3 = = 14,67 и т.д., причем полученные результаты скользящей средней отнесены к середине периода скольжения.

Для *четного* периода (четырехлетнего) первое значение скользящей средней есть (15 + 13 + 15 + 16)/4 = 14,75, второе — (13 + 15 + 16 + 18)/4 = 15,50 и т.д. Однако рассчитанные усредненные значения нельзя сопоставить каким-либо определенным значениям t, поэтому применяют *процедуру центрирования* (вычисляют среднее значение из двух последовательных скользящих средних). Первое значение центрированной скользящей средней есть (14,75 + 15,50)/2 = 15,125, второе — (15,50 + 16,50)/2 = 16 и т.д., причем первая центрированная средняя будет отнесена к третьему году, т.е. к 1998-му.

Замечание. *Используя пакет анализа Excel (программа «Скользящее среднее»), можно также получить результаты сглаживания.*

Метод аналитического выравнивания: фактические уровни ряда заменяются плавно изменяющимися уровнями, полученными из уравнения регрессии $y_t = f(t) + \varepsilon$. При аналитическом выравнивании используются различные виды трендовых моделей.

Упражнение 4.3. Имеются данные о розничном товарообороте региона (усл. ед.) за 10 лет:

Год	1	2	3	4	5	6	7	8	9	10
Товарооборот	11	13	22	18,5	20	19	25	23	24,5	35

Постройте следующие трендовые модели товарооборота и выберите из них наиболее подходящую:

Вид уравнения	Уравнение	R^2
Линейное	$y = 1{,}94t + 10{,}43$	0,767
Полином второго порядка	$y = 0{,}074t^2 + 1{,}127t + 12{,}06$	0,774
Полином третьего порядка	$y = 0{,}121t^3 - 1{,}921t^2 + 10{,}33t + 1{,}68$	0,886
Логарифмическое	$y = 7{,}70 \ln t + 9{,}46$	0,709
Степенное	$y = 10{,}88t^{0{,}407}$	0,815
Экспоненциальное	$y = 11{,}84\,e^{0{,}096t}$	0,781

Замечание. *Для нахождения наиболее адекватного уравнения тренда в Excel используется инструмент «Подбор линии тренда» из Мастера диаграмм.*

АНАЛИЗ АДДИТИВНОЙ МОДЕЛИ

Общий вид аддитивной модели таков:

$Y = T + S + \varepsilon$.

Построение модели включает в себя следующие шаги:

1) выравнивание исходного ряда методом скользящей средней;

2) расчет значений сезонной компоненты S;

3) устранение сезонной компоненты из исходных уровней ряда ($Y - S$) и получение выравненных данных ($T + \varepsilon$);

4) аналитическое выравнивание уровней ($T + \varepsilon$) и расчет значений T с использованием полученного уравнения тренда;

5) расчет полученных по модели значений ($T + S$);

6) расчет абсолютных ошибок.

Пример 4.6. Имеются поквартальные данные об объеме потребления электроэнергии y в некотором районе за четыре года (усл. ед.):

Год / Квартал	1	2	3	4
I	6,0	7,2	8,0	9,0
II	4,4	4,8	5,6	6,6
III	5,0	6,0	6,4	7,0
IV	9,0	10,0	11,0	10,8

В качестве *зависимой* переменной при анализе временного ряда выступают фактические уровни ряда y_t, а в качестве *независимой* переменной — время (сквозной номер квартала) $t = 1, 2, ..., 16$.

По графику ряда можно установить наличие приблизительно линейного тренда и сезонных колебаний (период равен 4) *одинаковой* амплитуды, поэтому используется *аддитивная модель.* Определим ее компоненты, сезонную и трендовую.

Для исключения влияния сезонной компоненты проведем выравнивание исходного ряда методом скользящей средней за четыре квартала и процедуру центрирования. Результаты расчетов представлены в таблице:

Сквозной номер квартала	Потребление электро-энергии y_t	Скользящая средняя за четыре квартала	Центриро-ванная скользящая средняя	Оценка сезонной вариации
1	6,0	–	–	–
2	4,4	6,10	–	–
3	5,0	6,40	6,250	−1,250
4	9,0	6,50	6,450	2,550
5	7,2	6,75	6,625	0,575
6	4,8	7,00	6,875	−2,075
7	6,0	7,20	7,100	−1,100
8	10,0	7,40	7,300	2,700
9	8,0	7,50	7,450	0,550
10	5,6	7,75	7,625	−2,025
11	6,4	8,00	7,875	−1,475
12	11,0	8,25	8,125	2,875
13	9,0	8,40	8,325	0,675
14	6,6	8,35	8,375	−1,775
15	7,0	–	–	–
16	10,8	–	–	–

На рис. 13 представлены графики фактических уровней ряда и центрированной скользящей средней (сглаженные уровни).

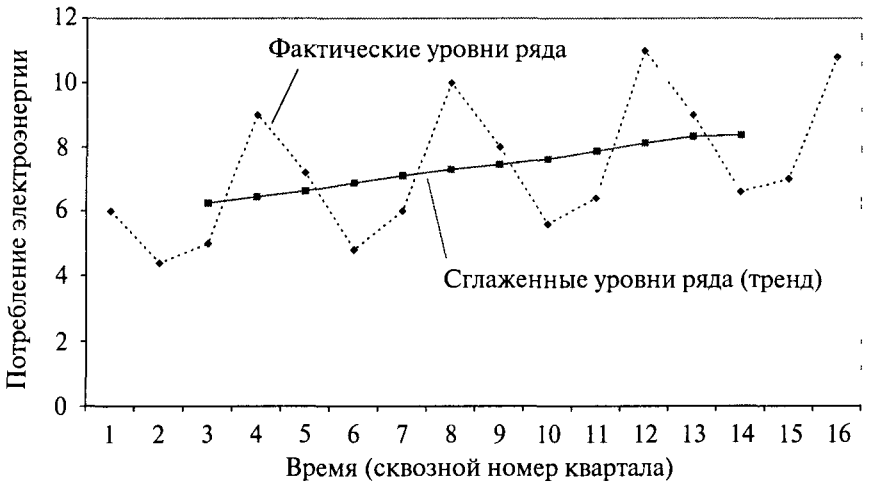


Рис. 13

Оценки сезонной вариации определяются как разность между фактическими уровнями ряда y_t и центрированными скользящими средними.

Расчет сезонной компоненты выполним в следующей расчетной таблице, в которой оценки сезонной вариации записываются под соответствующим номером квартала в году:

Показатели	Год	Номер квартала в году				
		I	II	III	IV	
	1	-	-	-1,250	2,550	
	2	0,575	-2,075	-1,100	2,700	
	3	0,550	-2,025	-1,475	2,875	
	4	0,675	-1,775	-	-	
Итого		1,800	-5,875	-3,825	8,125	Сумма
Среднее		0,600	-1,958	-1,275	2,708	0,075
Скорректированное S_1		0,581	-1,977	-1,294	2,690	0

В строке **Среднее** рассчитаны средние сезонной вариации по годам за каждый квартал и их сумма, равная 0,075.

В аддитивной модели предполагается, что сумма всех сезонных компонент по всем кварталам должна быть равна *нулю* (*условие взаимопогашаемости сезонных воздействий*).

В строке **Скорректированное S_1** рассчитаны значения сезонных компонент S_i как разность между средней сезонной вариацией и корректирующим коэффициентом 0,075/4, при этом $\sum S_i = 0$.

Расчет трендовой компоненты и ошибок выполним в следующей таблице:

t	Y	S	$Y-S=T+\varepsilon$	T	Ошибка e
1	6,0	0,581	5,419	5,893	–0,474
2	4,4	–1,977	6,337	6,088	0,256
3	5,0	–1,294	6,294	6,268	0,025
4	9,0	2,690	6,310	6,455	–0,145
5	7,2	0,581	6,619	6,642	–0,023
6	4,8	–1,977	6,777	6,829	–0,052
7	6,0	–1,294	7,294	7,016	0,277
8	10,0	2,690	7,310	7,204	0,106
9	8,0	0,581	7,419	7,391	0,027
10	5,6	–1,977	7,577	7,578	–0,001
11	6,4	–1,294	7,694	7,765	–0,071

Окончание таблицы

t	Y	S	$Y-S=T+\varepsilon$	T	Ошибка e
12	11,0	2,690	8,310	7,952	0,357
13	9,0	0,581	8,419	8,139	0,279
14	6,6	−1,977	8,577	8,326	0,250
15	7,0	−1,294	8,294	8,514	−0,220
16	10,8	2,690	8,110	8,701	−0.591

В столбце «$Y-S=T+\varepsilon$» исключается влияние сезонной компоненты: вычитая ее значение из каждого уровня исходного ряда, получим только тенденцию и случайную компоненту.

Проводя аналитическое выравнивание ряда ($T+\varepsilon$) с помощью линейного тренда, получим следующее уравнение линии тренда:

$$T = 5{,}706 + 0{,}187t.$$

Уровни ряда T для каждого $t = 1, 2, ..., 16$ указаны в вышеприведенной таблице.

Расчет ошибки в аддитивной модели осуществляется по формуле

$$e = Y - (T + S).$$

Дисперсии фактического ряда и ошибки, рассчитанные в Excel с помощью функции **ДИСПР**, составляют: var(y) = 4,196; var(e) = 0,0684.

Для оценки качества построенной модели по аналогии с моделью регрессии можно использовать выражение

$$1 - \frac{\text{var}(e)}{\text{var}(y)} = 1 - \frac{0{,}0684}{4{,}196} = 0{,}984,$$

т.е. аддитивная модель объясняет 98,4% общей вариации уровней исходного временного ряда.

ПРИМЕНЕНИЕ ФИКТИВНЫХ ПЕРЕМЕННЫХ ПРИ МОДЕЛИРОВАНИИ ВРЕМЕННЫХ РЯДОВ

По данным примера 4.6 рассмотрим построение аддитивной модели исходного временного ряда с использованием фактора времени и фиктивных переменных.

Произвольно возьмем I квартал каждого года в качестве эталонного и будем использовать фиктивные переменные для оценки разности между ним и другими кварталами.

Запишем модель как

$$y = \alpha + \beta t + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \varepsilon,$$

где z_1, z_2, z_3 — **фиктивные переменные**, определяемые следующим образом:

$$z_1 = \begin{cases} 1 \text{ (II квартал)}, \\ 0 \text{ (остальные)}, \end{cases} \quad z_2 = \begin{cases} 1 \text{ (III квартал)}, \\ 0 \text{ (остальные)}, \end{cases} \quad z_3 = \begin{cases} 1 \text{ (IV квартал)}, \\ 0 \text{ (остальные)}. \end{cases}$$

Коэффициенты δ_1, δ_2, δ_3 дают численную величину эффекта, вызываемого сменой года.

Исходные данные (усл. ед.) для расчета параметров уравнения с фиктивными переменными:

y	t	z_1	z_2	z_3	y	t	z_1	z_2	z_3
6,0	1	0	0	0	8,0	9	0	0	0
4,4	2	1	0	0	5,6	10	1	0	0
5,0	3	0	1	0	6,4	11	0	1	0
9,0	4	0	0	1	11,0	12	0	0	1
7,2	5	0	0	0	9,0	13	0	0	0
4,8	6	1	0	0	6,6	14	1	0	0
6,0	7	0	1	0	7,0	15	0	1	0
10,0	8	0	0	1	10,8	16	0	0	1

Используя компьютерную программу «Регрессия», получим следующее оцененное уравнение регрессии:

$$\hat{y} = 6{,}237 + 0{,}187t - 2{,}387z_1 - 1{,}825z_2 + 2{,}087z_3, \quad R^2 = 0{,}985.$$

Отдельные уравнения для каждого квартала таковы:

$\hat{y} = 6{,}237 + 0{,}187t$ (I квартал);

$\hat{y} = 3{,}850 + 0{,}187t$ (II квартал);

$\hat{y} = 4{,}412 + 0{,}187t$ (III квартал);

$\hat{y} = 8{,}324 + 0{,}187t$ (IV квартал).

Усредняя эти уравнения, получим линейный тренд

$$T = \frac{1}{4}(6{,}237 + 3{,}850 + 4{,}412 + 8{,}324) + 0{,}187t = 5{,}706 + 0{,}187t.$$

Расстояние между линией регрессии каждого квартала и трендом дает оценку сезонной компоненты в данном квартале:

$S_1 = 6{,}237 - 5{,}706 = 0{,}531$ (I квартал);

$S_2 = 3{,}850 - 5{,}706 = -1{,}856$ (II квартал);

$S_3 = 4{,}412 - 5{,}706 = -1{,}294$ (III квартал);

$S_4 = 8{,}324 - 5{,}706 = 2{,}618$ (IV квартал),

причем сумма сезонных отклонений должна равняться нулю, т.е.

$S_1 + S_2 + S_3 + S_4 = 0$.

АНАЛИЗ МУЛЬТИПЛИКАТИВНОЙ МОДЕЛИ

Общий вид мультипликативной модели таков:

$Y = TS\varepsilon$.

Построение модели включает в себя следующие шаги:

1) выравнивание исходного ряда методом скользящей средней;

2) расчет значений сезонной компоненты;

3) устранение сезонной компоненты из исходных уровней ряда (Y/S) и получение выравненных данных ($T\varepsilon$);

4) аналитическое выравнивание уровней ($T\varepsilon$) и расчет значений T с использованием полученного уравнения тренда;

5) расчет полученных по модели значений (TS);

6) расчет ошибок.

Пример 4.7. Имеются поквартальные данные о выплате доходов компании акционерам в форме дивидендов за последние четыре года (усл. ед.):

Год / Квартал	1	2	3	4
I	40	60	50	30
II	50	80	70	50
III	60	100	80	60
IV	70	110	130	70

По графику ряда можно установить наличие приблизительно линейного тренда и сезонных колебаний (период равен 4) *возрастающей* амплитуды, поэтому используется *мультипликативная модель*. Определим ее компоненты, сезонную и трендовую.

Для исключения влияния сезонной компоненты проведем выравнивание исходного ряда методом скользящей средней за четыре квартала и процедуру центрирования. Результаты расчетов представлены в таблице:

Сквозной номер квартала	Размер дивидендов y_t	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной вариации
1	40	–	–	–
2	60	45	–	–
3	50	47,5	46,25	1,081
4	30	52,5	50	0,600
5	50	57,5	55	0,909
6	80	62,5	60	1,333
7	70	65	63,75	1,098
8	50	70	67,5	0,741
9	60	72,5	71,25	0,842
10	100	75	73,75	1,356
11	80	77,5	76,25	1,049
12	60	80	78,75	0,762
13	70	92,5	86,25	0,811
14	110	95	93,75	1,173
15	130	–	–	–
16	70	–	–	–

Оценки сезонной вариации для мультипликативной модели определяются как частное от деления фактических уровней ряда y_t на центрированные скользящие средние.

