Министерство образования и науки Российской Федерации

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

ФАКУЛЬТЕТ ДИСТАНЦИОННОГО ОБУЧЕНИЯ (ФДО)

И. В. Потахова

ЭКОНОМЕТРИКА

Учебное пособие

Томск
2015

УДК 330.43(075.8)
ББК 65в631я73
П 640

Рецензенты:
Тарасенко В. Ф., докт. техн. наук, профессор кафедры теоретической кибернетики Томского государственного университета;
Лепихина З. П., канд. техн. наук, доцент кафедры автоматизации обработки информации ТУСУРа.

Потахова И. В.
П 640 Эконометрика : учебное пособие / И. В. Потахова. — Томск : факультет дистанционного обучения ТУСУРа, 2015. — 110 с.

Эконометрика как учебная дисциплина включена в основную образовательную программу подготовки экономистов, определяемую Федеральным государственным образовательным стандартом. Это определено тем, что современные экономические теории и исследования, требуют от экономистов свободного владения математическим аппаратом изучения статистических данных.

Целью изучения учебной дисциплины «Эконометрика» является овладение современными эконометрическими методами анализа конкретных экономических данных.

Данное пособие рассчитано в первую очередь на студентов экономических специальностей, которые изучают эконометрику. В пособии рассмотрены вопросы по основным разделам эконометрики: парная и множественная регрессия, системы эконометрических уравнений и временные ряды.

Для изучения эконометрики необходимо знание статистики и математики. Особенно важно владение корреляционным, регрессионным и дисперсионным анализом, а также методами проверки статистических гипотез.

Пособие предназначено для самостоятельного изучения дисциплины «Эконометрика» студентами заочной формы обучения, а также студентами, обучающихся с применением дистанционных образовательных технологий.

УДК 330.43(075.8)
ББК 65в631я73

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Эконометрика — наука, изучающая количественные и качественные экономические взаимосвязи с помощью математических и статистических методов и моделей.

Эконометрика как научная дисциплина зародилась и получила развитие на основе слияния экономической теории, математической экономики, экономической и математической статистики.

В настоящее время эконометрические методы широко применяются в количественном анализе фирм, фондового и товарного рынков, макроэкономических моделей. Применение эконометрики поднимает получаемые результаты на качественно новый уровень, поскольку в этом случае каждый вывод подтверждается точными количественными расчетами и конкретным значением критерия, который оценивает справедливость выдвинутой гипотезы.

Мощным инструментом эконометрических исследований является аппарат математической статистики. Как следствие, эконометрические методы — это, прежде всего, методы статистического анализа конкретных экономических данных.

Центральной проблемой любого эконометрического исследования является построение эконометрической модели.

В общем случае процедуру построения эконометрической модели можно разделить на несколько взаимосвязанных между собой этапов. Основные среди них имеют следующее содержание [3].

1. Анализ специфических свойств рассматриваемых явлений и процессов и обоснование класса моделей, наиболее подходящих для их описания.

   Целями этого этапа являются:

   1.1 Выбор рационального состава включаемых в модель переменных и определение количественных характеристик, отражающих их уровни в прошлые периоды времени (на однородных объектах некоторой совокупности — территориях, предприятиях и т. п.).

   1.2 Обоснование типа и формы модели, выражаемой математическим уравнением (системой уравнений), связывающим включенные в модель переменные.

2. Оценка параметров выбранного варианта модели на основании исходных данных, выражающих уровни показателей (переменных) в различные моменты времени или на совокупности однородных объектов.

3. Проверка качества построенной модели и обоснование вывода о целесообразности ее использования в ходе дальнейшего эконометрического исследования.
4. При выводе о нецелесообразности использования построенной эконометрической модели в дальнейших исследованиях следует вернуться к первому (или какому-либо другому этапу) и попытаться построить более качественную модификацию модели (другой вариант модели).

В данном пособии рассматриваются базовые эконометрические методы и модели. Пособие состоит из введения, основного учебного материала (гл. 1–6), приложений.

Во введении дается определение эконометрики, показано ее место в образовательной программе подготовки бакалавра.

В первой главе рассматриваются классические модели парной регрессии. На примере парной линейной регрессии показывается фундаментальный метод оценки параметров регрессии — метод наименьших квадратов. Вводятся понятия точности оценок коэффициентов регрессии, качества уравнения регрессии.

Во второй главе рассматриваются модели линейной множественной регрессии. Описывается применение метода наименьших квадратов для нахождения параметров уравнения множественной линейной регрессии. Рассматривается схема выполнения анализа уравнения множественной регрессии: оценка качества уравнения регрессии, оценка значимости коэффициентов регрессии и уравнения в целом.

В третьей главе исследуются причины и последствия невыполнимости предпосылок применения метода наименьших остатков: постоянство дисперсии остатков, отсутствие зависимости между случайными отклонениями (автокорреляция остатков). Приводятся способы обнаружения нарушений данных предпосылок. Рассматривается схема нахождения оценок регрессионной модели с помощью обобщенного метода наименьших квадратов.

В четвертой главе разбираются вопросы построения регрессионных моделей, в которых используются неколичественные факторы.

В пятой главе анализируются системы эконометрических уравнений. Рассматриваются методы оценки параметров систем эконометрических уравнений. Приводится схема исследования систем уравнений на возможность идентификации.

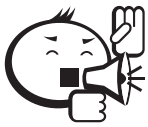
В шестой главе вводится понятие эконометрических моделей, построенных на основе временных рядов. Выделяются основные составляющие временного ряда. Приводится пример построения модели временного ряда.

Для изучения и освоения материала, изложенного в пособии, студенту достаточно знаний курсов математики, теории вероятностей и математической статистики. При изложении материала приводятся задачи с решениями. В заключении каждой главы имеются вопросы для самопроверки. В приложениях приводятся таблицы, необходимые для выполнения практических расчетов.

Соглашения, принятые в книге

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

Эта пиктограмма означает определение или новое понятие.



Эта пиктограмма означает внимание. Здесь выделена важная информация, требующая акцента на ней. Автор здесь может поделиться с читателем опытом, чтобы помочь избежать некоторых ошибок.

Пример

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

Контрольные вопросы по главе

Глава 1

ПАРНАЯ РЕГРЕССИЯ

1.1 Понятие парной регрессии

В эконометрике широко используется регрессионный анализ как метод выявления уравнения связи между зависимыми и независимыми переменными, наилучшим способом дающий оценку истинного соотношения между этими переменными.

Если переменные обозначить X и Y, то зависимость вида:

$$f(x) = M\left(\frac{Y}{X}\right)$$

называется функцией регрессии X на Y, где X — независимые (объясняющие) переменные (регрессоры, факторы); Y — зависимая (объясняемая) переменная.

Регрессия *— зависимость между независимыми (объясняющими) переменными и условным математическим ожиданием зависимой (объясняемой) переменной.*

При рассмотрении двух случайных величин говорят о парной регрессии. Зависимость такого типа в общем случае выражается уравнением:

$$y = f(x) + \varepsilon.$$

В уравнении можно выделить две части:

- систематическую $\widehat{y}_x = f(x)$, где $\widehat{y}_x$ характеризует некоторое среднее значение y для данного значения x;
- случайную ε, характеризующую отклонение реального значения результативного признака y от теоретического $\widehat{y}_x$, найденного по уравнению регрессии для данного значения x.

Среди причин обязательного присутствия в регрессионных моделях случайного фактора (отклонения) можно выделить следующие.

1. Невключение в модель всех объясняющих переменных. Любая регрессионная (в частности, эконометрическая) модель является упрощением реальной ситуации. Последняя всегда представляет собой сложнейшее переплетение различных факторов, многие из которых в модели не учитываются, что порождает отклонение реальных значений зависимой переменной от ее модельных значений. Например, спрос (Q) на товар определяется его ценой (P), ценой на товары-заменители (P_s), ценой на дополняющие товары (P_c), доходом (I) потребителей, их количеством (N), вкусами (T), ожиданиями (W) и т. д. Безусловно, перечислить все объясняющие переменные здесь практически невозможно. К примеру, не учтены такие факторы, как традиции, национальные или религиозные особенности, географическое положение региона, погода и многие другие, влияние которых приведет к некоторым отклонениям реальных наблюдений от модельных. Эти влияния выражаются через случайный член ε. Тогда модель спроса можно записать в виде функции:

$$Q = f(P, P_s, P_c, I, N, T, W, \varepsilon).$$

Проблема еще и в том, что никогда заранее неизвестно, какие факторы при создавшихся условиях действительно являются определяющими, а какими можно пренебречь. Например, в ряде случаев учесть непосредственно какой-то фактор нельзя в силу невозможности получения по нему статистических данных.

2. Неправильный выбор функциональной формы модели. Из-за слабой изученности исследуемого процесса либо из-за его переменчивости может быть неверно подобрана функция, его моделирующая. Это, безусловно, скажется на отклонении модели от реальности, что отразится на величине случайного члена. Кроме того, неверным может быть подбор объясняющих переменных.

3. Агрегирование переменных. Во многих моделях рассматриваются зависимости между факторами, которые сами представляют сложную комбинацию других, более простых переменных. Например, при рассмотрении в качестве зависимой переменной совокупного спроса проводится анализ зависимости, в которой объясняемая переменная является сложной композицией индивидуальных спросов, оказывающих на нее определенное влияние помимо факторов, учитываемых в модели. Это может оказаться причиной отклонения реальных значений от модельных.

4. Ошибки измерений. Какой бы качественной ни была модель, ошибки измерений переменных отразятся на несоответствии модельных значений эмпирическим данным, что также отразится на величине случайного члена.

5. Ограниченность статистических данных. Зачастую строятся модели, выражаемые непрерывными функциями. Но для этого используется набор данных, имеющих дискретную структуру. Это несоответствие находит свое выражение в случайном отклонении.

6. Непредсказуемость человеческого фактора. Эта причина может «испортить» самую качественную модель. Действительно, при правильном выборе формы модели, скрупулезном подборе объясняющих переменных все равно невозможно спрогнозировать поведение каждого индивидуума.

Решение задачи построения качественного уравнения регрессии, соответствующего эмпирическим данным и целям исследования, является сложным и многоступенчатым процессом. Его можно разбить на три этапа:

1) выбор формулы уравнения регрессии;
2) определение параметров выбранного уравнения;
3) анализ качества уравнения и проверка адекватности уравнения эмпирическим данным, совершенствование уравнения.

Выбор формулы связи переменных называется *спецификацией* уравнения регрессии. В парной регрессии выбор вида математической функции $\widehat{y}_x = f(x)$ может быть осуществлен тремя методами:

1) графическим;
2) аналитическим, т. е. исходя из теории изучаемой взаимосвязи;
3) экспериментальным.

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Он основан на визуальном анализе поля корреляции (рис. 1.1).

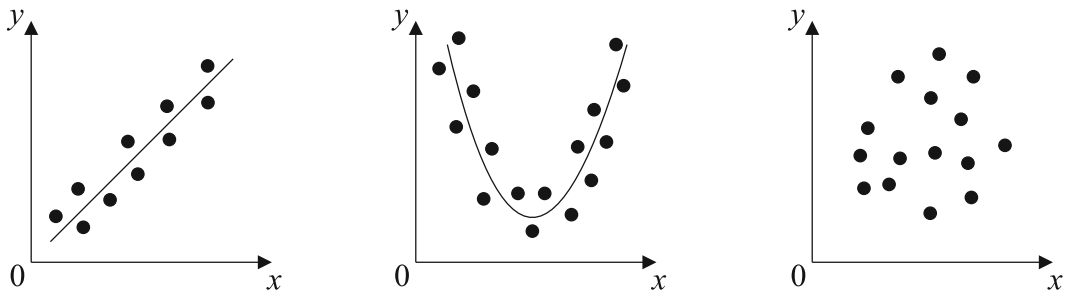


Рис. 1.1 – Поле корреляции

Значительный интерес представляет аналитический метод выбора типа уравнения регрессии. Он основан на изучении материальной природы связи исследуемых признаков.

При обработке информации на компьютере выбор вида уравнения регрессии обычно осуществляется экспериментальным методом, т. е. путем сравнения величины остаточной дисперсии $\sigma^2_{\text{ост}}$, рассчитанной при разных моделях.

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии $\widehat{y}_x = f(x)$, то фактические значения результативного признака совпадают с теоретическими $y = \widehat{y}_x$, т. е. они полностью обусловлены влиянием фактора x. В этом случае остаточная дисперсия $\sigma^2_{\text{ост}} = 0$.

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии факторов. Иными словами, наблюдаются отклонения фактических данных от теоретических $(y - \widehat{y}_x)$. Величина этих отклонений лежит в основе расчета остаточной дисперсии:

$$\sigma^2_{\text{ост}} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Чем меньше величина остаточной дисперсии, тем меньше влияние не учитываемых в уравнении регрессии факторов и тем лучше уравнение регрессии подходит к исходным данным.

Считается, что число наблюдений должно в 7–8 раз превышать число рассчитываемых параметров при переменной x. Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений, ибо каждый параметр при x должен рассчитываться хотя бы по 7 наблюдениям. Следовательно, если мы выбираем параболу второй степени $\widehat{y}_x = a + b \cdot x + c \cdot x^2$, то требуется объем информации уже не менее 14 наблюдений.

1.2 Линейная модель парной регрессии

Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров. Кроме того, построенное линейное уравнение может служить начальной точкой эконометрического анализа.

Линейная регрессия сводится к нахождению уравнения вида:

$$y = a + b \cdot x + \varepsilon.$$

На практике построение линейной регрессии сводится к оценке параметров уравнения $\widehat{y}_x = a + b \cdot x$.

1.2.1 Вычисление коэффициентов уравнения линейной регрессии

Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\widehat{y}_x$ минимальна:

$$\sum_{i=1}^{n} (y_i - \widehat{y}_{x_i})^2 = \sum_{i=1}^{n} \varepsilon_i^2 \to \min.$$

То есть из всего множества линий на графике линия регрессии выбирается так, чтобы сумма квадратов расстояний по вертикали между точками наблюдений и линией регрессии была бы минимальной (рис. 1.2).

Как известно из курса математического анализа, чтобы найти минимум функции, необходимо вычислить частные производные по каждому из параметров (в нашем случае это a и b) и приравнять их к нулю. Если обозначить $\sum_i \varepsilon_i^2$ через $S(a, b)$, то можно записать:

$$S(a, b) = \sum (y - a - b \cdot x)^2;$$

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2 \cdot \sum (y - a - b \cdot x) = 0, \\ \dfrac{\partial S}{\partial b} = -2 \cdot \sum x \cdot (y - a - b \cdot x) = 0. \end{cases}$$

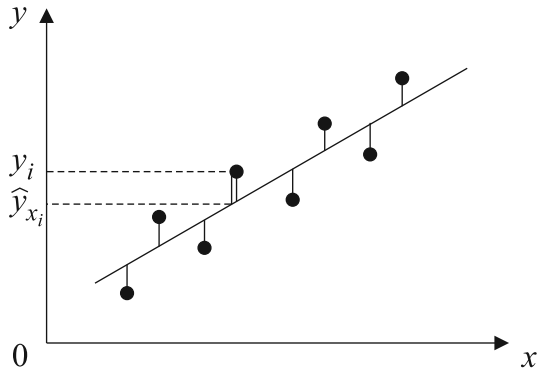


Рис. 1.2 – Линия регрессии с минимальной дисперсией остатков

После несложных преобразований получается следующая система линейных уравнений для оценки параметров a и b:

$$\begin{cases} a \cdot n + + b \cdot \sum x = \sum y, \\ a \cdot \sum x + + b \cdot \sum x^2 = \sum x \cdot y. \end{cases}$$

Решение системы уравнений позволяет найти оценки параметров a и b:

$$b = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\operatorname{cov}(x, y)}{\sigma_x^2},$$

$$a = \overline{y} - b \cdot \overline{x},$$

где $\operatorname{cov}(x, y)$ — выборочное значение корреляционного момента (ковариация), определенного по формуле $\operatorname{cov}(x, y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$; σ_x^2 — выборочное значение дисперсии величины x, определяемой по формуле $\sigma_x^2 = \overline{x^2} - (\overline{x})^2$;

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \quad \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i; \quad \overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i; \quad \overline{x_i^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2.$$

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу.

Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально a — значение y при $x = 0$. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка свободного члена a не имеет смысла, т. е. параметр a может не иметь экономического содержания.

1.2.2 Исследование уравнения линейной регрессии

Оценка тесноты связи и качества модели линейной регрессии.

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции r_{xy}, который можно рассчитать по следующей формуле:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y}.$$

Линейный коэффициент корреляции находится в пределах: $-1 \leqslant r_{xy} \leqslant 1$. Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ наблюдается строгая функциональная зависимость). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2, называемый коэффициентом детерминации:

$$r_{xy}^2 = \frac{\sigma_{\text{объясн}}^2}{\sigma_y^2} = 1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2},$$

где $\sigma_y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 = \overline{y^2} - \bar{y}^2$; $\sigma_{\text{объясн}}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$; $\sigma_{\text{ост}}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака.

Соответственно величина $1 - r_{xy}^2$ характеризует долю дисперсии y, вызванную влиянием остальных, не учтенных в модели факторов.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$\overline{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_{x_i}}{y_i} \right| \cdot 100\%.$$

Средняя ошибка аппроксимации не должна превышать 8–10%. Чем выше показатель детерминации или чем ниже средняя ошибка аппроксимации, тем лучше модель описывает исходные данные.

Оценка значимости уравнения линейной регрессии и существенности параметров линейной регрессии.

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

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

Оценка значимости уравнения регрессии в целом производится на основе F-критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

Согласно основной идее дисперсионного анализа общая сумма квадратов отклонений переменной y от среднего значения $\bar{y}$ раскладывается на две части — «объясненную» и «необъясненную»:

$$\sum (y-\bar{y})^2 = \sum (\hat{y}_x-\bar{y})^2 + \sum (y-\hat{y})^2,$$

где $\sum (y-\bar{y})^2$ — общая сумма квадратов отклонений; $\sum (\hat{y}_x-\bar{y})^2$ — сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum (y-\hat{y}_x)^2$ — остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n — число наблюдений; m — число параметров при переменной x).

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}.$$

Таблица 1.1 – Схема дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum (y-\bar{y})^2$	$n-1$	$S^2_{\text{общ}} = \frac{\sum_{i=1}^{n}(y-\bar{y})^2}{n-1}$
Факторная	$\sum (\hat{y}_x-\bar{y})^2$	m	$S^2_{\text{факт}} = \frac{\sum_{i=1}^{n}(\hat{y}_x-\bar{y})^2}{m}$
Остаточная	$\sum (y-\hat{y}_x)^2$	$n-m-1$	$S^2_{\text{ост}} = \frac{\sum_{i=1}^{n}(y-\hat{y}_x)^2}{n-m-1}$

Фактическое (вычисленное) значение F-критерия Фишера сравнивается с табличным значением $F_{\text{табл}}(\alpha, k_1, k_2)$ при уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n-m-1$. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m = 1$, поэтому

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}} = \frac{\sum (\hat{y}_x-\bar{y})^2}{\sum (y-\hat{y}_x)^2} \cdot (n-2).$$

Величина F-критерия связана с коэффициентом детерминации r^2_{xy} следующим соотношением:

$$F = \frac{r^2_{xy}}{1-r^2_{xy}} \cdot (n-2).$$

В регрессионном анализе оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{S^2_{\text{ост}}}{\sum (x - \bar{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}},$$

где $S^2_{\text{ост}} = \dfrac{\sum\limits_{i=1}^{n} (y - \hat{y}_x)^2}{n - m - 1}$ — остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с t-распределением Стьюдента при $n - 2$ степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительного интервала.

Для оценки существенности коэффициента регрессии определяется фактическое значение t-критерия Стьюдента: $t_b = b/m_b$. Вычисленное значение (t_b) сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы $(n-2)$. Здесь проверяется нулевая гипотеза H_0: $b = 0$, предполагающая несущественность статистической связи между y и x. Если $t_b > t_{\text{табл}}(\alpha, n-2)$, то гипотеза H_0: $b = 0$ должна быть отклонена, а статистическая связь y и x считается установленной. В случае $t_b < t_{\text{табл}}(\alpha, n - 2)$ нулевая гипотеза не может быть отклонена, и влияние y на x признается несущественным.

Интервальная оценка (доверительный интервал) для коэффициента b с надежностью (доверительной вероятностью), равной γ, определяется выражением: $b \pm t_{\text{табл}} \cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора $x(b > 0)$, уменьшение результативного признака y при увеличении признака-фактора $x(b < 0)$ или его независимость от независимой переменной $x(b = 0)$ (рис. 1.3), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например $-1{,}5 \leqslant b \leqslant 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит как положительные, так и отрицательные величины, что противоречит виду рассматриваемой зависимости между двумя переменными.

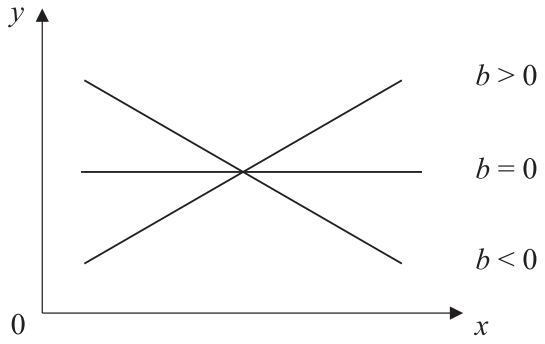


Рис. 1.3 – Наклон линии регрессии в зависимости от значения параметра b

Процедура оценивания существенности параметра a не отличается от рассмотренной выше для коэффициента регрессии. Стандартная ошибка параметра a определяется по формуле:

$$m_a = \frac{S_{\text{ост}} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}}{\sigma_x \cdot n}.$$

Вычисляется t-критерий: $t_a = a/m_a$. Его величина сравнивается с табличным значением критерия Стьюдента при $n - 2$ степенях свободы.

Интервальная оценка (доверительный интервал) для коэффициента a с надежностью (доверительной вероятностью), равной $\gamma = 1 - \alpha$, определяется выражением: $a \pm t_{\text{табл}} \cdot m_a$.

Построение интервальных оценок для функции парной линейной регрессии.

Интервальная оценка (доверительный интервал) для вычисленного значения $\widehat{y}_i$ при заданном значении x_i с надежностью (доверительной вероятностью), равной $\gamma = 1 - \alpha$, определяется выражением

$$\widehat{y}_i \pm t_{\text{табл}} \cdot m_{\widehat{y}_i}.$$

Стандартная ошибка вычисленного значения $\widehat{y}_i$ определяется по формуле:

$$m_{\widehat{y}_i} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \cdot \sigma_x^2}}.$$

В прогнозных расчетах по уравнению регрессии определяется предсказываемое $\widehat{y}_p$ значение как точечный прогноз $\widehat{y}_x$ при $x = x_p$, т. е. путем подстановки в уравнение регрессии $\widehat{y}_x = a + b \cdot x$ соответствующего значения x. Прогнозный расчет дополняется вычислением средней ошибки прогнозируемого значения $\widehat{y}_p$, т. е. $m_{\widehat{y}_p}$, и соответственно интервальной оценкой прогнозируемого значения $\widehat{y}_p$: $\widehat{y}_i \pm t_{\text{табл}} \cdot m_{\widehat{y}_i}$.

Рассмотрим пример построения парной линейной регрессии.

Пример 1.1

Известны данные об уровне механизации работ X (%) и производительности труда Y (т/час) для 14 однотипных предприятий (табл. 1.2). Требуется оценить регрессию X на Y [2].

Таблица 1.2 – Данные наблюдений

y_i	20	24	28	30	31	33	34	37	38	40	41	43	45	48
x_i	32	30	36	40	41	47	56	54	60	55	61	67	69	76

Предположим, что связь между механизацией работ и производительностью труда линейная. Для подтверждения нашего предположения построим поле корреляции (рис. 1.4).

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим вспомогательную таблицу 1.3.

Рассчитаем параметры линейного уравнения парной регрессии.

$$\widehat{y}_x = a + b \cdot x.$$

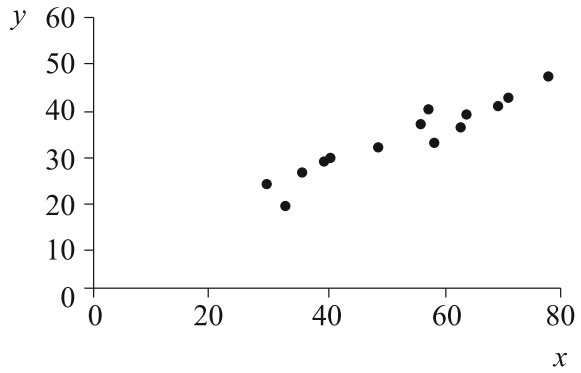


Рис. 1.4 – Поле корреляции по данным наблюдений

Таблица 1.3 – Вспомогательная таблица

	x	y	$x \cdot y$	x^2	y^2	$\widehat{y}_x$	$y - \widehat{y}_x$	$(y - \widehat{y}_x)^2$	A_i, %
1	32	20	640	1024	400	24,43	–4,43	19,61	22,14
2	30	24	720	900	576	23,34	0,66	0,43	2,75
3	36	28	1008	1296	784	26,60	1,40	1,95	4,99
4	40	30	1200	1600	900	28,78	1,22	1,50	4,08
5	41	31	1271	1681	961	29,32	1,68	2,82	5,42
6	47	33	1551	2209	1089	32,58	0,42	0,18	1,27
7	56	34	1904	3136	1156	37,47	–3,47	12,06	10,21
8	54	37	1998	2916	1369	36,39	0,61	0,38	1,66
9	60	38	2280	3600	1444	39,65	–1,65	2,71	4,33
10	55	40	2200	3025	1600	36,93	3,07	9,43	7,68
11	61	41	2501	3721	1681	40,19	0,81	0,66	1,98
12	67	43	2881	4489	1849	43,45	–0,45	0,20	1,05
13	69	45	3105	4761	2025	44,54	0,46	0,21	1,03
14	76	48	3648	5776	2304	48,34	–0,34	0,12	0,71
Итого	724	492	26907	40134	18138	492	0	52,26	69,3
Средн. знач-е	51,71	35,14	1921,93	2866,71	1295,57	—	—	—	4,95
σ	13,87	7,78	—	—	—	—	—	—	—
σ^2	192,35	60,55	—	—	—	—	—	—	—

Для этого воспользуемся формулами вычисления параметров уравнения регрессии:

$$b = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{1921{,}93 - 51{,}71 \cdot 35{,}14}{192{,}35} = 0{,}55;$$

$$a = \overline{y} - b \cdot \overline{x} = 35{,}14 - 0{,}54 \cdot 51{,}71 = 6{,}96.$$

Запишем уравнение: $\widehat{y}_x = 6{,}96 + 0{,}55 \cdot x$.

Коэффициент b уравнения регрессии показывает, что с увеличением уровня механизации работ на 1% производительность труда увеличится на 0,55 т/час.

Оценим качество уравнения регрессии.

Как было указано выше, уравнение линейной регрессии всегда дополняется показателем тесноты связи — линейным коэффициентом корреляции r_{xy}:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0{,}55 \cdot \frac{13{,}87}{7{,}78} = 0{,}98.$$

Близость коэффициента корреляции к единице указывает на тесную линейную связь между признаками.

Коэффициент детерминации $r_{xy}^2 = 0{,}961$ показывает, что уравнением регрессии объясняется 96,1% дисперсии результативного признака, а на долю прочих факторов приходится лишь 3,9%.

Средняя ошибка аппроксимации $\overline{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y}_{x_i}}{y_i} \right| \cdot 100\%$, $\overline{A} = 4{,}95\%$ говорит о хорошем качестве уравнения регрессии, т. е. свидетельствует о хорошем подборе модели к исходным данным.

Оценим значимость уравнения регрессии в целом с помощью F-критерия Фишера. Вычислим фактическое значение F-критерия:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2) = \frac{0{,}961}{1 - 0{,}961} \cdot 12 = 295{,}69.$$

Табличное значение F-критерия с числом степеней свободы $k_1 = 1$, $k_2 = 12$ и уровнем значимости $\alpha = 0{,}05$ равно 4,75. Так как $F_{\text{факт}} > F_{\text{табл}}$, то признается статистическая значимость уравнения в целом.

Для оценки *статистической значимости коэффициентов регрессии* рассчитаем t-критерий Стьюдента и доверительные интервалы каждого из показателей.