Расчет сезонной компоненты выполним в следующей расчетной таблице, в которой оценки сезонной вариации записываются под соответствующим номером квартала в году:

Показатели	Год	Номер квартала в году				
		I	II	III	IV	
	1	–	–	1,081	0,600	
	2	0,909	1,333	1,098	0,741	
	3	0,842	1,356	1,049	0,762	
	4	0,811	1,173	–	–	
Итого		2,562	3,862	3,228	2,103	Сумма
Среднее		0,854	1,287	1,076	0,701	3,918
Скорректированное S_i		0,872	1,314	1,099	0,715	4

В строке **Среднее** рассчитаны средние сезонной вариации по годам за каждый квартал и их сумма, равная 3,918.

В мультипликативной модели предполагается, что сумма всех сезонных компонент по всем кварталам должна быть равна четырем — числу сезонов в году (*условие взаимопогашаемости сезонных воздействий*).

В строке **Скорректированное** S_1 рассчитаны значения сезонных компонент S_i как произведение соответствующей средней сезонной вариации на корректирующий коэффициент $4/3{,}918 = 1{,}021$, при этом $\sum S_i = 4$.

Расчет трендовой компоненты и ошибок выполним в следующей таблице:

t	Y	S	$Y/S = Te$	T	$e = Y/(TS)$	$e = Y - TS$
1	40	0,872	45,871	38,97	1,18	6,02
2	60	1,314	45,662	43,046	1,06	3.44
3	50	1,099	45,496	47,122	0,96	–1,79
4	30	0,715	41,958	51,197	0,82	–6,60
5	50	0,872	57,335	55,273	1,04	1,80
6	80	1,314	60,883	59,349	1,02	2,01
7	70	1,099	63,694	63,425	1,00	0,29
8	50	0,715	69,930	67,501	1,03	1,73
9	60	0,872	68,807	71,577	0,96	–2,41
10	100	1,314	76,103	75,652	1,00	0,59
11	80	1,099	72,793	79,728	0,91	–7,62
12	60	0,715	83,916	83,804	1,00	0,08
13	70	0,872	80,275	87,880	0,91	–6,63
14	110	1,314	83,713	91,956	0,91	–10,83
15	130	1,099	118,289	96,032	1,23	24,46
16	70	0,715	97,902	100,108	0,98	–1,58

Разделив каждый уровень исходного ряда на соответствующие значения сезонной компоненты (столбец «$Y/S = Te$»), исключим влияние сезонной компоненты и в результате получим только тенденцию и случайную компоненту.

Проводя аналитическое выравнивание ряда (Te) с помощью линейного тренда, получим следующее уравнение линии тренда:

$$T = 34{,}89 + 4{,}087t.$$

Уровни ряда T для каждого $t = 1, 2, ..., 16$ указаны в вышеприведенной таблице.

Графики исходного ряда и его тренда приведены на рис. 14.

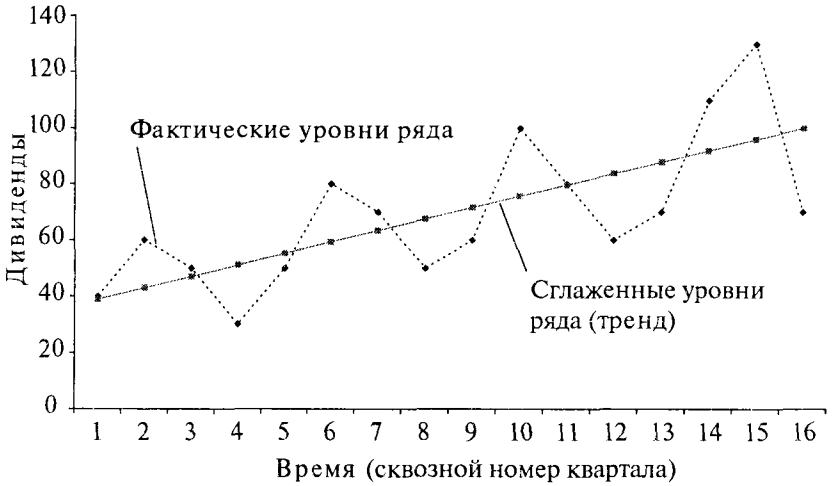


Рис. 14

Расчет ошибки в мультипликативной модели осуществляется по формуле

$$e = Y/(TS).$$

Чтобы сравнивать мультипликативную модель с другими моделями временного ряда, ошибки в мультипликативной модели определяются как

$$e = Y - TS.$$

Дисперсии фактического ряда и ошибки, рассчитанные в Excel с помощью функции **ДИСПР**, составляют: var(y) = 643,36; var(e) = 58,18.

Для оценки качества построенной модели можно по аналогии с аддитивной моделью использовать выражение

$$1 - \frac{\text{var}(e)}{\text{var}(y)} = 1 - \frac{58{,}18}{643{,}36} = 0{,}91,$$

т.е. мультипликативная модель объясняет 91% общей вариации уровней исходного временного ряда.

АВТОКОРРЕЛЯЦИЯ УРОВНЕЙ ВРЕМЕННОГО РЯДА

Корреляционная зависимость между последовательными уровнями временного ряда называется **автокорреляцией** уровней ряда.

Коэффициент автокорреляции порядка τ определяется как коэффициент корреляции между рядами y_t и $y_{t-\tau}$:

$$r_\tau = \frac{\operatorname{cov}(y_t, y_{t-\tau})}{\sqrt{\operatorname{var}(y_t)\operatorname{var}(y_{t-\tau})}}.$$

Число периодов τ, по которым рассчитывается коэффициент автокорреляции, называется **лагом**.

С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается.

Пример 4.8. Определим коэффициенты автокорреляции до четвертого порядка по данным примера 4.6 об объеме потребления электроэнергии.

Последовательно вводя данные (y_t, y_{t-1}), (y_t, y_{t-2}), (y_t, y_{t-3}), (y_t, y_{t-4}) и используя программу «Корреляция», получим следующие коэффициенты автокорреляции:

$$r_1 = 0{,}165, \quad r_2 = -0{,}566, \quad r_3 = 0{,}113, \quad r_4 = 0{,}983.$$

Последовательность коэффициентов автокорреляции первого, второго и более высоких порядков называется **автокорреляционной функцией** временного ряда. Автокорреляционную функцию обычно используют для выявления во временном ряде наличия или отсутствия трендовой и сезонной компонент.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, то исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка $\tau > 1$, то ряд содержит также сезонные колебания с периодом τ.

Анализ значений автокорреляционной функции рассматриваемого примера позволяет сделать вывод о наличии во временном ряде линейной тенденции и сезонных колебаний с периодичностью четыре квартала.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем состоит спецификация модели множественной регрессии?
2. Каковы основные предпосылки применения МНК для построения регрессионной модели?
3. Какие требования предъявляются к факторам для включения их в модель множественной регрессии?
4. К каким трудностям приводит мультиколлинеарность факторов, включенных в модель, и как они могут быть устранены?
5. Каковы методы устранения мультиколлинеарности факторов?
6. Какие коэффициенты используются для оценки сравнительной силы воздействия факторов на результат?
7. Как оценить качество модели в целом?
8. В каких случаях рассчитывается скорректированный коэффициент детерминации?
9. При каких условиях строится уравнение множественной регрессии с фиктивными переменными?
10. В чем состоит специфика построения моделей регрессии по временным рядам данных?
11. Каковы основные методы выявления тренда?
12. Какова интерпретация параметра при факторе времени в модели регрессии с включением фактора времени?
13. Как охарактеризовать понятие автокорреляции в остатках?

Глава 5
Гетероскедастичность и автокоррелированность случайного члена

5.1. ОБНАРУЖЕНИЕ ГЕТЕРОСКЕДАСТИЧНОСТИ

Одной из предпосылок регрессионного анализа является предположение о *постоянстве* дисперсии случайного члена для всех наблюдений (*гомоскедастичность*). Это значит, что для каждого значения объясняющей переменной случайные члены имеют одинаковые дисперсии. Если это условие не соблюдается, то имеет место *гетероскедастичность*.

При отсутствии гетероскедастичности коэффициенты регрессии имеют *наименьшую* дисперсию среди всех несмещенных оценок, являющихся линейными функциями от наблюдений y.

Если наблюдается *гетероскедастичность*, то МНК-оценки будут *неэффективными* (они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра).

Оценки стандартных ошибок коэффициентов регрессии вычисляются в предположении, что распределение случайного члена гомоскедастично; если это не так, то оценки неверны (занижены), а, следовательно, t-статистика завышена. Это может привести к статистически значимым коэффициентам регрессии, тогда как в действительности это неверно.

Проблема гетероскедастичности характерна для пространственных данных, полученных от неоднородных объектов. Например, если исследуется зависимость прибыли предприятий от размера основного фонда, то можно ожидать, что для больших предприятий размах колебаний прибыли будет больше, чем для малых.

Предложено большое количество тестов для обнаружения гетероскедастичности, в которых делаются различные предположения о зависимости между дисперсией случайного члена и величиной объясняющей переменной (или объясняющих переменных), на-

пример тест ранговой корреляции Спирмена, тест Голдфельда — Квандта и тест Глейзера.

ТЕСТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Выдвигается нулевая гипотеза об отсутствии гетероскедастичности случайного члена. При выполнении теста ранговой корреляции Спирмена предполагается, что дисперсия случайного члена будет либо увеличиваться, либо уменьшаться по мере увеличения x, и поэтому в регрессии, оцениваемой с помощью МНК, абсолютные величины остатков $|e|$ и значения x будут коррелированы.

Данные по x и остатки $|e|$ ранжируются по переменной x, и определяются их ранги.

Ранг — это порядковый номер значений переменной в ранжированном ряду.

Коэффициент ранговой корреляции Спирмена определяется по формуле

$$r = 1 - \frac{6\sum D_i^2}{n(n^2-1)},$$

где D_i — разность между рангами x и $|e|$.

Если предположить, что коэффициент корреляции для генеральной совокупности равен нулю, то коэффициент ранговой корреляции имеет нормальное распределение с нулевым математическим ожиданием и дисперсией $1/(n-1)$ в больших выборках. Соответствующая тестовая статистика $r\sqrt{n-1}$ сравнивается с критическим значением $t_{кр}$ при заданном уровне значимости ($t_{кр} = 1{,}96$ при $\alpha = 5\%$, $t_{кр} = 2{,}58$ при $\alpha = 1\%$).

Если $r\sqrt{n-1} > t_{кр}$, то нулевая гипотеза об отсутствии гетероскедастичности будет отклонена.

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

Пример 5.1. Оценим регрессионную зависимость выпуска продукции обрабатывающей промышленности на душу населения y от валового внутреннего продукта на душу населения x в том же году для 17 стран. Исходные данные (усл. ед.):

n	y	x	n	y	x
1	18	3	10	100	24
2	27	6	11	63	25
3	18	7	12	130	26
4	45	9	13	135	27
5	55	13	14	60	28
6	68	15	15	70	35
7	51	18	16	80	37
8	84	21	17	180	44
9	85	22			

Пусть модель описывается выражением $y = \alpha + \beta x + \varepsilon$. По исходным данным с помощью МНК получена следующая регрессионная зависимость:

$$\hat{y} = \underset{(14,5)}{12,84} + \underset{(0,6)}{2,92}x, \quad R^2 = 0,608 \tag{5.1}$$

(в скобках указаны стандартные ошибки).

На рис. 15 представлен график остатков e_i.

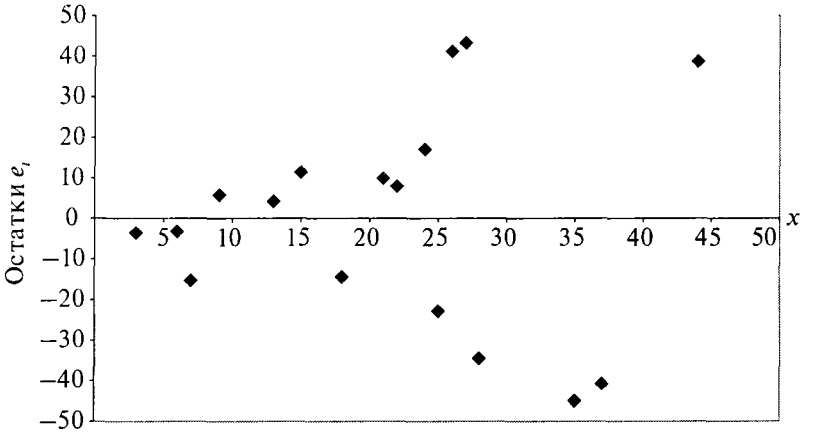


Рис. 15

Из рисунка видно, что с увеличением переменной x размах колебаний остатков e_i тоже возрастает, поэтому есть предположение о зависимости ошибки регрессии от независимой переменной (гетероскедастичность).

Для установления гетероскедастичности применим тест Спирмена. Выдвигается нулевая гипотеза об отсутствии гетероскедастичности.

Отклонения от линии регрессии (остатки e) и данные по x в порядке возрастания приведены в следующей таблице:

x	Ранг	$\lvert e_i \rvert$	Ранг	D_i	D_i^2	x	Ранг	$\lvert e_i \rvert$	Ранг	D_i	D_i^2
3	1	3,6	2	−1	1	24	10	17,1	10	0	0
6	2	3,3	1	1	1	25	11	22,8	11	0	0
7	3	15,2	9	−6	36	26	12	41,2	15	−3	9
9	4	5,9	4	0	0	27	13	43,3	16	−3	9
13	5	4,2	3	2	4	28	14	34,5	12	2	4
15	6	11,4	7	−1	1	35	15	45,0	17	−2	4
18	7	14,4	8	−1	1	37	16	40,8	14	2	4
21	8	9,8	6	2	4	44	17	38,7	13	4	16
22	9	7,9	5	4	16	*Итого*					110

Замечание. *При расчете ранга переменной можно воспользоваться статистической функцией* **РАНГ** *пакета Excel.*

На основе этих данных вычислен коэффициент ранговой корреляции:

$$r = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 110}{17 \cdot 288} = 0{,}866.$$

Тестовая статистика составляет $r\sqrt{n-1} = 0{,}866\sqrt{17-1} = 3{,}5$. Это больше, чем $t_{кр}$, и, следовательно, нулевая гипотеза об отсутствии гетероскедастичности отклоняется.

ТЕСТ ГОЛДФЕЛЬДА — КВАНДТА

При проведении проверки по этому тесту предполагается, что стандартное отклонение σ случайного члена пропорционально значению независимой переменной x.

Тест включает следующие шаги:

1. Все n наблюдений в выборке упорядочиваются по возрастанию переменной x.
2. Оцениваются отдельные регрессии для первых n_0 и для последних n_0 наблюдений. Средние $(n - 2n_0)$ наблюдений отбрасываются.

3. Составляется статистика: $F = RSS_2 / RSS_1$, где RSS_1, RSS_2 — суммы квадратов остатков для первых и последних n_0 наблюдений соответственно. Если верна гипотеза H_0 об отсутствии гетероскедастичности, то F имеет распределение Фишера с $\nu_1 = n_0 - k - 1$, $\nu_2 = n_0 - k - 1$ степенями свободы, где k — число объясняющих переменных.
   По таблице определяется критическое значение критерия $F_{кр}$. Если $F > F_{кр}$, то нулевая гипотеза об отсутствии гетероскедастичности отклоняется.

Замечание. *Тест Голдфельда — Квандта можно также использовать для проверки на гетероскедастичность при предположении, что σ_i обратно пропорционально x_i. В этом случае тестовой статистикой является величина $F = RSS_1 / RSS_2$.*

Пример 5.2. На основе данных примера 5.1 с помощью обычного МНК оценим регрессии для шести стран с наименьшими значениями показателя x и для шести стран с наибольшими значениями этого показателя.

Получены соответственно уравнения:

$$\hat{y}_1 = -0{,}18 + 4{,}38x;$$

$$\hat{y}_2 = 39{,}9 + 2{,}11x.$$

Суммы квадратов отклонений составляют $RSS_1 = 229$, $RSS_2 = 9804$, при этом $F = 9804/229 = 42{,}8$. Критическое значение $F_{кр} = 6{,}39$ при 5%-ном уровне значимости. Поскольку $F = 42{,}8 > F_{кр} = 6{,}39$, то нулевая гипотеза об отсутствии гетероскедастичности отклоняется.

ТЕСТ ГЛЕЙЗЕРА

Тест Глейзера основывается на более общих представлениях о зависимости стандартной ошибки случайного члена от значений объясняющей переменной. Например, зависимость может быть представлена в виде

$$\sigma_i = \alpha + \beta x_i^{\gamma} + \varepsilon_i. \qquad (5.2)$$

Используя абсолютные значения остатков в качестве оценки σ_i, оценивают данную регрессионную зависимость при различных значениях γ и выбирают наилучшую из них.

Таким образом, гетероскедастичность аппроксимируется уравнением

$$s_i = a + bx_i^{\gamma},$$

где $s_i = |e_i|$ — оценка σ_i.

Нулевая гипотеза об отсутствии гетероскедастичности отклоняется, если оценка b значимо отличается от нуля.

Пример 5.3. На основе данных $|e_i|$ и x примера 5.1 с использованием различных значений γ были оценены уравнения (5.2):

$\gamma = 1/2$, $s = -20{,}2 + 9{,}4\sqrt{x}$, $R^2 = 0{,}655$, $t_b = 5{,}34$;

$\gamma = 1$, $s = -3{,}15 + 1{,}146x$, $R^2 = 0{,}698$, $t_b = 5{,}9$;

$\gamma = 2$, $s = 7{,}32 + 0{,}024x^2$, $R^2 = 0{,}654$, $t_b = 5{,}3$;

$\gamma = 3$, $s = 12{,}12 + 0{,}0005x^3$, $R^2 = 0{,}538$, $t_b = 4{,}18$.

Наилучший результат (по R^2) соответствует значению $\gamma = 1$, при этом оценкой σ является величина

$$\underset{(4,7)}{s = -3{,}15} + \underset{(0{,}19)}{1{,}146x}. \qquad (5.3)$$

Коэффициент $b = 1{,}146$ значимо отличается от нуля, следовательно, имеет место гетероскедастичность.

5.2. МЕТОД ВЗВЕШЕННЫХ НАИМЕНЬШИХ КВАДРАТОВ

При наличии гетероскедастичности используют **обобщенный (взвешенный) МНК**.

Суть метода заключается в *уменьшении вклада данных наблюдений, имеющих бóльшую дисперсию в результате расчета.*

Пусть в исходной модели регрессии

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

случайные члены гетероскедастичны, т.е. $M(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma_i^2$ $(i = \overline{1, n})$.

Допустим, что дисперсии σ_i^2 в каждом наблюдении известны. Разделив каждое наблюдение на соответствующее ему значение σ_i, получим преобразованную модель

$$\frac{y_i}{\sigma_i} = \frac{\alpha}{\sigma_i} + \beta\frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i},$$

для которой выполнены условия гомоскедастичности, так как

$$M\left(\frac{\varepsilon_i}{\sigma_i}\right)=0,\quad D\left(\frac{\varepsilon_i}{\sigma_i}\right)=\frac{D(\varepsilon_i)}{\sigma_i^2}=1\quad (i=\overline{1,n}).$$

Оценка параметров преобразованного уравнения обычным МНК приводит к оценке параметров исходного уравнения взвешенным МНК. Для каждого из этих методов необходимо минимизировать сумму квадратов отклонений вида

$$Q=\sum\left(\frac{y_i}{\sigma_i}-\frac{a}{\sigma_i}-b\frac{x_i}{\sigma_i}\right)^2=\sum\frac{1}{\sigma_i^2}(y_i-a-bx_i)^2.$$

При минимизации этой суммы квадратов отдельные ее слагаемые взвешиваются: наблюдениям с большей дисперсией придается меньший вес $1/\sigma_i^2$. Оценки МНК коэффициентов преобразованной модели дают непосредственно оценки исходной модели.

На практике дисперсии σ_i^2 неизвестны, поэтому их заменяют оценками s_i^2, т.е. оценивается обычным МНК преобразованная модель

$$\frac{y}{s}=\alpha\frac{1}{s}+\beta\frac{x}{s}+\frac{\varepsilon}{s},$$

в которой отсутствует постоянный член.

Для экономических данных σ_i часто пропорциональны значениям объясняющей переменной x. При этом оценивается обычным МНК преобразованная модель

$$\frac{y}{x}=\alpha\frac{1}{x}+\beta+\frac{\varepsilon}{x}.$$

Коэффициент при $1/x$ будет *эффективной оценкой* α, а постоянный член — *эффективной оценкой* β.

Замечание. *Коэффициент детерминации не может служить удовлетворительной мерой качества подгонки при использовании взвешенного МНК.*

При применении взвешенного МНК оценки параметров будут *несмещенными*, кроме того, они имеют меньшую дисперсию, чем невзвешенные оценки.

Пример 5.4. По данным примера 5.1 обычным МНК получена регрессионная зависимость (5.1) и с помощью различных тестов

установлена гетероскедастичность случайного члена. Проведем коррекцию на гетероскедастичность путем использования взвешенного МНК.

Регрессионные зависимости, полученные взвешенным МНК в двух вариантах, следующие:

а) $\frac{\hat{y}}{s} = \underset{(1,12)}{8,64}\frac{1}{s} + \underset{(0,34)}{3,11}\frac{x}{s}, \quad \alpha = 8,64, \quad \beta = 3,11;$

б) $\frac{\hat{y}}{x} = \underset{(3,47)}{7,78}\frac{1}{x} + \underset{(0,37)}{3,19}, \quad \alpha = 7,78, \quad \beta = 3,19.$

Здесь вариант «а» — с делением на s, где s определяется выражением (5.3), а вариант «б» — с делением на x.

Можно показать, что колебания остатков полученных регрессионных зависимостей имеют неупорядоченный характер, и рассмотренные тесты не обнаруживают гетероскедастичности, т.е. коррекция на гетероскедастичность прошла успешно.

Полученные зависимости дают более эффективные оценки коэффициентов регрессии, чем уравнение (5.1) с использованием обычного МНК.

5.3. ОБНАРУЖЕНИЕ АВТОКОРРЕЛЯЦИИ

ОБНАРУЖЕНИЕ АВТОКОРРЕЛЯЦИИ ПЕРВОГО ПОРЯДКА

Одной из предпосылок регрессионного анализа является *независимость* случайного члена в любом наблюдении от его значений во всех других наблюдениях, т.е. $M(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)$.

Если данное условие не выполняется, то говорят, что случайный член подвержен автокорреляции. В этом случае коэффициенты регрессии, получаемые по МНК, оказываются *неэффективными*, хотя и *несмещенными*, а их стандартные ошибки рассчитываются некорректно (занижаются).

Причиной автокорреляции может быть либо неверная спецификация модели, либо наличие неучтенных факторов. Устранение этих причин не всегда дает нужные результаты. Автокорреляция имеет собственные внутренние причины, связанные с автокорреляционной зависимостью.

Автокорреляция обычно встречается в регрессионном анализе при использовании данных временного ряда. В силу этого в даль-

нейшем вместо символа i (порядковый номер наблюдения) будем использовать символ t (момент наблюдения).

Необходимым условием независимости случайных членов является их некоррелированность для каждых двух соседних значений.

Пусть ρ — коэффициент корреляции между двумя соседними случайными членами ε_t и ε_{t-1}:

- если $\rho > 0$, то автокорреляция *положительная*;
- если $\rho < 0$, то автокорреляция *отрицательная*;
- если $\rho = 0$, то автокорреляция *отсутствует* и третье условие Гаусса — Маркова удовлетворяется.

Поскольку значения случайных членов ε_t неизвестны, то проверяется статистическая некоррелированность остатков e_t и e_{t-1} с использованием обычного МНК.

Соответствующей оценкой коэффициента корреляции ρ является *коэффициент автокорреляции остатков первого порядка*, который при достаточно большом числе наблюдений имеет вид

$$r = \frac{\operatorname{cov}(e_t, e_{t-1})}{\sqrt{\operatorname{var}(e_t)\operatorname{var}(e_{t-1})}} \approx \frac{\sum e_t e_{t-1}}{\sum e_t^2}.$$

Считается, что $\bar{e}_t = \bar{e}_{t-1} = 0$, $\sum e_t^2 = \sum e_{t-1}^2$.

Выдвигается нулевая гипотеза об отсутствии корреляции первого порядка, т.е. H_0: $\rho = 0$. В качестве альтернативной гипотезы может выступать либо H_1: $\rho > 0$, либо H_2: $\rho < 0$.

Для проверки нулевой гипотезы используют **статистику Дарбина — Уотсона**, рассчитываемую по формуле

$$DW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} \approx 2(1 - r), \quad 0 \le DW \le 4.$$

Если автокорреляция остатков *отсутствует* ($r = 0$), то $DW \approx 2$.

При *положительной* автокорреляции ($r > 0$) имеем $0 \le DW < 2$, а при *отрицательной* ($r < 0$) — соответственно, $2 < DW \le 4$.

По таблице определяются критические значения критерия Дарбина — Уотсона d_1 и d_2 для заданного числа наблюдений, числа объясняющих переменных и уровня значимости. По этим значениям отрезок [0; 4] разбивается на пять зон (рис. 16). В зависимости от того, в какую зону попадает расчетное значение критерия, принимают или отвергают соответствующую гипотезу.

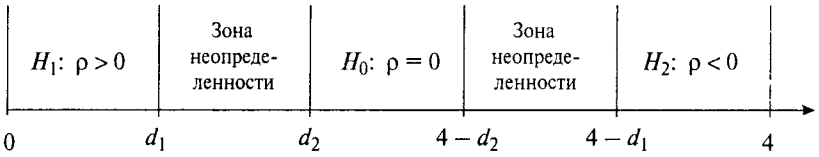


Рис. 16

Наличие зоны неопределенности связано с тем, что распределение *DW*-статистики зависит не только от числа наблюдений и числа объясняющих переменных, но и от значений объясняющих переменных.

Пример 5.5. Имеются данные об объеме предложения товара y, его цены x_1 и зарплаты сотрудников x_2 за 10 месяцев. Выявим на уровне значимости 0,05 наличие автокорреляции остатков в модели регрессии

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Исходные данные и результаты промежуточных расчетов (усл. ед.) представлены в следующей таблице:

t	x_1	x_2	y	e_t	e_{t-1}
1	10	12	20	8,30	–
2	15	10	35	4,26	8,30
3	20	9	30	–12,46	4,26
4	25	9	45	–1,86	–12,46
5	40	8	60	–7,38	–1,86
6	37	8	70	5,26	–7,38
7	43	6	75	–9,66	5,26
8	35	4	90	–2,26	–9,66
9	40	4	105	8,34	–2,26
10	55	5	110	7,46	8,34

Выборочная регрессия для этой модели:

$$\hat{y} = 90{,}72 + 0{,}88x_1 - 7{,}32x_2.$$

Коэффициент автокорреляции остатков первого порядка $r = -0{,}02512$, следовательно, значение критерия Дарбина — Уотсона для этой модели составляет $DW = 2{,}05$. По таблице распределения Дарбина — Уотсона (см. Приложение) находим $d_1 = 0{,}70$ и

$d_2 = 1{,}64$. Поскольку $d_2 < DW < 4 - d_2$, то нет оснований отклонять гипотезу H_0 об отсутствии автокорреляции в остатках.

Замечание. *Тест Дарбина — Уотсона построен в предположении, что объясняющие переменные некоррелированы со случайным членом. Поэтому этот тест неприменим к моделям, включающим в качестве объясняющих переменных лаговые значения зависимой переменной y.*

ОБНАРУЖЕНИЕ АВТОКОРРЕЛЯЦИИ В МОДЕЛИ С ЛАГОВОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

В случае когда уравнение регрессии включает лаговую зависимую переменную, например y_{t-1}, можно использовать ***h*-статистику Дарбина**, которая также вычисляется на основе остатков:

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{n}{1 - n\,\mathrm{var}(b)}},$$

где DW — значение статистики Дарбина — Уотсона, n — число наблюдений в выборке, $\mathrm{var}(b)$ — оцененная дисперсия коэффициента при лаговой зависимой переменной.

Значение h можно вычислить на основе обычных результатов оценивания регрессии. Этот тест предназначен только для проверки на наличие автокорреляции первого порядка.

При больших выборках h распределена как $N(0; 1)$ по нулевой гипотезе об отсутствии автокорреляции. Следовательно, при применении двустороннего критерия и большой выборке гипотеза об отсутствии автокорреляции может быть отклонена:

- если $|h| > 1{,}96$ при уровне значимости 5%;
- если $|h| > 2{,}58$ при уровне значимости 1%.

Тест Дарбина неприменим, если $n\,\mathrm{var}(b) \geq 1$.