1. Вычислим остаточную дисперсию на одну степень свободы:

$$S_{\text{ост}}^2 = \frac{\sum_{i=1}^{n} (y - \widehat{y}_x)^2}{n - m - 1} = \frac{52{,}26}{14 - 1 - 1} = 4{,}36.$$

2. Вычислим значения ошибок вычисления параметров регрессии:

$$m_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{4{,}36}}{13{,}87 \cdot \sqrt{14}} = 0{,}04.$$

$$m_a = \frac{S_{\text{ост}} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}}{\sigma_x \cdot n} = \frac{\sqrt{4{,}36 \cdot 40\ 134}}{13{,}87 \cdot 14} = 2{,}15.$$

3. Вычислим фактические значения t-статистик:

$$t_b = \frac{b}{m_b} = \frac{0{,}55}{0{,}04} = 13{,}75; \quad t_a = \frac{a}{m_a} = \frac{6{,}96}{2{,}15} = 3{,}24.$$

Табличное значение t-критерия Стьюдента при $\alpha = 0{,}05$ и числе степеней свободы $k = 14 - 2$ равно 2,18. Так как $t_a > t_{\text{табл}}(\alpha, n-2)$, $t_b > t_{\text{табл}}(\alpha, n-2)$, то признаем статистическую значимость параметров регрессии.

4. Рассчитаем доверительные интервалы для параметров регрессии a и b:

$$a \pm t_{\text{табл}} \cdot m_a \quad \text{и} \quad b \pm t_{\text{табл}} \cdot m_b.$$

В результате получим:

$$a \in [2{,}27;\ 11{,}64], \quad b \in [0{,}46;\ 0{,}64].$$

Найдем *прогнозное значение* результативного фактора $\widehat{y}_p$ при значении признака-фактора, составляющем 150% от среднего уровня $x_p = 1{,}5 \cdot \bar{x} = 1{,}5 \cdot 51{,}71 = 77{,}57$, т. е. найдем производительность труда, если уровень механизации составит 77,57%:

$$\widehat{y}_p = 6{,}96 + 0{,}55 \cdot 77{,}57 = 49{,}6 \text{ (т/час)}.$$

Результат вычисления показывает, что если уровень механизации составит 77,57%, то производительность труда будет равна 49,6 т/час.

Найдем доверительный интервал прогноза. Средняя ошибка прогноза:

$$m_{\widehat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sigma_x^2}} = \sqrt{4{,}36 + \frac{1}{14} + \frac{(77{,}57 - 51{,}71)^2}{14 \cdot 192{,}35}} = 2{,}16,$$

а доверительный интервал $\widehat{y}_p \pm t_{\text{табл}} \cdot m_{\widehat{y}_p}$:

$$44{,}89 < \widehat{y}_p < 54{,}31.$$

Отобразим на одном графике исходные данные и линию регрессии (рис. 1.5).

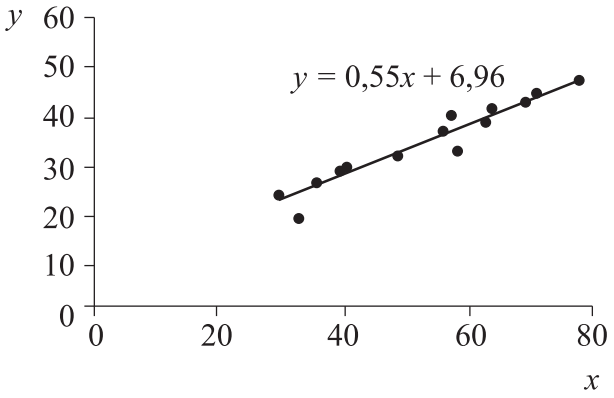


Рис. 1.5 – Графическое представление уравнения регрессии

1.2.3 Нелинейные модели регрессии

Зависимости между экономическими показателями не всегда можно выразить линейными функциями. Например, жизненный цикл товара можно описать в виде выпуклой параболы, ремаркетинг (оживление спроса в случае его падения) — в виде вогнутой параболы, а зависимость между объемом выпуска продукции и затратами капитала и труда — степенной функцией. Широко применяются обратная и экспоненциальные модели. Равносторонняя гипербола может быть использована для характеристики связи удельных расходов сырья, материалов, топлива от объема выпускаемой продукции; времени обращения товаров от величины товарооборота; процента прироста заработной платы от уровня безработицы (кривая Филлипса); расходов на непродовольственные товары от доходов или общей суммы расходов (кривая Э. Энгеля) и в других случаях.

Оценка параметров нелинейной модели.

Различают два класса нелинейных регрессий:

1. Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам.
2. Регрессии, нелинейные по оцениваемым параметрам.

Построение и анализ нелинейных моделей имеют свою специфику. В рамках данного курса ограничимся рассмотрением нелинейных моделей, допускающих сведение их к линейному типу.

Регрессии, нелинейные относительно объясняющих (независимых) переменных.

Примерами регрессий, нелинейных по объясняющим переменным, но линейных по оцениваемым параметрам, могут служить:

- полиномы различных степеней:

$$y = a + b_1 \cdot x + b_2 \cdot x^2 + \ldots + b_n \cdot x^n + \varepsilon;$$

- равносторонние гиперболы (обратная): $y = a + (b/x) + \varepsilon$.
- полулогарифмическая: $y = a + b \cdot \ln(x) + \varepsilon$.

Среди полиномов наиболее часто используются полиномы низших порядков (второго, реже — третьего). Это связано с тем, что:

- во-первых, чем выше степень полинома, тем больше изгибов имеет кривая и, следовательно, тем менее однородны исходные данные;
- во-вторых, увеличение степени полинома означает включение в модель дополнительной переменной и, следовательно, возможные сложности при статистической оценке модели.

При оценке параметров регрессий, нелинейных по объясняющим переменным, используется метод замены переменных. Суть его состоит в замене нелинейных объясняющих переменных, в результате чего нелинейные функции регрессии сводятся к линейным. К новой, преобразованной функции регрессии может быть применен обычный метод наименьших квадратов (МНК). Для рассмотренных функций ниже приведены примеры таких замен.

Пример 1.2

Полином второй степени: $\widehat{y}_x = a + b_1 \cdot x + b_2 \cdot x^2$.

Линеаризующее преобразование: $x_1 = x$, $x_2 = x^2$. В результате приходим к двухфакторному линейному уравнению $\widehat{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2$, параметры которого определяются из системы следующих нормальных уравнений:

$$\begin{cases} n \cdot a + b_1 \cdot \sum x_1 + b_2 \cdot \sum x_2 = \sum y, \\ a \cdot \sum x_1 + b_1 \cdot \sum x_1^2 + b_2 \cdot \sum x_1 \cdot x_2 = \sum x_1 \cdot y, \\ a \cdot \sum x_2 + b_1 \cdot \sum x_1 \cdot x_2 + b_2 \cdot \sum x_2^2 = \sum x_2 \cdot y. \end{cases}$$

После обратной замены переменных получим:

$$\begin{cases} n \cdot a + b_1 \cdot \sum x + b_2 \cdot \sum x^2 = \sum y, \\ a \cdot \sum x_1 + b_1 \cdot \sum x^2 + b_2 \cdot \sum x^3 = \sum x \cdot y, \\ a \cdot \sum x^2 + b_1 \cdot \sum x^3 + b_2 \cdot \sum x^4 = \sum x^2 \cdot y. \end{cases}$$

Пример 1.3

Равносторонняя гипербола: $\widehat{y}_x = a + (b/x)$.

Линеаризующее преобразование: $x_1 = 1/x$. Система линейных уравнений при применении МНК будет выглядеть следующим образом:

$$\begin{cases} n \cdot a + b \cdot \sum \frac{1}{x} = \sum y, \\ a \cdot \sum \frac{1}{x} + b \cdot \sum \frac{1}{x^2} = \sum \frac{1}{x} \cdot y. \end{cases}$$

Регрессии, нелинейные относительно оцениваемых параметров.

Регрессии, нелинейные относительно оцениваемых параметров, представляют собой более сложный случай. Эти модели, в свою очередь, могут быть разделены на две группы: *нелинейные модели внутренне линейные* и *нелинейные модели внутренне нелинейные*.

Если нелинейная модель внутренне линейна, то при помощи определенных процедур она может быть сведена к линейной и оценена при помощи метода наименьших квадратов. Если же модель внутренне нелинейна по параметрам, ее нельзя свести к линейной, а для оценки параметров используются итеративные процедуры, успешность которых зависит от вида уравнений и особенностей применяемого итеративного подхода.

Примерами внутренне линейных регрессий могут служить следующие модели:

- логарифмическая модель (степенная): $y = a \cdot x^b + \varepsilon$;
- показательная: $y = a \cdot b^x + \varepsilon$;
- экспоненциальная: $y = e^{a+b \cdot x} + \varepsilon$.

В следующем примере показано решение задачи построения логарифмической (степенной) модели $\widehat{y} = a \cdot x^b$.

Пример 1.4

По статистическим данным (табл. 1.4), описывающим зависимость значения рентабельности производства синтетического каучука от индекса Лернера, построить логарифмическую модель парной регрессии $\widehat{y} = a \cdot x^b$.

Таблица 1.4 – Исходные данные к примеру 1.4

Индекс Лернера	0,14	0,33	0,21	0,14	0,22	0,25	0,28
Рентабельность, %	15,8	49	26,2	15,7	27,4	30	35

Решение:

1. Логарифмируем обе части уравнения: $\ln y = \ln a + b \ln x$.
2. Вводим линеаризующие преобразования: $Y = \ln y$, $X = \ln x$, $A = \ln a$.
3. Выполняем оценку параметров вновь полученного уравнения $Y = A + b \cdot X$, используя метод наименьших квадратов. Полученное уравнение регрессии имеет вид $Y = 5{,}19 + 1{,}24 \cdot X$.
4. Определяем искомое уравнение $\widehat{y} = a \cdot x^b$. Для этого:
   - выполняем потенцирование уравнения $\widehat{y} = e^{5{,}19} \cdot x^{1{,}24}$;
   - вычислив $e^{5{,}19} = 179{,}47$, записываем уравнение логарифмической (степенной) модели регрессии: $\widehat{y} = 179{,}47 \cdot x^{1{,}24}$.

. . . . . .

Возможные замены переменных для рассмотренных функций приведены в таблице 1.5.

Таблица 1.5 – Линеаризующие преобразования

№	**Вид модели**	**Линеаризующие преобразования**	**Ограничения**	**Обратная замена переменных**	
1	$\widehat{y}_x = a \cdot x^b$	$Y = \ln y$, $X = \ln x$, $A = \ln a$	$y > 0$, $x > 0$, $a > 0$	$a = e^A$	$b = b$
2	$\widehat{y}_x = a \cdot b^x$	$Y = \ln y$, $B = \ln b$, $A = \ln a$	$y > 0$, $b > 0$, $a > 0$	$a = e^A$	$b = e^B$
3	$\widehat{y}_x = e^{a+b \cdot x}$	$Y = \ln y$	$y > 0$	$a = a$	$b = b$
4	$\widehat{y}_x = a + b \cdot \ln x$	$X = \ln x$	$x > 0$	$a = a$	$b = b$
5	$\widehat{y}_x = a + b \cdot \frac{1}{x}$	$X = \frac{1}{x}$	$x \neq 0$	$a = a$	$b = b$

Анализ представленной таблицы показывает, что линеаризация функций 1–3 выполняется логарифмированием с последующей заменой переменных. Линеаризация функций 4–5 выполняется простой заменой переменных.

Исследование нелинейных регрессионных моделей.

Для выявления тесноты связи между переменными в случае нелинейной зависимости используется ***индекс корреляции***. Он показывает тесноту связи между фактором x и зависимой переменной y:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}},$$

где $\sigma_y^2 = (1/n) \cdot \sum (y - \bar{y})^2$ — общая дисперсия результативного признака y; $\sigma_{ост}^2 = (1/n) \cdot \sum (y - \hat{y}_x)^2$ — остаточная дисперсия.

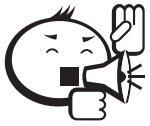
Индекс корреляции есть неотрицательная величина, не превосходящая единицу: $0 \leqslant \rho_{xy} \leqslant 1$. Чем ближе значение индекса корреляции к единице, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

В случае линейной зависимости $\rho_{xy} = |r_{xy}|$. Расхождение между значением индекса корреляции ρ_{xy} и значением коэффициента линейной корреляции r_{xy} может быть использовано для проверки линейности корреляционной зависимости. Чем больше кривизна линии регрессии, тем величина r_{xy} меньше ρ_{xy}. Близость этих показателей указывает на то, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейную функцию.

Квадрат индекса корреляции носит название ***индекса детерминации*** и характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака и вычисляется по формуле:

$$\rho_{xy}^2 = \frac{\sigma_{объясн}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2},$$

где $\sigma_{объясн}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{y}_{xi} - \bar{y})^2$.



Следует обратить внимание на то, что разности в соответствующих суммах $\sum (y - \bar{y})^2$, $\sum (\hat{y} - \bar{y})^2$ и $\sum (y - \hat{y}_x)^2$ берутся не в преобразованных, а в исходных значениях результативного признака, если уравнение регрессии нелинейно относительно объясняющих переменных. Иначе говоря, при вычислении этих сумм следует использовать не преобразованные (линеаризованные) зависимости, а исходные нелинейные уравнения регрессии. В случае регрессий, нелинейных относительно параметров, соответствующие суммы вычисляются в преобразованных данных.

Индекс детерминации используется для проверки существенности в целом уравнения регрессии по F-критерию Фишера:

$$F = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \cdot \frac{n - m - 1}{m},$$

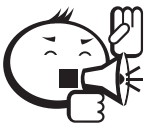
где ρ_{xy}^2 — индекс детерминации; n — число наблюдений; m — число параметров при переменной x. Фактическое значение F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы $k_2 = n - m - 1$ (для остаточной суммы квадратов) и $k_1 = m$ (для факторной суммы квадратов). Вычисленное значение F-критерия признается достоверным, если оно больше табличного при заданном уровне значимости α. В этом случае делается вывод о существенности уравнения регрессии в целом.

Степень аппроксимации данных выборки полученной регрессией оценивается с помощью средней относительной ошибки аппроксимации:

$$\overline{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y}_{x_i}}{y_i} \right| \cdot 100\%.$$

Как и в случае линейной регрессии, значение $\overline{A}$ не должно превышать 10%.

Наряду с индексами корреляции и детерминации в случае нелинейных форм связи для характеристики зависимости между результативной переменной и факторными переменными применяются ***коэффициенты эластичности***.



Коэффициент эластичности показывает, на сколько процентов изменится в среднем результат, если фактор изменится на 1%.

Формула для расчета коэффициента эластичности имеет вид:

$$Э = f'(x) \cdot \frac{x}{y}.$$

Так как для большинства функций коэффициент эластичности не является постоянной величиной, а зависит от соответствующего значения фактора x, то обычно рассчитывается средний коэффициент эластичности:

$$\overline{Э} = f'(x) \cdot \frac{\overline{x}}{\overline{y}}.$$

В таблице 1.6 приведены формулы для расчета средних коэффициентов эластичности ($\overline{Э}$) для наиболее часто используемых типов уравнений регрессии.

Таблица 1.6 – Средний коэффициент эластичности

Вид модели, y	**Первая производная,** y'	**Средний коэффициент эластичности,** $\overline{Э}$
$y = a + b \cdot x + \varepsilon$	b	$\frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$
$y = a + b \cdot x + c \cdot x^2 + \varepsilon$	$b + 2 \cdot c \cdot x$	$\frac{(b + 2 \cdot c \cdot \overline{x}) \cdot \overline{x}}{a + b \cdot \overline{x} + c \cdot \overline{x^2}}$
$y = a + \frac{b}{x} + \varepsilon$	$-\frac{b}{x^2}$	$-\frac{b}{a \cdot \overline{x} + b}$
$y = a \cdot x^b \cdot \varepsilon$	$y = a \cdot b \cdot x^{b-1}$	b
$y = a \cdot b^x \cdot \varepsilon$	$a \cdot \ln b \cdot b^x$	$\overline{x} \cdot \ln b$
$y = a + b \cdot \ln x + \varepsilon$	$\frac{b}{x}$	$\frac{b}{a + b \cdot \ln \overline{x}}$
$y = \frac{1}{a + b \cdot x + \varepsilon}$	$-\frac{b}{(a + b \cdot x)^2}$	$-\frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$

Возможны случаи, когда расчет коэффициента эластичности не имеет смысла. Это происходит тогда, когда для рассматриваемых признаков бессмысленно определение изменения в процентах. Например, изучая соотношение ставок межбанковского кредита Y (в % годовых) и срока его предоставления X (в днях), было получено степенное уравнение с очень высоким значением индекса корреляции (0,98). При этом коэффициент эластичности, равный 0,4%, лишен смысла, так как срок предоставления кредита не измеряется в процентах.

Рассмотрим пример оценки качества уравнения регрессии для модели $\hat{y} = 179{,}9383 \cdot x^{1{,}244552}$, построенной по таблице 1.4.

Пример 1.5

Поскольку построенная модель относится к виду моделей, нелинейных относительно параметров, качества модели проведем по линеаризованному уравнению $Y = 5{,}192614 + 1{,}244552 \cdot X$. Здесь $Y = \ln(y)$, $X = \ln(x)$. Составляем вспомогательную таблицу 1.7.

Таблица 1.7 – Вспомогательная таблица к примеру 1.5

	X	Y	$\hat{Y}$	$(Y - \hat{Y})^2$	$(Y - \overline{Y})^2$	$(\hat{Y} - \overline{Y})^2$
1	**4**	**5**	**6**	**7**	**8**	**9**
1	−1,9661	2,7600	2,7456	0,0002	0,2672	0,2822
2	−1,1087	3,8918	3,8127	0,0063	0,3781	0,2872
3	−1,5606	3,2658	3,2503	0,0002	0,0001	0,0007
4	−1,9661	2,7537	2,7456	0,0001	0,2737	0,2822
5	−1,5141	3,3105	3,3082	0	0,0011	0,0010
6	−1,3863	3,4012	3,4672	0,0044	0,0155	0,0362
7	−1,2730	3,5553	3,6082	0,0028	0,0775	0,1098
Итого	—	22,9383	22,9378	0,0140	1,0132	0,9993
Среднее значение	—	3,2769	3,2768	0,0020	0,1447	0,1427

Вычислим индекс корреляции по формуле:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{0{,}002}{0{,}1447}} = 0{,}993.$$

Вычислим индекс детерминации по формуле:

$$\rho^2_{xy} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y} = 1 - \frac{0{,}002}{1447} = 0{,}986,$$

который показывает, что 98,6% вариации результативного признака объясняется вариацией признака-фактора, а 1,4% приходится на долю прочих факторов.

Вычислим среднюю ошибку аппроксимации, используя данные исходного нелинейного уравнения регрессии:

$$\overline{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y}_{x_i}}{y_i} \right| \cdot 100\% = \frac{0{,}2388}{7} \cdot 100\% = 3{,}4\%.$$

Вычисленное значение средней ошибки аппроксимации не превышает 10%, следовательно, кривая регрессии хорошо приближает исходные данные.

F-критерий Фишера:

$$F = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}986 \cdot 6}{1 - 0{,}986} = 422{,}57,$$

значительно превышает табличное $F_{0{,}05,\,1,\,6} = 5{,}99$, что говорит о существенности модели в целом.

Контрольные вопросы по главе 1

1. Дайте определение парной регрессии.
2. Поясните экономическую сущность параметров уравнения парной регрессии.
3. Назовите основные причины наличия в регрессионной модели случайного отклонения.
4. Назовите основные этапы регрессионного анализа.
5. Что понимается под спецификацией модели?
6. Какие требования предъявляются к объему наблюдений, необходимых для построения уравнения регрессии?
7. По каким вычислениям можно судить о значимости модели в целом?
8. Объясните суть коэффициента детерминации.
9. В каких пределах должна находиться ошибка аппроксимации, чтобы можно было сделать вывод о хорошем подборе модели к исходным данным?
10. Какие значения может принимать коэффициент детерминации r_{xy}^2?
11. Назовите классы нелинейных моделей.
12. Объясните суть среднего коэффициента эластичности.

Глава 2

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

2.1 Понятие множественной регрессии

В настоящее время множественная регрессия — один из наиболее распространенных методов в эконометрике. Например, исследуется урожайность зерновых культур. Она определяется набором таких факторов, как число колесных тракторов, число зерноуборочных комбайнов, число орудий поверхностной обработки почвы, количество удобрений, расходуемых на гектар, количество химических средств оздоровления растений, расходуемых на гектар. Следовательно, модели парной регрессии не пригодны для того, чтобы в полной мере охарактеризовать поведение исследуемого показателя. В этом случае рассматривается множественная регрессия:

$$y = f(x_1, x_2, \ldots, x_m) + \varepsilon,$$

где y — зависимая переменная (результат); $x_1, x_2, \ldots, x_m$ — независимые переменные (факторы); ε — случайная ошибка регрессионной зависимости; f — некоторая математическая функция.

Основной целью множественной регрессии является построение модели с большим числом факторов и определение влияния каждого фактора в отдельности, а также их совместного воздействия на моделируемый показатель (результат).

Возможны разные виды уравнений множественной регрессии: линейные и нелинейные. Ввиду четкой интерпретации параметров наиболее широко используется *линейная* модель множественной регрессии:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m + \varepsilon,$$

где $a, b_1, b_2, \ldots, b_m$ — параметры функции.

Параметр a называется свободным членом и определяет значение результирующей переменной y в случае, когда все объясняющие переменные $x_1, x_2, \ldots, x_m$

равны нулю. Если же факторы по своему экономическому содержанию не могут принимать нулевых значений, то значение параметра a может не иметь экономического смысла.

Параметры b_j называются ***коэффициентами «чистой» регрессии***. Они характеризуют среднее изменение результата y с изменением соответствующего фактора x_j на единицу при неизменном значении других факторов, закрепленном на среднем уровне.

2.2 Спецификация модели. Отбор факторов при построении уравнения множественной регрессии

Вопрос о спецификации модели включает в себя две задачи: отбор факторов и выбор вида уравнения регрессии. В данном пособии рассматривается линейная множественная регрессия, поэтому остановимся на задаче отбора факторов.

Отбор факторов обычно осуществляется в два этапа:

1) отбор факторов на основе теоретического анализа природы взаимосвязи моделируемого показателя с другими экономическими показателями;
2) проверка на статистическую значимость отобранных факторов и решение о включении того или иного фактора в модель, основанное на количественной оценке степени влияния соответствующего фактора на изучаемый показатель.

Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям [1]:

1) быть количественно измеримы. Если необходимо учесть влияние качественного фактора (не имеющего количественной оценки), то в модель включается соответствующая ему «фиктивная» переменная, имеющая конечное количество формально численных значений, соответствующих градациям качественного фактора. Например, если нужно учесть влияние уровня образования на размер заработной платы, то в уравнение регрессии можно включить переменную z, принимающую значения: 0 — при начальном образовании, 1 — при среднем, 2 — при высшем;
2) должны объяснить вариацию результативного признака. Так как данная величина характеризуется таким показателем, как коэффициент детерминации R^2, включение фактора в модель должно приводить к заметному изменению последнего. Например, если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов оценивается как $1 - R^2$ с соответствующей остаточной дисперсией $S^2_{\text{ост}}$. При дополнительном включении в регрессию $m + 1$ фактора, оказывающего существенное влияние на результат, коэффициент детерминации должен возрастать ($R^2_{m+1} > R^2_m$), а остаточная дисперсия уменьшаться ($S^2_{\text{ост}_{m+1}} < S^2_{\text{ост}_m}$). В противном случае включаемый в анализ фактор x_{m+1} не улучшает модель и практически является лишним фактором. Насыщение модели лишними факторами не только не снижает величину остаточной

дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по критерию Стьюдента;

3) не должны быть взаимно коррелированы, тем более, находиться в точной функциональной связи. Наличие высокой степени коррелированности определяется по значению коэффициента парной корреляции $r_{x_i x_j} \geqslant \geqslant 0{,}8$ и может привести к нежелательным последствиям:

   - затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы, то есть находятся в линейной зависимости; параметры линейной регрессии теряют экономический смысл;
   - оценки параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Взаимная корреляция факторов

Исследование взаимной коррелированности объясняющих переменных позволяет исключать из модели дублирующие факторы. Проверка наличия взаимной корреляции двух факторов (парной корреляции) основывается на анализе матрицы парных коэффициентов корреляции. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами [1].

Например, при изучении зависимости $\widehat{y} = f(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей (табл. 2.1).

Таблица 2.1 – Матрица парных коэффициентов корреляции

	y	x_1	x_2	x_3
y	1	0,8	0,7	0,6
x_1	0,8	1	0,8	0,5
x_2	0,7	0,8	1	0,2
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга, поскольку значение коэффициента парной корреляции $r_{x_1 x_2} = 0{,}8$. Однако в анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с результирующей переменной y $(r_{yx_2} = 0{,}7 < r_{yx_1} = 0{,}8)$, но зато значительно слабее корреляция фактора x_2 с объясняющей переменной x_3 $(r_{x_2 x_3} = 0{,}2 < r_{x_1 x_3} = 0{,}5)$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

По величине парных коэффициентов корреляции обнаруживается лишь явная линейная зависимость факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторов, когда более чем два фактора связаны между собой линейной зависимостью,

т. е. имеет место совокупное воздействие факторов друг на друга. Наличие мультиколлинеарности факторов может означать, что некоторые факторы будут всегда действовать в унисон. В результате вариация в исходных данных перестает быть полностью независимой и нельзя оценить воздействие каждого фактора в отдельности.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами (межфакторная корреляция).

$$R = \begin{bmatrix} r_{x_1}r_{x_1} & r_{x_1}r_{x_2} & \dots & r_{x_1}r_{x_m} \\ r_{x_2}r_{x_1} & r_{x_2}r_{x_2} & \dots & r_{x_2}r_{x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_m}r_{x_1} & r_{x_m}r_{x_2} & \dots & r_{x_m}r_{x_m} \end{bmatrix} = \begin{bmatrix} 1 & r_{x_1}r_{x_2} & \dots & r_{x_1}r_{x_m} \\ r_{x_2}r_{x_1} & 1 & \dots & r_{x_2}r_{x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_m}r_{x_1} & r_{x_m}r_{x_2} & \dots & 1 \end{bmatrix}.$$

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность между факторами и тем ненадежнее результаты множественной регрессии. С другой стороны, чем ближе к единице определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Для преодоления явления линейной зависимости между факторами используются такие подходы, как:

- исключение из модели одного или нескольких коррелированных факторов;
- увеличение объема выборки;
- преобразование факторов, при котором уменьшается корреляция между ними.

Например, для модели $\widehat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$ возможным путем учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т. е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие.

$$\widehat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3.$$

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость, но, как правило, взаимодействия третьего и более высоких порядков оказываются статистически незначимыми.

Отбор факторов, включаемых в регрессию, является одним из важнейших этапов практического использования методов регрессии. Подходы к отбору факторов на основе показателей корреляции могут быть разные. Они приводят построение уравнения множественной регрессии соответственно к разным методикам. В зависимости от того, какая методика построения уравнения регрессии принята, меняется алгоритм ее решения на ЭВМ. Следует также учитывать ограничение, накладываемое на количество факторов, имеющимся числом наблюдений. Количество наблюдений должно превышать количество факторов более чем в 6–7 раз.