5.4. АВТОРЕГРЕССИОННОЕ ПРЕОБРАЗОВАНИЕ

Пусть исходное уравнение регрессии

$$y_t = \alpha + \beta x_t + \varepsilon_t \qquad (5.4)$$

содержит автокорреляцию случайных членов.

Допустим, что автокорреляция подчиняется *авторегрессионной схеме первого порядка*:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

где ρ — коэффициент авторегрессии, а u_t — случайный член, удовлетворяющий предпосылкам МНК.

Данная схема оказывается *авторегрессионной*, поскольку ε определяется значениями этой же величины с запаздыванием, и *схемой первого порядка*, потому что в этом случае запаздывание равно единице.

Величина ρ есть коэффициент корреляции между двумя соседними ошибками. Пусть ρ известно.

Обратимся к исходной модели (5.4). Для момента времени $t-1$ эта модель примет вид

$$y_{t-1} = \alpha + \beta x_{t-1} + \varepsilon_{t-1}. \qquad (5.5)$$

Вычтем из обеих частей исходного уравнения (5.4) умноженное на ρ соотношение (5.5):

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}).$$

Обозначим:

$$\begin{cases} y_t' = y_t - \rho y_{t-1}, \\ x_t' = x_t - \rho x_{t-1}, \\ \alpha' = (1-\rho)\alpha. \end{cases}$$

Это преобразование переменных называется **авторегрессионным** (*AR*), или **преобразованием Бокса — Дженкинса**.

Тогда преобразованное уравнение

$$y_t' = \alpha' + \beta x_t' + u_t,$$

где $t \geq 2$, не содержит автокорреляцию, и для оценки его параметров (α', β) используется обычный МНК.

Оценка коэффициента β из этой зависимости непосредственно используется и для исходного уравнения, а коэффициент α рассчитывается по формуле $\alpha = \alpha'/(1-\rho)$.

На практике величина ρ неизвестна, ее оценка получается одновременно с оценками (α, β) в результате следующих итеративных процедур.

Процедура Кохрейна — Оркатта. Процедура включает следующие этапы:

1) применяя МНК к исходному уравнению регрессии, получают первоначальные оценки параметров α, β;

2) вычисляют остатки e и в качестве оценки ρ используют коэффициент автокорреляции остатков первого порядка, т.е. полагают $\rho = r_1$;
3) применяя МНК к преобразованному уравнению, получают новые оценки параметров α, β.

Процесс обычно заканчивается, когда очередное приближение ρ мало отличается от предыдущего. Процедура Кохрейна — Оркатта реализована в большинстве эконометрических компьютерных программ.

Процедура Хильдрата — Лу. Эта процедура, также широко применяемая в регрессионных пакетах, основана на тех же самых принципах, но использует другой алгоритм вычислений:
1) преобразованное уравнение оценивают для каждого значения ρ из интервала (–1, 1) с заданным шагом внутри его;
2) выбирают то значение ρ, для которого сумма квадратов остатков в преобразованном уравнении минимальна, а коэффициенты регрессии определяют при оценивании преобразованного уравнения с использованием этого значения.

Пример 5.6. Пусть изучается зависимость среднедушевых расходов на конечное потребление y от среднедушевого дохода x по данным некоторой страны за 16 лет.

Исходные (y_t, x_t) и расчетные (e_t, y'_t, x'_t) данные (усл. ед.) представлены в следующей таблице:

t	y_t	x_t	e_t	y'_t	x'_t
1	70	73	0,18	–	–
2	73	76	0,76	37,51	38,99
3	78	83	0,12	40,99	44,47
4	83	89	0,28	43,45	46,92
5	86	95	–1,55	43,92	49,88
6	89	100	–2,58	45,40	51,83
7	96	107	–1,22	50,88	56,30
8	96	108	–2,03	47,33	53,75
9	103	113	0,94	54,33	58,24
10	109	119	2,10	56,78	61,71
11	112	121	3,49	56,74	60,66
12	114	122	4,69	57,22	60,65
13	115	131	–1,56	57,20	69,14
14	118	135	–1,79	59,70	68,58
15	122	139	–1,01	62,17	70,55
16	123	140	–0,82	61,15	69.53

Пусть исходная модель имеет вид $y_t = \alpha + \beta x_t + \varepsilon_t$.

По исходным данным с использованием МНК получено следующее оцененное уравнение регрессии:

$$\hat{y}_t = \underset{(2,8)}{11,0} + \underset{(0,025)}{0,80} x_t, \quad R^2 = 0,986$$

(в скобках указаны стандартные ошибки).

Коэффициент автокорреляции остатков первого порядка составляет $r = 0,507$, следовательно, $DW = 2(1 - r) = 0,986$. При уровне значимости 5% табличное значение $d_1 = 1,10$ и $d_2 = 1,37$. Поскольку $0 < DW < d_1$, то имеется положительная автокорреляция остатков.

Применяя МНК к преобразованным данным

$$y'_t = y_t - 0,507 y_{t-1},$$

$$x'_t = x_t - 0,507 x_{t-1} \quad (t \geq 2),$$

получим оценку преобразованного уравнения

$$(\hat{y}_t)' = \underset{(3,0)}{6,2} + \underset{(0,05)}{0,79} x'_t, \quad R^2 = 0,95$$

(в скобках указаны стандартные ошибки).

Коэффициент автокорреляции остатков первого порядка составляет $r = 0,145$, следовательно, $DW = 2(1 - r) = 1,71$. Поскольку $d_2 < DW < 4 - d_2$, то автокорреляция остатков отсутствует.

Пересчитывая оценку $\alpha = 6,2/(1 - 0,507) = 12,62$, получим следующую оценку исходной модели:

$$\hat{y}_t = 12,62 + 0,79 x_t, \quad R^2 = 0,993.$$

Это уравнение отличается от полученного ранее уравнения, оцененного обычным МНК.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Как можно проверить наличие гомо- или гетероскедастичности остатков?
2. Какова суть взвешенного МНК?
3. Что такое критерий Дарбина — Уотсона? Каков алгоритм его применения для тестирования модели регрессии на автокорреляцию в остатках?
4. Какие преобразования переменных используются при наличии автокорреляции в остатках?

Глава 6
Динамические эконометрические модели

Во многих экономических задачах встречаются лагированные (взятые в предыдущий момент времени) переменные. Например, y_t — выпуск предприятия за год t — может зависеть не только от инвестиций I_t в этот год, но и от инвестиций в предыдущие годы.

Эконометрическая модель, содержащая в качестве факторов не только текущие переменные, но и лаговые их значения, называется **динамической.**

Выделим два основных типа динамических эконометрических моделей:

1) модели с распределенным лагом;
2) модели авторегрессии.

Моделями с распределенным лагом называются модели, содержащие в качестве факторов лаговые значения *факторных* переменных, например модель вида

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t.$$

Моделями авторегрессии называются модели, содержащие в качестве факторов лаговые значения *зависимой* переменной, например модель вида

$$y_t = \alpha + \beta_0 x_t + \beta_1 y_{t-1} + \varepsilon_t.$$

Обе модели включают в себя лаговые значения переменных, но существенно различаются с точки зрения статистического оценивания параметров.

6.1. МОДЕЛИ С РАСПРЕДЕЛЕННЫМ ЛАГОМ

Модель с распределенным лагом в предположении, что максимальная величина лага конечна, имеет вид

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \varepsilon_t.$$

В этой модели влияние x на y сохраняется в течение времени p.

В краткосрочном (текущем) периоде влияние x на y отражается величиной β_0, называемой **краткосрочным мультипликатором**. Он

характеризует среднее абсолютное изменение y_t при изменении x_t на единицу в некоторый фиксированный момент t без учета воздействия лаговых значений фактора x.

В долгосрочном периоде (через p моментов времени) суммарное влияние x на y отражается величиной $\beta = \beta_0 + \beta_1 + ... + \beta_p$, называемой **долгосрочным мультипликатором**. Он характеризует общее изменение результата y в долгосрочном периоде $(t + p)$ под влиянием изменения на единицу фактора x.

В моделях с распределенным лагом объясняющие переменные *некоррелированы* со случайным членом, поэтому модель можно оценивать с помощью обычного МНК. Однако на практике оценка параметров модели затруднительна из-за высокой мультиколлинеарности факторов.

Для уменьшения числа объясняющих переменных и уменьшения эффекта мультиколлинеарности разработан ряд подходов, например *модель геометрических лагов* и *модель полиномиальных лагов*.

МОДЕЛЬ ГЕОМЕТРИЧЕСКИХ ЛАГОВ (МОДЕЛЬ КОЙКА)

Предположим, что в модели с бесконечным лагом коэффициенты при лаговых значениях объясняющих переменных убывают в геометрической прогрессии. Модель имеет вид

$$y_t = \alpha + \beta_0 x_t + \beta_0 \delta x_{t-1} + \beta_0 \delta^2 x_{t-2} + \beta_0 \delta^3 x_{t-3} + ... + \varepsilon_t,$$

где $\delta \in (0; 1)$.

В этой модели влияние x на y продолжается бесконечно.

В краткосрочном (текущем) периоде влияние x на y отражается коэффициентом β_0.

В долгосрочном периоде суммарное влияние x на y равно

$$\sum_{k=0}^{\infty} \beta_0 \delta^k = \frac{\beta_0}{1 - \delta}.$$

Модель содержит только три параметра (α, β_0, δ) и является нелинейной.

Процедура оценивания нелинейной модели такова:

1) перебирается с некоторым шагом значение δ из интервала (0; 1);
2) для каждого δ рассчитывается $z_t = x_t + \delta x_{t-1} + \delta^2 x_{t-2} + \delta^3 x_{t-3} + ... + \delta^p x_{t-p}$ с таким значением p, при котором дальнейшие лаговые значения x не оказывают существенного воздействия на z;
3) оценивается уравнение регрессии $y_t = \alpha + \beta_0 z_t + \varepsilon_t$;

4) выбирается такое значение δ, которое обеспечивает наибольший коэффициент детерминации R^2 при оценке уравнения. Выбранному δ соответствуют вычисленные значения α, β_0 этого уравнения.

Использование этого метода при оценке параметров позволяет избежать проблему мультиколлинеарности объясняющих переменных.

МОДЕЛЬ ПОЛИНОМИАЛЬНЫХ ЛАГОВ (МЕТОД АЛМОНА)

В модели полиномиальных лагов предполагается, что зависимость коэффициентов при лаговых значениях объясняющей переменной от величины лага описывается полиномом m-й степени. Модель имеет вид

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \varepsilon_t,$$

где

$$\beta_S = \gamma_0 + \gamma_1 S + \gamma_2 S^2 + \ldots + \gamma_m S^m, \quad m \le p.$$

Предположим, что величина лага p известна. Кроме того, необходимо установить степень полинома m. Обычно на практике ограничиваются рассмотрением полиномов второй и третьей степени.

Пусть, например, $p = 3$, $m = 2$, тогда исходная модель есть

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \varepsilon_t,$$

где

$$\beta_0 = \gamma_0,$$

$$\beta_1 = \gamma_0 + \gamma_1 + \gamma_2,$$

$$\beta_2 = \gamma_0 + 2\gamma_1 + 4\gamma_2,$$

$$\beta_3 = \gamma_0 + 3\gamma_1 + 9\gamma_2.$$

Преобразованная модель имеет вид

$$y_t = \alpha + \gamma_0 z_0 + \gamma_1 z_1 + \gamma_2 z_2,$$

где

$$z_0 = x_t + x_{t-1} + x_{t-2} + x_{t-3},$$

$$z_1 = x_{t-1} + 2x_{t-2} + 3x_{t-3},$$

$$z_2 = x_{t-1} + 4x_{t-2} + 9x_{t-3}.$$

Используя МНК, оцениваем параметры преобразованной модели и затем рассчитываем параметры исходной модели с распределенным лагом.

Пример 6.1. Имеются данные об объеме валового внутреннего продукта y некоторой страны в зависимости от инвестиций x в ее экономику за 25 лет. Построим модель с распределенным лагом для $p = 3$ в предположении, что структура лага описывается полиномом второй степени.

Общий вид исходной модели:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \varepsilon_t.$$

Исходные (y_t, x_t) и преобразованные (z_0, z_1, z_2) данные (усл. ед.) представлены в следующей таблице:

t	y_t	x_t	z_0	z_1	z_2
1	193	30	–	–	–
2	197	29	–	–	–
3	202	29	–	–	–
4	213	32	120	177	415
5	222	34	124	177	409
6	234	37	132	185	423
7	247	41	144	201	461
8	262	44	156	217	495
9	269	42	164	237	541
10	280	44	171	253	587
11	287	46	176	260	608
12	287	43	175	260	600
13	296	48	181	267	623
14	310	53	190	272	634
15	326	59	203	278	632
16	325	54	214	309	703
17	322	44	210	331	767
18	338	52	209	329	791
19	353	60	210	302	714
20	370	66	222	296	664
21	380	67	245	342	774
22	377	54	247	379	871
23	384	63	250	386	916
24	376	54	238	372	882
25	390	60	231	342	792

Оцененная исходная модель имеет вид

$$\hat{y}_t = \underset{(9,4)}{41,73} + \underset{(0,33)}{2,42}x_t + \underset{(0,39)}{0,71}x_{t-1} + \underset{(0,41)}{0,99}x_{t-2} + \underset{(0,34)}{1,47}x_{t-3}, \quad R^2 = 0,981,$$

в которой коэффициент 0,71 при переменной x_{t-1} незначим (в скобках указаны стандартные ошибки).

Оцененная преобразованная модель имеет вид

$$\hat{y}_t = \underset{(9,27)}{41,6} + \underset{(0,30)}{2,32}z_0 - \underset{(0,66)}{1,92}z_1 + \underset{(0,21)}{0,56}z_2, \quad R^2 = 0,981,$$

и все коэффициенты при переменных значимы.

Получили следующие оценки параметров преобразованной модели:

$$\gamma_0 = 2,32, \quad \gamma_1 = -1,92, \quad \gamma_2 = 0,56.$$

Коэффициенты регрессии исходной модели:

$$\beta_0 = 2,32,$$

$$\beta_1 = 2,32 - 1,92 + 0,56 = 0,96,$$

$$\beta_2 = 2,32 - 2 \cdot 1,92 + 4 \cdot 0,56 = 0,72,$$

$$\beta_3 = 2,32 - 3 \cdot 1,92 + 9 \cdot 0,56 = 1,6.$$

Таким образом, модель с распределенным лагом имеет вид

$$\hat{y}_t = 41,6 + 2,32x_t + 0,96x_{t-1} + 0,72x_{t-2} + 1,6x_{t-3}.$$

Краткосрочный мультипликатор равен 2,32, а долгосрочный мультипликатор равен 2,32 + 0,96 + 0,72 + 1,6 = 5,6. Это означает, что увеличение инвестиций в экономику страны на 1 усл. ед. приведет к росту валового внутреннего продукта в среднем на 2,32 усл. ед. в текущем периоде и на 5,6 усл. ед. через 3 года.