2.3 Оценка параметров уравнения множественной линейной регрессии

Рассмотрим линейную модель множественной регрессии:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m + \varepsilon.$$

Для оценки параметров уравнения множественной линейной регрессии применяется метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных $\widehat{y}$ минимальна:

$$Q = \sum_{i=1}^{n} (y_i - \widehat{y}_{x_i})^2 \to \min.$$

С учетом величина Q определена как функция неизвестных параметров a и b_i.

$$Q(a, b_1, b_2, \ldots, b_m) = \sum_{i=1}^{n} (y_i - a - b_1 \cdot x_1 - b_2 \cdot x_2 - \ldots - b_m \cdot x_m)^2.$$

Необходимым условием минимизации функции Q является равенство нулю частных производных первого порядка по каждому из параметров b_i. Результатом является следующая система уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum\limits_{i=1}^{n}(y_i - a - b_1 \cdot x_1 - b_2 \cdot x_2 - \ldots - b_m \cdot x_m), \\ \dfrac{\partial Q}{\partial b_1} = -2b_1\sum\limits_{i=1}^{n}(y_i - a - b_1 \cdot x_1 - b_2 \cdot x_2 - \ldots - b_m \cdot x_m), \\ \ldots, \\ \dfrac{\partial Q}{\partial b_m} = -2b_m\sum\limits_{i=1}^{n}(y_i - a - b_1 \cdot x_1 - b_2 \cdot x_2 - \ldots - b_m \cdot x_m). \end{cases}$$

После выполнения преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 + \ldots + b_m\sum x_m = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + \ldots + b_m\sum x_1x_m = \sum yx_1, \\ \ldots, \\ a\sum x_m + b_1\sum x_1x_m + b_2\sum x_2x_m + \ldots + b_m\sum x_m^2 = \sum yx_m. \end{cases}$$

Пример 2.1

Построить уравнение множественной регрессии, выражающее оценку стоимости группы небольших офисных зданий в деловом районе. Данные представлены в таблице 2.2.

Таблица 2.2 – Исходные данные для примера 2.1

Общая площадь в квадратных метрах (x_1)	**Количество офисов (x_2)**	**Количество входов (0,5 входа означает вход только для доставки корреспонденции) (x_3)**	**Время эксплуатации здания в годах (x_4)**	**Оценочная цена здания под офис (y)**
2310	2	2	20	142 000
2333	2	2	12	144 000
2356	3	1,5	33	151 000
2379	3	2	43	150 000
2402	2	3	53	139 000
2425	4	2	23	169 000
2448	2	1,5	99	126 000
2471	2	2	34	142 900
2494	3	3	23	163 000
2517	4	4	55	169 000
2540	2	3	22	149 000

Запишем систему нормальных уравнений для четырехфакторной модели:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 + b_3\sum x_3 + b_4\sum x_4 = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + b_3\sum x_1x_3 + b_4\sum x_1x_4 = \sum yx_1, \\ a\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 + b_3\sum x_1x_3 + b_4\sum x_1x_4 = \sum yx_2, \\ a\sum x_3 + b_1\sum x_1x_3 + b_2\sum x_2x_3 + b_3\sum x_3^2 + b_4\sum x_1x_4 = \sum yx_3, \\ a\sum x_4 + b_1\sum x_1x_4 + b_2\sum x_2x_4 + b_3\sum x_3x_4 + b_4\sum x_4^2 = \sum yx_4. \end{cases}$$

Вычислив соответствующие суммы, получаем:

$$\begin{cases} 11 \cdot a + 26\ 675 \cdot b_1 + 29 \cdot b_2 + 26 \cdot b_3 + 417 \cdot b_4 = 1\ 644\ 900, \\ 26\ 675 \cdot a + 64\ 745\ 065 \cdot b_1 + 70\ 463 \cdot b_2 + 63\ 418 \cdot b_3 + \\ \qquad +1\ 015\ 365 \cdot b_4 = 3\ 992\ 189\ 900, \\ 29 \cdot a + 70\ 463 \cdot b_1 + 83 \cdot b_2 + 70{,}5 \cdot b_3 + 1089 \cdot b_4 = 4\ 429\ 800, \\ 26 \cdot a + 63\ 418 \cdot b_1 + 70{,}5 \cdot b_2 + 67{,}5 \cdot b_3 + 976 \cdot b_4 = 3\ 940\ 300, \\ 417 \cdot a + 1\ 015\ 365 \cdot b_1 + 1089 \cdot b_2 + 976 \cdot b_3 + 21\ 815 \cdot b_4 = 60\ 909\ 600. \end{cases}$$

Решение данной системы уравнений можно выполнить различными способами.

1. Вычислим оценки параметров модели $\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4$, решая систему методом определителей:

$$a = \frac{\Delta_a}{\Delta} = \frac{3{,}93346 \cdot 10^{15}}{75\ 183\ 958\ 894} = 52\ 317{,}8;$$

$$b_1 = \frac{\Delta_{b_1}}{\Delta} = \frac{2{,}07819 \cdot 10^{12}}{75\ 183\ 958\ 894} = 27{,}64;$$

$$b_2 = \frac{\Delta_{b_2}}{\Delta} = \frac{9{,}42038 \cdot 10^{14}}{75\ 183\ 958\ 894} = 12\ 529{,}8;$$
$$b_3 = \frac{\Delta_{b_3}}{\Delta} = \frac{1{,}9196 \cdot 10^{14}}{75\ 183\ 958\ 894} = 2553{,}21;$$
$$b_4 = \frac{\Delta_{b_4}}{\Delta} = \frac{-1{,}76109 \cdot 10^{13}}{75\ 183\ 958\ 894} = -234{,}24.$$

2. Оценим параметры модели $\widehat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4$ с помощью матричных операций. Введем обозначения:

$$B = \begin{bmatrix} a \\ b_1 \\ \dots \\ b_m \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{m1} \\ 1 & x_{12} & \dots & x_{m2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{mn} \end{bmatrix},$$

где B — матрица-столбец, размерностью $(m+1\times 1)$ параметров уравнения регрессии; Y — матрица-столбец размерностью $(n\times 1)$ наблюдений зависимой переменной; X — матрица размерностью $(m+1\times n)$ исходных значений независимых переменных x_{ji}, в которой первый столбец из единиц можно рассматривать как значение «фиктивной» переменной при коэффициенте a.

В этих обозначениях уравнение регрессии записывается следующим образом:

$$Y = XB + \varepsilon,$$

где $\varepsilon = Y - XB$ — вектор-столбец остатков регрессии.

По условию применения метода наименьших квадратов минимизируется функционал $Q = \sum \varepsilon_i^2$, который можно записать как произведение вектора-строки ε' на вектор-столбец ε:

$$Q = \varepsilon' \cdot \varepsilon = (Y - X \cdot B)' \cdot (Y - X \cdot B).$$

Дифференцирование Q по вектору B приводит к выражению:

$$\frac{\partial Q}{\partial B} = -2 \cdot X' \cdot Y + 2 \cdot (X' \cdot X)^{-1} \cdot B,$$

которое приравнивается к нулю. В результате последующих преобразований получаем выражение для вычисления параметров уравнения регрессии:

$$B = (X' \cdot X)^{-1} X' \cdot Y.$$

Здесь X' — транспонированная матрица X; $(X' \cdot X)^{-1}$ — матрица, обратная к $X' \cdot X$.

Для таблицы 2.1 определим матрицы:

$$B = \begin{bmatrix} a \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \quad Y = \begin{bmatrix} 142\ 000 \\ 144\ 000 \\ 151\ 000 \\ 150\ 000 \\ 139\ 000 \\ 169\ 000 \\ 126\ 000 \\ 142\ 900 \\ 163\ 000 \\ 169\ 000 \\ 149\ 000 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 2310 & 2 & 2 & 20 \\ 1 & 233 & 2 & 2 & 12 \\ 1 & 2356 & 3 & 1{,}5 & 33 \\ 1 & 2379 & 3 & 2 & 43 \\ 1 & 2402 & 2 & 3 & 53 \\ 1 & 2425 & 4 & 2 & 23 \\ 1 & 2448 & 2 & 1{,}5 & 99 \\ 1 & 2471 & 2 & 2 & 34 \\ 1 & 2494 & 3 & 3 & 23 \\ 1 & 2517 & 4 & 4 & 55 \\ 1 & 2540 & 2 & 3 & 22 \end{bmatrix}.$$

С использованием матричных операций вычисляем:

$$X' \cdot X = \begin{bmatrix} 11 & 26\ 675 & 29 & 26 & 417 \\ 26\ 675 & 6{,}5 \cdot 10^7 & 70\ 463 & 63\ 418 & 10\ 115\ 365 \\ 29 & 70\ 463 & 83 & 70{,}5 & 1089 \\ 26 & 63\ 418 & 70{,}5 & 67{,}5 & 976 \\ 417 & 1\ 015\ 365 & 1089 & 976 & 21\ 815 \end{bmatrix},$$

$$(X' \cdot X)^{-1} = \begin{bmatrix} 158{,}97 & -0{,}0701 & -0{,}0319 & 3{,}96351 & 0{,}04831 \\ -0{,}0701 & 3{,}1 \cdot 10^{-5} & -0{,}0001 & -0{,}0019 & -2 \cdot 10^{-5} \\ -0{,}0319 & -0{,}0001 & 0{,}1699 & -0{,}0465 & 0{,}00031 \\ 3{,}96351 & -0{,}0019 & -0{,}0465 & 0{,}29894 & 0{,}00171 \\ 0{,}04831 & -2 \cdot 10^{-5} & 0{,}00031 & 0{,}00171 & 0{,}00019 \end{bmatrix},$$

$$B = (X' \cdot X)^{-1} X' \cdot Y = \begin{bmatrix} 52\ 317{,}8 \\ 27{,}64 \\ 12\ 529{,}8 \\ 2553{,}21 \\ -234{,}24 \end{bmatrix}.$$

В итоге получаем уравнение регрессионной модели:

$$\widehat{y} = 52\ 317{,}8 + 27{,}64 \cdot x_1 + 12\ 529{,}8 \cdot x_2 + 2553{,}21 \cdot x_3 - 234{,}24 \cdot x_4.$$

. . . . . .

2.4 Регрессионная модель в стандартизованном масштабе

Независимые переменные x_i имеют различный экономический смысл, разные единицы измерения и масштаб. Если требуется определить степень относительного влияния отдельных факторов x_i на изменение результативной переменной y, то переменные x_i следует привести к сопоставимому виду. Это можно осуществить,

вводя так называемые «стандартизованные» переменные t_y, t_{x_1}, $t_{x_2}, \ldots, t_{x_m}$ с помощью соотношений:

$$t_y = \frac{y - \bar{y}}{\sigma_y}, \quad t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}, \quad (i = 1, 2, \ldots, m).$$

Стандартизованные переменные обладают следующими свойствами:

1) средние значения равны нулю ($\bar{t}_y$, $\bar{t}_{x_i} = 0$);
2) среднеквадратичные отклонения равны единице ($\sigma_{t_y} = \sigma_{t_{xi}} = 1$).

Уравнение регрессии в стандартизованных переменных принимает вид:

$$t_y = \beta_1 \cdot t_{x_1} + \beta_2 \cdot t_{x_2} + \ldots + \beta_m \cdot t_{x_m} + \varepsilon.$$

Величины β_i называются стандартизованными коэффициентами. Их связь с коэффициентами «чистой» регрессии b_i задается соотношениями:

$$b_i = \beta_i \cdot \frac{\sigma_y}{\sigma_{x_i}} \quad \text{или} \quad \beta_i = b_i \cdot \frac{\sigma_{x_i}}{\sigma_y}.$$

Стандартизованные коэффициенты регрессии показывают, на сколько с.к.о. (средних квадратичных отклонений) изменится в среднем результат y, если соответствующий фактор x_i изменится на одно с.к.о. при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой, что позволяет ранжировать факторы по силе их воздействия на результат. Большее относительное влияние на изменение результативной переменной y оказывает тот фактор, которому соответствует большее по модулю значение коэффициента β_i. Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет использовать их при отсеве факторов: из модели исключаются факторы с наименьшим значением β_i. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Метод наименьших квадратов можно применять и для вычисления стандартизованных коэффициентов β_i. При этом система нормальных уравнений МНК принимает вид:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 \cdot r_{x_2x_1} + \beta_3 \cdot r_{x_3x_1} + \ldots + \beta_m \cdot r_{x_mx_1}, \\ r_{yx_2} = \beta_1 \cdot r_{x_1x_2} + \beta_2 \cdot + \beta_3 \cdot r_{x_3x_2} + \ldots + \beta_m \cdot r_{x_mx_2}, \\ \ldots, \\ r_{yx_m} = \beta_1 \cdot r_{x_1x_m} + \beta_2 \cdot r_{x_2x_m} + \beta_3 \cdot r_{x_3x_m} + \ldots + \beta_m, \end{cases}$$

где r_{yx_i} и $r_{x_ix_j}$ — коэффициенты парной и межфакторной корреляции.

Пример 2.2

Пусть имеются следующие данные (условные) о прибыли от реализации продукции y, постоянных затратах на производство продукции x_1 и объеме выпускаемой продукции x_2, представленные в таблице 2.3.

Таблица 2.3 – Исходные данные к примеру 2.2

№	1	2	3	4	5	6	7	8	9	10
x_1	100	120	90	94	91	80	93	95	103	101
x_2	12	8	7,5	7,9	13	7	7,7	6	5	8,3
y	20	30	21	25	23	18	22	24	29	27

Предполагая, что между переменными y, x_1, x_2 существует линейная корреляционная зависимость, найдем уравнение регрессии y по x_1 и x_2.

Для удобства дальнейших вычислений составляем вспомогательную таблицу 2.4.

Для нахождения параметров уравнения регрессии в данном случае необходимо решить следующую систему нормальных уравнений:

$$\begin{cases} 10 \cdot a + 967 \cdot b_1 + 82{,}4 \cdot b_2 = 239, \\ 967 \cdot a + 94\ 501 \cdot b_1 + 7960 \cdot b_2 = 23\ 413, \\ 82{,}4 \cdot a + 7960 \cdot b_1 + 733{,}84 \cdot b_2 = 1942{,}5. \end{cases}$$

Получаем: $a = -1{,}487$, $b_1 = 0{,}3005$, $b_2 = -0{,}445$. Запишем уравнение множественной регрессии:

$$\widehat{y} = -1{,}487 + 0{,}3005 \cdot x_1 + (-0{,}445) \cdot x_2.$$

Оно показывает, что при увеличении постоянных затрат на производство продукции x_1 (при неизменном x_2) на 1 условную единицу затрат прибыль y увеличится в среднем на 0,3005, а при увеличении объема выпускаемой продукции x_2 (при неизменном x_1) на 1 условную единицу объема прибыль уменьшается в среднем на 0,445 условных единиц прибыли.

Найдем уравнение множественной регрессии в стандартизованном масштабе:

$$\widehat{t}_y = \beta_1 \cdot t_{x_1} + \beta_2 \cdot t_{x_2}.$$

Вычисляем стандартизованные коэффициенты регрессии:

$$\beta_1 = b_1 \cdot \frac{\sigma_{x_1}}{\sigma_y} = 0{,}3005 \cdot \frac{9{,}96}{3{,}7} = 0{,}8089,$$

$$\beta_2 = b_2 \cdot \frac{\sigma_{x_2}}{\sigma_y} = -0{,}445 \cdot \frac{2{,}34}{3{,}7} = -0{,}2819.$$

В результате уравнение регрессии в стандартизованном масштабе будет выглядеть следующим образом:

$$\widehat{t}_y = 0{,}8089 \cdot t_{x_1} - 0{,}2819 \cdot t_{x_2}.$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что постоянные затраты оказывают большее влияние на рост цены на продукцию.

Таблица 2.4 – Вспомогательная таблица

№	x_1	x_2	y	x_1^2	x_2^2	y_2	$x_1 \cdot x_2$	$x_1 \cdot y$	$x_2 \cdot y$	$\widehat{y}_x$	ε^2
1	100	12	20	10 000	144	400	1200	2000	240	23,22	10,35
2	120	8	30	14 400	64	900	960	3600	240	31,01	1,02
3	90	7,5	21	8100	56,25	441	675	1890	157,5	22,22	1,48
4	94	7,9	25	8836	62,41	625	742,6	2350	197,5	23,24	3,10
5	91	13	23	8281	169	529	1183	2093	299	20,07	8,60
6	80	7	18	6400	49	324	560	1440	126	19,43	2,06
7	93	7,7	22	8649	59,29	484	716,1	2046	169,4	23,03	1,06
8	95	6	24	9025	36	576	570	2280	144	24,39	0,15
9	103	5	29	10 609	25	841	515	2987	145	27,24	3,11
10	101	8,3	27	10 201	68,89	729	838,3	2727	224,1	25,17	3,37
Сумма	967	82,4	239	94 501	733,84	5849	7960	23 413	1942,5	239	34,29
Среднее значение	96,7	8,24	23,9	9450,1	73,38	584,9	796	2341,3	194,25	23,90	3,43
σ^2	99,21	5,49	13,69	—	—	—	—	—	—	—	—
σ	9,960	2,34	3,7	—	—	—	—	—	—	—	—

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности:

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_x} \quad (i = 1, 2, \ldots, m).$$

Вычисляем:

$$\overline{Э}_1 = 0{,}3005 \cdot \frac{96{,}7}{23{,}9} = 1{,}22, \quad \overline{Э}_1 = -0{,}2819 \cdot \frac{8{,}24}{23{,}9} = -0{,}15.$$

То есть увеличение только постоянных затрат (от своего среднего значения) на 1% увеличивает в среднем прибыль на 1,22%. В то же время увеличение объема производства на 1% уменьшает в среднем прибыль на 0,15%. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

. . . . . .

2.5 Частные уравнения регрессии

Частные уравнения регрессии характеризуют изолированное влияние одного из факторов x_i на результативную переменную y при исключении влияния остальных факторов, включенных в уравнение регрессии. Частные уравнения регрессии получаются из общего уравнения линейной множественной регрессии при закреплении всех факторов, кроме фактора x_i, на их среднем уровне:

$$\widehat{y}_{x_i \cdot x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m} = a + b_1 \cdot \overline{x}_1 + \ldots + b_{i-1} \cdot \overline{x}_{i-1} + b_i \cdot x_i + b_{i+1} \cdot \overline{x}_{i+1} + \ldots + b_m \cdot x_m, \quad (i = 1, 2, \ldots, m).$$

При подстановке в эти уравнения средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии, т. е. имеем:

$$\widehat{y}_{x_i \cdot x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m} = A_i + b_i \cdot x_i, \quad (i = 1, 2, \ldots, m),$$

где $A_i = a + b_1 \cdot \overline{x}_1 + \ldots + b_{i-1} \cdot \overline{x}_{i-1} + b_{i+1} \cdot \overline{x}_{i+1} + \ldots + b_m \cdot x_m$.

На основе частных уравнений регрессии определяют частные коэффициенты эластичности:

$$Э_{y x_i} = b_i \cdot \frac{x_i}{\widehat{y}_{x_i \cdot x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m}}, \quad (i = 1, 2, \ldots, m),$$

где b_i — коэффициент регрессии для фактора x_i в уравнении множественной регрессии; $\widehat{y}_{x_i \cdot x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m}$ — частное уравнение регрессии.

Средние частные коэффициенты эластичности

$$\overline{Э}_{y x_i} = b_i \cdot \frac{\overline{x}_i}{\overline{\widehat{y}}_{x_i \cdot x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m}}, \quad (i = 1, 2, \ldots, m)$$

показывают, на сколько процентов в среднем по совокупности изменится результат y при изменении фактора x от своего среднего значения на 1% при неизменных значениях других факторов, и могут использоваться для выделения факторов, наиболее влияющих на результат.

Если факторы x_i, x_j, находятся в корреляционной связи, то с помощью частных коэффициентов корреляции $r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}$ можно оценить тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии. Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель.

В общем виде при наличии m факторов для уравнения:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m + \varepsilon$$

коэффициент частной корреляции, измеряющий влияние фактора x_i на результат y при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 \ldots x_i \ldots x_m}}{1 - R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}}},$$

где $R^2_{yx_1 x_2 \ldots x_i \ldots x_m}$ — множественный коэффициент детерминации всех m факторов с результатом y; $R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}$ — тот же показатель детерминации, но без введения в модель фактора x_i.

При двух факторах формулы частных коэффициентов корреляции принимают вид:

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_2}}}; \quad r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_1}}}.$$

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. Например, $r_{yx_1 \cdot x_2}$ — коэффициент частной корреляции первого порядка. Соответственно коэффициенты парной корреляции называются коэффициентами нулевого порядка. Коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_p} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \ldots x_{p-1}} \cdot r_{x_i x_p \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{p-1}}}{\sqrt{(1 - r^2_{yx_p \cdot x_1 x_2 \ldots x_{p-1}}) \cdot (1 - r^2_{x_i x_p \cdot x_1 x_2 \ldots x_{p-1}})}},$$

где x_i — фактор, дополнительно включаемый в модель; $x_1, x_2, x_3, \ldots, x_p$ — факторы, включенные в модель до фактора x_i.

В случае двух факторов x_1 и x_2 формула принимает вид:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r^2_{yx_2}) \cdot (1 - r^2_{x_1 x_2})}};$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{(1 - r^2_{yx_1}) \cdot (1 - r^2_{x_1 x_2})}}.$$

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции

первого порядка. Так, по уравнению $\widehat{y}_x = a + b_1x_1 + b_2x_2 + b_3x_4$ возможно вычисление трех частных коэффициентов корреляции второго порядка:

$$r_{yx_1 \cdot x_2x_3},\ r_{yx_2 \cdot x_1x_3},\ r_{yx_3 \cdot x_1x_2}.$$

$$r_{yx_1 \cdot x_2x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_3 \cdot x_2} \cdot r_{x_1x_3 \cdot x_2}}{\sqrt{(1 - r^2_{yx_3 \cdot x_2}) \cdot (1 - r^2_{x_1x_3 \cdot x_2})}};$$

$$r_{yx_2 \cdot x_1x_3} = \frac{r_{yx_2 \cdot x_1} - r_{yx_3 \cdot x_1} \cdot r_{x_2x_3 \cdot x_1}}{\sqrt{(1 - r^2_{yx_3 \cdot x_1}) \cdot (1 - r^2_{x_2x_3 \cdot x_1})}};$$

$$r_{yx_3 \cdot x_1x_2} = \frac{r_{yx_3 \cdot x_1} - r_{yx_2 \cdot x_1} \cdot r_{x_2x_3 \cdot x_1}}{\sqrt{(1 - r^2_{yx_2 \cdot x_1}) \cdot (1 - r^2_{x_2x_3 \cdot x_1})}}.$$

Частные коэффициенты корреляции позволяют ранжировать факторы по степени влияния на результативный признак и находят применение в процедуре отбора факторов для включения их в уравнение регрессии (учитываются факторы, которым соответствуют значимые коэффициенты частной корреляции).

2.6 Анализ качества эмпирического уравнения регрессии

Построение эмпирического уравнения регрессии является начальным этапом эконометрического анализа. Следующей важнейшей задачей является анализ качества уравнения регрессии. Здесь можно выделить два направления:

- оценка значимости параметров модели множественной регрессии;
- оценка значимости модели множественной регрессии.

2.6.1 Оценка статистической значимости параметров модели множественной регрессии

Оценки коэффициентов регрессии зависят от используемой выборки значений переменных x, y и являются случайными величинами. Как и в парной регрессии, статистическая значимость параметров оценивается двумя способами: с помощью сравнения фактического и табличного значений критерия Стьюдента (t-критерия) и с помощью доверительных интервалов.

Фактическое значение критерия Стьюдента для параметров уравнения множественной регрессии рассчитывается по формулам:

$$t_{b_i} = \frac{b_i}{m_{b_i}}, \quad (i = 1, 2, \ldots, m), \quad t_a = \frac{a}{m_a}.$$

Здесь m_{b_i}, m_a — стандартные ошибки параметров уравнения регрессии.

Стандартные ошибки параметров уравнения множественной регрессии определяются соотношениями:

$$m_{b_i} = \sqrt{S^2_{\text{ост}} \cdot \left[(X' \cdot X)^{-1}\right]_{ii}}, \quad (i = 0, 1, 2, \ldots, m),$$

где $\left[(X' \cdot X)^{-1}\right]_{ii}$ — элемент (ii) матрицы $(X' \cdot X)^{-1}$. Значение $i = 0$ соответствует номеру элемента матрицы $(X' \cdot X)^{-1}$ для вычисления стандартной ошибки параметра a.

$S^2_{\text{ост}} = \dfrac{\sum\limits_{i=1}^{n} (y_i - \widehat{y}_{x_i})^2}{n - m - 1}$ — несмещенная оценка остаточной дисперсии.

Формулируется «нулевая» гипотеза о статистической незначимости параметров уравнения регрессии H_0: $b_i = 0$ или H_0: $a = 0$. Если для анализируемого параметра регрессии выполняется условие $t > t_{\text{табл}}(\alpha; n - m - 1)$, то он считается статистически значимым и «нулевая» гипотеза отвергается.

t-критерий Стьюдента применяется в процедуре принятия решения о целесообразности включения фактора в модель. Если коэффициент при факторе в уравнении регрессии оказывается незначимым, то включать данный фактор в модель не рекомендуется. Отметим, что это правило не является абсолютным и бывают ситуации, когда включение в модель статистически незначимого фактора определяется экономической целесообразностью.

Доверительные интервалы для параметров b_i уравнения линейной множественной регрессии указывают границы, в которых с заданной долей вероятности находятся значения соответствующих параметров и определяются соотношениями:

$$b_i - t(\alpha, n - m - 1) \cdot m_{b_i} < b_i < b_i + t(\alpha, n - m - 1) \cdot m_{b_i};$$
$$a - t(\alpha, n - m - 1) \cdot m_a < a < a + t(\alpha, n - m - 1) \cdot m_a.$$

Величина $t(\alpha; n - m - 1)$ представляет собой табличное значение t-критерия Стьюдента на уровне значимости α при степени свободы $n - m - 1$. Оцениваемый параметр значим, если в границы доверительного интервала не попадает нуль.

2.6.2 Оценка статистической значимости уравнения множественной регрессии

Общее качество уравнения множественной регрессии можно проверить с помощью показателя множественной детерминации R^2, который в общем случае рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} (y_i - \widehat{y}_{x_i})^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}.$$

Коэффициент детерминации R^2 принимает значения в диапазоне от нуля до единицы $0 \leqslant R^2 \leqslant 1$ и показывает, какая часть дисперсии результативного признака y объяснена уравнением регрессии. Чем выше значение R^2, тем лучше данная модель согласуется с данными наблюдений. Однако величина R^2, как правило, увеличивается при добавлении объясняющей переменной к уравнению регрессии, так как при этом уменьшается величина остаточной дисперсии. Это происходит даже при слабой связи факторов с результатом. Чтобы скомпенсировать это увеличение и получить несмещенные оценки при расчете коэффициента детерминации, в числителе и знаменателе вычитаемой из единицы дроби делается поправка на число степеней свободы. Вводится скорректированный коэффициент детерминации:

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_{x_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \cdot \frac{n-1}{n-m-1}.$$

Соотношение может быть представлено следующим выражением:

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

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

При сравнении разных моделей предпочтение при прочих равных условиях отдается той, у которой больше скорректированный коэффициент детерминации. Этот факт может использоваться при отборе факторов в модель. Добавление в модель новых факторов осуществляется до тех пор, пока растет скорректированный коэффициент детерминации.

Для оценки тесноты связи факторов с исследуемым признаком, задаваемой построенным уравнением регрессии $y = f(x_1, x_2 \ldots, x_m) + \varepsilon$, используется коэффициент множественной корреляции R:

$$R = \sqrt{R^2} = \sqrt{1 - \frac{S^2_{\text{ост}}}{S^2_y}} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - y^2_{x_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Коэффициент множественной корреляции R принимает значения в диапазоне $0 \leqslant R \leqslant 1$. Чем ближе величина R к единице, тем лучше зависимость $\widehat{y} = f(x_1, x_2, \ldots, x_m)$ согласуется с данными наблюдений. При $R = 1$ $(R^2 = 1)$ связь становится функциональной, т. е. соотношение $\widehat{y} = f(x_1, x_2, \ldots, x_m)$ точно выполняется для всех наблюдений. Коэффициент множественной корреляции может использоваться как характеристика качества построенного уравнения регрессии $\widehat{y} = f(x_1, x_2, \ldots, x_m)$, точности построенной модели. При правильном включении факторов в регрессионную модель должно выполняться соотношение $R > \max(r_{yx})$, то есть величина коэффициента множественной корреляции должна существенно отличаться от максимального парного коэффициента корреляции. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны, то коэффициент множественной корреляции может практически совпадать с коэффициентом парной корреляции (различия в третьем, четвертом знаках). Отсюда ясно, что сравнивая коэффициенты множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

В случае линейной зависимости коэффициент корреляции R связан с парными коэффициентами корреляции r_{yx} соотношением:

$$R = \sqrt{\sum_{i=1}^{m} \beta_i \cdot x_i},$$

где β_i — стандартизованные коэффициенты регрессии:

$$t_y = \beta_1 \cdot t_{x_1} + \beta_2 \cdot t_{x_2} + \ldots + \beta_m \cdot t_{x_m} + \varepsilon.$$

Значимость уравнения множественной регрессии в целом (а также коэффициента детерминации R^2) оценивается с помощью F-критерия Фишера:

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

где $S^2_{\text{факт}}$ — факторная сумма квадратов на одну степень свободы; $S^2_{\text{ост}}$ — остаточная сумма квадратов на одну степень свободы; R^2 — коэффициент множественной детерминации; m — число параметров при переменных x (в линейной регрессии совпадает с числом включенных в модель факторов); n — число наблюдений.

Согласно F-критерию Фишера выдвигаемая «нулевая» гипотеза H_0 о статистической незначимости уравнения регрессии отвергается при выполнении условия $F > F_{\text{табл}}$, где $F_{\text{табл}}$ определяется по таблицам F-критерия Фишера по двум степеням свободы $k_1 = m$, $k_2 = (n - m - 1)$ и заданному уровню значимости α.

Во множественном регрессионном анализе оценивается значимость не только уравнения в целом, но и фактора, дополнительно включенного в регрессионную модель. Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель. Мерой для оценки включения фактора в модель служит частный F-критерий (F_{x_i}).

В общем виде для фактора x_i частный F-критерий определяется следующим выражением:

$$F_{x_i} = \frac{R^2_{yx_1x_2\ldots x_m} - R^2_{yx_1\ldots x_{i-1}x_{i+1}\ldots x_m}}{1 - R^2_{yx_1x_2\ldots x_m}} \cdot \frac{n - m - 1}{1},$$

где $R^2_{yx_1x_2\ldots x_m}$ — коэффициент множественной детерминации для модели с полным набором факторов; $R^2_{yx_1\ldots x_{i-1}x_{i+1}\ldots x_m}$ — тот же показатель, но без включения в модель фактора x_i; n — число наблюдений; m — число параметров в модели (без свободного члена).

Для модели с двумя факторами частные F-критерии вычисляются по формулам:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n - 3); \quad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n - 3).$$

С помощью частного F-критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним.

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и степенях свободы $k_1 = 1$, $k_2 = n - m - 1$. Если для фактора x_i выполняется условие $F_{x_i} > F_{\text{табл}}$, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i

статистически значим. Если же фактическое значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y, следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Пример 2.3

Оценим качество уравнения, полученного в примере 2.2.

$$\hat{y} = -1{,}487 + 0{,}3005 \cdot x_1 + (-0{,}445) \cdot x_2.$$

При вычислении используем данные вспомогательной таблицы 2.3.

1. Оценим тесноту связи факторов с исследуемым признаком. Для этого вычислим парные коэффициенты корреляции и коэффициент множественной корреляции.

$$r_{yx_1} = \frac{\overline{y \cdot x_1} - \bar{y} \cdot \bar{x}_1}{\sigma_y \cdot \sigma_{x_1}} = \frac{2341{,}3 - 23{,}9 \cdot 96{,}7}{3{,}7 \cdot 9{,}96} = 0{,}819;$$

$$r_{yx_2} = \frac{\overline{y \cdot x_2} - \bar{y} \cdot \bar{x}_2}{\sigma_y \cdot \sigma_{x_2}} = \frac{194{,}25 - 23{,}9 \cdot 8{,}24}{3{,}7 \cdot 2{,}34} = -0{,}3099;$$

$$r_{x_1x_2} = \frac{\overline{x_1 \cdot x_2} - \bar{x}_1 \cdot \bar{x}_2}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{796 - 96{,}7 \cdot 8{,}24}{9{,}96 \cdot 2{,}34} = -0{,}035;$$

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{3{,}43}{13{,}6}} = 0{,}866.$$

Значения парных коэффициентов корреляции указывают на достаточно тесную связь доходов от продажи продукции y с постоянными затратами на ее производство x_1 и на умеренную связь с объемом выпуска продукции x_2. В то же время межфакторная связь $r_{x_1x_2}$ слабая ($r_{x_1x_2} = |-0{,}035| < 0{,}7$), что говорит о том, что оба фактора являются информативными, т. е. и x_1, и x_2 необходимо включить в модель.

2. Выполним оценку коэффициента множественной детерминации.

Коэффициент множественной детерминации ($R^2_{yx_1x_2}$) как квадрат совокупного коэффициента множественной корреляции равен 0,7496. Следовательно, 74,96% вариации результата объясняется вариацией представленных в уравнении признаков.

3. Выполним оценку частных коэффициентов корреляции.

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2}) \cdot (1 - r^2_{x_1x_2})}} = \frac{0{,}819 - 0{,}3099 \cdot 0{,}0346}{\sqrt{(1 - 0{,}096) \cdot (1 - 0{,}0012)}} = 0{,}85;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_1}) \cdot (1 - r^2_{x_1x_2})}} = \frac{-0{,}3099 + 0{,}8186 \cdot 0{,}0346}{\sqrt{(1 - 0{,}67) \cdot (1 - 0{,}0012)}} = -0{,}49.$$

То есть можно сделать вывод, что фактор x_1 оказывает более сильное влияние на результат, чем фактор x_2.

4. Оценим надежность уравнения регрессии в целом с помощью F-критерия Фишера.

Вычислим фактическое значение F-критерия:

$$F_{\text{факт}} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = \frac{0{,}7496}{1-0{,}7496} \cdot \frac{10-2-1}{2} = 10{,}48.$$

Табличное значение F-критерия при уровне значимости $\alpha = 0{,}05$ и степенях свободы $k_1 = 2$, $k_2 = 10 - 2 - 1 = 7$ равно 4,74. Так как $F_{\text{факт}} = 10{,}48 > F_{\text{табл}} = 4{,}74$, то уравнение признается статистически значимым.

5. Оценим целесообразность включения фактора x_1 после фактора x_2 и x_2 после x_1 с помощью частного F-критерия Фишера:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1-R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}7496 - 0{,}096}{1-0{,}7496} \cdot 7 = 18{,}266;$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1-R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}7496 - 0{,}67}{1-0{,}7496} \cdot 7 = 2{,}2187.$$

Табличное значение F-критерия при уровне значимости $\alpha = 0{,}05$ и степенях свободы $k_1 = 1$, $k_2 = 10 - 2 - 1 = 7$ равно 5,59. Так как $F_{x_1} = 18{,}266 > F_{\text{табл}} = 5{,}59$, то включение фактора x_1 в модель статистически оправдано и коэффициент чистой регрессии b_1 статистически значим. Для фактора x_2 выполняется условие $F_{x_2} = 2{,}2187 < F_{\text{табл}} = 5{,}59$, что говорит о нецелесообразности включения в модель фактора x_2, после того, как уже введен фактор x_1.

Контрольные вопросы по главе 2

1. Дайте определение множественной линейной регрессионной модели. Дайте краткую характеристику ее элементов.
2. В чем отличие целей построения модели парной регрессии и модели множественной регрессии?
3. Каким требованиям должны отвечать факторы модели множественной регрессии?
4. Что такое мультиколлинеарность факторов и как ее выявить?
5. Назовите основные алгоритмы построения модели множественной линейной регрессии.
6. Каковы свойства стандартизованных коэффициентов регрессии?
7. Как связаны между собой коэффициенты «чистой» регрессии и регрессии в стандартизованном масштабе?
8. В чем заключается смысл расчета скорректированного индекса корреляции и какова связь его с индексом корреляции при различном числе вводимых в модель факторов?
9. Какой критерий используется для оценки значимости коэффициентов регрессии?
10. Для чего используется частный F-критерий?

Глава 3

ГЕТЕРОСКЕДАСТИЧНОСТЬ И АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ

3.1 Предпосылки МНК

В задачу практического регрессионного анализа входит получение качественных оценок параметров уравнения регрессии. Качество оценок параметров определяется свойствами: несмещенность, состоятельность и эффективность.

Несмещенность оценки параметра означает, что ее математическое ожидание равно оцениваемому параметру:

$$M(b_j) = b_j^{\text{ген}},$$

где b_j — оценка параметра; $b_j^{\text{ген}}$ — значение параметра в генеральной совокупности.

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

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

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

$$b_j \xrightarrow[n \to \infty]{} b_j^{\text{ген}}.$$

Перечисленные свойства оценок параметров имеют чрезвычайно важное практическое значение в использовании результатов регрессии и обязательно учитываются при разных способах оценивания. МНК строит оценки регрессионной модели на основе минимизации суммы квадратов остатков, поэтому их свойства напрямую зависят от свойств случайной составляющей ε.

В модели

$$y = f(x_1, x_2, \ldots, x_m) + \varepsilon$$

случайная составляющая ε представляет собой ненаблюдаемую величину. На практике оценки случайной составляющей ε рассчитываются как разности фактических и теоретических значений результативного признака y:

$$\varepsilon_i = y_i - \widehat{y}_{x_i}.$$

Исследование остатков ε_i предполагает проверку выполнения пяти предпосылок МНК:

1) случайный характер остатков ε_i;
2) нулевая средняя величина остатков, не зависящая от x_j;
3) гомоскедастичность — дисперсия каждого отклонения ε_i одинакова для всех значений x;
4) отсутствие автокорреляции остатков — значения остатков ε_i распределены независимо друг от друга;
5) остатки подчиняются нормальному распределению.

Если распределение случайных остатков ε_i не соответствует некоторым предпосылкам МНК, то следует корректировать модель.

Проверка первой предпосылки МНК о случайном характере остатков ε_i выполняется визуально на основе графика зависимости остатков ε_i от теоретических значений результативного признака (рис. 3.1). Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и МНК оправдан, теоретические значения $\widehat{y}_x$ хорошо аппроксимируют фактические значения y.

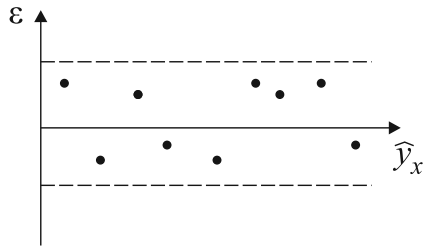


Рис. 3.1 – Зависимость случайных остатков ε от теоретических значений $\widehat{y}_x$

Проверка второй предпосылки МНК относительно нулевой средней величины остатков, не зависящей от x, означает что $\sum(y - \widehat{y}_x) = 0$. Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных.

В рамках соблюдения второй предпосылки МНК также исследуется независимость случайных остатков и величины x. С этой целью строится график зависимости случайных остатков ε от факторов, включенных в регрессию x_j (рис. 3.2).

Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_j. Если же график показывает наличие зависимости ε_i и x_j, то модель не может быть принята. Причины могут быть разные:

- нарушение третьей предпосылки МНК;
- неправильная спецификация модели и в нее необходимо ввести дополнительные члены от x_j, например x_j^2;

- наличие систематической погрешности модели, что отражается скоплением точек в определенных участках значений фактора x_j.

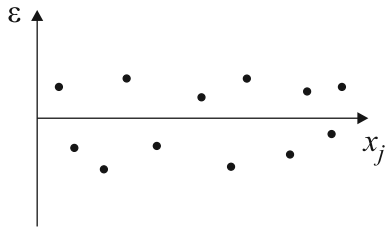
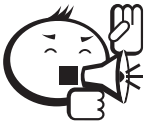


Рис. 3.2 – Зависимость величины остатков от величины фактора x_j



Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений. Это значит, что для каждого значения фактора x_j остатки ε_i имеют одинаковую дисперсию. Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсий отклонений)*. Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством дисперсий отклонений)*.

При невыполнимости предпосылки постоянства дисперсий отклонений последствия применения МНК будут следующими.

1. Оценки коэффициентов по-прежнему останутся несмещенными и линейными.
2. Оценки не будут эффективными (т. е. они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра). Они не будут даже асимптотически эффективными. Увеличение дисперсии оценок снижает вероятность получения максимально точных оценок.
3. Дисперсии оценок будут рассчитываться со смещением.
4. Вследствие вышесказанного все выводы, получаемые на основе соответствующих *t*- и *F*-статистик, а также интервальные оценки будут ненадежными. Следовательно, статистические выводы, получаемые при стандартных проверках качества оценок, могут быть ошибочными и приводить к неверным заключениям по построенной модели. Вполне вероятно, что стандартные ошибки коэффициентов будут занижены, а следовательно, *t*-статистики будут завышены. Это может привести к признанию статистически значимыми коэффициентов, таковыми на самом деле не являющихся [4].

Обнаружение гетероскедастичности дисперсии остатков может быть выполнено различными методами. К настоящему времени разработано большое число тестов и критериев для них. Наиболее популярные из них: графический анализ остатков, тест ранговой корреляции Спирмена, тест Парка, тест Голдфелда—Квандта.

При построении регрессионных моделей чрезвычайно важно соблюдение четвертой предпосылки МНК — отсутствие автокорреляции остатков, т. е. значения

остатков ε_i распределены независимо друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений. Коэффициент корреляции между ε_i и ε_{i-1}, где ε_i — остатки текущих наблюдений; ε_{i-1} — остатки предыдущих наблюдений, может быть определен как:

$$r_{\varepsilon_i\varepsilon_{i-1}} = \frac{\operatorname{cov}(\varepsilon_i, \varepsilon_{i-1})}{\sigma_{\varepsilon_i} \cdot \sigma_{\varepsilon_{i-1}}},$$

т. е. по обычной формуле линейного коэффициента корреляции. Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированы и функция плотности вероятности $F(\varepsilon)$ зависит от i-й точки наблюдения и от распределения значений остатков в других точках наблюдения.

Отсутствие автокорреляции остаточных величин обеспечивает состоятельность и эффективность оценок коэффициентов регрессии. Особенно актуально соблюдение данной предпосылки МНК при построении регрессионных моделей по рядам динамики, где ввиду наличия тенденции последующие уровни динамического ряда, как правило, зависят от своих предыдущих уровней.

Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью F- и t-критериев. Вместе с тем оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т. е. при нарушении пятой предпосылки МНК.

3.2 Гетероскедастичность. Обнаружение гетероскедастичности

Гомоскедастичность — постоянство дисперсии остатков.

Гетероскедастичность — непостоянство дисперсии остатков.

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

3.2.1 Графический анализ остатков

Для проведения визуального анализа дисперсии остатков необходимо построить график зависимости дисперсии остатков ε_i^2 от значений переменной x_j. В случае гомоскедастичности дисперсии остатков все отклонения ε_i^2 находятся внутри полосы постоянной ширины, параллельной оси абсцисс (рис. 3.3). Все прочие случаи соответствуют гетероскедастичности остатков. При множественной регрессии графический анализ возможен как для каждой объясняющей переменной x_j, так и для выровненного значения результата $\widehat{y}_x$. В этом случае по оси абсцисс откладываются значения $\widehat{y}_{x_i}$.

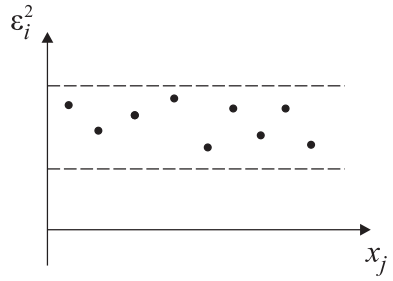


Рис. 3.3 – Зависимость ε_i^2 от величины фактора x_j

3.2.2 Тест ранговой корреляции Спирмена

Использование данного теста основано на предположении, что дисперсии отклонений увеличиваются либо уменьшаются с ростом значения какого-либо фактора. Поэтому для регрессии, построенной по МНК, абсолютные величины отклонений ε_i и значения x_i будут коррелированы. Для проведения проверки по этому тесту выполняются следующие действия:

1) ранжируются (упорядочиваются по величинам) значения модулей остатков ε_i и значения выбранного фактора x_i;
2) определяется коэффициент ранговой корреляции Спирмена:

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

где x — одна из объясняющих переменных; d_i — разность между рангом i-го наблюдения x и рангом модуля остатка в i-м наблюдении, $i = 1, 2, \ldots, n$; n — число наблюдений;

3) проверяется значимость вычисленного коэффициента ранговой корреляции.

Гипотеза H_0: $r_{x,\varepsilon} = 0$ — гомоскедастичность остатков.
Гипотеза H_1: $r_{x,\varepsilon} \neq 0$ — гетероскедастичность остатков.
Для проверки гипотезы H_0 рассчитывается фактическое значение t-критерия:

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

Если значение, рассчитанное по указанной формуле, превышает табличное $t_{\text{табл}} = t_{\alpha, n-2}$, гипотеза H_0 о гомоскедастичности остатков отклоняется. В противном случае гипотеза о гомоскедастичности принимается.

Если в модели регрессии больше чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью t-статистики для каждой из них отдельно.

Пример 3.1

Изучим зависимость спроса на товар от его цены (столбцы 2, 3 табл. 3.1). После построения регрессии вычислим остатки (столбец 4). Для анализа остатков с помощью теста ранговой корреляции Спирмена выполним следующие действия:

1) отсортируем данные в таблице 3.1 по возрастанию значений x;
2) присвоим каждому наблюдению ранг, для чего необходимо добавить новый столбец и в нем задать числа от 1 до n (столбец 1);
3) отсортируем данные по возрастанию модулей остатков и добавим новый столбец (столбец 5) рангов остатков, задав значения от 1 до n;
4) в дополнительном столбце вычислим значения разности между двумя полученными рангами (это и будет значение d_i);

Таблица 3.1 – Тест ранговой корреляции Спирмена

Ранг по x	**Цена x (р.)**	**Спрос y (тыс. шт.)**	**Остатки**	**Ранг по остаткам**	**Разность рангов d_i**	$d_i \cdot d_i$
1	2	3	4	5	6	7
8	15,91	117,088	–0,32	1	7	49
5	15,54	119,864	–0,396	2	3	9
15	16,76	110,023	–0,84	3	12	144
2	15,21	123,809	1,006	4	–2	4
3	15,28	121,175	–1,088	5	–2	4
9	15,92	116,17	–1,163	6	3	9
10	15,95	118,344	1,241	7	3	9
14	16,69	110,106	–1,296	8	6	36
1	15,09	125,178	1,450	9	–8	64
6	15,62	118,068	–1,576	10	–4	16
11	16,31	116,201	1,87	11	0	0
12	16,33	111,457	–2,72	12	0	0
13	16,60	115,103	3,01	13	0	0
4	15,49	116,914	–3,73	14	–10	100
7	15,70	123,589	4,56	15	–8	64
					Сумма	515

5) вычислим коэффициент ранговой корреляции и t-статистику и проверим гипотезу о гомоскедастичности остатков.

$$r_{x,\varepsilon} = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 515}{15 \cdot (225 - 1)} = 0{,}0804.$$

$$t = \frac{0{,}0804 \cdot \sqrt{15 - 2}}{\sqrt{1 - 0{,}0065^2}} = 0{,}29.$$

Выбрав уровень значимости 5%, получим критическую точку $t_{0,05,\,13} = 2{,}16$.

Поскольку условие $|t| > t_{\alpha,\,n-2}$ не выполняется, то гипотеза о наличии гомоскедастичности будет принята.

3.2.3 Тест Парка

Тест Парка основан на предположении, что дисперсия $\sigma_i^2 = \sigma^2(\varepsilon_i)$ является функцией i-го значения объясняющей переменной X. Парк предложил следующую зависимость:

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

где x_{ij} — i-е значение j-го фактора; ν_i — случайный остаток.

Выдвигаются гипотеза H_0: $b = 0$, что соответствует гомоскедастичности остатков, и гипотеза H_1: $b \neq 0$, которая выявляет наличие связи между $\ln \varepsilon_i^2$ и $\ln x_{ij}$. Отсюда следует, что гетероскедастичность остатков имеет место.

Условие принятия гипотезы H_1: $t_b > t_{\alpha,\,n-2}$.

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

Пример 3.2

Проверим гипотезу о гетероскедастичности остатков с помощью теста Парка для данных из примера 3.1.

Зависимость между остаточной дисперсией и объясняющим фактором имеет вид:

$$\ln \varepsilon^2 = -3{,}36 + 1{,}45 \cdot \ln(x);$$
$$t_b = 0{,}12.$$

Табличное значение критерия Стьюдента равно $t_{0,05,\,13} = 2{,}16$.

Поскольку условие $t_b < t_{\alpha,\,n-2}$ выполняется, то гипотеза о наличии гетероскедастичности отклоняется.

3.2.4 Тест Голдфелда—Квандта

Тест Голдфелда—Квандта применяется, если случайные остатки предполагаются нормально распределенными случайными величинами и стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению x_i переменной X в этом наблюдении, т. е. $\sigma_i^2 = \sigma^2 x_i^2$, $i = 1, 2, \ldots, n$.

Процедура проверки состоит в следующем.

1. Все наблюдения упорядочиваются по возрастанию фактора X.

2. Упорядоченная совокупность разбивается на три группы размерностей k, $(n - 2 \cdot k)$, k соответственно. Причем k должно быть больше чем число параметров модели. Для парной регрессии Голдфелд и Квандт предлагают следующие пропорции: $n = 30$, $k = 11$; $n = 60$, $k = 22$.

3. Оцениваются отдельные регрессии для первой группы (k первых наблюдений) и для третьей группы (k последних наблюдений). Если предположение о пропорциональности дисперсий отклонений значениям фактора X верно, то дисперсия регрессии по первой группе (рассчитываемая как $S_1 = \sum_{i=1}^{k} \varepsilon_i^2$) будет существенно меньше дисперсии регрессии по третьей группе (рассчитываемой как $S_3 = \sum_{i=n-k+1}^{n} \varepsilon_i^2$).

4. Формулируются:

- основная гипотеза, предполагающая постоянство дисперсий случайных ошибок модели регрессии, т. е. присутствие в модели условия гомоскедастичности: H_0: $S_1 = S_3$;
- альтернативная гипотеза, предполагающая непостоянство дисперсий случайных ошибок в различных наблюдениях, т. е. присутствие в модели условия гетероскедастичности: H_1: $S_1 \neq S_3$.

5. Для сравнения соответствующих дисперсий вычисляется фактическое значение F-критерия:

$$F_{\text{факт}} = \frac{S_3/(k - m - 1)}{S_1/(k - m - 1)} = \frac{S_3}{S_1}.$$

Здесь $(k - m - 1)$ — число степеней свободы соответствующих выборочных дисперсий (m — количество объясняющих переменных в уравнении регрессии).

Если $(F_{\text{факт}} = S_3/S_1) > F_{\text{табл}}$ (где $F_{\text{табл}} = F_{\alpha, k_1, k_2}$, α — выбранный уровень значимости), то гипотеза H_0 об отсутствии гетероскедастичности отклоняется.

Этот же тест может использоваться при предположении об обратной пропорциональности между σ_i и значениями объясняющей переменной. При этом статистика Фишера имеет вид:

$$F = \frac{S_1}{S_3}.$$

Для множественной регрессии данный тест обычно проводится для той объясняющей переменной, которая в наибольшей степени связана с σ_i. При этом k должно быть больше, чем $(m + 1)$. Если нет уверенности относительно выбора переменной X_j, то данный тест может осуществляться для каждой из объясняющих переменных.

Пример 3.3

Проверим гипотезу о гетероскедастичности остатков с помощью теста Гольдфелда—Квандта для данных из примера 3.1.

1. Данные таблицы 3.1 упорядочим по значению фактора x (табл. 3.2).

Таблица 3.2 – Тест Гольдфелда–Квандта

№	Цена x (р.)	Спрос y (тыс. шт.)
1	15,09	125,178
2	15,21	123,809
3	15,28	121,175
4	15,49	116,914
5	15,54	119,864
6	15,62	118,068
7	15,70	123,589
8	15,91	117,088
9	15,92	116,170
10	15,95	118,344
11	16,31	116,201
12	16,33	111,457
13	16,60	115,103
14	16,69	110,106
15	16,76	110,023

2. Определим значение $k = 6$.
3. Оценим регрессии по первой и третьей группе данных:

$$\widehat{y}_1 = 335{,}992 - 13{,}998 \cdot x, \quad S_1 = 9{,}1,$$
$$\widehat{y}_3 = 263{,}939 - 9{,}148 \cdot x, \quad S_3 = 22{,}63.$$

4. Вычислим фактическое значение F-критерия:

$$F_{\text{факт}} = \frac{S_3}{S_1} = \frac{22{,}63}{9{,}1} = 2{,}486.$$

Табличное значение критерия Фишера $F_{0{,}05,4,4} = 6{,}39$.

Поскольку условие $F_{\text{факт}} < F_{\alpha,k_1,k_2}$ выполняется, то гипотеза о наличии гетероскедастичности отклоняется.

3.3 Методы устранения гетероскедастичности

При нарушении гомоскедастичности и наличии автокорреляции ошибок рекомендуется традиционный метод наименьших квадратов (известный в английской терминологии как метод OLS — Ordinary Least Squares) заменять *обобщенным методом*, т. е. *методом GLS* (Generalized Least Squares).

Применение обычного МНК к модели, в которой нарушены эти предпосылки, ведет к тому, что найденные параметры уравнения регрессии не будут эффективными оценками генеральных параметров. Кроме того, их дисперсии будут рассчитаны со смещением, что приведет к ложным выводам при оценке качества модели и при проведении прогнозирования по ней.

Для случая гетероскедастичности остатков обобщенный метод наименьших квадратов (ОМНК) называют еще методом взвешенных наименьших квадратов (ВМНК). ОМНК используется для корректировки гетероскедастичности за счет преобразования данных, позволяющего получать оценки, которые обладают не только свойством несмещенности, но и имеют меньшие выборочные дисперсии.

Пусть σ_{ε_i} — стандартное отклонение случайной ошибки ε_i в i-м наблюдении. В случае если σ_{ε_i} известно, гетероскедастичность можно корректировать, разделив каждое наблюдение на соответствующее ему значение σ_{ε_i}. Так, для парной регрессии $y_i = a + b \cdot x_i + \varepsilon_i$ соответствующее преобразование данных будет иметь вид:

$$\frac{y_i}{\sigma_{\varepsilon_i}} = \frac{a}{\sigma_{\varepsilon_i}} + b \cdot \frac{x_i}{\sigma_{\varepsilon_i}} + \frac{\varepsilon_i}{\sigma_{\varepsilon_i}}.$$

Тогда дисперсия остатков представляется в виде:

$$D\left(\frac{\varepsilon_i}{\sigma_{\varepsilon_i}}\right) = \frac{1}{\sigma^2_{\varepsilon_i}} \cdot \sigma^2(\varepsilon_i) = \frac{\sigma^2_{\varepsilon_i}}{\sigma^2_{\varepsilon_i}} = 1.$$

В результате этой процедуры каждое наблюдение будет иметь случайную ошибку с единичной дисперсией. Следовательно, для преобразованной модели выполняется предпосылка МНК о гомоскедастичности дисперсии остатков, а оценки параметров регрессии, полученные по МНК, будут наилучшими несмещенными оценками.

Применение вышеописанного метода в значительной степени ограничено тем, что на практике фактические значения σ_{ε_i} чаще всего неизвестны. В этом случае применение ОМНК основано на предположении, что среднее значение остаточных величин равно нулю, а вот дисперсия их представлена в виде произведения некоторой величины K_i на постоянную величину σ^2:

$$\sigma^2_{\varepsilon_i} = \sigma^2 \cdot K_i.$$

При этом в отношении величин K_i выдвигаются определенные гипотезы, характеризующие структуру гетероскедастичности.