6.2. МОДЕЛИ АВТОРЕГРЕССИИ

Пусть имеется модель авторегрессии вида

$$y_t = \alpha + \beta_0 x_t + \beta_1 y_{t-1} + \varepsilon_t.$$

Для интерпретации коэффициентов модели авторегрессии сделаем предположение о наличии бесконечного лага в воздействии текущего значения зависимой переменной на ее последующие значения и о выполнении неравенства $|\beta_1| < 1$ (**условие устойчивости**).

В краткосрочном (текущем) периоде влияние x на y отражается величиной β_0 (**краткосрочный мультипликатор**). Он характеризует краткосрочное изменение y под влиянием изменения x на единицу.

В долгосрочной перспективе суммарное влияние x на y отражается величиной β (**долгосрочный мультипликатор**):

$$\beta = \beta_0 + \beta_0\beta_1 + \beta_0\beta_1^2 + \beta_0\beta_1^3 + \ldots = \frac{\beta_0}{1-\beta_1}.$$

В модели авторегрессии $y_t = \alpha + \beta_0 x_t + \beta_1 y_{t-1} + \varepsilon_t$ объясняющая переменная y_{t-1} находится непосредственно под воздействием ε_{t-1} и косвенно под влиянием всех предшествующих значений случайного члена. Следовательно, объясняющая переменная y_{t-1} и случайный член ε_t зависимы, и МНК не дает несмещенных оценок. В этом случае одним из возможных методов оценки параметров уравнения авторегрессии является *метод инструментальных переменных*.

В качестве инструментальной переменной можно взять переменную x_{t-1}, которая коррелирована с y_{t-1} и некоррелирована с ε_t. Практически в качестве инструментальной переменной можно взять оценку

$$\hat{y}_{t-1} = \gamma_0 + \gamma_1 x_{t-1},$$

полученную из предполагаемой линейной зависимости y_{t-1} от x_{t-1}. Тогда оценку параметров модели автокорреляции можно найти обычным МНК из соотношения

$$y_t = \alpha + \beta_0 x_t + \beta_1 \hat{y}_{t-1} + \varepsilon_t,$$

где x_t, y_t — исходные, а $\hat{y}_{t-1}$ — расчетные данные.

Практическая реализация метода инструментальных переменных осложняется появлением мультиколлинеарности факторов x_t и $\hat{y}_{t-1}$ в модели.

Пример 6.2. Построим модель авторегрессии по данным о среднедушевом располагаемом доходе x и среднедушевых расходах на конечное потребление y за 32 года. Исходные и расчетные данные (усл. ед.) представлены в следующей таблице:

t	y_t	x_t	$\hat{y}_{t-1}$	e_t
1	67	73	–	–
2	67	74	65,58	0,57
3	69	76	66,5	0,85

Окончание таблицы

t	y_t	x_t	$\hat{y}_{t-1}$	e_t
4	71	77	68,34	1,79
5	74	81	69,26	1,47
6	77	85	72,94	0,76
7	80	88	76,62	0,85
8	82	91	79,38	0,078
9	85	94	82,15	0,3
10	87	96	84,91	0,31
11	88	99	86,75	−1,34
12	90	101	89,51	−1,32
13	94	104	91,35	0,03
14	97	110	94,11	−2,14
15	96	108	99,63	−2,32
16	97	108	97,79	−1,06
17	101	112	97,79	−0,25
18	104	114	101,47	0,63
19	107	118	103,31	0,18
20	108	120	106,99	−0,93
21	107	120	108,84	−2,19
22	108	121	108,84	−1,98
23	108	121	109,76	−2,11
24	112	123	109,76	0,29
25	116	130	111,6	−1,54
26	120	133	118,04	−0,84
27	123	135	120,8	0,17
28	126	136	122,64	2,11
29	129	139	123,56	2,59
30	130	140	126,32	2,40
31	130	141	127,24	1,48
32	129	140	128,16	1,14

Определим по этим данным параметры модели авторегрессии вида

$$y_t = \alpha + \beta_0 x_t + \beta_1 y_{t-1} + \varepsilon_t.$$

Применение обычного МНК для оценки параметров этой модели приводит к следующим результатам:

$$\hat{y}_t = \underset{(1,25)}{-0,215} + \underset{(0,097)}{0,504}\, x_t + \underset{(0,104)}{0,455}\, y_{t-1}, \quad R^2 = 0,997$$

(в скобках указаны стандартные ошибки).

Однако, как уже было отмечено, оценка параметра $\beta_1 = 0{,}455$ является смещенной. Для получения несмещенных оценок параметров этого уравнения воспользуемся методом инструментальных переменных.

Оценка уравнения регрессии $y_{t-1} = \gamma_0 + \gamma_1 x_{t-1} + \varepsilon_t$ обычным МНК дает следующие результаты:

$$\hat{y}_{t-1} = \underset{(1{,}398)}{-1{,}609} + \underset{(0{,}012)}{0{,}920}\, x_{t-1}, \quad R^2 = 0{,}994.$$

Тогда в результате оценки модели авторегрессии $y_t = \alpha + \beta_0 x_t + \beta_1 \hat{y}_{t-1} + \varepsilon_t$ обычным МНК получаем

$$\hat{y}_t = \underset{(1{,}63)}{-1{,}801} + \underset{(0{,}146)}{0{,}797}\, x_t + \underset{(0{,}157)}{0{,}141}\, \hat{y}_{t-1}, \quad R^2 = 0{,}995.$$

Применение метода инструментальных переменных привело к статистической незначимости оценки параметра $\beta_1 = 0{,}141$ при переменной $\hat{y}_{t-1}$. Это произошло ввиду высокой мультиколлинеарности переменных x_t и $\hat{y}_{t-1}$.

Поскольку ни один из методов оценок параметров модели авторегрессии не привел к достоверным результатам, следует использовать другие методы оценок.

Заметим, что для данной модели авторегрессии при наличии автокорреляции остатков не существует состоятельного метода оценивания.

В качестве примера проверим гипотезу о наличии автокорреляции в модели регрессии, полученной методом инструментальных переменных. Проверку осуществим по h-критерию Дарбина.

Коэффициент автокорреляции остатков первого порядка $r = 0{,}722$, следовательно, значение $DW = 2(1 - r) = 0{,}556$.

Значение h-критерия Дарбина, определяемое выражением

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{n}{1 - n\,\mathrm{var}(b)}} = \left(1 - \frac{0{,}556}{2}\right)\sqrt{\frac{32}{1 - 32 \cdot 0{,}157^2}} = 8{,}9,$$

указывает на положительную автокорреляцию остатков.

6.3. ПРИМЕРЫ МОДЕЛЕЙ С ЛАГИРОВАННЫМИ ПЕРЕМЕННЫМИ

МОДЕЛЬ ЧАСТИЧНОЙ КОРРЕКТИРОВКИ

В модели частичной корректировки предполагается, что поведенческое уравнение определяет не фактическое значение зависимой переменной y_t, а ее *желаемый* (целевой) уровень:

$$y_t^* = \alpha + \beta x_t + \varepsilon_t, \quad \varepsilon_t \sim N(0; \sigma^2). \tag{6.1}$$

Предполагается также, что фактическое значение зависимой переменной не выходит мгновенно на желаемый уровень, а изменяется только на долю λ в нужном направлении:

$$y_t - y_{t-1} = \lambda(y_t^* - y_{t-1}) \quad (0 \le \lambda \le 1). \tag{6.2}$$

Это выражение можно переписать следующим образом:

$$y_t = \lambda y_t^* + (1 - \lambda) y_{t-1},$$

откуда видно, что y_t получается как взвешенное среднее желаемого уровня и фактического значения этой переменной в предыдущем периоде.

Параметр λ называется **корректирующим коэффициентом**. Чем больше λ, тем быстрее происходит процесс корректировки.

Если $\lambda = 1$, то $y_t = y_t^*$ и полная корректировка происходит за один период.

Если $\lambda = 0$, то корректировка y_t не происходит совсем.

Подставляя y_t^* в выражение для y_t, получим

$$y_t = \alpha\lambda + \beta\lambda x_t + (1 - \lambda) y_{t-1} + \lambda\varepsilon_t. \tag{6.3}$$

Полученное уравнение включает только фактические значения переменных.

Поскольку случайные члены некоррелированы, состоятельные оценки параметров можно получить, применяя МНК к оцениванию составных параметров $\alpha\lambda$, $\beta\lambda$ и $(1 - \lambda)$ в уравнении (6.3).

Пример 6.3. Производственные компании распределяют прибыль П, оставшуюся после уплаты налогов: одну часть на выплату доходов акционерам в форме дивидендов D, другую — на финансирование инвестиций.

Известны данные о деятельности производственных компаний за ряд предыдущих лет (усл. ед.):

t	D	П	t	D	П
1	100	400	6	800	1100
2	300	600	7	900	1300
3	450	700	8	1000	1400
4	550	800	9	1100	1500
5	700	1000	10	1200	1700

Предположим, что у фирмы имеется целевая долгосрочная доля выплат γ и что желаемый объем дивидендов D_t^* соотносится с текущей прибылью Π_t как $D_t^* = \gamma\Pi_t + \varepsilon_t$. Однако реальный объем дивидендов подвержен процессу частичной корректировки:

$$D_t - D_{t-1} = \lambda(D_t^* - D_{t-1}),$$

или

$$D_t = \gamma\lambda\Pi_t + (1-\lambda)D_{t-1} + \lambda\varepsilon_t \quad (0 \le \lambda \le 1).$$

На основе данных о деятельности производственных компаний за ряд лет построено уравнение регрессии

$$D_t = 68 + 0{,}29\Pi_t + 0{,}58D_{t-1},$$

где все коэффициенты значимы.

Из соотношения $1 - \lambda = 0{,}58$ определяется корректирующий коэффициент $\lambda = 0{,}42$, а из соотношения $\gamma\lambda = 0{,}29$ — оценка доли выплат $\gamma = 0{,}69$.

МОДЕЛЬ АДАПТИВНЫХ ОЖИДАНИЙ

Предположим, что зависимая переменная y_t связана с ожидаемым значением объясняющей переменной x в $(t+1)$-м периоде соотношением

$$y_t = \alpha + \beta x_{t+1}^* + \varepsilon_t, \tag{6.4}$$

где x_{t+1}^* — ожидаемое значение ненаблюдаемой объясняющей переменной, которую необходимо заменить наблюдаемыми переменными.

Такая модель возникает, например, в следующем случае: фирма принимает решение об объеме производимой в период t продукции y_t до того, как станет известной цена x_{t+1}, по которой эта продукция может быть продана в следующем периоде. Поскольку цена x_{t+1} неизвестна в период t, то решение принимается на основе ожидаемого значения x_{t+1}^*.

8*-2337

Процесс формирования ожиданий таков:

$$x_{t+1}^* - x_t^* = \lambda(x_t - x_t^*), \tag{6.5}$$

или

$$x_{t+1}^* = \lambda x_t + (1-\lambda)x_t^* \quad (0 \le \lambda \le 1),$$

т.е. ожидаемое значение переменной x_t^* в следующем периоде является взвешенным средним ее фактического и ожидаемого значений в текущем периоде.

Параметр λ называется **коэффициентом ожидания**. Величину x_t^* можно выразить через x_{t-1}^* и т.д. Повторяя эту процедуру бесконечное число раз, получим

$$x_{t+1}^* = \lambda[x_t + (1-\lambda)x_{t-1} + (1-\lambda)^2 x_{t-2} + \ldots]. \tag{6.6}$$

В итоге получаем *модель адаптивных ожиданий*, в которой ожидаемое значение переменной является взвешенным средним ее прошлых значений с геометрически убывающим весом.

Подставим выражение (6.6) для x_{t+1}^* в исходную модель (6.4) и заменим $(1-\lambda)$ на δ:

$$y_t = \alpha + \beta\lambda(x_t + \delta x_{t-1} + \delta^2 x_{t-2} + \ldots) + \varepsilon_t,$$

т.е. получим *модель геометрических лагов*. Параметры уравнения можно оценить *методом нелинейного оценивания*.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Для каких экономических задач требуется применение моделей с распределенным лагом и моделей авторегрессии?
2. Какова интерпретация параметров модели с распределенным лагом?
3. Какова интерпретация параметров модели авторегрессии? В чем специфика долгосрочного лага в этой модели?
4. В чем заключается метод Алмона?
5. В чем заключается подход Койка к построению модели с распределенным лагом?
6. В чем сущность модели адаптивных ожиданий? Какова методика оценки ее параметров?
7. В чем сущность модели частичной корректировки? Какова методика оценки ее параметров?
8. В чем заключается метод инструментальных переменных для оценки параметров модели авторегрессии?

Глава 7
Системы одновременных уравнений

7.1. СТРУКТУРНАЯ И ПРИВЕДЕННАЯ ФОРМЫ УРАВНЕНИЙ

Сложные экономические процессы описываются с помощью системы взаимосвязанных (одновременных) уравнений.

Различают следующие **виды эконометрических систем**:

- системы независимых уравнений;
- системы рекурсивных уравнений;
- системы взаимозависимых уравнений.

Система независимых уравнений — каждая зависимая переменная y рассматривается как функция одного и того же набора фактора x:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Каждое уравнение такой системы может рассматриваться самостоятельно. Для нахождения его параметров используется метод наименьших квадратов.

Система рекурсивных уравнений — зависимая переменная y включает в каждое последующее уравнение в качестве факторов все зависимые переменные предшествующих уравнений и набор фактора x:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + \varepsilon_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

В таких моделях уравнения оцениваются последовательно (от первого уравнения к последнему) с использованием МНК.

Система взаимозависимых уравнений — одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других — в правую часть системы:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Структурной формой модели (системой одновременных уравнений) называется система уравнений, в каждом из которых аргументы содержат не только объясняющие переменные, но и объясняемые переменные из других уравнений.

Уравнения, составляющие исходную модель, называются **структурными уравнениями модели**.

Простейшая структурная форма модели имеет вид

$$\begin{cases} y_1 = \alpha_1 + \beta_{12}y_2 + \alpha_{11}x_1 + \alpha_{12}x_2 + \varepsilon_1, \\ y_2 = \alpha_2 + \beta_{21}y_1 + \alpha_{21}x_1 + \alpha_{22}x_2 + \varepsilon_2, \end{cases} \qquad (7.1)$$

где y и x — зависимая и независимая переменные, ε_1 и ε_2 — случайные члены, а (α, β) — параметры модели.