Тогда уравнение $y_i = a + b \cdot x_i + \varepsilon_i$ преобразуется к виду:

$$\frac{y_i}{\sqrt{K_i}} = \frac{a}{\sqrt{K_i}} + b \cdot \frac{x_i}{\sqrt{K_i}} + \frac{\varepsilon_i}{\sqrt{K_i}}.$$

В данной модели остаточные величины гетероскедастичны, следовательно для регрессии применим обычный МНК. Действительно, в силу выполнимости предпосылки $\sigma^2_{\varepsilon_i} = \sigma(\varepsilon_i) = \sigma^2 \cdot K_i$ имеем:

$$\sigma^2\left(\frac{\varepsilon_i}{\sqrt{K_i}}\right) = \frac{1}{K_i} \cdot \sigma^2(\varepsilon_i) = \frac{1}{K_i} \cdot \sigma^2 \cdot K_i = \sigma^2 = \text{const}.$$

Оценка параметров нового уравнения с преобразованными переменными основана на минимизации суммы квадратов отклонений вида $S = \sum_{i=1}^{n} \frac{1}{K_i} \cdot (y_i - a - b \cdot x_i)^2$ и последующего решения системы уравнений:

$$\begin{cases} \sum \frac{y}{K} = a \cdot \sum \frac{1}{K} + b \cdot \sum \frac{x}{K}, \\ \sum \frac{y \cdot x}{K} = a \cdot \sum \frac{x}{K} + b \cdot \sum \frac{x^2}{K}. \end{cases}$$

Аналогичный подход возможен не только для уравнения парной, но и для множественной регрессии. Например, рассматривается модель вида:

$$y_i = a + b_1 \cdot x_{1_i} + b_2 \cdot x_{2_i} + \varepsilon_i,$$

для которой дисперсия остаточных величин оказалась пропорциональна K_i^2. Коэффициент пропорциональности K принимает различные значения для соответствующих i значений факторов x_1 и x_2. Ввиду того, что

$$\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i^2,$$

для корректировки гетероскедастичности выполняется переход к уравнению с новым преобразованным переменным:

$$\frac{y_i}{K_i} = \frac{a}{K_i} + b_1 \cdot \frac{x_{1_i}}{K_i} + b_2 \cdot \frac{x_{2_i}}{K_i} + \frac{\varepsilon_i}{K_i}.$$

Параметры такой модели зависят от концепции, принятой для коэффициента пропорциональности K. В эконометрических исследованиях довольно часто выдвигается гипотеза, что остатки ε_i пропорциональны значениям какого-либо фактора. Так, если в уравнении

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m + \varepsilon$$

предположить, что $K_i = x_{1_i}$ и $\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot x_{1_i}^2$, то ОМНК предполагает оценку параметров следующего трансформированного уравнения:

$$\frac{y_i}{x_{1_i}} = b_1 + b_2 \cdot \frac{x_{2_i}}{x_{1_i}} + \ldots + b_m \cdot \frac{x_{m_i}}{x_{1i}} + \frac{\varepsilon_i}{x_{1_i}}.$$

Таким образом, «взвешивая» каждый остаток $\varepsilon_i = (\widehat{y}_i - y_i)$ с помощью коэффициента $1/K_i$, можно добиться равномерного вклада остатков в общую сумму, что приводит в конечном итоге к получению наиболее эффективных оценок параметров регрессии. Вместе с тем следует иметь в виду, что новые преобразованные переменные получают новое экономическое содержание и их регрессия имеет иной смысл, чем регрессия по исходным данным.

Пример 3.4

Пусть y — издержки производства, x_1 — объем продукции, x_2 — основные производственные фонды, x_3 — численность работников, тогда уравнение

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \varepsilon$$

является моделью издержек производства с объемными факторами. Предполагая, что дисперсия остатков пропорциональна квадрату численности работников $\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot x_3^2$, мы получим в качестве результативного признака затраты на одного работника y/x_3, а в качестве факторов следующие показатели: производительность

труда x_1/x_3 и фондовооруженность труда x_2/x_3. Соответственно трансформированная модель примет вид:

$$\frac{y}{x_3} = b_3 + b_1 \cdot \frac{x_1}{x_3} + b_2 \cdot \frac{x_2}{x_3} + \frac{\varepsilon}{x_3},$$

где параметры b_1, b_2, b_3 численно не совпадают с аналогичными параметрами предыдущей модели. Кроме этого, коэффициенты регрессии меняют экономическое содержание: из показателей силы связи, характеризующих среднее абсолютное изменение издержек производства с изменением абсолютной величины соответствующего фактора на единицу, они фиксируют при обобщенном МНК среднее изменение затрат на работника; с изменением производительности труда на единицу при неизменном уровне фондовооруженности труда; и с изменением фондовооруженности труда на единицу при неизменном уровне производительности труда.

Переход к относительным величинам существенно снижает вариацию фактора и соответственно уменьшает дисперсию ошибки. Он представляет собой наиболее простой случай учета гетероскедастичности в регрессионных моделях с помощью ОМНК. Процесс перехода к относительным величинам может быть осложнен выдвижением иных гипотез о пропорциональности ошибок относительно включенных в модель факторов. Использование той или иной гипотезы предполагает специальные исследования остаточных величин для соответствующих регрессионных моделей. Применение ОМНК позволяет получить оценки параметров модели, обладающие меньшей дисперсией.

. . . . . .

3.4 Автокорреляция в остатках

Автокорреляция (последовательная корреляция) определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные данные).

Важной предпосылкой построения качественной регрессионной модели по МНК является независимость значений случайных отклонений ε_i от значений отклонений во всех других наблюдениях.

Автокорреляция в остатках может быть вызвана несколькими причинами, имеющими различную природу.

1. Автокорреляция может быть связана с исходными данными и вызвана наличием ошибок измерения в значениях результативного признака.
2. В ряде случаев автокорреляция может быть следствием неправильной спецификации модели. Модель может не включать фактор, который оказывает существенное воздействие на результат и влияние которого отражается в остатках, вследствие чего последние могут оказаться автокоррелированными. Очень часто этим фактором является фактор времени t.

Автокорреляция остатков чаще встречается в регрессионном анализе при использовании данных временных рядов (глава 6).

Последствия автокорреляции схожи с последствиями гетероскедастичности: выводы по t- и F-статистикам, определяющие значимость коэффициента регрессии и коэффициента детерминации, возможно, будут неверными.

Обнаружение автокорреляции. Критерий Дарбина—Уотсона

Как было отмечено в параграфе 3.1, проверка некоррелированности остатков может быть выполнена с помощью коэффициента корреляции, называемого в этом случае *коэффициентом автокорреляции первого порядка*:

$$r_{\varepsilon_i\varepsilon_{i-1}} = \frac{\operatorname{cov}(\varepsilon_i, \varepsilon_{i-1})}{\sigma_{\varepsilon_i} \cdot \sigma_{\varepsilon_{i-1}}}.$$

При этом проверяется некоррелированность только соседних величин ε_i. Соседними обычно считаются соседние во времени (при рассмотрении временных рядов) или по возрастанию объясняющей переменной X (в случае перекрестной выборки) значения ε_i.

На практике для анализа коррелированности отклонений вместо коэффициента корреляции используют статистику Дарбина—Уотсона (DW), рассчитываемую по формуле:

$$DW = \frac{\sum (\varepsilon_i - \varepsilon_{i-1})^2}{\sum \varepsilon_i^2}.$$

Критерий Дарбина—Уотсона и коэффициент автокорреляции связаны между собой соотношением:

$$DW \cong 2 \cdot (1 - r_{\varepsilon_i\varepsilon_{i-1}}).$$

Из этого соотношения видно, что при положительной автокорреляции остатков ($r_{\varepsilon_i\varepsilon_{i-1}} = 1$) критерий $DW = 0$. При полной отрицательной автокорреляции ($r_{\varepsilon_i\varepsilon_{i-1}} = -1$) критерий $DW = 4$. Если автокорреляция в остатках отсутствует ($r_{\varepsilon_i\varepsilon_{i-1}} = 0$), критерий $DW = 2$. Следовательно, критерий Дарбина—Уотсона изменяется в пределах: $0 \leqslant DW \leqslant 4$.

Алгоритм выявления автокорреляции остатков с использованием критерия Дарбина—Уотсона следующий.

1. Выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Альтернативные гипотезы H_1 и H_1^* состоят, соответственно, в наличии положительной или отрицательной автокорреляции в остатках.

2. Определяются критические значения критерия Дарбина—Уотсона d_L и d_U для заданного числа наблюдений n, числа независимых переменных модели m и уровня значимости α (приложение А).

3. По этим значениям числовой промежуток $[0; 4]$ разбивают на пять отрезков (рис. 3.4).

4. Вычисляется фактическое значение критерия Дарбина—Уотсона:

$$DW = \frac{\sum (\varepsilon_i - \varepsilon_{i-1})^2}{\sum \varepsilon_i^2}.$$

5. Принятие или отклонение каждой из гипотез с вероятностью $1 - \alpha$ осуществляется следующим образом:

- $0 < DW < d_L$ — наблюдается положительная автокорреляция остатков, H_0 отклоняется, с вероятностью $P = 1 - \alpha$ принимается H_1;
- $d_L \leqslant DW \leqslant d_U$ — зона неопределенности;
- $d_U < DW < 4 - d_U$ — нет оснований отклонять H_0, т. е. автокорреляция остатков отсутствует;
- $4 - d_U \leqslant DW \leqslant 4 - d_L$ — зона неопределенности;
- $4 - d_L < DW < 4$ — наблюдается отрицательная автокорреляция остатков, H_0 отклоняется, с вероятностью $P = 1 - \alpha$ принимается H_1^*.

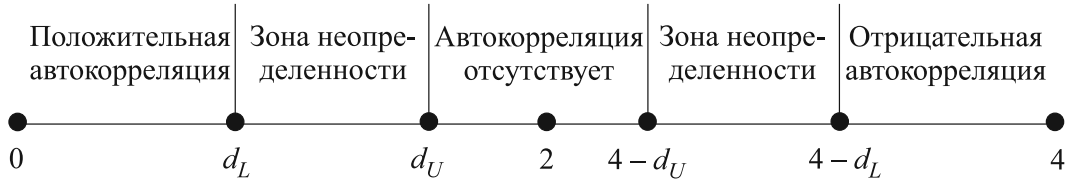


Рис. 3.4 – Схема применения критерия Дарбина—Уотсона

Если фактическое значение критерия Дарбина—Уотсона попадает в зону неопределенности, то на практике либо предполагают существование автокорреляции остатков и отклоняют гипотезу H_0, либо проводят дополнительные исследования, например по большему числу наблюдений.

Контрольные вопросы по главе 3

1. В чем суть гетероскедастичности?
2. Каковы последствия автокорреляции и гетероскедастичности остатков?
3. В чем суть обобщенного МНК (ОМНК)?
4. Приведите схему теста Голдфелда–Квандта.
5. Приведите схему теста Парка.
6. Приведите схему теста Спирмена.
7. Опишите алгоритм выявления автокорреляции остатков с использованием критерия Дарбина—Уотсона.

Глава 4

РЕГРЕССИОННЫЕ МОДЕЛИ С ПЕРЕМЕННОЙ СТРУКТУРОЙ

4.1 Понятие фиктивных переменных

Экономические величины складываются под влиянием множества различных факторов, как количественных, так и качественных по своей природе. Это могут быть разного рода атрибутивные признаки, такие, например, как профессия, пол, образование и пр., а также факторы, оказывающие косвенное воздействие (во времени и/или пространстве) на изучаемый процесс, что приводит к неоднородной выборке рассматриваемых показателей. Иногда представляет интерес включение этих факторов в эконометрическую модель и исследование их влияния на изучаемую зависимость. Например, влияние пола или образования на уровень заработной платы или влияние дефолта на величину основных макроэкономических показателей.

Возможным решением было бы разбить имеющиеся исходные статистические данные на заведомо однородные группы и строить модели для каждой однородной выборки с последующим выяснением различия в моделях. Например, построить модели зависимости заработной платы от стажа отдельно для мужчин и женщин или изучать поведение макроэкономических показателей отдельно на временном интервале до дефолта и после.

Другой возможный подход состоит в построении и оценивании одной модели для всей совокупности наблюдений и измерении влияния качественного фактора, явившегося причиной появления неоднородной выборки. Чтобы ввести качественные факторы в регрессионную модель, им должны быть присвоены те или иные *цифровые метки*, т. е. качественные переменные преобразованы в количественные. Такого вида сконструированные переменные в эконометрике принято называть *фиктивными переменными или дамми-переменными*.

Построение модели с фиктивными переменными обладает следующими преимуществами:

- имеется простой способ проверки, является ли воздействие качественного фактора значимым (путем проверки на статистическую значимость коэффициента перед фиктивной переменной);
- при условии выполнения определенных предположений оценки модели оказываются более эффективными (вследствие большей выборки).

Регрессионные модели могут содержать одновременно как количественные, так и качественные переменные, и даже только качественные. В эконометрике рассматриваются различные варианты моделей регрессии с фиктивными переменными.

4.2 Модели регрессии с фиктивными переменными сдвига

Рассмотрим применение фиктивных переменных для функции спроса. Предположим, что по группе лиц мужского и женского пола изучается линейная зависимость потребления кофе от цены [1]. В общем виде для общей совокупности наблюдений уравнение регрессии имеет вид:

$$y = a + b \cdot x + \varepsilon,$$

где y — количество потребляемого кофе; x — цена.

Аналогичные уравнения могут быть найдены, если рассматривать отдельно потребление кофе для категории «лица мужского пола»: $y_1 = a_1 + b_1 \cdot x_1 + \varepsilon_1$ и категории «лица женского пола»: $y_2 = a_2 + b_2 \cdot x_2 + \varepsilon_2$.

Различия в потреблении кофе проявятся в различии средних $\bar{y}_1$ и $\bar{y}_2$. Вместе с тем сила влияния x на y может быть одинаковой, т. е. $b \approx b_1 \approx b_2$. Для построения общего уравнения регрессии, учитывающего различия в потреблении кофе мужчинами и женщинами, возможно включение в него фактора «пол» в виде фиктивной переменной:

$$y = a + bx + \delta \cdot z + \varepsilon.$$

В этом случае зависимая переменная y рассматривается как функция не только цены x, но и пола z. Переменная z рассматривается как бинарная переменная, принимающая всего два значения: 1 и 0.

$$z = \begin{cases} 1 — \text{мужской пол}, \\ 0 — \text{женский пол}. \end{cases}$$

Тогда уравнение для лиц женского пола можно записать: $\hat{y} = a + bx$, а для лиц мужского пола: $\hat{y} = (a + \delta) + bx$.

Сравнивая два полученных уравнения, видим, что они различаются величиной свободного члена. То есть для одного уровня неколичественной переменной уровень результата в среднем будет на δ единиц выше или ниже другого. Иными словами δ показывает сдвиг в потреблении кофе мужчинами по сравнению с женщинами (рис. 4.1).

Если рассмотреть зависимость потребления кофе не только от цены, но и от региона проживания: северные регионы, центральные и южные, то в этом случае

все данные разбиваются на три категории. В модель вводятся две фиктивные переменные z_1 и z_2:

$$z_1 = \begin{cases} 1 — \text{проживание в северном регионе}, \\ 0 — \text{в остальных регионах}; \end{cases}$$

$$z_2 = \begin{cases} 1 — \text{проживание в южном регионе}, \\ 0 — \text{в остальных регионах}. \end{cases}$$

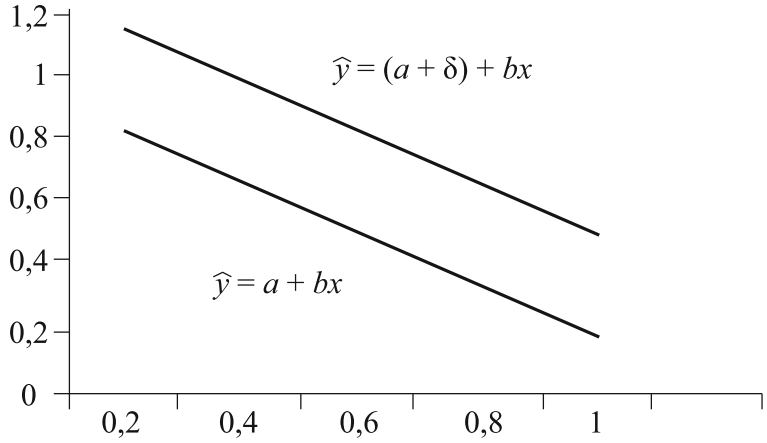


Рис. 4.1 – Модель регрессии с фиктивной переменной сдвига δ

Значение $z_1 = 0$ и $z_2 = 0$ принимается за эталонное и определяет среднее значение потребления кофе проживающих в центральном регионе.

Линейная регрессионная модель в этом случае определяется следующим уравнением:

$$\hat{y} = a + bx + \delta_1 \cdot z_1 + \delta_2 \cdot z_2,$$

где коэффициенты δ_1 и δ_2 показывают сдвиг в объеме потребления кофе в соответствующих регионах по отношению к потреблению кофе в центральных регионах.

Таким образом, построение модели с фиктивными переменными требует выполнения следующих этапов:

1) статистические данные разбиваются на категории, число которых определяется числом значений качественного признака. Одна из категорий принимается за эталонную (выбирается произвольно);

2) вводятся фиктивные переменные для всех категорий, кроме эталонной. Каждая из введенных фиктивных переменных принимает значение, равное единице для данных рассматриваемой категории и нуль для данных остальных категорий;

3) фиктивные переменные вводятся в уравнение с коэффициентом δ_i, $i = \overline{1, k-1}$, где k — число категорий. Каждый из коэффициентов δ_i характеризует сдвиг значения результативного показателя для данных i-ой категории относительно эталонной. Если δ_i оказывается статистически значимым, то фактор (событие), выражаемый этой фиктивной переменной, оказывает существенное влияние на результативный показатель.

Модель может содержать несколько качественных признаков. В этом случае фиктивные переменные для каждого признака вводятся в соответствии с вышеприведенной методикой.

Пример 4.1

Предположим, что изучается потребление кофе в зависимости от цены, пола и региона проживания: северные регионы, центральные и южные (табл. 4.1).

Таблица 4.1 – Данные к примеру 4.1

N	**Потребл. (кг)**	**Цена (тыс. руб.)**	**Пол**	**Северн. регион**	**Южн. регион**	N	**Потребл. (кг)**	**Цена (тыс. руб.)**	**Пол**	**Северн. регион**	**Южн. регион**
	Y	x	z	z_1	z_2		Y	x	z	z_1	z_2
1	0,2	1	1	0	0	16	0,6	0,6	0	1	0
2	0,4	1	0	0	0	17	0,6	0,5	0	1	0
3	0,4	0,8	1	0	0	18	0,65	0,5	1	1	0
4	0,6	0,8	0	0	0	19	0,6	0,3	1	1	0
5	0,6	0,6	1	0	0	20	0,7	0,3	0	1	0
6	0,8	0,6	0	0	0	21	0,5	1	1	0	1
7	0,75	0,5	0	0	0	22	0,6	1	0	0	1
8	0,9	0,5	1	0	0	23	0,7	0,8	1	0	1
9	0,9	0,3	1	0	0	24	0,9	0,8	0	0	1
10	1,1	0,3	0	0	0	25	0,9	0,6	1	0	1
11	0,2	1	1	1	0	26	1,1	0,6	0	0	1
12	0,45	1	0	1	0	27	1	0,5	0	0	1
13	0,45	0,8	1	1	0	28	1,2	0,5	1	0	1
14	0,6	0,8	0	1	0	29	1,2	0,3	1	0	1
15	0,5	0,6	1	1	0	30	1,4	0,3	0	0	1

Вводим фиктивную бинарную переменную z для признака «пол» и две бинарные переменные z_1 и z_2 для регионов проживания.

Линейная регрессионная модель запишется:

$$\hat{y} = a + bx + \delta \cdot z + \delta_1 \cdot z_1 + \delta_2 \cdot z_2.$$

Коэффициент δ показывает сдвиг в потреблении кофе мужчинами относительно женщин, а коэффициенты δ_1 и δ_2 соответственно показывают сдвиг в объеме потребления кофе в северных и южных регионах относительно центрального региона. После вычисления коэффициентов уравнение регрессии имеет вид:

$$\hat{y} = 1{,}26 - 0{,}84x - 0{,}11z - 0{,}13z_1 + 0{,}29z_2.$$

Коэффициент детерминации $R^2 = 0{,}88$, что говорит о хорошем качестве модели. Статистическая значимость модели в целом подтверждается значением критерия Фишера $F = 47{,}951 > F_{\text{табл}} = 2{,}76$.

Оценка значимости коэффициентов регрессии выполняется на основе анализа следующих величин (табл. 4.2).

Таблица 4.2 – Оценка значимости коэффициентов регрессии

	Значения коэффициентов	**Стат. ошибка**	**Значения *t*-статистики**	***P*-значение**
a	1,26	0,07	19,26	$1{,}64E - 16$
b	–0,84	0,08	–10,30	$1{,}77E - 10$
δ	–0,11	0,04	–2,88	0,008
δ_1	–0,13	0,05	–2,70	0,012
δ_2	0,29	0,05	5,91	$3{,}63E - 0{,}6$

Так как расчетные значения t-статистики для всех коэффициентов по модулю превышают табличное значение $t_{0{,}05,\,10} = 2{,}228$, то они статистически значимы. Следовательно, потребление кофе существенно зависит от цены, пола и проживания в определенном регионе.

Можно построить отдельные уравнения для мужчин и женщин каждого региона (табл. 4.3).

Таблица 4.3 – Уравнения регрессии

Тип категории	**Уравнение**
Женщины (северные регионы)	$\widehat{Y} = 1{,}13 - 0{,}84x$
Мужчины (северные регионы)	$\widehat{Y} = 1{,}02 - 0{,}84x$
Женщины (центральные регионы)	$\widehat{Y} = 1{,}26 - 0{,}84x$
Мужчины (центральные регионы)	$\widehat{Y} = 1{,}15 - 0{,}84x$
Женщины (южные регионы)	$\widehat{Y} = 1{,}55 - 0{,}84x$
Мужчины (южные регионы)	$\widehat{Y} = 1{,}44 - 0{,}84x$

В этих уравнениях различны только свободные члены, угол наклона всех прямых одинаков (одинаковый коэффициент перед переменной «цена»).

. . . . . .

4.3 Модели регрессии с фиктивными переменными наклона

Модели регрессии с фиктивными переменными наклона отражают зависимость количественного фактора от значения фиктивной переменной [1]. Это может быть представлено следующими выражениями:

$$\widehat{y} = a + b_1 \cdot x, \text{ если } z = 0;$$
$$\widehat{y} = a + b_2 \cdot x, \text{ если } z = 1;$$
$$b_1 \neq b_2.$$

В таком случае говорят, что имеют место структурные изменения в исследуемой зависимости. Для их учета в уравнение регрессии вводят фиктивную переменную z как сомножитель при количественной переменной:

$$\hat{y} = a + b \cdot x + \delta \cdot x \cdot z.$$

Тогда для значения фиктивной переменной $z = 0$ уравнение запишется:

$$\hat{y} = a + b \cdot x.$$

А для значения фиктивной переменной $z = 1$:

$$\hat{y} = a + (b + \delta) \cdot x.$$

Следовательно, коэффициент b_2 будет равен $(b + \delta)$. Графически эта модель может быть представлена в виде двух прямых разным углом наклона (рис. 4.2).

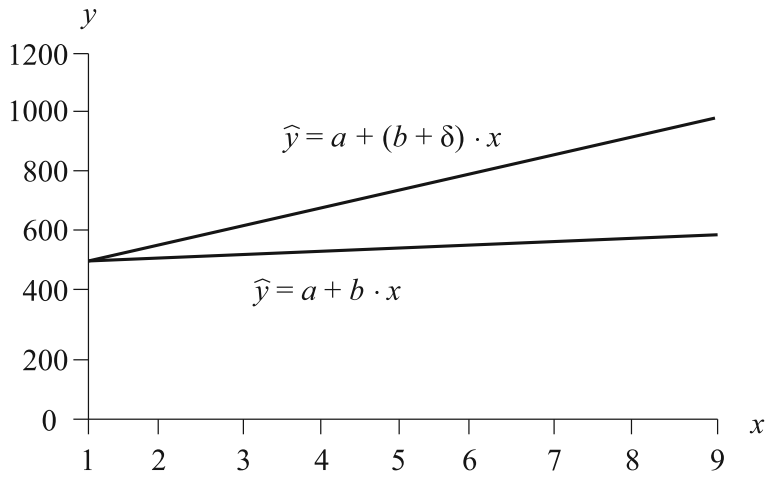


Рис. 4.2 – Модель регрессии с фиктивной переменной наклона

4.4 Общий вид модели регрессии с фиктивными переменными

Общий вид модели с фиктивными переменными объединяет модель с фиктивными переменными наклона и модель с фиктивными переменными сдвига. Рассмотрим в качестве примера зависимость изменения заработной платы y от стажа x и пола.

Можно предположить, что фактор «пол» будет оказывать влияние не только на разницу в заработной плате мужчин и женщин, но и скорость ее изменения (наклон линии регрессии). Чтобы учесть этот факт, вводим фиктивную бинарную переменную z для признака «пол»:

$$z = \begin{cases} 0 \text{ — мужчины,} \\ 1 \text{ — женщины,} \end{cases}$$

а также переменную для коэффициента наклона — $(z \cdot x)$.

Получаем уравнение с тремя факторными переменными:

$$\widehat{y} = a + bx + \delta \cdot z + \gamma \cdot (z \cdot x).$$

Пример 4.2

По статистическим данным (табл. 4.4) построить уравнение регрессии зависимости заработной платы (y) от стажа работы (x) с учетом пола работающего.

Таблица 4.4 – Данные к примеру 4.2

№ набл.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y	9	6	10	7	12	8,5	13	9	15	9	18	9,5	20	12	22	15	25	16
x	2	2	3	3	4	4	5	5	7	7	8	8	10	10	12	12	15	15
z	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
$z \cdot x$	0	2	0	3	0	4	0	5	0	7	0	8	0	10	0	12	0	15

Найдем параметры уравнения:

$$\widehat{y} = 6{,}69 + 1{,}27 \cdot x - 2{,}08 \cdot z - 0{,}5 \cdot (z \cdot x).$$

Коэффициент детерминации $R^2 = 0{,}98$. Критерий Фишера $F = 283{,}92 > F_{\text{табл}} = 3{,}33$. Следовательно, уравнение статистически значимо в целом с вероятностью 95%.

Проверку статистической значимости коэффициентов уравнения регрессии выполним с учетом данных таблицы 4.5.

Таблица 4.5 – Оценка статистической значимости коэффициентов регрессии

	Значения коэффициентов	**Ст. ошибка**	**Значения t-статистики**	**P-значение**
a	6,69	0,51	13,01	$3{,}3E-09$
b	1,27	0,06	20,76	$6{,}48E-12$
δ	–2,08	0,73	–2,86	0,01262
γ	–0,50	0,09	–5,83	$4{,}36E-05$

Все параметры модели статистически значимы, следовательно, как стаж, так и пол оказывают существенное влияние на уровень заработной платы, причем не только на ее общее изменение, но и на скорость изменения.

Уравнение для мужчин запишется: $\widehat{y} = 6{,}69 + 1{,}27 \cdot x$, а для женщин: $\widehat{y} = 4{,}61 + 0{,}77 \cdot x$. В этом случае имеется различие не только свободных членов, но и коэффициентов наклона, что и подтверждает рисунок 4.3.

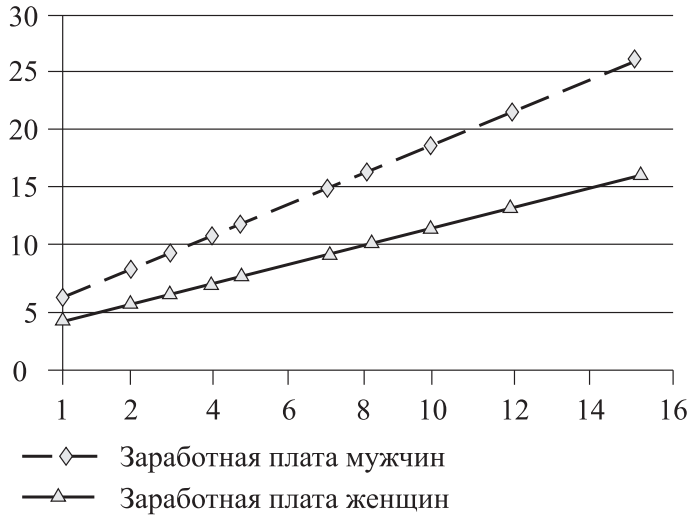


Рис. 4.3 – Изменение заработной платы мужчин и женщин в зависимости от стажа

. . . . . .

4.5 Исследование структурных изменений с помощью теста Чоу

В практике эконометрических исследований нередки случаи, когда для выявления зависимости между показателями имеются выборки их значений, полученных при разных условиях. Необходимо выяснить, действительно ли две выборки однородны в регрессионном смысле. Другими словами, можно ли объединить две выборки в одну и рассматривать единую модель регрессии.

При достаточных объемах выборок можно, например, построить интервальные оценки параметров регрессии по каждой из выборок и в случае пересечения соответствующих доверительных интервалов сделать вывод о единой модели регрессии. Возможны и другие подходы.

В случае если объем хотя бы одной из выборок незначителен, то возможности такого подхода резко сужаются из-за невозможности построения регрессии с достаточно надежными оценками.

Для проверки возможности объединения выборок в одну можно использовать тест Чоу.

Алгоритм теста.

Пусть имеется две подвыборки: одна объемом n_1, а другая объемом n_2.

1. По каждой подвыборке строятся линейные регрессионные модели с m переменными:

- для первой подвыборки — $y_1 = b_{1_0} + \sum_{j=1}^{m} b_{1_j} \cdot x_j + \varepsilon_1$;
- для второй подвыборки — $y_2 = b_{2_0} + \sum_{j=1}^{m} b_{2_j} \cdot x_j + \varepsilon_2$.

Рассчитываются суммы квадратов остатков для этих регрессий SS_1 и SS_2.

2. Строится линейная регрессия по объединенной выборке:

$$y = b_0 + \sum_{j=1}^{m} b_j \cdot x_j + \varepsilon.$$

Вычисляется ее сумма квадратов остатков SS.

3. Формулируется нулевая гипотеза:

$$H_0: b_{1_j} = b_{2_j},\ j = \overline{0, m},\ \text{где } b_{1_j},\ b_{2_j} \text{ — параметры моделей.}$$

Очевидно, что при совпадении параметров регрессии выполняется равенство $SS = SS_1 + SS_2$. Чем сильнее различие в поведении y для двух подвыборок, тем больше значение SS будет превосходить значение суммы $SS_1 + SS_2$.

4. Для проверки гипотезы вычисляется фактическое значение F-статистики по формуле:

$$F = \frac{SS - (SS_1 + SS_2)}{(SS_1 + SS_2)} \cdot \frac{n - 2 \cdot m - 2}{m + 1}.$$

Здесь m — количество параметров уравнений регрессий, n — число наблюдений по всей совокупности.

В случае если $F < F_{\text{табл}}\big(\alpha, (m + 1), (n - 2 \cdot m - 2)\big)$, считается, что различие между SS и $SS_1 + SS_2$ статистически незначимо и возможно построение уравнения регрессии по объединенной выборке объема $n = n_1 + n_2$.

Если $F > F_{\text{табл}}\big(\alpha, (m+1), (n-2\cdot m-2)\big)$, то различие между SS и $SS_1 + SS_2$ статистически значимо, что определяет существенность различия поведения наблюдаемой переменной y для двух подвыборок. В случае регрессионного анализа с фиктивными переменными это означает необходимость введения в уравнение регрессии соответствующей фиктивной переменной.

Пример 4.3

Используем тест Чоу для выявления целесообразности рассмотрения общей выборки и введения фиктивной переменной на примере данных предыдущего параграфа, выделив две подвыборки: ($z = 0$) и ($z = 1$). Данные представлены в таблицах 4.6 и 4.7.

Таблица 4.6 – Изменение заработной платы мужчин в зависимости от стажа

№ набл.	1	2	3	4	5	6	7	8	9
y_1	9	10	12	13	15	18	20	22	25
x	2	3	4	5	7	8	10	12	15

Построим по каждой из таблиц линейные модели зависимости заработной платы (y) от стажа (x):

$$\hat{y}_1 = 6{,}689 + 1{,}269 \cdot x; \quad R^2 = 0{,}99; \quad SS_1 = 2{,}94;$$
$$\hat{y}_2 = 4{,}61 + 0{,}765 \cdot x; \quad R^2 = 0{,}95; \quad SS_2 = 5{,}019.$$

Таблица 4.7 – Изменение заработной платы женщин в зависимости от стажа

№ набл.	1	2	3	4	5	6	7	8	9
y_2	6	7	8,5	9	9	9,5	12	15	16
x	2	3	4	5	7	8	10	12	15

Построим линейную модель по объединенной выборке:

$$\widehat{y} = 5{,}649 + 1{,}018 \cdot x; \quad R^2 = 0{,}639; \quad SS = 177{,}52.$$

Рассчитаем статистику:

$$F = \frac{(177{,}52 - 2{,}94 - 5{,}019) \cdot 14}{(2{,}94 + 5{,}019) \cdot 2} = 80{,}4.$$

$F_{\text{табл}}(\alpha = 0{,}05;\ 2{,}14) = 3{,}74$. Так как вычисленное значение $F > F_{\text{табл}}$, то следует признать существенность различия роста заработной платы от стажа в зависимости от пола. Следовательно, для построения общего уравнения регрессии целесообразно ввести фиктивную переменную, определяющую пол работника, что и показано в примере предыдущего параграфа.

Контрольные вопросы по главе 4

1. Можно ли учесть в уравнении регрессии неколичественные факторы? Каким образом?
2. Дайте определение фиктивной переменной.
3. В чем суть основного правила использования фиктивных переменных?
4. Приведите примеры использования фиктивных переменных.
5. Каков общий вид модели регрессии с одной количественной и одной фиктивной переменными?
6. Какова область применения теста Чоу?
7. Какие показатели сравниваются между собой по тесту Чоу? Какой статистический критерий используется в этом тесте?
8. Опишите методику применения теста Чоу.

Глава 5

СИСТЕМЫ ЭКОНОМЕТРИЧЕСКИХ УРАВНЕНИЙ

5.1 Общие положения

Многие экономические взаимосвязи в силу своей многогранности не могут быть в полной мере описаны с помощью отдельных изолированных уравнений регрессии. Данное описание предполагает, что факторы можно изменять независимо друг от друга. Однако ряд экономических процессов моделируется в условиях, когда изменение одной переменной, как правило, не может происходить при абсолютной неизменности других, а повлечет за собой изменения во всей совокупности взаимосвязанных признаков. Следовательно, отдельно взятое уравнение множественной регрессии уже не может характеризовать истинные влияния отдельных признаков на вариацию результирующей переменной. В силу этого возникает необходимость использования системы уравнений, которая в матричном виде может быть представлена как

$$AY + BX = E,$$

где A — матрица коэффициентов при зависимых переменных; Y — вектор зависимых переменных; B — матрица коэффициентов при объясняющих переменных; X — вектор объясняющих переменных; E — вектор ошибок.

Наиболее широко этот подход применяется в макроэкономических исследованиях, а также в исследованиях спроса и предложения. Приведем примеры таких систем.

1. *Модель «спрос-предложение»*.

Данная модель является одной из простейших систем одновременных уравнений. В этом случае, предполагая, что объем спроса Q_t^d и объем предложения Q_t^S в момент времени t являются линейными функциями от цены P_t, получаем систему:

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + U_t, & a_1 < 0, \\ Q_t^S = b_0 + b_1 P_t + V_t, & b_1 > 0, \\ Q_t^S = Q_t^d. \end{cases}$$

2. *Кейнсианская модель формирования доходов.*

Простейшая модель данного типа в предположении, что рассматривается закрытая экономика без государственных расходов:

$$\begin{cases} C_t = b_0 + b_1 Y_t + \varepsilon_t, \\ Y_t = C_t + I_t. \end{cases}$$

Здесь Y_t, C_t, I_t — совокупный выпуск, объемы потребления и инвестиций соответственно, t — текущий момент времени.

Система уравнений в эконометрических исследованиях может быть построена по-разному.

1. ***Система независимых уравнений*** — это система, в которой каждая зависимая переменная $y_i (i = \overline{1, n})$ рассматривается как функция одного и того же набора факторов $x_j (j = \overline{1, m})$:

$$\begin{cases} y_1 = b_{10} + b_{11}x_1 + b_{12}x_2 + \ldots + b_{1m}x_m + \varepsilon_1, \\ y_2 = b_{20} + b_{21}x_1 + b_{22}x_2 + \ldots + b_{2m}x_m + \varepsilon_2, \\ \ldots, \\ y_n = b_{n0} + b_{n1}x_1 + b_{n2}x_2 + \ldots + b_{nm}x_m + \varepsilon_n. \end{cases}$$

Набор факторов x_j в каждом уравнении может варьировать. Каждое уравнение системы независимых уравнений может рассматриваться самостоятельно. Для нахождения его параметров используется метод наименьших квадратов. По существу, каждое уравнение этой системы является уравнением множественной регрессии. Так как фактические значения зависимой переменной отличаются от теоретических на величину случайной ошибки, то в каждом уравнении присутствует величина случайной ошибки ε_i.

2. ***Система рекурсивных уравнений*** — это система, в которой зависимая переменная y_i одного уравнения выступает в виде фактора в другом уравнении:

$$\begin{cases} y_1 = b_{10} + b_{11}x_1 + b_{12}x_2 + \ldots + b_{1m}x_m + \ldots + b_{2m}x_m + \varepsilon_2, \\ \ldots, \\ y_n = b_{n0} + a_{n1}y_1 + \ldots + a_{n,n-1}y_{n-1} + b_{n1}x_1 + \ldots + b_{nm}x_m + \varepsilon_n. \end{cases}$$

В данной системе зависимая переменная $y_i (i = \overline{1, n})$ включает в каждое последующее уравнение в качестве факторов все зависимые переменные предшествующих уравнений наряду с набором собственно факторов $x_j (j = \overline{1, m})$. Каждое уравнение этой системы может рассматриваться самостоятельно, и его параметры определяются методом наименьших квадратов (МНК).

3. Наибольшее распространение в эконометрических исследованиях получила ***система взаимозависимых уравнений***. В ней одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях — в правую часть системы:

$$\begin{cases} y_1 = b_{10} + a_{12}y_2 + a_{13}y_3 + \ldots + a_{1n}y_n + b_{11}x_1 + \ldots + b_{1m}x_m + \varepsilon_1, \\ y_2 = b_{20} + a_{21}y_1 + a_{23}y_3 + \ldots + a_{2n}y_n + b_{21}x_1 + \ldots + b_{2m}x_m + \varepsilon_2, \\ \ldots, \\ y_n = b_{n0} + a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{n,n-1}y_{n-1} + b_{n1}x_1 + \ldots + b_{nm}x_m + \varepsilon_n. \end{cases}$$

Система взаимозависимых уравнений получила название *системы совместных, одновременных уравнений*. Тем самым подчеркивается, что в системе одни и те же переменные одновременно рассматриваются как зависимые в одних уравнениях и как независимые в других. В эконометрике эта система уравнений называется также *структурной формой модели*. В отличие от предыдущих систем каждое уравнение системы одновременных уравнений не может рассматриваться самостоятельно, и для нахождения его параметров традиционный МНК неприменим. С этой целью используются специальные приемы оценивания.

5.2 Составляющие систем одновременных уравнений

При рассмотрении систем одновременных уравнений переменные делятся на два класса — эндогенные и экзогенные переменные.

Эндогенные переменные — это зависимые переменные, число которых равно числу уравнений в системе. Обозначаются через y.

Экзогенные переменные — это предопределенные переменные, влияющие на эндогенные переменные, но не зависящие от них. Обозначаются через x.

Классификация переменных на эндогенные и экзогенные зависит от теоретической концепции принятой модели. Экономические переменные могут выступать в одних моделях как эндогенные, а в других — как экзогенные переменные. Внеэкономические переменные (например, климатические условия, социальное положение, пол, возрастная категория) входят в систему только как экзогенные переменные. В качестве экзогенных переменных могут рассматриваться значения эндогенных переменных за предшествующий период времени (*лаговые переменные*).

Структурная форма модели позволяет увидеть влияние изменений любой экзогенной переменной на значения эндогенной переменной. Целесообразно в качестве экзогенных переменных выбирать такие переменные, которые могут быть объектом регулирования. Меняя их и управляя ими, можно заранее иметь целевые значения эндогенных переменных.

Структурная форма модели в правой части содержит при эндогенных переменных коэффициенты a_{ik} и при экзогенных переменных — коэффициенты b_{ij}, которые называются *структурными коэффициентами* модели. Для оценки параметров структурной формы модели используется *приведенная форма модели*.

Приведенная форма модели представляет собой систему линейных функций эндогенных переменных от экзогенных:

$$\begin{cases} y_1 = A_1 + \delta_{11}x_1 + \ldots + \delta_{1m}x_m + u_1, \\ y_2 = A_2 + \delta_{21}x_1 + \ldots + \delta_{2m}x_m + u_2, \\ \ldots, \\ y_n = A_n + \delta_{n1}x_1 + \ldots + \delta_{nm}x + u_n. \end{cases}$$

где δ_{ij} — коэффициенты приведенной формы модели; u_i — остаточная величина для приведенной формы.

По своему виду приведенная форма модели ничем не отличается от системы независимых уравнений, параметры которой оцениваются традиционным МНК. Применяя МНК, можно оценить δ_{ij}, а затем оценить значения эндогенных переменных через экзогенные.

Коэффициенты приведенной формы модели представляют собой нелинейные функции коэффициентов структурной формы модели. Рассмотрим это положение на примере простейшей структурной модели, выразив коэффициенты приведенной формы модели через коэффициенты структурной модели.

Для структурной модели вида:

$$\begin{cases} y_1 = a_{12}y_2 + b_{11}x_1 + b_{10} + \varepsilon_1, \\ y_2 = a_{21}y_1 + b_{22}x_2 + b_{20} + \varepsilon_2 \end{cases}$$

приведенная форма модели имеет вид:

$$\begin{cases} y_1 = A_1 + \delta_{11}x_1 + \delta_{12}x_2 + u_1, \\ y_2 = A_2 + \delta_{21}x_1 + \delta_{22}x_2 + u_2. \end{cases}$$

Из первого уравнения структурной модели выразим y_2. Тогда структурная модель примет вид (ради упрощения опускаем случайную величину):

$$\begin{cases} y_2 = \dfrac{y_1 - b_{11}x_1 - b_{10}}{a_{12}}, \\ y_2 = a_{21}y_1 + b_{22}x_2 + b_{20}. \end{cases}$$

Приравнивая правые части этой системы, после соответствующих преобразований будем иметь:

$$y_1 = \frac{b_{11}}{1 - a_{12}a_{21}} \cdot x_1 + \frac{b_{22}a_{12}}{1 - a_{12}a_{21}} \cdot x_2 + \frac{b_{10} + b_{20}a_{12}}{1 - a_{12}a_{21}}.$$

Поступая аналогично со вторым уравнением структурной модели, получим:

$$y_2 = \frac{b_{11}a_{21}}{1 - a_{12}a_{21}} \cdot x_1 + \frac{b_{22}}{1 - a_{12}a_{21}} \cdot x_2 + \frac{b_{20} + b_{10}a_{21}}{1 - a_{12}a_{21}}.$$

В итоге получаем систему уравнений, по структуре совпадающую с приведенной формой модели:

$$\begin{cases} y_1 = \dfrac{b_{11}}{1 - a_{12}a_{21}} \cdot x_1 + \dfrac{b_{22}a_{12}}{1 - a_{12}a_{21}} \cdot x_2 + \dfrac{b_{10} + b_{20}a_{12}}{1 - a_{12}a_{21}}, \\ y_2 = \dfrac{b_{11}a_{21}}{1 - a_{12}a_{21}} \cdot x_1 + \dfrac{b_{22}}{1 - a_{12}a_{21}} \cdot x_2 + \dfrac{b_{20} + b_{10}a_{21}}{1 - a_{12}a_{21}}. \end{cases}$$

Таким образом, можно сделать вывод о том, что коэффициенты приведенной формы модели будут выражаться через коэффициенты структурной формы:

$$\begin{cases} A_1 = \dfrac{b_{10} + b_{20}a_{12}}{1 - a_{12}a_{21}}; \quad \delta_{11} = \dfrac{b_{11}}{1 - a_{12}a_{21}}; \quad \delta_{12} = \dfrac{b_{10} + b_{20}a_{12}}{1 - a_{12}a_{21}}; \\ A_2 = \dfrac{b_{10} + b_{20}a_{12}}{1 - a_{12}a_{21}}; \quad \delta_{21} = \dfrac{b_{11}a_{21}}{1 - a_{12}a_{21}}; \quad \delta_{22} = \dfrac{b_{22}}{1 - a_{12}a_{21}}. \end{cases}$$

Взаимосвязь коэффициентов приведенной и структурной форм моделей имеет реальное практическое применение. Коэффициенты приведенной формы модели могут быть оценены обычным МНК, и на их основе может быть произведена оценка структурных коэффициентов модели.

5.3 Идентификация структурной модели

При переходе от приведенной формы модели к структурной эконометрист сталкивается с проблемой идентификации.

Идентификация — это единственность соответствия между приведенной и структурной формами модели, позволяющая однозначно оценить структурные коэффициенты модели.

С позиции идентифицируемости структурные модели можно подразделить на три вида:

1) идентифицируемые;
2) неидентифицируемые;
3) сверхидентифицируемые.

Модель *идентифицируема*, если все структурные ее коэффициенты определяются однозначно, единственным образом по коэффициентам приведенной формы модели, т. е. если число параметров структурной модели равно числу параметров приведенной формы модели. В этом случае структурные коэффициенты модели оцениваются через параметры приведенной формы модели и модель идентифицируема.

Модель *неидентифицируема*, если число приведенных коэффициентов меньше числа структурных коэффициентов, и в результате структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы модели.

Модель *сверхидентифицируема*, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае на основе коэффициентов приведенной формы можно получить два или более значений одного структурного коэффициента. В этой модели число структурных коэффициентов меньше числа коэффициентов приведенной формы. Сверхидентифицируемая модель в отличие от неидентифицируемой модели практически решаема, но требует для этого специальных методов исчисления параметров.

Структурная модель всегда представляет собой систему совместных уравнений, каждое из которых требуется проверять на идентификацию. Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Сверхидентифицируемая модель содержит хотя бы одно сверхидентифицируемое уравнение.

Выполнение условия идентифицируемости модели проверяется для каждого уравнения системы. Чтобы уравнение было идентифицируемо, необходимо, чтобы число предопределенных переменных, отсутствующих в данном уравнении, но присутствующих в системе, было равно числу эндогенных переменных в данном уравнении без одного.

Если обозначить число эндогенных переменных в i-м уравнении системы через H, а число экзогенных (предопределенных) переменных, которые содержатся в системе, но не входят в данное уравнение, — через D, то условие идентифицируемости модели может быть записано в виде следующего счетного правила (табл. 5.1).

Таблица 5.1 – Условие идентифицируемости модели

$D + 1 = H$	уравнение идентифицируемо
$D + 1 < H$	уравнение неидентифицируемо
$D + 1 > H$	уравнение сверхидентифицируемо

Для оценки параметров структурной модели система должна быть идентифицируема или сверхидентифицируема.

Рассмотренное счетное правило отражает необходимое, но недостаточное условие идентификации. Более точно условия идентификации определяются, если накладывать ограничения на коэффициенты матриц параметров структурной модели. Уравнение идентифицируемо, если по отсутствующим в нем переменным (эндогенным и экзогенным) можно из коэффициентов при них в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы был на единицу меньше числа эндогенных переменных в системе.

Целесообразность проверки условия идентификации модели через определитель матрицы коэффициентов, отсутствующих в данном уравнении, но присутствующих в других уравнениях системы, объясняется тем, что возможна ситуация, когда для каждого уравнения системы выполнено счетное правило, а определитель матрицы названных коэффициентов равен нулю. В этом случае соблюдается лишь необходимое, но недостаточное условие идентификации.

В эконометрических моделях часто наряду с уравнениями, параметры которых должны быть статистически оценены, используются балансовые тождества переменных, коэффициенты при которых равны ±1. В этом случае, хотя само тождество и не требует проверки на идентификацию, ибо коэффициенты при переменных в тождестве известны, в проверке на идентификацию собственно структурных уравнений системы тождества участвуют.

Пример 5.1

Изучается модель вида:

$$\begin{cases} C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot C_{t-1} + \varepsilon_1, \\ I_t = a_2 + b_{21} \cdot r_t + b_{22} \cdot I_{t-1} + \varepsilon_2, \\ r_t = a_3 + b_{31} \cdot Y_t + b_{32} \cdot M_t + \varepsilon_3, \\ Y_t = C_t + I_t + G_t, \end{cases}$$

где C_t — расходы на потребление в период t; Y_t — совокупный доход в период t; I_t — инвестиции в период t; r_t — процентная ставка в период t; M_t — денежная масса в период t; G_t — государственные расходы в период t; C_{t-1} — расходы на потребление в период $t - 1$; I_{t-1} — инвестиции в период $t - 1$. Первое уравнение — функция

потребления, второе уравнение — функция инвестиций, третье уравнение — функция денежного рынка, четвертое уравнение — тождество дохода. Модель представляет собой систему одновременных уравнений. Проверим каждое ее уравнение на идентификацию.

Модель включает четыре эндогенные переменные (C_t, I_t, Y_t, r_t) и четыре предопределенные переменные (две экзогенные переменные — M_t и G_t и две лаговые переменные — C_{t-1} и I_{t-1}).

Проверим необходимое условие идентификации для каждого из уравнений модели.

Первое уравнение: $C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot C_{t-1} + \varepsilon$. Это уравнение содержит две эндогенные переменные C_t и Y_t и одну предопределенную переменную C_{t-1}. Таким образом, $H = 2$, а $D = 4 - 1 = 3$, т. е. выполняется условие $D + 1 > H$. Уравнение сверхидентифицируемо.

Второе уравнение: $I_t = a_2 + b_{21} \cdot r_t + b_{22} \cdot I_{t-1} + \varepsilon_2$. Оно включает две эндогенные переменные I_{t-1} и r_t и одну экзогенную переменную I_{t-1}. Выполняется условие $D + 1 = 3 + 1 > H = 2$. Уравнение сверхидентифицируемо.

Третье уравнение: $r_t = a_3 + b_{31} \cdot Y_t + b_{32} \cdot M_t + \varepsilon_3$. Оно включает две эндогенные переменные Y_t и r_t и одну экзогенную переменную M_t. Выполняется условие $D+1 = 3 + 1 > H = 2$. Уравнение сверхидентифицируемо.

Четвертое уравнение: $Y_t = C_t + I_t + G_t$. Оно представляет собой тождество, параметры которого известны. Необходимости в идентификации нет.

Проверим для каждого уравнения достаточное условие идентификации. Для этого составим матрицу коэффициентов при переменных модели (табл. 5.2).

Таблица 5.2 – Матрица коэффициентов при переменных модели

	C_t	I_t	r_t	Y_t	C_{t-1}	I_{t-1}	M_t	G_t
I уравнение	−1	0	0	b_{11}	b_{12}	0	0	0
II уравнение	0	−1	b_{21}	0	0	b_{22}	0	0
III уравнение	0	0	−1	b_{31}	0	0	b_{32}	0
Тождество	1	1	0	−1	0	0	0	1

В соответствии с достаточным условием идентификации ранг матрицы коэффициентов при переменных, не входящих в исследуемое уравнение, должен быть равен числу эндогенных переменных модели без одного.

Первое уравнение. Матрица коэффициентов при переменных, не входящих в уравнение, представлена в таблице 5.3.

Таблица 5.3 – Матрица коэффициентов, не входящих в первое уравнение

	I_t	r_t	I_{t-1}	M_t	G_t
II уравнение	−1	b_{21}	b_{22}	0	0
III уравнение	0	−1	0	b_{32}	0
Тождество	1	0	0	0	1

Ранг данной матрицы равен трем, так как определитель квадратной подматрицы 3×3 не равен нулю:

$$\begin{vmatrix} b_{22} & 0 & 0 \\ 0 & b_{32} & 0 \\ 0 & 0 & 1 \end{vmatrix} = b_{22}b_{32} \neq 0.$$

Достаточное условие идентификации для данного уравнения выполняется.

Второе уравнение. Матрица коэффициентов при переменных, не входящих в уравнение, представлена в таблице 5.4.

Таблица 5.4 – Матрица коэффициентов, не входящих во второе уравнение

	C_t	Y_t	C_{t-1}	M_t	G_t
I уравнение	−1	b_{11}	b_{12}	0	0
III уравнение	0	b_{31}	0	b_{32}	0
Тождество	1	−1	0	0	1

Ранг данной матрицы равен трем, так как определитель квадратной подматрицы 3 × 3 не равен нулю:

$$\begin{vmatrix} b_{12} & 0 & 0 \\ 0 & b_{32} & 0 \\ 0 & 0 & 1 \end{vmatrix} = b_{12}b_{32} \neq 0.$$

Достаточное условие идентификации для данного уравнения выполняется.

Третье уравнение. Матрица коэффициентов при переменных, не входящих в уравнение, представлена в таблице 5.5.

Таблица 5.5 – Матрица коэффициентов, не входящих в третье уравнение

	C_t	I_t	C_{t-1}	I_{t-1}	G_t
I уравнение	−1	0	b_{12}	0	0
II уравнение	0	−1	0	$b_{22}b_{22}$	0
Тождество	1	1	0	0	1

Ранг данной матрицы равен трем, так как определитель квадратной подматрицы 3 × 3 не равен нулю:

$$\begin{vmatrix} b_{12} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & 1 \end{vmatrix} = b_{12}b_{22} \neq 0.$$

Достаточное условие идентификации для данного уравнения выполняется.

Таким образом, все уравнения модели сверхидентифицируемы. Приведенная форма модели в общем виде будет выглядеть следующим образом:

$$\begin{cases} C_t = A_1 + \delta_{11}C_{t-1} + \delta_{12}I_{t-1} + \delta_{13}M_t + +\delta_{14}G_t + u_1, \\ I_t = A_2 + \delta_{21}C_{t-1} + \delta_{22}I_{t-1} + \delta_{23}M_t + +\delta_{24}G_t + u_2, \\ r_t = A_3 + \delta_{31}C_{t-1} + \delta_{32}I_{t-1} + \delta_{33}M_t + +\delta_{34}G_t + u_3, \\ Y_t = A_4 + \delta_{41}C_{t-1} + \delta_{42}I_{t-1} + \delta_{43}M_t + +\delta_{44}G_t + u_4. \end{cases}$$

. . . . . .

5.4 Оценивание параметров системы одновременных уравнений

Структурные коэффициенты системы одновременных уравнений могут быть оценены разными способами в зависимости от вида модели с точки зрения ее идентификации. Наибольшее распространение получили следующие методы оценивания коэффициентов системы одновременных уравнений:

1) косвенный метод наименьших квадратов (КМНК);
2) двухшаговый метод наименьших квадратов (ДМНК);
3) трехшаговый метод наименьших квадратов (ТНМК);
4) метод максимального правдоподобия (ММП).

Косвенный метод наименьших квадратов (КМНК) применяется в случае точно идентифицируемой структурной модели. Если система сверхидентифицируема, то КМНК не используется, ибо он не дает однозначных оценок для параметров структурной модели. В этом случае могут использоваться разные методы оценивания, среди которых наиболее распространенным и простым является двухшаговый метод наименьших квадратов (ДМНК).

5.4.1 Косвенный метод наименьших квадратов

Процедура применения КМНК предполагает выполнение следующих этапов работы.

1) структурная модель преобразовывается в приведенную форму модели;
2) для каждого уравнения приведенной формы модели обычным МНК оцениваются приведенные коэффициенты δ_{ij};
3) коэффициенты приведенной формы модели трансформируются в параметры структурной модели.

Пример 5.2

Исследуется зависимость спроса и предложения некоторого товара от его цены (P), дохода на душу населения и инвестиций в производство (I). Модель спроса и предложения имеет вид:

$$\begin{cases} Q_t^d = a_0 + a_1 P_t + a_2 y_t + U_t, \\ Q_t^S = b_0 + b_1 P_t + b_2 I_t + V_t, \\ Q_t^S = Q_t^d, \end{cases}$$

где Q_t^d — спрос в момент времени t; Q_t^S — предложение в момент времени t.

Учитывая вид третьего уравнения системы, обозначим $Q_t = Q_t^S = Q_t^d$.

В данной модели Q_t и P_t — эндогенные переменные, причем переменная P_t является эндогенной как по экономическому смыслу (цена зависит от спроса и предложения), так и в силу наличия тождества $Q_t^S = Q_t^d$. Переменные y_t и I_t являются

экзогенными переменными. Каждое уравнение системы точно идентифицировано. Применим КМНК, используя следующую информацию (табл. 5.6).