Параметры структурной формы модели называются **структурными коэффициентами**.

Структурная форма модели обычно включает в систему не только уравнения, отражающие взаимосвязи между отдельными переменными, но и уравнения, отражающие тенденцию развития явления, а также разного рода уравнения-тождества. Тождества не содержат каких-либо подлежащих оценке параметров, а также не включают случайного члена.

В процессе оценивания параметров одновременных уравнений следует различать *эндогенные* и *экзогенные* переменные. Приставки «эндо» и «экзо» означают соответственно внутреннее и внешнее.

Эндогенными считаются переменные, значения которых определяются *внутри* модели. Это зависимые переменные, число которых равно числу уравнений системы.

Экзогенными считаются переменные, значения которых определяются *вне* модели. Это заданные переменные, влияющие на эндогенные переменные, но не зависящие от них.

В качестве экзогенных могут рассматриваться значения эндогенных переменных за предшествующий период времени (лаговые переменные).

Предполагается, что в каждом уравнении экзогенные переменные некоррелированы со случайным членом.

В общем случае эндогенные переменные коррелированы со случайным членом, поэтому применение МНК к структурной форме модели приводит к *смещенным* и *несостоятельным* оценкам структурных коэффициентов.

Для определения структурных коэффициентов структурная форма модели преобразуется в приведенную форму.

Приведенной формой модели называется система уравнений, в каждом из которых эндогенные переменные выражены только через экзогенные переменные и случайные составляющие.

Например, приведенная форма исходной модели (7.1) имеет вид

$$\begin{cases} y_1 = \alpha_1' + \alpha_{11}' x_1 + \alpha_{12}' x_2 + \nu_1, \\ y_2 = \alpha_2' + \alpha_{21}' x_1 + \alpha_{22}' x_2 + \nu_2, \end{cases} \tag{7.2}$$

где α' — параметры приведенной формы, а ν_1 и ν_2 — случайные члены.

Параметры приведенной формы модели называются **коэффициентами приведенной формы** (приведенными коэффициентами). Коэффициенты приведенной формы оцениваются обычным МНК, поскольку экзогенные переменные некоррелированы со случайным членом.

Оцененные коэффициенты приведенной формы могут быть использованы для оценивания структурных коэффициентов. Такой способ оценивания структурных коэффициентов называется **косвенным методом наименьших квадратов**.

Приведенная форма модели аналитически уступает структурной форме, так как в ней отсутствуют оценки взаимосвязи между эндогенными переменными. При переходе от приведенной формы к структурной возникает проблема идентификации. *Идентификация* — это единственность соответствия между приведенной и структурной формами модели.

Тот или иной структурный коэффициент может либо однозначно выражаться через приведенные коэффициенты, либо иметь несколько разных оценок, либо совсем не выражаться через них.

Структурный коэффициент называется **идентифицируемым**, если его можно вычислить на основе приведенных коэффициентов, причем **точно идентифицируемым**, если он единствен, и **сверхидентифицируемым**, если он имеет несколько разных оценок; в противном случае он называется **неидентифицируемым**.

Какое-либо структурное уравнение является идентифицируемым, если идентифицируемы все его коэффициенты. Если хотя бы один структурный коэффициент неидентифицируем, то и всё уравнение является неидентифицируемым.

Модель считается идентифицируемой, если каждое ее уравнение идентифицируемо. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель неидентифицируема.

7.2. ОЦЕНИВАНИЕ ПАРАМЕТРОВ СТРУКТУРНОЙ МОДЕЛИ

Коэффициенты структурной модели могут быть оценены различными способами в зависимости от вида системы одновременных уравнений. Наибольшее распространение получили следующие методы:

- метод инструментальных переменных (ИП);
- косвенный метод наименьших квадратов (КМНК);
- двухшаговый метод наименьших квадратов (ДМНК).

МЕТОДЫ ОЦЕНИВАНИЯ СТРУКТУРНЫХ УРАВНЕНИЙ РАЗЛИЧНЫХ ВИДОВ

1. Точная идентифицируемость

Допустим, требуется оценить параметры уравнения функции потребления в простой модели Кейнса формирования доходов:

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t & \text{(функция потребления)}, \\ Y_t = C_t + I_t & \text{(тождество дохода)}, \end{cases} \tag{7.3}$$

где C_t, Y_t, I_t — объем потребления, совокупный доход и инвестиции соответственно, а ε_t — случайный член.

Структурный коэффициент β характеризует предельную склонность к потреблению.

В исходной модели C_t, Y_t — эндогенные переменные, а I_t — экзогенная. Непосредственное оценивание параметров (α, β) в струк-

турном уравнении функции потребления дает *смещенные* и *несостоятельные* оценки, так как объясняющая переменная Y_t является эндогенной.

Разрешая структурную систему относительно эндогенных переменных, получим приведенную систему

$$\begin{cases} C_t = \dfrac{\alpha}{1-\beta} + \dfrac{\beta}{1-\beta} I_t + \dfrac{\varepsilon_t}{1-\beta}, \\ Y_t = \dfrac{\alpha}{1-\beta} + \dfrac{1}{1-\beta} I_t + \dfrac{\varepsilon_t}{1-\beta}. \end{cases} \tag{7.4}$$

В приведенной системе коэффициенты при переменной I_t, равные $\beta/(1-\beta)$ и $1/(1-\beta)$, — это **инвестиционные мультипликаторы потребления** и **дохода** соответственно. Это значит, что если объем инвестиций возрастет на единицу, то объем потребления увеличится на $\beta/(1-\beta)$, а совокупный доход — на $1/(1-\beta)$.

Рассмотрим различные методы оценивания структурных коэффициентов (α, β).

Косвенный метод наименьших квадратов. Уравнение для C_t в приведенной форме можно также представить в виде

$$C_t = \alpha' + \beta' I_t + \varepsilon'_t, \tag{7.5}$$

где

$$\alpha' = \frac{\alpha}{1-\beta}, \quad \beta' = \frac{\beta}{1-\beta}, \quad \varepsilon'_t = \frac{\varepsilon_t}{1-\beta}. \tag{7.6}$$

В этом уравнении экзогенная переменная I_t некоррелирована со случайным членом ε'_t, поэтому для оценки параметров (α', β') можно использовать обычный МНК.

З а м е ч а н и е. *Для удобства рассмотрения оценку параметра и сам параметр будем в дальнейшем обозначать одним символом (параметром).*

Оцененное уравнение (7.5), полученное по выборочным данным с помощью МНК,

$$\hat{C}_t = \alpha' + \beta' I_t$$

дает *несмещенные* и *состоятельные* оценки параметров.

Из выражения (7.6) получаем оценки (α, β) структурных коэффициентов:

$$\alpha = \frac{\alpha'}{1+\beta'}, \quad \beta = \frac{\beta'}{1+\beta'}. \tag{7.7}$$

Поскольку получены единственные оценки (α, β) структурных коэффициентов через оценки (α', β') приведенных коэффициентов, то структурное уравнение функции потребления является *однозначно определенным* (*точно идентифицируемым*).

Метод инструментальных переменных. Проблема коррелированности объясняющей переменной Y_t со случайным членом ε_t в структурном уравнении (7.3) для C_t может быть разрешена с помощью метода ИП.

Для применения метода ИП необходимо найти такую инструментальную переменную, которая обладает следующими свойствами:

1) коррелирует с неудачно объясняющей переменной Y_t;
2) не коррелирует со случайным членом ε_t.

В данном случае модель сама предоставляет такую переменную. Величина I_t коррелирует с Y_t, поскольку Y_t зависит от I_t в уравнении (7.4), и I_t не коррелирует с ε_t, поскольку является экзогенной переменной.

Оценка β с помощью инструментальной переменной I_t определяется как

$$\beta_{\text{ИП}} = \frac{\text{cov}(I, C)}{\text{cov}(I, Y)}.$$

Полученная оценка $\beta_{\text{ИП}}$ эквивалентна $\beta_{\text{КМНК}}$ — оценке β с помощью КМНК. Действительно, из соотношения (7.7) и учитывая, что β' рассчитывается как $\text{cov}(I, C)/\text{var}(I)$, получим

$$\beta_{\text{КМНК}} = \frac{\beta'}{1+\beta'} = \frac{\text{cov}(I,C)/\text{var}(I)}{1+\text{cov}(I,C)/\text{var}(I)} = \frac{\text{cov}(I,C)}{\text{var}(I)+\text{cov}(I,C)} =$$

$$= \frac{\text{cov}(I,C)}{\text{var}(I,Y)} = \beta_{\text{ИП}},$$

поскольку $\text{cov}(I, Y) = \text{cov}(I, I+C) = \text{var}(I) + \text{cov}(I, C)$.

В общем случае, когда оценка, полученная косвенным методом, единственна, она совпадает с оценкой, полученной методом ИП, т.е. КМНК можно рассматривать как частный случай метода ИП.

Пример 7.1. Для некоторой страны имеются данные о совокупном доходе Y, объеме потребления C и инвестициях I, полученные за 10 лет (усл. ед.):

C_t	190	198	200	180	200	210	220	210	205	210
I_t	10	20	30	20	10	20	30	20	15	30
Y_t	200	218	230	200	210	230	250	230	220	240

Построим функцию потребления, используя модель Кейнса формирования доходов (7.3).

Непосредственное оценивание структурного уравнения функции потребления обычным МНК приводит к следующим результатам:

$$\hat{C} = 60{,}9 + 0{,}635Y,$$

т.е. оценки $\alpha = 60{,}9, \quad \beta = 0{,}635$.

Было показано, что исходная модель (7.3) точно идентифицируема, поэтому для оценки ее структурных коэффициентов используем КМНК.

Оценка для C в приведенной форме

$$\hat{C} = 188 + 0{,}695I,$$

т.е. $\alpha' = 188, \quad \beta' = 0{,}695$.

Из выражения (7.7) получим оценки структурных коэффициентов:

$$\alpha = 188/(1 + 0{,}695) = 110{,}9, \quad \beta = 0{,}695/(1 + 0{,}695) = 0{,}41,$$

т.е. $\hat{C} = 110{,}9 + 0{,}41Y$.

Оценки структурных коэффициентов функции потребления, полученные КМНК, являются *несмещенными* и *состоятельными*.

II. Сверхидентифицируемость

Рассмотрим следующую простую модель Кейнса формирования доходов:

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t & \text{(функция потребления)}, \\ Y_t = C_t + I_t + G_t & \text{(тождество дохода)}, \end{cases} \tag{7.8}$$

где C_t, Y_t, I_t, G_t — объем потребления, совокупный доход, инвестиции и государственные расходы соответственно, а ε_t — случайный член.

В исходной модели C_t, Y_t — эндогенные переменные, а I_t, G_t — экзогенные.

Разрешая структурную систему относительно эндогенных переменных, получим приведенную систему

$$\begin{cases} C_t = \dfrac{\alpha}{1-\beta} + \dfrac{\beta}{1-\beta} I_t + \dfrac{\beta}{1-\beta} G_t + \dfrac{\varepsilon_t}{1-\beta}, \\ Y_t = \dfrac{\alpha}{1-\beta} + \dfrac{1}{1-\beta} I_t + \dfrac{1}{1-\beta} G_t + \dfrac{\varepsilon_t}{1-\beta}. \end{cases} \tag{7.9}$$

Рассмотрим различные методы оценивания структурных коэффициентов (α, β).

Метод инструментальных переменных. В структурном уравнении функции потребления в качестве инструментальных переменных для Y_t можно использовать как I_t, так и G_t. Полученные при этом оценки (α, β) будут различаться, но в обоих случаях они *состоятельны*.

Наилучшее решение в данном случае — применение инструментальной переменной, которая является комбинацией I_t и G_t.

Структурное уравнение с избыточным числом экзогенных переменных, которые можно использовать как инструментальные, является *переопределенным* (*сверхидентифицируемым*).

Двухшаговый метод наименьших квадратов. Двухшаговый МНК можно рассматривать как частный случай инструментальных переменных. В методе ИП было показано, что структурное уравнение функции потребления оказалось переопределенным и сразу две переменные I_t и G_t можно использовать для Y_t.

Однако вместо их раздельного применения можно предложить их комбинацию $z_t = \gamma_0 + \gamma_1 I_t + \gamma_2 G_t$. В этом случае требуется оценить значения коэффициентов γ_0, γ_1, γ_2.

Фактически вместо z_t можно использовать оценку $\hat{Y}_t$ приведенного уравнения Y_t, т.е. $\hat{Y}_t = \gamma_0 + \gamma_1 I_t + \gamma_2 G_t$.

Подставляя теоретические значения $\hat{Y}_t$ вместо фактических значений в структурное уравнение функции потребления, получим уравнение

$$C_t = \alpha + \beta \hat{Y}_t + \varepsilon_t,$$

которое оценивается обычным МНК. При этом оценки структурных коэффициентов будут *состоятельными*.

Двухшаговый МНК можно рассматривать как способ конструирования наилучшей из возможных комбинаций инструментальных переменных, если в уравнении имеется избыток экзогенных переменных, которые можно использовать как инструментальные.

Пример 7.2. Для некоторой страны имеются данные о совокупном доходе Y, объеме потребления C, инвестициях I и государственных расходах G, полученные за 10 лет (усл. ед.):

C_t	195	203	210	200	215	215	210	215	225	220
I_t	10	20	30	20	10	20	30	20	15	30
G_t	20	10	20	40	30	10	20	10	40	20
Y_t	225	233	260	260	255	245	260	245	280	270

Построим функцию потребления, используя модель Кейнса формирования доходов (7.8).

Непосредственное оценивание структурного уравнения функции потребления обычным МНК приводит к следующим результатам:

$$\hat{C} = 109{,}8 + 0{,}4Y,$$

т.е. оценки $\alpha = 109{,}8, \quad \beta = 0{,}4$.

Было показано, что исходная модель (7.8) сверхидентифицируема, поэтому для оценки ее структурных коэффициентов используем ДМНК.

Расчетные значения эндогенной переменной Y, полученные МНК:

$$\hat{Y} = 201{,}7 + 1{,}29I + 1{,}14G.$$

Подставим расчетные значения $\hat{Y}$ вместо фактических значений в структурное уравнение функции потребления и оценим полученное уравнение МНК:

$$\hat{C} = 171{,}3 + 0{,}156Y,$$

т.е. оценки $\alpha = 171{,}3, \quad \beta = 0{,}156$.

Оценки структурных коэффициентов функции потребления, полученные ДМНК, являются *состоятельными*.