Таблица 5.6 – Исходные данные для примера 5.2

Q_t	20	33	28	41	40	36	42	38	51
P_t	3	3	5	4	5	6	6	7	7
y_t	34	43	51	49	55	62	70	68	78
I_t	5	6	6	7	7	6	8	8	12

Запишем приведенную форму модели:

$$\begin{cases} Q_t = A_1 + \delta_{11} y_t + \delta_{12} I_t + u_1, \\ P_t = A_2 + \delta_{21} y_t + \delta_{22} I_t + u_2. \end{cases}$$

Оценку параметров каждого уравнения приведенной формы модели выполним, используя обычный МНК. В результате получим:

$$\begin{cases} Q_t = 6{,}022 + 0{,}234 y_t + 2{,}393 I_t + u_1, \\ P_t = -0{,}692 + 0{,}126 y_t - 0{,}189 I_t + u_2. \end{cases}$$

Выполним переход к структурной форме модели. Найдем уравнение спроса:

$$Q_t^d = a_0 + a_1 P_t + a_2 y_t + U_t.$$

Для этого из второго уравнения приведенной модели выразим I_t и подставим его в первое уравнение:

$$I_t = \frac{-P_t - 0{,}692 + 0{,}127 y_t}{0{,}189};$$

$$Q_t = 6{,}022 + 0{,}234 y_t + 2{,}393 \cdot \frac{-P_t - 0{,}692 + 0{,}126 y_t}{0{,}189} = -2{,}743 - 12{,}667 P_t + 1{,}829 y_t.$$

Получили первое уравнение структурной формы модели. Аналогично найдем и второе уравнение структурной формы модели, характеризующее функцию предложения: $Q_t^S = b_0 + b_1 P_t + b_2 I_t + V_t$. Для этого из первого уравнения приведенной формы модели исключим y_t, выразив его через второе уравнение и подставив в первое.

$$y_t = \frac{P_t + 0{,}692 + 0{,}189 I_t}{0{,}127};$$

$$Q_t = 6{,}022 + 0{,}234 \cdot \frac{P_t + 0{,}692 + 0{,}189 I_t}{0{,}127} + 2{,}393 I_t = 7{,}297 + 1{,}843 P_t + 2{,}741 I_t.$$

В итоге модель спроса и предложения имеет вид:

$$\begin{cases} Q_t^d = -2{,}743 - 12{,}667 P_t + 1{,}829 y_t + U_t, \\ Q_t^S = 7{,}297 + 1{,}843 P_t + 2{,}741 I_t + V_t, \\ Q_t^S = Q_t^d. \end{cases}$$

5.4.2 Двухшаговый метод наименьших квадратов

Двухшаговый метод наименьших квадратов (ДМНК) применяется как к точно идентифицированной, так и к сверхидентифицированной системе.

Применение метода предусматривает следующие шаги:

1) построение приведенной формы модели (ПФМ);
2) для каждого уравнения структурной формы модели выполняются следующие действия:
   - находят эндогенные переменные, являющиеся факторными признаками (стоят в правой части уравнения);
   - для этих переменных определяют их выровненные значения $\widehat{y}_i = A_i + \delta_{i1}x_1 + \delta_{i2}x_2 + \ldots + \delta_{im}x_m$, используя соответствующее уравнение ПФМ;
   - находят параметры рассматриваемого уравнения структурной формы модели обычным МНК, заменяя исходные значения эндогенных переменных-факторов их выровненными значениями.

Сверхидентифицируемая структурная модель может быть двух типов:

1) все уравнения системы сверхидентифицируемы;
2) система содержит наряду со сверхидентифицируемыми точно идентифицируемые уравнения.

Если все уравнения системы сверхидентифицируемые, то для оценки структурных коэффициентов каждого уравнения используется ДМНК. Если в системе есть точно идентифицируемые уравнения, то вычисление структурных коэффициентов возможно из системы приведенных уравнений.

Пример 5.3

Имеются квартальные данные об объемах валового внутреннего продукта ($ВВП_t$, трлн руб.), расходов на конечное потребление ($КП_t$, трлн руб.), валового накопления ($ВН_t$, трлн руб.) и чистого экспорта ($Э_t$, трлн руб.). Данные представлены в таблице 5.7.

Таблица 5.7 – Динамика внутреннего валового продукта по кварталам

Номер наблюдения	**Год, квартал**	**$ВВП_t$**	**$КП_t$**	**$ВН_t$**	**$Э_t$**
1	2003, 1 квартал	330,10	260,15	43,40	26,55
2	2003, 2 квартал	341,60	248,27	73,10	20,22
3	2003, 3 квартал	395,70	266,59	124,23	4,89
4	2003, 4 квартал	361,10	251,42	105,49	4,19
5	2004, 1 квартал	322,80	257,22	55,71	9,87
6	2004, 2 квартал	330,10	254,48	64,34	11,28
7	2004, 3 квартал	374,00	244,54	116,24	13,22
8	2004, 4 квартал	350,10	240,90	86,66	22,55
продолжение на следующей странице					

Таблица 5.7 — Продолжение

Номер наблюдения	Год, квартал	$ВВП_t$	$КП_t$	$ВН_t$	$Э_t$
9	2005, 1 квартал	321,40	255,54	52,31	13,55
10	2005, 2 квартал	327,30	257,72	63,56	6,01
11	2005, 3 квартал	384,70	266,72	112,83	5,16
12	2005, 4 квартал	362,60	278,34	77,86	6,39
13	2006, 1 квартал	316,70	247,37	73,12	–3,79
14	2006, 2 квартал	324,20	237,30	87,60	–0,70
15	2006, 3 квартал	350,80	270,15	62,16	18,49
16	2006, 4 квартал	329,70	275,55	–6,67	60,82
17	2007, 1 квартал	310,80	244,14	23,93	42,73
18	2007, 2 квартал	334,20	232,50	51,61	50,09
19	2007, 3 квартал	390,90	242,69	89,33	58,87
20	2007, 4 квартал	369,40	244,74	41,35	83,31
21	2008, 1 квартал	346,30	221,76	40,13	84,41
22	2008, 2 квартал	368,40	222,53	61,49	84,38
23	2008, 3 квартал	432,00	253,44	99,66	78,90
24	2008, 4 квартал	399,80	250,08	85,17	64,55
25	2009, 1 квартал	347,10	235,57	46,33	65,21
26	2009, 2 квартал	364,15	242,31	70,86	50,98
27	2009, 3 квартал	424,67	261,28	116,53	46,86
28	2009, 4 квартал	399,80	265,24	100,71	33,85
29	2010, 1 квартал	363,31	271,79	53,62	37,90
30	2010, 2 квартал	382,50	269,36	69,12	44,02
31	2010, 3 квартал	450,21	289,35	111,53	49,33
32	2010, 4 квартал	417,90	287,06	87,79	43,05
33	2011, 1 квартал	377,14	271,76	50,60	54,77
34	2011, 2 квартал	399,38	279,64	74,57	45,17
35	2011, 3 квартал	470,24	295,94	127,63	46,66
36	2011, 4 квартал	443,71	300,05	97,98	45,68
37	2012, 1 квартал	405,83	298,01	58,24	49,58
38	2012, 2 квартал	431,29	296,59	80,61	54,09
39	2012, 3 квартал	499,50	316,51	122,58	60,42
40	2012, 4 квартал	478,06	306,31	112,89	58,86
41	2013, 1 квартал	435,47	314,10	60,92	60,45
42	2013, 2 квартал	463,66	309,69	84,13	69,83
43	2013, 3 квартал	535,01	336,24	127,61	71,15
44	2013, 4 квартал	510,18	318,43	128,26	63,49

На основе данных была построена модель:

$$\begin{cases} КП_t = a_1 + b_{11} ВВП_t + \varepsilon_1, \\ ВН_t = a_2 + b_{21} ВВП_{t-4} + \varepsilon_2, \\ ВВП_t = КП_t + ВН_t + Э_t, \end{cases}$$

где $ВВП_{t-4}$ — объем ВВП за аналогичный квартал предыдущего года.

Выполним проверку системы на идентификацию с помощью необходимого условия.

Для первого уравнения имеем:

- количество эндогенных переменных, входящих в это уравнение, равно двум ($КП_t$ и $ВВП_t$), $H = 2$;
- количество предопределенных переменных, не входящих в это уравнение, равно двум ($Э_t$ и $ВВП_{t-4}$), $D = 2$.

$H < D + 1$, следовательно, уравнение сверхидентифицировано.

Для второго уравнения имеем:

- количество эндогенных переменных, входящих в это уравнение, — одна ($ВН_t$), $H = 1$;
- количество предопределенных переменных, не входящих в это уравнение, — одна ($Э_t$), $D = 1$.

$H < D + 1$, следовательно, уравнение сверхидентифицировано.

Третье уравнение является тождеством и на идентификацию не проверяется.

Поскольку оба уравнения системы сверхидентифицированы, то и система сверхидентифицирована.

Выполним проверку системы на идентификацию с помощью достаточного условия.

Определим матрицу коэффициентов при переменных, отсутствующих в первом уравнении ($ВН_t$, $ВВП_{t-4}$, $Э_t$):

$$\begin{pmatrix} -1 & b_{21} & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Определитель данной матрицы не равен нулю, ранг матрицы равен двум, что на единицу меньше количества эндогенных переменных в системе. Таким образом, достаточное условие соблюдается.

Определим матрицу коэффициентов при переменных, отсутствующих во втором уравнении ($КП_t$, $ВВП_t$, $Э_t$):

$$\begin{pmatrix} -1 & b_{11} & 0 \\ 1 & -1 & 1 \end{pmatrix}.$$

Ранг данной матрицы равен 2, следовательно, достаточное условие соблюдается.

Обобщая выводы относительно идентификации системы, получаем, что система имеет решение и является сверхидентифицированной.

Для нахождения ее коэффициентов применим двухшаговый метод наименьших квадратов.

1 шаг. Запишем приведенную форму модели:

$$\begin{cases} ВВП_t = A_1 + \delta_{11} \cdot ВВП_{t-4} + \delta_{12} \cdot Э_t + u_1, \\ КП_t = A_2 + \delta_{21} \cdot ВВП_{t-4} + \delta_{22} \cdot Э_t + u_2, \\ ВН_t = A_3 + \delta_{31} \cdot ВВП_{t-4} + \delta_{23} \cdot Э_t + u_3. \end{cases}$$

Вычисление параметров каждого уравнения этой модели с помощью МНК проведем по укороченным рядам: все показатели, кроме $ВВП_{t-4}$, берутся с пятого по 44 наблюдение, а показатель $ВВП_{t-4}$ — с первого по 40-е наблюдение (табл. 5.7).

В результате система уравнений примет вид:

$$\begin{cases} \text{ВВП}_t = -31{,}53 + 1{,}06 \cdot \text{ВВП}_{t-4} + 0{,}51 \cdot \text{Э}_t + u_1, \\ \text{КП}_t = 79{,}06 + 0{,}52 \cdot \text{ВВП}_{t-4} - 0{,}11 \cdot \text{Э}_t + u_2, \\ \text{ВН}_t = -110{,}59 + 0{,}54 \cdot \text{ВВП}_{t-4} - 0{,}38 \cdot \text{Э}_t + u_3. \end{cases}$$

2 шаг. Рассмотрим уравнения структурной формы модели (СФМ).

В первом уравнении СФМ эндогенной переменной-фактором является переменная ВВП_t. Найдем ее выровненные значения по первому уравнению ПФМ (табл. 5.8).

Таблица 5.8 – Расчет выровненных значений ВВП_t

Номер наблюдения	**ВВП_{t-4}**	**Э_t**	**Выровненные значения ВВП_t**
5	330,10	9,87	323
6	341,60	11,28	336
7	395,70	13,22	395
8	361,10	22,55	363
9	322,80	13,55	318
10	330,10	6,01	321
11	374,00	5,16	368
12	350,10	6,39	343
13	321,40	–3,79	307
14	327,30	–0,70	315
15	384,70	18,49	386
16	362,60	60,82	384
17	316,70	42,73	326
18	324,20	50,09	338
19	350,80	58,87	370
20	329,70	83,31	360
21	310,80	84,41	341
22	334,20	84,38	366
23	390,90	78,90	423
24	369,40	64,55	393
25	346,30	65,21	369
26	368,40	50,98	385
27	432,00	46,86	450
28	399,80	33,85	410
29	347,10	37,90	356
30	364,15	44,02	377
31	424,67	49,33	444
32	399,80	43,05	414
33	363,31	54,77	382

продолжение на следующей странице

Таблица 5.8 — Продолжение

Номер наблюдения	$ВВП_{t-4}$	$Э_t$	Выровненные значения $ВВП_t$
34	382,50	45,17	397
35	450,21	46,66	469
36	417,90	45,68	435
37	377,14	49,58	394
38	399,38	54,09	419
39	470,24	60,42	498
40	443,71	58,86	469
41	405,83	60,45	429
42	431,29	69,83	461
43	499,50	71,15	534
44	478,06	63,49	508

После этого применим к первому уравнению СФМ метод наименьших квадратов, используя в качестве исходной информации значения эндогенной переменной-результата ($КП_t$) и выровненные значения эндогенной переменной фактора ($ВВП_t$). В результате получим:

$$КП_t = 110{,}48 + 0{,}4 \cdot ВВП_t + \varepsilon_1.$$

Фактическое значение t-критерия Стьюдента для коэффициента регрессии в этом уравнении равно 7,95, табличное — 2,0244 ($df = 40 - 2 = 38$, $\alpha = 0{,}05$), следовательно коэффициент при переменной $ВВП_t$ значим. Значимо и уравнение в целом ($R^2 = 0{,}62$; $F = 63{,}19$; $F_{табл} = 4{,}1$).

Второе уравнение не содержит эндогенных переменных-факторов, поэтому его параметры можно найти, применяя к нему обычный МНК. В результате получим следующее уравнение регрессии:

$$ВН_t = -103{,}89 + 0{,}48 \cdot ВВП_{t-4} + \varepsilon_2.$$

Это уравнение также имеет значимый коэффициент ($t_{факт} = 7{,}05$) и значимо в целом ($R^2 = 0{,}57$; $F = 49{,}73$).

Третье уравнение СФМ не является уравнением регрессии и не имеет неизвестных параметров (все параметры равны единице).

Таким образом, получена следующая система уравнений:

$$\begin{cases} КП_t = 110{,}48 + 0{,}4 \cdot ВВП_t + \varepsilon_1, \\ ВН_t = -103{,}89 + 0{,}48 \cdot ВВП_{t-4} + \varepsilon_2, \\ ВВП_t = КП_t + ВН_t + Э_t. \end{cases}$$

. . . . . .

Контрольные вопросы по главе 5

1. В чем сходство и различие моделей систем эконометрических уравнений с простыми моделями множественной регрессий?
2. Приведите примеры экономических процессов и явлений, которые могут быть описаны системами независимых, рекурсивных и взаимозависимых уравнений.
3. Опишите систему эконометрических уравнений в общем виде.
4. Какие типы переменных принято выделять в системах эконометрических уравнений?
5. Основные виды систем эконометрических уравнений.
6. Что называется структурной формой модели?
7. Для чего необходима приведенная форма модели? Какой вид она имеет? Что такое идентификация модели?
8. Какие классы моделей можно определить с точки зрения их идентификации?
9. В чем состоит необходимое и достаточное условия идентификации?
10. В каком случае вся модель является идентифицируемой и сверхидентифицируемой?
11. Дайте краткое описание методики косвенного метода наименьших квадратов.

Глава 6

ВРЕМЕННЫЕ РЯДЫ

6.1 Составляющие временного ряда

При построении эконометрической модели используются два типа данных:

1) данные, характеризующие совокупность различных объектов в определенный момент времени;
2) данные, характеризующие один объект за ряд последовательных моментов времени.

Модели, построенные по данным первого типа, называются *пространственными моделями*. Модели, построенные на основе второго типа данных, называются *моделями временных рядов*.

Временной ряд (ряд динамики) — это совокупность значений какого-либо показателя (y_i) за несколько последовательных моментов или периодов времени (t_i).

Величины y_i называются уровнями временного ряда, а t_i — временными метками (моменты или интервалы наблюдения). Обычно рассматриваются временные ряды с равными интервалами между наблюдениями, в качестве значений t_i берутся порядковые номера наблюдений и временной ряд представляется в виде последовательности $y_1, y_2, y_3, \ldots, y_n$, где n — количество наблюдений (табл. 6.1).

Таблица 6.1 – Динамика ВВП Российской Федерации

	2009 г.	**2010 г.**	**2011 г.**	**2012 г.**	**2013 г.**
Номинальный ВВП, млрд руб.	38 797,2	44 491,4	54 370,1	58 496,2	66 755,3,0

Целью исследования временного ряда является выявление закономерностей в изменении уровней ряда и построении его модели в целях прогнозирования и исследования взаимосвязей между явлениями [1]. Каждый уровень временного ряда

формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

1) факторы, формирующие тенденцию ряда (T);
2) факторы, формирующие периодические колебания ряда (S);
3) случайные факторы (E).

Большинство временных рядов экономических показателей имеют тенденцию, характеризующую совокупное долговременное воздействие множества факторов на динамику изучаемого показателя. Все эти факторы, взятые в отдельности, могут оказывать разнонаправленное воздействие на исследуемый показатель. Однако в совокупности они формируют его возрастающую или убывающую тенденцию (рис. 6.1).

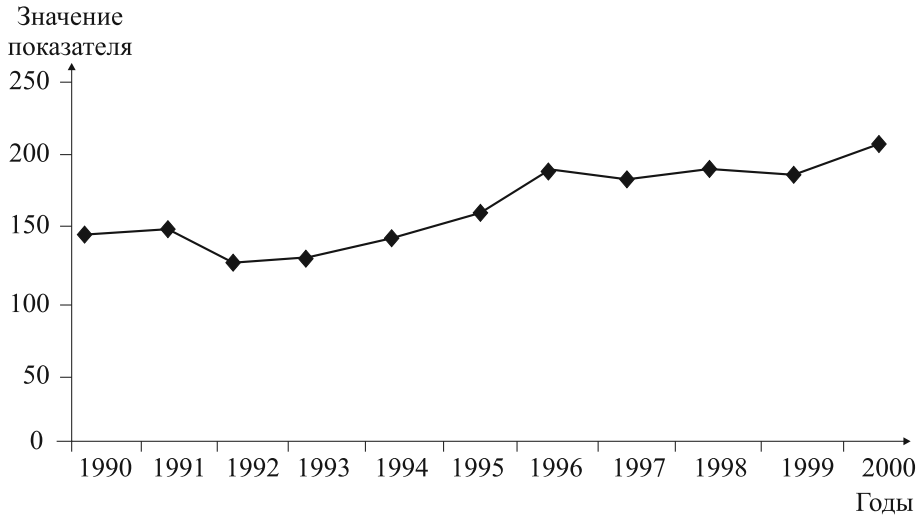


Рис. 6.1 – Временной ряд, содержащий возрастающую тенденцию

Также изучаемый показатель может быть подвержен периодическим колебаниям. В экономике периодические колебания принято подразделять на сезонные, у которых период колебаний не превышает одного года, и циклические с периодом колебаний несколько лет, связанные с циклами деловой активности. Сезонный характер колебаний характерен для ряда отраслей экономики, деятельность которых зависит от времени года (например, цены на сельскохозяйственную продукцию в летний период выше, чем в зимний; уровень безработицы в курортных городах в зимний период выше по сравнению с летним). При наличии больших массивов данных за длительные промежутки времени можно выявить циклические колебания. На рисунке 6.2 показан временной ряд, который помимо тенденции содержит сезонную компоненту.

Некоторые временные ряды не содержат тенденции и циклической компоненты, а каждый следующий их уровень образуется как сумма среднего уровня ряда и некоторой (положительной или отрицательной) случайной компоненты (рис. 6.3).

Очевидно, что на практике каждый уровень временного ряда формируется под воздействием тенденции, сезонных колебаний и случайной компоненты. Объединяя различным образом эти компоненты, можно получить различные модели временного ряда (Y):

- аддитивную — модель, в которой временной ряд представлен как сумма перечисленных компонент:

$$Y_t = T_t + S_t + E_t;$$

- мультипликативную — модель, в которой временной ряд представлен как произведение перечисленных компонент:

$$Y_t = T_t \cdot S_t \cdot E_t;$$

- смешанную:

$$Y_t = T_t \cdot S_t + E_t.$$

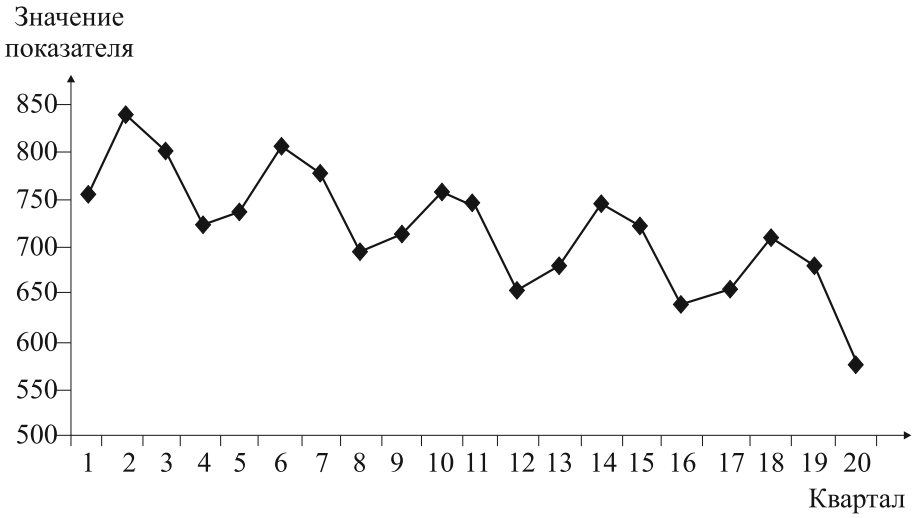


Рис. 6.2 – Временной ряд, содержащий убывающую тенденцию и циклическую компоненту

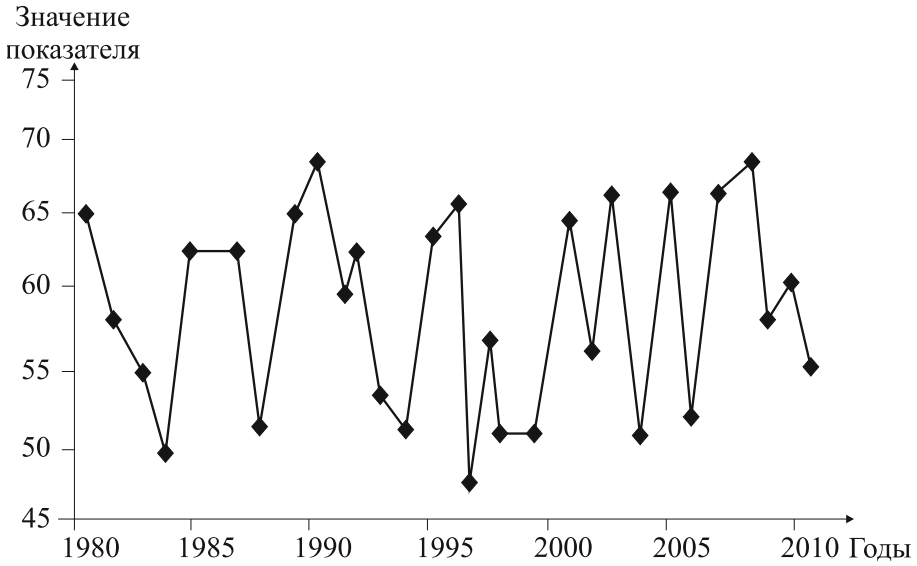


Рис. 6.3 – Временной ряд, содержащий случайную компоненту

Основная задача эконометрического исследования отдельного временного ряда — выявление и придание количественного выражения каждой из перечисленных выше компонент с тем, чтобы использовать полученную информацию для прогнозирования будущих значений ряда или при построении моделей взаимосвязи двух или более временных рядов.

Перед построением модели исходные данные проверяются на сопоставимость (применение одинаковой методики получения или расчета данных), однородность (отсутствие случайных выбросов), устойчивость (наличие закономерности в изменении уровней ряда) и достаточность (число наблюдений должно в 7–10 превосходить число параметров модели).

6.2 Автокорреляция уровней временного ряда

При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда называют автокорреляцией уровней ряда.

Степень тесноты автокорреляционной связи между уровнями ряда может быть определена с помощью коэффициентов автокорреляции, т. е. коэффициентов линейной корреляции между уровнями исходного временного ряда и уровнями ряда, сдвинутыми на несколько шагов назад во времени.

$$r_\tau = \frac{\sum_{t=\tau+1}^{n} (y_t - \bar{y}_{1\tau}) \cdot (y_{t-\tau} - \bar{y}_{2\tau})}{\sqrt{\sum_{t=\tau+1}^{n} (y_t - \bar{y}_{1\tau})^2 \cdot \sum_{t=\tau+1}^{n} (y_{t-\tau} - \bar{y}_{2\tau})^2}},$$

где τ — величина сдвига, называемая лагом, определяет порядок коэффициента автокорреляции,

$$\bar{y}_{1\tau} = \frac{\sum_{t=\tau+1}^{n} y_t}{n-\tau}; \quad \bar{y}_{2\tau} = \frac{\sum_{t=\tau+1}^{n} y_{t-\tau}}{n-\tau}.$$

При $\tau = 1$ эту величину называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда y_t и y_{t-1}.

$$r_1 = \frac{\sum_{t=2}^{n} (y_t - \bar{y}_1) \cdot (y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n} (y_t - \bar{y}_1)^2 \cdot \sum_{t=2}^{n} (y_{t-1} - \bar{y}_2)^2}}.$$

Аналогично можно определить коэффициенты автокорреляции второго и более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)\cdot(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}},$$

где

$$\bar{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}, \quad \bar{y}_4 = \frac{\sum_{t=3}^{n} y_{t-2}}{n-2}.$$

С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Считается целесообразным для обеспечения статистической достоверности коэффициентов автокорреляции использовать правило — максимальный лаг должен быть не больше $n/4$.

Коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции и таким образом характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для проверки ряда на наличие нелинейной тенденции рекомендуется вычислить линейные коэффициенты автокорреляции для временного ряда, состоящего из логарифмов исходных уровней. Отличные от нуля значения коэффициентов автокорреляции будут свидетельствовать о наличии нелинейной тенденции.

По знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных содержат положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию.

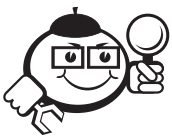
Последовательность коэффициентов автокорреляции уровней первого, второго и т. д. порядков называют *автокорреляционной функцией* временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции) называется *коррелограммой*.

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная. Таким образом, при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию.

Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то ряд содержит циклические колебания с периодичностью в τ моментов времени.

Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из двух предположений относительно структуры этого ряда: либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ. Поэтому коэффициент автокорреляции уровней и автокорреляционную функцию целесообразно использовать для выявления во временном ряде наличия или отсутствия трендовой компоненты и циклической (сезонной) компоненты.



Пример 6.1

Допустим, за указанный период (2002–2004 гг.) необходимо выявить основную тенденцию изменения фактического объема выпуска продукции и характер сезонных колебаний этого показателя. Данные для примера представлены в таблице 6.2.

Таблица 6.2 – Динамика выпуска продукции

Год	**Квартал**	t	**Объем выпуска продукции (тыс. руб.),** y_t
2002	I	1	15
	II	2	21
	III	3	9
	IV	4	18
2003	I	5	17
	II	6	20
	III	7	10
	IV	8	18
2004	I	9	17
	II	10	24
	III	11	13
	IV	12	22
2005	I	13	16
	II	14	25
	III	15	11
	IV	16	21

Построим поле корреляции.

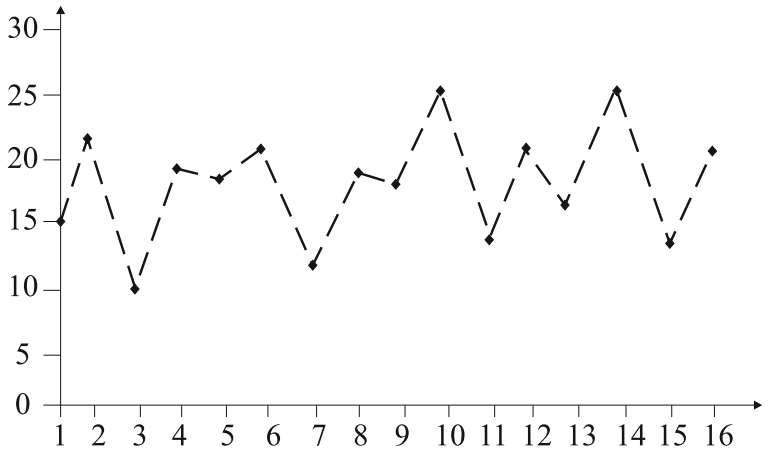


Рис. 6.4 – Поле корреляции

Уже исходя из графика (рис. 6.4) видно, что значения y_t образуют пилообразную фигуру. Рассчитаем несколько последовательных коэффициентов автокорреляции. Для этого составляем первую вспомогательную таблицу 6.3.

Таблица 6.3 – Вспомогательная таблица

t	y_t	y_{t-1}	$y_t - \bar{y}_1$	$y_{t-1} - \bar{y}_2$	$(y_t - \bar{y}_1) \times (y_{t-1} - \bar{y}_2)$	$(y_t - \bar{y}_1)^2$	$(y_{t-1} - \bar{y}_2)^2$
1	15	—	—	—	—	—	—
2	21	15	3,533	–2,067	–7,302	12,484	4,271
3	9	21	–8,467	3,933	–33,302	71,684	15,471
4	18	9	0,533	–8,067	–4,302	0,284	65,071
5	17	18	–0,467	0,933	–0,436	0,218	0,871
6	20	17	2,533	–0,067	–0,169	6,418	0,004
7	10	20	–7,467	2,933	–21,902	55,751	8,604
8	18	10	0,533	–7,067	–3,769	0,284	49,938
9	17	18	–0,467	0,933	–0,436	0,218	0,871
10	24	17	6,533	–0,067	–0,436	42,684	0,004
11	13	24	–4,467	6,933	–30,969	19,951	48,071
12	22	13	4,533	–4,067	–18,436	20,551	16,538
13	16	22	–1,467	4,933	–7,236	2,151	24,338
14	25	16	7,533	–1,067	–8,036	56,751	1,138
15	11	25	–6,467	7,933	–51,302	41,818	62,938
16	21	11	3,533	–6,067	–21,436	12,484	36,804
Сумма	262	256	$2{,}5E-14$	0	–209,467	343,733	334,933
Среднее значение	17,467	17,067	—	—	—	—	—

Следует заметить, что среднее значение получается путем деления не на 16, а на 15, т. к. у нас теперь на одно наблюдение меньше.

Теперь вычисляем коэффициент автокорреляции первого порядка по формуле:

$$r_1 = \frac{-209{,}467}{\sqrt{343{,}733 \cdot 334{,}933}} = -0{,}617.$$

Составляем вспомогательную таблицу для расчета коэффициента автокорреляции второго порядка (табл. 6.4).

Вычислим коэффициент автокорреляции второго порядка:

$$r_2 = \frac{148{,}5}{\sqrt{330{,}357 \cdot 295{,}5}} = 0{,}475.$$

Аналогично находим коэффициенты автокорреляции более высоких порядков, а все полученные значения заносим в сводную таблицу (табл. 6.5).

Построим график зависимости значений коэффициентов автокорреляции уровней ряда от величины лага.

Анализ коррелограммы (рис. 6.5) и графика исходных уровней временного ряда позволяет сделать вывод о наличии в изучаемом временном ряде сезонных колебаний периодичностью в четыре квартала.

Таблица 6.4 – Таблица для расчета коэффициента автокорреляции второго порядка

t	y_t	y_{t-2}	$y_t - \bar{y}_3$	$y_{t-2} - \bar{y}_4$	$(y_t - \bar{y}_3) \times (y_{t-2} - \bar{y}_4)$	$(y_t - \bar{y}_3)^2$	$(y_{t-2} - \bar{y}_4)^2$
1	15	—	—	—	—	—	—
2	21	—	—	—	—	—	—
3	9	15	–8,214	–2,500	20,536	67,474	6,250
4	18	21	0,786	3,500	2,750	0,617	12,250
5	17	9	–0,214	–8,500	1,821	0,046	72,250
6	20	18	2,786	0,500	1,393	7,760	0,250
7	10	17	–7,214	–0,500	3,607	52,046	0,250
8	18	20	0,786	2,500	1,964	0,617	6,250
9	17	10	–0,214	–7,500	1,607	0,046	56,250
10	24	18	6,786	0,500	3,393	46,046	0,250
11	13	17	–4,214	–0,500	2,107	17,760	0,250
12	22	24	4,786	6,500	31,107	22,903	42,250
13	16	13	–1,214	–4,500	5,464	1,474	20,250
14	25	22	7,786	4,500	35,036	60,617	20,250
15	11	16	–6,214	–1,500	9,321	38,617	2,250
16	21	25	3,786	7,500	28,393	14,332	56,250
Сумма	241	245	$-1{,}4E-14$	0	148,5	330,357	295,5
Среднее значение	17,214	17,5	—	—	—	—	—

Таблица 6.5

Лаг	**Коэффициент автокорреляции уровней**
1	–0,617
2	0,475
3	–0,643
4	**0,908**
5	–0,622
6	0,380
7	–0,709
8	**0,931**
9	–0,647
10	0,487
11	–0,749
12	**0,988**

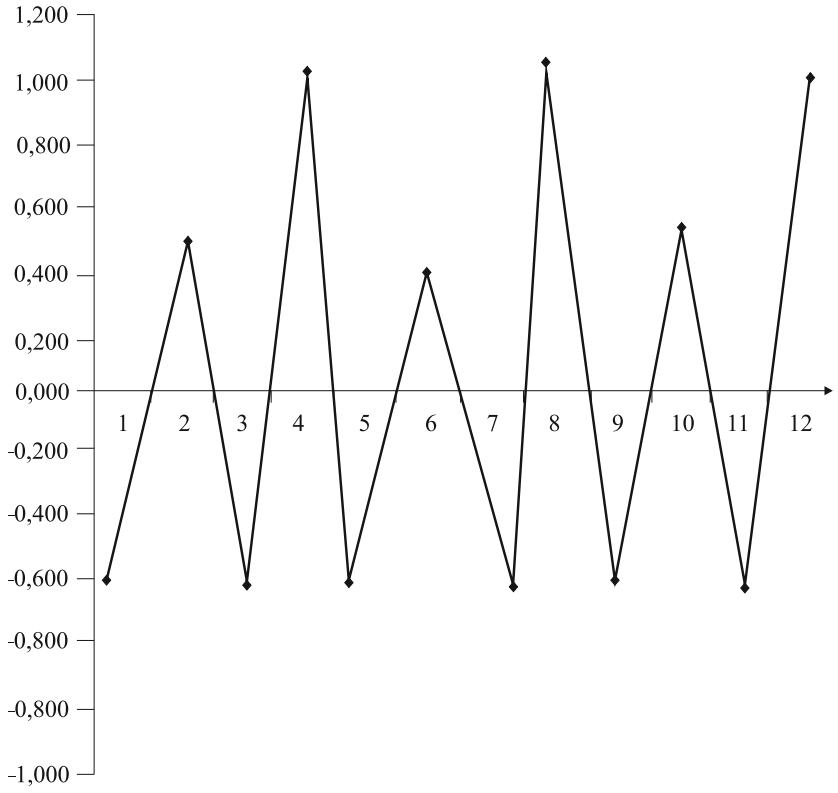


Рис. 6.5 – Коррелограмма

6.3 Моделирование тенденции временного ряда

Важнейшей задачей анализа временных рядов является моделирование тенденции временного ряда. Распространенным способом моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени или тренда. Этот способ называют *аналитическим выравниванием временного ряда*.

Зависимость от времени может принимать разные формы, поэтому для ее формализации используют различные виды функций:

- линейная: $\widehat{y}_t = a + b \cdot t$;
- гипербола: $\widehat{y}_t = a + b/t$;
- экспонента: $\widehat{y}_t = e^{a+b \cdot t}$;
- степенная функция: $\widehat{y}_t = a \cdot t^b$;
- полином: $\widehat{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + \ldots + b_k \cdot t^k$.

Параметры каждого из перечисленных выше трендов можно определить обычным МНК, используя в качестве независимой переменной время $t = 1, 2, \ldots, n$, а в качестве зависимой переменной — фактические уровни временного ряда y_t. Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

Определение типа тенденции можно выполнить различными способами.

1. Построение и визуальный анализ графика зависимости уровней ряда от времени. В этих же целях можно использовать и коэффициенты автокорреляции уровней ряда.
2. Сравнительный анализ коэффициентов автокорреляции первого порядка, рассчитанных по исходным и преобразованным уровням ряда. Если временной ряд имеет линейную тенденцию, то коэффициент автокорреляции первого порядка уровней y_t и y_{t-1} исходного ряда должен быть высоким. Если временной ряд содержит нелинейную тенденцию, например в форме экспоненты, то коэффициент автокорреляции первого порядка по логарифмам уровней исходного ряда будет выше, чем соответствующий коэффициент, рассчитанный по уровням исходного ряда. Чем сильнее выражена нелинейная тенденция в изучаемом временном ряде, тем в большей степени будут различаться значения указанных коэффициентов.
3. Перебор основных форм тренда с расчетом по каждому уравнению скорректированного коэффициента детерминации и средней ошибки аппроксимации. Выбор наилучшего уравнения посредством анализа вычисленных показателей. Этот метод легко реализуется при компьютерной обработке данных.

6.4 Моделирование сезонных колебаний

Простейший прием моделирования периодической компоненты основан на использовании сглаживания временного ряда по методу простой скользящей средней. Предварительно следует определиться с видом модели временного ряда — аддитивной или мультипликативной. Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда $Y = S + T + E$, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда $Y = S \cdot T \cdot E$.

Построение аддитивной и мультипликативной моделей сводится к расчету значений T, S и E для каждого уровня ряда.

Процесс построения 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 и получение выровненных данных в аддитивной $(T+E)$ или в мультипликативной $(T \cdot E)$ модели.
4. Аналитическое выравнивание уровней $(T+E)$ или $(T \cdot E)$ и расчет значений T с использованием полученного уравнения тренда.
5. Расчет полученных по модели значений $(T+E)$ или $(T \cdot E)$.
6. Расчет абсолютных и/или относительных ошибок. Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные

уровни ряда и в дальнейшем использовать временной ряд ошибок E для анализа взаимосвязи исходного ряда и других временных рядов.

Методику построения каждой из моделей рассмотрим на примерах.

Пример 6.2

Построение аддитивной модели временного ряда.

Обратимся к данным, представленным в таблице 6.2.

Было показано, что данный временной ряд содержит сезонные колебания периодичностью 4. Рассчитаем компоненты аддитивной модели временного ряда.

Шаг 1. Проведем выравнивание исходных уровней ряда методом скользящей средней. Для этого:

1.1 Просуммируем уровни ряда последовательно за каждые четыре квартала со сдвигом на один момент времени и определим условные годовые объемы потребления электроэнергии (гр. 3 табл. 6.6).

1.2 Разделив полученные суммы на 4, найдем скользящие средние (гр. 4 табл. 6.6). Полученные таким образом выровненные значения уже не содержат сезонной компоненты.

1.3 Приведем эти значения в соответствие с фактическими моментами времени, для чего найдем средние значения из двух последовательных скользящих средних — центрированные скользящие средние (гр. 5 табл. 6.6).

Шаг 2. Найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (гр. 6 табл. 6.6). Используем эти оценки для расчета значений сезонной компоненты S (табл. 6.7). Для этого найдем средние за каждый квартал (по всем годам) оценки сезонной компоненты S_i. В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопогашаются. В аддитивной модели это выражается в том, что сумма значений сезонной компоненты по всем кварталам должна быть равна нулю.

Для данной модели сумма средних оценок сезонной компоненты равна:

$$-0{,}83 + 5{,}29 - 6{,}38 + 2{,}08 = 0{,}16.$$

Эта сумма оказалась не равной нулю, поэтому каждую оценку уменьшим на величину поправки, равной одной четверти полученного значения:

$$\Delta = \frac{0{,}16}{4} = 0{,}04.$$

Рассчитываем скорректированные значения сезонной компоненты ($S_i = \overline{S}_i - \Delta$) и заносим полученные данные в таблицу 6.7.

Проверим равенство нулю суммы значений сезонной компоненты:

$$-0{,}87 + 5{,}25 - 6{,}42 + 2{,}04 = 0.$$

Таблица 6.6 – Выравнивание исходных уровней ряда

t	y_t	Итого за 4 квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	**2**	**3**	**4**	**5**	**6**
1	15	—	—	—	—
2	21	63	15,75	—	—
3	9	65	16,25	16,00	–7,00
4	18	64	16,00	16,13	1,88
5	17	65	16,25	16,13	0,88
6	20	65	16,25	16,25	3,75
7	10	65	16,25	16,25	–6,25
8	18	69	17,25	16,75	1,25
9	17	72	18,00	17,63	–0,63
10	24	76	19,00	18,50	5,50
11	13	75	18,75	18,88	–5,88
12	22	76	19,00	18,88	3,13
13	16	74	18,50	18,75	–2,75
14	25	73	18,25	18,38	6,63
15	11	—	—	—	—
16	21	—	—	—	—

Таблица 6.7 – Оценка сезонной компоненты

Показатели	Год	№ квартала, i			
		I	II	III	IV
	2002	—	—	–7,00	1,88
	2003	0,88	3,75	–6,25	1,25
	2004	–0,63	5,50	–5,88	3,13
	2005	–2,75	6,63	—	—
Всего за i-й квартал		–2,50	15,88	–19,13	6,25
Средняя оценка сезонной компоненты для i-го квартала, $\overline{S}_i$		–0,83	5,29	–6,38	2,08
Скорректированная сезонная компонента, S_i		–0,87	5,25	–6,42	2,04

Шаг 3. Исключим влияние сезонной компоненты, вычитая ее значение из каждого уровня исходного временного ряда. Получим величины $T + E = Y - S$ (гр. 4 табл. 6.8). Эти значения рассчитываются за каждый момент времени и содержат только тенденцию и случайную компоненту.

Шаг 4. Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $(T + E)$ с помощью линейного тренда. Результаты аналитического выравнивания следующие:

$$T = 0{,}251 \cdot t + 15{,}179.$$

Подставляя в это уравнение значения $t = 1, 2, 3, \ldots, 16$, найдем уровни T для каждого момента времени (гр. 5 табл. 6.8).

Таблица 6.8 – Выделение тенденции ряда

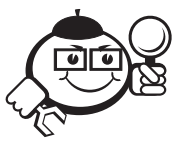
t	y_t	S_i	$y_t - S_i$	T	$T + S$	$E = y_t - (T + S)$	E_2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	15	–0,88	15,88	15,43	14,56	0,45	0,20
2	21	5,25	15,75	15,68	20,93	0,07	0,00
3	9	–6,42	15,42	15,93	9,52	–0,52	0,27
4	18	2,04	15,96	16,18	18,22	–0,22	0,05
5	17	–0,88	17,88	16,43	15,56	1,44	2,08
6	20	5,25	14,75	16,69	21,94	–1,94	3,74
7	10	–6,42	16,42	16,94	10,52	–0,52	0,27
8	18	2,04	15,96	17,19	19,23	–1,23	1,51
9	17	–0,88	17,88	17,44	16,56	0,44	0,19
10	24	5,25	18,75	17,69	22,94	1,06	1,13
11	13	–6,42	19,42	17,94	11,52	1,48	2,18
12	22	2,04	19,96	18,19	20,23	1,77	3,12
13	16	–0,88	16,88	18,44	17,57	–1,57	2,46
14	25	5,25	19,75	18,69	23,94	1,06	1,12
15	11	–6,42	17,42	18,94	12,53	–1,53	2,33
16	21	2,04	18,96	19,20	21,24	–0,24	0,06

Шаг 5. Найдем значения уровней ряда, полученные по аддитивной модели. Для этого прибавим к уровням T значения сезонной компоненты для соответствующих кварталов (гр. 6 табл. 6.8).

Для оценки качества построенной модели вычислим коэффициент детерминации.

$$R^2 = 1 - \frac{\sum E^2}{\sum (y_t - \bar{y})^2} = 1 - \frac{20{,}70}{349{,}44} = 0{,}94.$$

Следовательно, можно сказать, что аддитивная модель объясняет 94% общей вариации уровней временного ряда, отражающего изменения фактического объема выпуска продукции по кварталам за 4 года.



Пример 6.3

Построение мультипликативной модели.

Данные возьмем из таблицы 6.2.

Шаг 1. Методика, применяемая на этом шаге, полностью совпадает с методикой построения аддитивной модели.

Шаг 2. Найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированные скользящие средние (гр. 6 табл. 6.9). Эти

оценки используются для расчета сезонной компоненты S (табл. 6.10). Для этого найдем средние за каждый квартал оценки сезонной компоненты S_i. Так же, как и в аддитивной модели, считается, что сезонные воздействия за период взаимопогашаются. В мультипликативной модели это выражается в том, что сумма значений сезонной компоненты по всем кварталам должна быть равна числу периодов в цикле. В нашем случае число периодов одного цикла равно 4.

Таблица 6.9 – Выравнивание исходных уровней ряда

t	y_t	**Итого за 4 квартала**	**Скользящая средняя за четыре квартала**	**Центрированная скользящая средняя**	**Оценка сезонной компоненты**
1	**2**	**3**	**4**	**5**	**6**
1	15	—	—	—	—
2	21	63	15,75	—	—
3	9	65	16,25	16,00	0,56
4	18	64	16,00	16,13	1,12
5	17	65	16,25	16,13	1,05
6	20	65	16,25	16,25	1,23
7	10	65	16,25	16,25	0,62
8	18	69	17,25	16,75	1,07
9	17	72	18,00	17,63	0,96
10	24	76	19,00	18,50	1,30
11	13	75	18,75	18,88	0,69
12	22	76	19,00	18,88	1,17
13	16	74	18,50	18,75	0,85
14	25	73	18,25	18,38	1,36
15	11	—	—	—	—
16	21	—	—	—	—

Таблица 6.10 – Оценка сезонной компоненты

Показатели	**Год**	**№ квартала, i**			
		I	**II**	**III**	**IV**
	2002	—	—	0,563	1,116
	2003	1,054	1,231	0,615	1,075
	2004	0,965	1,297	0,689	1,166
	2005	0,853	1,361	—	—
Всего за i-й квартал		2,872	3,889	1,867	3,356
Средняя оценка сезонной компоненты для i-го квартала, $\overline{S}_i$		0,957	1,296	0,622	1,119
Скорректированная сезонная компонента, S_i		0,959	1,298	0,623	1,120

Для данной модели сумма средних оценок сезонной компоненты равна:

$$0{,}957 + 1{,}296 + 0{,}622 + 1{,}119 = 3{,}994.$$

Эта сумма оказалась не равной четырем, поэтому для корректировки оценки сезонной компоненты вычислим значение поправки по формуле:

$$\Delta = \frac{4}{3{,}994} = 1{,}0015.$$

Скорректированные значения сезонной компоненты S_i получаются при умножении ее средней оценки $\overline{S}_i$ на корректирующий коэффициент Δ.

Проверяем условие равенства четырем суммы значений сезонной компоненты:

$$0{,}959 + 1{,}298 + 0{,}623 + 1{,}12 = 4.$$

Шаг 3. Разделим каждый уровень исходного ряда на соответствующие значения сезонной компоненты. В результате получим величины $T \cdot E = Y/S$ (гр. 4 табл. 6.11), которые содержат только тенденцию и случайную компоненту.

Таблица 6.11 – Выделение тенденции ряда

t	y_t	S_i	y_t/S_i	T	$T \cdot S$	$E = y_t/(T \cdot S)$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	15	0,96	15,64	15,37	14,74	1,02
2	21	1,30	16,18	15,63	20,29	1,04
3	9	0,62	14,45	15,88	9,90	0,91
4	18	1,12	16,07	16,14	18,08	1,00
5	17	0,96	17,73	16,40	15,72	1,08
6	20	1,30	15,41	16,65	21,61	0,93
7	10	0,62	16,05	16,91	10,53	0,95
8	18	1,12	16,07	17,16	19,22	0,94
9	17	0,96	17,73	17,42	16,71	1,02
10	24	1,30	18,49	17,68	22,94	1,05
11	13	0,62	20,87	17,93	11,17	1,16
12	22	1,12	19,64	18,19	20,37	1,08
13	16	0,96	16,68	18,44	17,69	0,90
14	25	1,30	19,26	18,70	24,27	1,03
15	11	0,62	17,66	18,96	11,81	0,93
16	21	1,12	18,75	19,21	21,52	0,98

Шаг 4. Определим компоненту T в мультипликативной модели. Для этого рассчитаем параметры линейного тренда, используя уровни $T \cdot E$. В результате получим уравнение тренда:

$$T = 0{,}2559 \cdot t + 15{,}117.$$

Подставляя в это уравнение значения $t = 1, 2, 3, \ldots, 16$, найдем уровни T для каждого момента времени (гр. 5 табл. 6.11).

Шаг 5. Найдем уровни ряда, умножив значения T на соответствующие значения сезонной компоненты (гр. 6 табл. 6.11). Расчет ошибки в мультипликативной модели производится по формуле: $E = y_t/(T \cdot S)$.

Для того чтобы сравнить мультипликативную модель и другие модели временного ряда, можно по аналогии с аддитивной моделью использовать сумму квадра-

тов абсолютных ошибок. Абсолютные ошибки в мультипликативной модели определяются как:

$$E = y_t - (T \cdot S).$$

Значение коэффициента детерминации для мультипликативной модели равно:

$$R^2 = 1 - \frac{\sum (y_t - T \cdot S)^2}{\sum (y_t - \bar{y})^2} = 1 - \frac{18{,}9}{349{,}44} = 0{,}9459.$$

Таким образом, доля объясненной дисперсии уровней ряда составляет 94,59%. Сравнивая показатели детерминации аддитивной и мультипликативной моделей, делаем вывод, что они примерно одинаково аппроксимируют исходные данные.

Прогнозирование по аддитивной или мультипликативной модели временного ряда сводится к расчету будущего значения временного ряда по уравнению модели в виде:

$$\hat{y}_t = T + S$$

для аддитивной или

$$y_t = T \cdot S$$

для мультипликативной модели.

Контрольные вопросы по главе 6

1. Что такое временной ряд? Основные составляющие временного ряда.
2. Дайте определение автокорреляции уровней и поясните, как она используется при моделировании динамического ряда.
3. Что такое автокорреляционная функция?
4. Запишите уравнение, определяющее аддитивную модель временного ряда.
5. Запишите уравнение, определяющее мультипликативную модель временного ряда.
6. Опишите шаги построения аддитивной модели ряда по методу скользящей средней.

ЗАКЛЮЧЕНИЕ

По словам Р. Фриша: «. . .каждая из трех отправных точек — статистика, экономическая теория и математика — необходимое, но недостаточное условие для понимания количественных соотношений в современной экономической жизни. Это единство всех трех составляющих. И это единство образует эконометрику».

В данном пособии рассмотрены основные вопросы курса «Эконометрика». Основное внимание уделено базовым вопросам эконометрического моделирования.

Автор постарался изложить материал в простой, доступной к пониманию форме и надеется, что изучение предложенного материала позволит читателю получить достаточно знаний и умений для освоения более сложных разделов эконометрики.

ЛИТЕРАТУРА

[1] Эконометрика : учебник / под ред. И. И. Елисеевой. — М. : Проспект, 2009. — 288 с.

[2] Кремер Н. Ш. Эконометрика : учебник для студентов вузов / Н. Ш. Кремер, Б. А. Путко ; под ред. Н. Ш. Кремера. — 2-е изд., стереотип. — М. : ЮНИТИ-ДАНА, 2008. — 311 с.

[3] Тихомиров Н. П. Эконометрика / Н. П. Тихомиров, Е. Ю. Дорохина. — 2-е изд., стереотип. — М. : Экзамен, 2007. — 512 с.

[4] Бородич С. А. Эконометрика : учеб. пособие / С. А. Бородич. — 3-е изд., стереотип. — Мн. : Новое знание, 2006. — 408 с.

Приложение А

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Таблица А.1 – Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84

продолжение на следующей странице

Таблица А.1 — Продолжение

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Таблица А.2 – Критические значения *t*-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01 (двухсторонний)

Число степеней свободы	α			**Число степеней свободы**	α		
d. f.	**0,10**	**0,05**	**0,01**	***d. f.***	**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	4,5041	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

продолжение на следующей странице

Таблица А.2 — Продолжение

Число степеней свободы ***d. f.***	α			**Число степеней свободы** ***d. f.***	α		
	0,10	**0,05**	**0,01**		**0,10**	**0,05**	**0,01**
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

Таблица А.3 – Значения статистик Дарбина– Уотсона $d_L d_U$ при 5%-ном уровне значимости

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
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	0,85	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	1,99
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

продолжение на следующей странице

Таблица А.3 — Продолжение

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
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

ГЛОССАРИЙ

Автокорреляционная функция временного ряда — последовательность коэффициентов автокорреляции первого, второго и других порядков.

Автокорреляция — корреляционная зависимость последовательных элементов временного или пространственного ряда данных.

Аппроксимация — процесс подбора эмпирической функции $f(x)$, которая была бы максимально близка к экспериментальным точкам, но в то же время была бы нечувствительна к случайным отклонениям измеряемой величины.

Временные ряды — данные о каких-либо показателях, характеризующих одни и те же объекты в различные моменты времени.

Гетероскедастичность — непостоянство дисперсии случайных ошибок регрессионной (эконометрической) модели.

Гомоскедастичность — постоянство дисперсии случайных ошибок регрессионной модели.

Доверительная вероятность — вероятность того, что значение оцениваемого параметра находится в доверительном интервале. Доверительная вероятность обычно выбирается из значений 0,9; 0,95; 0,99.

Доверительный интервал — интервал, который с заданной вероятностью накроет неизвестное значение оцениваемого параметра регрессионной модели. Границы доверительного интервала называют *доверительными границами*.

Качественная переменная — переменная, значение которой выражается, как правило, текстовым описанием.

Коррелограмма — график значений коэффициентов автокорреляции разных порядков.

Корреляционное поле (поле корреляции) — графическое представление, отражающее совокупность точек результативного и факторного признаков.

Корреляция — статистическая взаимосвязь двух или нескольких случайных величин, при которой изменения одной или нескольких из этих величин приводят к изменению другой или других величин.

Коэффициент детерминации — коэффициент, который характеризует долю дисперсии результативного признака (y), объясняемую регрессией, в общей дисперсии результативного признака.

Коэффициент регрессии — коэффициент при независимой переменной.

Лаг — величина интервала запаздывания.

Лаговая переменная — объясняющая переменная, значения которой взяты с запаздыванием во времени.

Математическое ожидание — среднее ожидаемое значение случайной величины.

Мультиколлинеарность — наличие линейной зависимости между независимыми переменными (факторами) регрессионной модели.

Несмещенная оценка — оценка, математическое ожидание которой равно оцениваемому параметру.

Остатки регрессии — разности между наблюдаемыми значениями и значениями, предсказанными изучаемой регрессионной моделью.

Пространственные данные — относящиеся к одному и тому же моменту времени данные о каком-либо экономическом показателе, характеризующем однотипные объекты.

Ранг матрицы — размер наибольшей ее квадратной подматрицы, определитель которой не равен нулю.

Регрессионная модель — это уравнение, в котором объясняемая переменная представляется в виде функции от объясняющих переменных.

Сезонная составляющая временного ряда — периодические колебания уровней временного ряда в течение не очень длительного периода (недели, месяца, максимум — года).

Состоятельная оценка — оценка, которая дает истинное значение при достаточно большом объеме выборки.

Спецификация модели (спецификация уравнения регрессии) — выбор формулы связи переменных.

Среднее квадратичное отклонение случайной величины — показатель рассеивания значений случайной величины относительно её математического ожидания. Определяется как квадратный корень из среднего арифметического всех квадратов разностей между данными величинами и их средним арифметическим.

Тренд временного ряда — последовательность значений объясняемой переменной y_t, соответствующей возрастающей последовательности моментов времени t.

Уровень значимости — вероятность ошибочного отклонения нулевой гипотезы. Величины 0,05, 0,01 и 0,001 — это так называемые стандартные уровни статистической значимости.

Экзогенная переменная — внешняя по отношению к модели переменная.

Эндогенная переменная — переменная, определяемая внутри модели.

Эффективная оценка — оценка, у которой дисперсия меньше дисперсии любой другой альтернативной оценки при фиксированном объеме выборки.

Учебное издание

Потахова Ирина Владимировна

ЭКОНОМЕТРИКА

Учебное пособие

Корректор Осипова Е. А.
Компьютерная верстка Мурзагулова Н. Е.

Издано в Томском государственном университете
систем управления и радиоэлектроники.
634050, г. Томск, пр. Ленина, 40
Тел. (3822) 533018.