III. Неидентифицируемость

Рассмотрим следующую модель спроса и предложения:

$$\begin{cases} y^D = \alpha + \beta P + u^D & \text{(спрос)}, \\ y^S = \delta + \varepsilon P + u^S & \text{(предложение)}, \\ y^D = y^S = y & \text{(равновесие)}, \end{cases}$$

где P — цена товара, а u^D, u^S — случайные члены.

Переменные y, P являются эндогенными, и их значения определяются в процессе установления равновесия.

В рассматриваемой модели нет экзогенных переменных, поэтому ни одно из этих уравнений *не является идентифицируемым*. Чтобы модель имела статистическое решение, в нее вводятся экзогенные переменные.

Предположим, что продавцы товара облагаются специальным налогом T, который они должны платить с выручки. При этом уравнение спроса останется неизменным, если переменная P означает рыночную цену, а уравнение предложения изменится:

$$\begin{cases} y^D = \alpha + \beta P + u^D & \text{(спрос)}, \\ y^S = \delta + \varepsilon P + \sigma T + u^S & \text{(предложение)}, \\ y^D = y^S = y & \text{(равновесие)}, \end{cases} \tag{7.10}$$

где T — экзогенная переменная.

Уравнение спроса будет *идентифицируемым*, поскольку переменная T не включена в него и может выступать как инструментальная для P, а *уравнение предложения — неидентифицируемым*.

Включим в уравнение спроса экзогенную переменную x — доход на душу населения:

$$\begin{cases} y^D = \alpha + \beta P + \gamma x + u^D & \text{(спрос)}, \\ y^S = \delta + \varepsilon P + \sigma T + u^S & \text{(предложение)}, \\ y^D = y^S = y & \text{(равновесие)}. \end{cases} \tag{7.11}$$

Экзогенную переменную x можно использовать как инструментальную вместо P для уравнения предложения.

В итоге получили в целом *точно идентифицируемую модель спроса и предложения*.

Пусть структурное уравнение спроса имеет временной тренд (скажем, потому что привычки медленно меняются со временем):

$$\begin{cases} y^D = \alpha + \beta P + \gamma x + \rho t + u^D & \text{(спрос)}, \\ y^S = \delta + \varepsilon P + \sigma T + u^S & \text{(предложение)}, \\ y^D = y^S = y & \text{(равновесие)}, \end{cases} \tag{7.12}$$

где t — переменная времени, а ρ — коэффициент при ней.

В модели спроса имеются две экзогенные переменные x, t, которые можно использовать в качестве инструментальных для P в уравнении предложения.

В итоге получили *сверхидентифицируемое уравнение предложения* и *точно идентифицируемое уравнение спроса*.

ПОРЯДКОВОЕ УСЛОВИЕ ДЛЯ ИДЕНТИФИКАЦИИ

В общем случае отдельное структурное уравнение системы является идентифицируемым, если имеется достаточное количество экзогенных переменных, не включенных в само уравнение, которые можно использовать как инструментальные для всех эндогенных объясняющих переменных уравнения.

В полностью определенной модели будет столько уравнений, сколько имеется эндогенных переменных.

Пусть D — число не включенных в уравнение, но присутствующих в системе экзогенных переменных, а G — число включенных в уравнение эндогенных переменных.

Необходимое условие идентификации. Уравнение в структурной модели может быть идентифицировано, если число не включенных в него экзогенных переменных не меньше числа включенных в него объясняющих эндогенных переменных, т.е.

$D \geq G - 1$ (**порядковое условие**).

Данное условие является *необходимым*, но *не достаточным* для идентификации.

В частности:

- **если $D = G - 1$, то уравнение точно идентифицируемо;**
- **если $D > G - 1$, то уравнение сверхидентифицируемо;**
- **если $D < G - 1$, то уравнение неидентифицируемо.**

Достаточное условие идентификации. Уравнение идентифицируемо, если ранг матрицы, составленной из коэффициентов при переменных (эндогенных и экзогенных), отсутствующих в исследуемом уравнении, не меньше $N - 1$, где N — число эндогенных переменных системы.

Пример 7.3. Проверим на идентификацию каждое уравнение модели

$$\begin{cases} y_1 = \alpha_{01} + \beta_{13} y_3 + \beta_{14} y_4 + \varepsilon_1, \\ y_2 = \alpha_{02} + \beta_{23} y_3 + \alpha_{21} x_1 + \varepsilon_2, \\ y_3 = \alpha_{03} + \beta_{34} y_4 + \alpha_{31} x_1 + \varepsilon_3, \\ y_4 = y_1 + y_2 + x_2, \end{cases}$$

где y_1 — расходы на потребление текущего года; y_2 — валовые инвестиции в текущем году; y_3 — расходы на заработную плату в текущем году; y_4 — валовой доход за текущий год; x_1 — валовой доход предыдущего года; x_2 — государственные расходы текущего года; ε — случайные ошибки.

В данной модели четыре эндогенные переменные (y_1, y_2, y_3, y_4), т.е. $N = 4$, и две экзогенные (x_1, x_2).

Для *первого* уравнения: $G = 3$ (y_1, y_3, y_4 присутствуют), $D = 2$ (x_1, x_2 отсутствуют) и $D = G - 1$, поэтому уравнение точно идентифицируемо (*необходимое условие*).

Для проверки на *достаточное условие* идентификации выпишем матрицу A коэффициентов при переменных, не входящих в первое уравнение:

Уравнение	y_2	x_1	x_2
2	-1	α_{21}	0
3	0	α_{31}	0
4	1	0	1

Определитель матрицы $\det A = -\alpha_{31} \neq 0$, следовательно, ранг матрицы равен $3 \geq N - 1$, т.е. достаточное условие идентификации выполняется, и первое уравнение точно идентифицируемо.

Второе уравнение системы также точно идентифицируемо: $G = 2$, $D = 1$ и $D = G - 1$.

Выпишем матрицу A коэффициентов при переменных, не входящих во второе уравнение:

Уравнение	y_1	y_4	x_2
1	-1	β_{14}	0
3	0	β_{34}	0
4	1	-1	1

Выполняется также достаточное условие идентификации: $\det A = -\beta_{34} \neq 0$, ранг матрицы равен $3 \geq N - 1$.

Аналогично *третье* уравнение системы точно идентифицируемо: $G = 2$, $D = 1$, $D = G - 1$.

Выпишем матрицу A коэффициентов при переменных, не входящих в третье уравнение:

Уравнение	y_1	y_2	x_2
1	-1	0	0
2	0	-1	0
4	1	1	1

Здесь также выполняется достаточное условие идентификации: $\det A = 1$, ранг матрицы равен $3 \geq N - 1$.

Четвертое уравнение представляет собой тождество, параметры которого известны, поэтому необходимости в его идентификации нет.

Таким образом, все уравнения модели точно идентифицированы.

Пример 7.4. Выполним идентификацию следующей модели:

$$\begin{cases} C_t = \alpha_{01} + \beta_{11} Y_t + \alpha_{12} C_{t-1} + \varepsilon_1 & \text{(функция потребления)}, \\ I_t = \alpha_{02} + \beta_{21} r_t + \alpha_{22} I_{t-1} + \varepsilon_2 & \text{(функция инвестиций)}, \\ r_t = \alpha_{03} + \beta_{31} Y_t + \alpha_{32} M_t + \varepsilon_3 & \text{(функция денежного рынка)}, \\ Y_t = C_t + I_t + G_t & \text{(тождество дохода)}, \end{cases}$$

где C — расходы на потребление; Y — совокупный доход; I — инвестиции; r — процентная ставка; M — денежная масса; G — государственные расходы; t — текущий период; $t-1$ — предыдущий период.

В данной модели четыре эндогенные переменные (C_t, I_t, r_t, Y_t), т.е. $N = 4$, и четыре экзогенные (M_t, G_t, C_{t-1}, I_{t-1}).

Для *первого* уравнения: $G = 2$ (C_t, Y_t присутствуют), $D = 3$ (M_t, G_t, I_{t-1} отсутствуют) и $D > G - 1$, поэтому уравнение сверхидентифицируемо (*необходимое условие*).

Для проверки на *достаточное условие* идентификации выпишем матрицу коэффициентов при переменных, не входящих в первое уравнение:

Уравнение	I_t	r_t	I_{t-1}	M_t	G_t
2	-1	β_{21}	α_{22}	0	0
3	0	-1	0	α_{32}	0
4	1	0	0	0	1

Минор третьего порядка данной матрицы $\begin{vmatrix} -1 & \beta_{21} & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{vmatrix} \neq 0$, следовательно, ранг матрицы равен $3 \geq N - 1$, т.е. достаточное условие идентификации выполняется.

Для *второго* уравнения: $G = 2$ (I_t, r_t присутствуют), $D = 3$ (M_t, G_t, C_{t-1} отсутствуют) и $D > G - 1$, поэтому уравнение сверхидентифицируемо.

Выпишем матрицу коэффициентов при переменных, не входящих во второе уравнение:

Уравнение	C_t	Y_t	C_{t-1}	M_t	G_t
1	-1	β_{11}	α_{12}	0	0
3	0	β_{31}	0	α_{32}	0
4	1	-1	0	0	1

Минор третьего порядка данной матрицы $\begin{vmatrix} -1 & 0 & 0 \\ 0 & \alpha_{32} & 0 \\ 1 & 0 & 1 \end{vmatrix} \neq 0$, следовательно, ранг матрицы равен $3 \geq N - 1$, т.е. достаточное условие идентификации выполняется.

Для *третьего* уравнения: $G = 2$ (Y_t, r_t присутствуют), $D = 3$ (G_t, C_{t-1}, I_{t-1} отсутствуют) и $D > G - 1$, поэтому уравнение сверхидентифицируемо.

Выпишем матрицу коэффициентов при переменных, не входящих в третье уравнение:

Уравнение	C_t	C_{t-1}	I_t	I_{t-1}	G_t
1	-1	α_{12}	0	0	0
2	0	0	-1	α_{22}	0
4	1	0	1	0	1

Минор третьего порядка данной матрицы $\begin{vmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \end{vmatrix} \neq 0$, следовательно, ранг матрицы равен $3 \geq N - 1$, т.е. достаточное условие идентификации выполняется.

Четвертое уравнение представляет собой тождество, параметры которого известны, поэтому необходимости в его идентификации нет.

Таким образом, все уравнения модели сверхидентифицированы.

НЕНУЛЕВОЕ ОГРАНИЧЕНИЕ

Добавление экзогенной переменной не единственный способ, который может привести к идентифицируемости уравнения. В некоторых случаях неидентифицируемая модель может быть идентифицируема путем задания соотношения между структурными коэффициентами.

Рассмотрим неидентифицируемую модель спроса и предложения (7.10). Улучшим спецификацию модели, введя ограничение на коэффициенты $\sigma = -\varepsilon$:

$$\begin{cases} y^D = \alpha + \beta P + u^D & \text{(спрос)}, \\ y^S = \delta + \varepsilon(P - T) + u^S & \text{(предложение)}, \\ y^D = y^S = y & \text{(равновесие)}. \end{cases} \tag{7.13}$$

Благодаря введению ограничения на коэффициенты *уравнение предложения* также стало *идентифицируемым*. Действительно, при использовании ИП можно рассмотреть новую версию модели как систему из четырех уравнений:

$$\begin{cases} y^D = \alpha + \beta P + u^D, \\ y^S = \delta + \varepsilon P_1 + u^S, \\ P_1 = P - T, \\ y^D = y^S, \end{cases} \tag{7.14}$$

где P_1 — цена товара для продавца (сумма, остающаяся у него после уплаты налога).

Последние два уравнения системы (7.14) являются уравнениями-тождествами и не требуют проверки на идентификацию. Переменная T не включена в уравнение спроса, поэтому она может использоваться как инструментальная для P. Точно так же эта переменная не включена в уравнение предложения, поэтому она может использоваться как инструментальная для P_1.

В итоге модель в целом является *точно определенной* (*точно идентифицируемой*).

В ы в о д. Ненулевое ограничение позволяет исключить одну объясняющую переменную из уравнения. Если эта переменная *э н д о г е н н а я*, для нее не нужно искать инструментальную переменную, если *э к з о г е н н а я*, то она освобождается на роль инструментальной для одной из эндогенных переменных, оставшихся в уравнении.

Пример 7.5. Опишем процедуру оценивания структурной модели (7.13). Модель имеет две эндогенные переменные (Y, P) и одну экзогенную (T).

Было показано, что исходная модель точно идентифицируема, и поэтому для оценки ее структурных коэффициентов используем КМНК.

Разрешая исходную систему относительно Y, P, получим приведенную систему

$$\begin{cases} P = \alpha' + \beta' T + \nu_P, \\ Y = \delta' + \varepsilon' T + \nu_Y, \end{cases}$$

где

$$\alpha' = \frac{\alpha - \delta}{\varepsilon - \beta}, \quad \beta' = \frac{\varepsilon}{\varepsilon - \beta} \quad \delta' = \frac{\varepsilon\alpha - \delta\beta}{\varepsilon - \beta}, \quad \varepsilon' = \frac{\varepsilon\beta}{\varepsilon - \beta}. \tag{7.15}$$

Пусть имеются следующие наблюдения:

T	0	2	5	8	10	12	14
P	40	42	43	44	45	48	49
Y	70	68	63	61	60	56	52

Оцененные уравнения приведенной системы, полученные по выборочным данным с использованием МНК, есть

$$\begin{cases} \hat{P} = 40 + 0{,}6T, \\ \hat{Y} = 70 - 1{,}2T, \end{cases}$$

т.е. оценки $\alpha' = 40$, $\beta' = 0{,}6$, $\delta' = 70$, $\varepsilon' = -1{,}2$.

Тогда соотношения (7.15) имеют вид

$$\frac{\alpha - \delta}{\varepsilon - \beta} = 40, \quad \frac{\varepsilon}{\varepsilon - \beta} = 0{,}6, \quad \frac{\varepsilon\alpha - \delta\beta}{\varepsilon - \beta} = 70, \quad \frac{\varepsilon\beta}{\varepsilon - \beta} = -1{,}2.$$

Отсюда получаем следующие оценки структурных коэффициентов:

$$\alpha = 150, \quad \beta = -2, \quad \delta = -50, \quad \varepsilon = 3.$$

Перейти от приведенной формы модели к структурной с учетом соотношения (7.14) можно также следующим образом.

Выразив T из первого уравнения приведенной формы в виде $T = \dfrac{P - 40}{0{,}6}$ и подставив его во второе, получим $\hat{Y} = 150 - 2P$, т.е. $\alpha = 150$, $\beta = -2$.

Выразив T из первого уравнения приведенной формы в виде $T = \frac{40 - P_1}{0,4}$, где $P_1 = P - T$, и подставив его во второе, получим $\hat{Y} = -50 + 3P_1$, т.е. $\delta = -50, \quad \varepsilon = 3$.

7.3. АНАЛИЗ МЕТОДОВ ОЦЕНИВАНИЯ

Приступать к оцениванию того или иного структурного уравнения системы имеет смысл после того, как установлена его идентифицируемость. Для установления идентифицируемости используется метод ИП.

Для решения *точно идентифицируемого* уравнения применяется КМНК, а для решения *сверхидентифицируемого* уравнения — ДМНК.

Сформулируем основные этапы указанных методов.

Этапы КМНК:

1. Структурная модель преобразуется в приведенную форму.
2. Для каждого приведенного уравнения обычным МНК оцениваются приведенные коэффициенты.
3. Оценки приведенных коэффициентов преобразуются в оценки параметров структурных уравнений.

Этапы ДМНК:

1. На основе приведенной формы модели получают для сверхидентифицируемого уравнения теоретические (расчетные) значения эндогенных переменных, содержащихся в правой части уравнения.
2. Подставляя теоретические значения эндогенных переменных вместо их фактических значений в сверхидентифицируемое уравнение и применяя обычный МНК, определяют его структурные коэффициенты.

Метод называется двухшаговым, так как МНК используется дважды: при нахождении теоретических значений эндогенных переменных из приведенной формы модели и при определении структурных коэффициентов по теоретическим значениям эндогенных переменных и исходным данным экзогенных переменных.

Сверхидентифицируемая структурная модель может быть двух типов:

- все уравнения системы сверхидентифицируемы;
- система содержит как сверхидентифицируемые, так и точно идентифицируемые уравнения.

Если все уравнения системы *сверхидентифицируемы*, то для оценки структурных коэффициентов каждого уравнения используется ДМНК. Если в системе есть *точно идентифицируемые* уравнения, то структурные коэффициенты по ним находятся из системы приведенных уравнений.

Для точно идентифицируемых уравнений ДМНК дает тот же результат, что и КМНК.

Пример 7.6. Рассмотрим следующую идентифицируемую эконометрическую модель с двумя эндогенными (y_1, y_2) и двумя экзогенными (x_1, x_2) переменными:

$$\begin{cases} y_1 = \alpha_1 + \beta_{12} y_2 + \alpha_{11} x_1 + \varepsilon_1, \\ y_2 = \alpha_2 + \beta_{21} y_1 + \alpha_{22} x_2 + \varepsilon_2. \end{cases} \tag{7.16}$$

Имеются следующие выборочные данные (усл. ед.):

y_1	y_2	x_1	x_2
2	5	1	3
3	6	2	1
4	7	3	2
5	8	2	5
6	5	4	6

Для точно идентифицируемой структурной модели применим КМНК.

Приведенная форма модели:

$$\begin{cases} y_1 = \delta_1 + \delta_{11} x_1 + \delta_{12} x_2 + \nu_1, \\ y_2 = \delta_2 + \delta_{21} x_1 + \delta_{22} x_2 + \nu_2. \end{cases}$$

Оцененные уравнения приведенной системы, полученные по выборочным данным с использованием МНК, есть

$$\begin{cases} \hat{y}_1 = 0{,}685 + 0{,}8524 x_1 + 0{,}3733 x_2, \\ \hat{y}_2 = 6{,}393 - 0{,}0724 x_1 - 0{,}0056 x_2. \end{cases} \tag{7.17}$$

Перейдем от приведенной формы модели к структурной. Для этой цели из первого уравнения приведенной формы надо исключить x_2, выразив его из второго уравнения:

$$x_2 = \frac{6{,}393 - 0{,}0724 x_1 - y_2}{0{,}0056},$$

и подставить в первое, а из второго уравнения следует исключить x_1, выразив его из первого уравнения:

$$x_1 = \frac{y_1 - 0{,}685 - 0{,}3733x_2}{0{,}8524},$$

и подставить во второе.

В результате получим следующую структурную форму модели:

$$\begin{cases} y_1 = 426{,}91 - 66{,}96y_2 - 3{,}97x_1 + \varepsilon_1, \\ y_2 = 6{,}45 - 0{,}085y_1 + 0{,}026x_2 + \varepsilon_2. \end{cases} \qquad (7.18)$$

Покажем, что для точно идентифицируемых уравнений ДМНК дает тот же результат, что и КМНК.

Из уравнений (7.17) можно найти расчетные значения эндогенных переменных $\hat{y}_1, \hat{y}_2$. Подставляя их вместо фактических y_1, y_2 в правую часть структурной модели (7.16) и применяя обычный МНК к каждому уравнению модели, получим тот же результат, что и при КМНК.

Расчетные данные для использования ДМНК приведены ниже:

y_1	$\hat{y}_2$	x_1
2	6,303	1
3	6,242	2
4	6,164	3
5	6,220	2
6	6,070	4

y_2	$\hat{y}_1$	x_2
5	2,657	3
6	2,763	1
7	3,989	2
8	4,256	5
5	6,334	6

Пример 7.7. В идентифицируемой модели (7.16) примера 7.6 наложим ограничение на ее параметры $\beta_{12} = \alpha_{11}$, тогда придем к модели

$$\begin{cases} y_1 = \alpha_1 + \beta_{12}(y_2 + x_1) + \varepsilon_1, \\ y_2 = \alpha_2 + \beta_{21}y_1 + \alpha_{22}x_2 + \varepsilon_2. \end{cases} \qquad (7.19)$$

В результате *первое* уравнение стало *сверхидентифицируемым*: $G = 1\ (y_1)$, $D = 1\ (x_2)$ и $D > G - 1$.

Второе уравнение не изменилось и является *точно идентифицируемым*: $G = 2\ (y_1, y_2)$, $D = 1\ (x_1)$ и $D = G - 1$.

При использовании тех же данных, что и в примере 7.6, получим ту же систему приведенных уравнений (7.17).

Для определения структурных коэффициентов второго, точно идентифицируемого уравнения системы (7.19) применяем КМНК.

Его структурная форма, найденная из системы приведенных уравнений, та же, что и в примере 7.6.

Для определения структурных коэффициентов первого, сверхидентифицируемого уравнения системы (7.19) используем ДМНК. На основе второго уравнения приведенной системы (7.17) находим расчетные значения $\hat{y}_2$ эндогенной переменной. Подставляя их вместо фактических y_2 в первое уравнение системы (7.19) и применяя обычный МНК, получим решение поставленной задачи.

Исходные данные для использования ДМНК следующие:

y_1	x_1	$\hat{y}_2$	$\hat{y}_2 + x_1$
2	1	6,303	7,303
3	2	6,242	8,242
4	3	6,164	9,164
5	2	6,220	8,220
6	4	6,070	10,070

Окончательно рассматриваемая система уравнений составит

$$\begin{cases} y_1 = -6,693 + 1,244(y_2 + x_1), \\ y_2 = 6,45 - 0,085y_1 + 0,026x_2. \end{cases}$$

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие существуют способы построения систем уравнений? Чем они отличаются друг от друга?
2. Как связаны между собой структурная и приведенная формы модели?
3. В чем состоят проблемы идентификации модели и какие условия идентификации (необходимые и достаточные) вы знаете?
4. Какова суть косвенного метода наименьших квадратов?
5. В каких случаях используется двухшаговый метод наименьших квадратов? Каково его содержание?
6. Что представляют собой мультипликаторные модели кейнсианского типа? Как интерпретируются коэффициенты приведенной формы такой модели?
7. Как строится структурная модель спроса и предложения?

ПРИЛОЖЕНИЕ
(математико-статистические таблицы)

КРИТИЧЕСКИЕ ЗНАЧЕНИЯ *t*-КРИТЕРИЯ СТЬЮДЕНТА ПРИ УРОВНЯХ ЗНАЧИМОСТИ 0,10; 0,05; 0,01

(*df* — число степеней свободы)

df	α			*df*	α		
	0,10	,0,05	0,01		0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

КРИТИЧЕСКИЕ ЗНАЧЕНИЯ F-КРИТЕРИЯ ФИШЕРА ПРИ УРОВНЕ ЗНАЧИМОСТИ 0,05

(*df* — число степеней свободы)

df_2	df_1										
	1	2	3	4	5	6	7	8	9	10	∞
1	161	200	216	225	230	234	237	239	241	242	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5
3	10,1	9,6	9,3	9,2	9,0	8,9	8,9	8,8	8,8	8,8	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	6,1	6,0	6,0	6,0	5,6
5	6,6	5,8	5,4	5,2	5,0	4,9	4,9	4,9	4,8	4,7	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,2	4,1	4,1	4,1	3,7
7	5,6	4,7	4,3	4,1	4,0	3,9	3,8	3,7	3,7	3,6	3,2
8	5,3	4,5	4,1	3,8	3,7	3,6	3,5	3,4	3,4	3,3	2,9
9	5,1	4,3	3,9	3,6	3,5	3,4	3,3	3,2	3,2	3,1	2,7
10	5,0	4,1	3,7	3,5	3,3	3,2	3,1	3,1	3,0	3,0	2,5
11	4,8	4,0	3,6	3,4	3,2	3,1	3,0	3,0	2,9	2,9	2,4
12	4,7	3,9	3,5	3,3	3,1	3,0	2,9	2,9	2,8	2,8	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,8	2,8	2,7	2,7	2,2
14	4,6	3,7	3,3	3,1	3,0	2,9	2,8	2,7	2,7	2,6	2,1
15	4,5	3,7	3,3	3,1	2,9	2,8	2,7	2,6	2,6	2,6	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,5	2,0
17	4,4	3,6	3,2	3,0	2,8	2,7	2,6	2,6	2,5	2,5	2,0
18	4,4	3,6	3,2	3,0	2,8	2,7	2,6	2,5	2,5	2,4	1,9
19	4,4	3,5	3,1	2,9	2,7	2,6	2,6	2,5	2,4	2,4	1,9
20	4,3	3,5	3,1	2,9	2,7	2,6	2,5	2,5	2,4	2,4	1,9
21	4,3	3,5	3,1	2,8	2,7	2,6	2,5	2,4	2,4	2,3	1,8
22	4,3	3,4	3,0	2,8	2,7	2,5	2,5	2,4	2,4	2,3	1,8
23	4,3	3,4	3,0	2,8	2,6	2,5	2,5	2,4	2,3	2,3	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3	1,8
25	4,2	3,4	3,0	2,8	2,6	2,5	2,4	2,3	2,3	2,2	1,7
26	4,2	3,4	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2	1,7
27	4,2	3,3	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2	1,7
28	4,2	3,3	3,0	2,7	2,6	2,4	2,4	2,3	2,2	2,2	1,7
29	4,2	3,3	2,9	2,7	2,5	2,4	2,4	2,3	2,2	2,2	1,6
30	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2	1,6
40	4,1	3,2	2,8	2,6	2,4	2,3	2,2	2,2	2,1	2,1	1,5
50	4,0	3,2	2,8	2,6	2,4	2,3	2,2	2,1	2,1	2,0	1,4
60	4,0	3,1	2,8	2,5	2,4	2,2	2,2	2,1	2,0	2,0	1,4
100	3,9	3,1	2,7	2,5	2,3	2,2	2,1	2,0	2,0	1,9	1,3
∞	3,8	3,0	2,6	2,4	2,2	2,1	2,0	1,9	1,9	1,8	1,0

КРИТИЧЕСКИЕ ЗНАЧЕНИЯ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ ПРИ УРОВНЯХ ЗНАЧИМОСТИ 0,05; 0,01

(*df* — число степеней свободы)

df	$\alpha = 0{,}05$	$\alpha = 0{,}01$	*df*	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	0,9969	0,9998	17	0,4555	0,5751
2	0,9950	0,9900	18	0,4438	0,5614
3	0,8783	0,9587	19	0,4329	0,5487
4	0,8114	0,9172	20	0,4227	0,5368
5	0,7545	0,8745	25	0,3809	0,4869
6	0,7067	0,8343	30	0,3494	0,4487
7	0,6664	0,7977	35	0,3246	0,4182
8	0,6319	0,7640	40	0,3044	0,3932
9	0,6021	0,7348	45	0,2875	0,3721
10	0,5760	0,7079	50	0,2732	0,3541
11	0,5529	0,6835	60	0,2500	0,3248
12	0,5324	0,6614	70	0,2319	0,3017
13	0,5139	0,6411	80	0,2172	0,2830
14	0,4973	0,6226	90	0,2050	0,2673
15	0,4821	0,6055	100	0,1946	0,2540
16	0,4683	0,5897			

КРИТИЧЕСКИЕ ЗНАЧЕНИЯ КОЭФФИЦИЕНТОВ АВТОКОРРЕЛЯЦИИ ПРИ УРОВНЯХ ЗНАЧИМОСТИ 0,05; 0,01

(*n* — объем выборки)

n	Положительные значения		Отрицательные значения	
	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	0,253	0,297	–0,753	–0,798
6	0,345	0,447	–0,708	–0,863
7	0,370	0,510	–0,674	–0,799
8	0,371	0,531	–0,625	–0,764
9	0,366	0,533	–0,593	–0,737
10	0,360	0,525	–0,564	–0,705
11	0,353	0,515	–0,539	–0,679
12	0,348	0,505	–0,516	–0,655
13	0,341	0,495	–0,497	–0,634
14	0,335	0,485	–0,479	–0,615
15	0,328	0,475	–0,462	–0,597
20	0,299	0,432	–0,399	–0,524

ЗНАЧЕНИЯ d_1 И d_2 КРИТЕРИЯ ДАРБИНА — УОТСОНА ПРИ УРОВНЕ ЗНАЧИМОСТИ 0,05

(n — число наблюдений,
m — число объясняющих переменных)

n	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
6	0,61	1,40								
7	0,70	1,36	0,47	1,90						
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1,05	1,35	0,91	1,55	0,77	1,78				
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Список рекомендуемой литературы

1. *Доугерти К.* Введение в эконометрику: Пер. с англ. — 2-е изд. — М.: ИНФРА-М, 2007.
2. *Замков О.О.* Эконометрические методы в макроэкономическом анализе: Курс лекций. — М.: ГУ ВШЭ, 2001.
3. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика: Начальный курс. — М.: Дело, 1998.
4. *Мандас А.Н.* Эконометрика: Учеб. пособие. — СПб.: Питер, 2001.
5. *Просветов Г.И.* Эконометрика: Учеб.-метод. пособие. — М.: РДЛ, 2005.
6. Эконометрика / Под ред. И.И. Елисеевой. — М.: Финансы и статистика, 2001.

Оглавление