ЭКОНОМЕТРИКА

Учебно-методическое пособие

Оглавление

Введение

Эконометрика – одна из базовых дисциплин экономического образования во всем мире. Однако до недавнего времени она не была признана в СССР и России. Это было связано с тем, что из трех основных составляющих эконометрики – экономической теории, экономической статистики и математики – две первые были представлены в нашей стране неудовлетворительно. Но теперь ситуация изменилась коренным образом.

Существуют различные варианты определения эконометрики:

1) расширенные, при которых к эконометрике относят все, что связано с измерениями в экономике;

2) узко инструментально ориентированные, при которых понимают определенный набор математико-статистических средств, позволяющих верифицировать модельные соотношения между анализируемыми экономическими показателями.

На наш взгляд, наиболее точно объяснил сущность эконометрики один из основателей этой науки Р.Фриш, который и ввел этот название в 1926 г.: «Эконометрика – это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией, хотя значительная часть этой теории носит количественный характер. Эконометрика не является синонимом приложений математики к экономике. Как показывает опыт, каждая из трех отправных точек – статистика, экономическая теория и математика – необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это единство всех трех составляющих. И это единство образует эконометрику»[1].

Эконометрика – это самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, приемов, методов и моделей, предназначенных для того, чтобы на базе экономической теории, экономической статистики и экономических измерений, математико-

[1] Frisch R. Editorial. Econometrica. – 1933. – № 1. – P. 2.

статистического инструментария придавать конкретное количественное выражение общим (качественным) закономерностям, обусловленным экономической теорией.

Эконометрический метод складывался в преодолении следующих трудностей, искажающих результаты применения классических статистических методов (сущность новых терминов будет раскрыта в дальнейшем):

1. асимметричности связей;
2. мультиколлинеарности связей;
3. эффекта гетероскедастичности;
4. автокорреляции;
5. ложной корреляции;
6. наличия лагов.

Для описания сущности эконометрической модели удобно разбить весь процесс моделирования на шесть основных этапов[2]:

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

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

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

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

[2] Подробнее см. [9], с. 31-37.

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

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

Эконометрическое моделирование реальных социально-экономических процессов и систем обычно преследует два типа конечных прикладных целей (или одну из них): 1) прогноз экономических и социально-экономических показателей, характеризующих состояние и развитие анализируемой системы; 2) имитацию различных возможных сценариев социально-экономического развития анализируемой системы (многовариантные сценарные расчеты, ситуационное моделирование).

При постановке задач эконометрического моделирования следует определить их иерархический уровень и профиль. Анализируемые задачи могут относиться к макро- (страна, межстрановой анализ), мезо- (регионы внутри страны) и микро- (предприятия, фирмы, семьи) уровням и быть направленными на решение вопросов различного профиля инвестиционной, финансовой или социальной политики, ценообразования, распределительных отношений и т.п.

Данное пособие написано на основе книг [1], [2] и с использованием других указанных источников. Учебный материал в пособии условно разбит на четыре части и приложения:

В первой части рассмотрены модели парной регрессии (линейная и нелинейные модели).

Во второй части достаточно подробно разбирается модель множественной линейной регрессии и кратко обсуждается проблемы гомоскедастичности и автокоррелированности остатков.

Третья часть посвящена системам одновременных эконометрических уравнений.

В четвертой части рассматриваются модели временных рядов.

Приложение A содержит краткие сведения из теории вероятностей и математической статистики.

В приложениях B, C и D содержатся тестовые задания, варианты контрольных работ по всем темам и экзаменационные вопросы.

Приложение E содержит статистико-математические таблицы распределений Фишера, Стьюдента и Дарбина-Уотсона.

1. Парная регрессия и корреляция

Парная регрессия представляет собой регрессию между двумя переменными – y и x, т. е. модель вида:

$$y = \hat{f}(x),$$

где y – зависимая переменная (результативный признак); x – независимая, или объясняющая, переменная (признак-фактор). Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости, поэтому практически в каждом отдельном случае величина y складывается из двух слагаемых:

$$y = \hat{y}_x + \varepsilon,$$

где y – фактическое значение результативного признака; $\hat{y}_x$ – теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

От правильно выбранной спецификации модели зависит величина случайных ошибок: они тем меньше, чем в большей мере теоретические значения результативного признака $\hat{y}_x$, подходят к фактическим данным y.

К ошибкам спецификации относятся неправильный выбор той или иной математической функции для $\hat{y}_x$ и недоучет в уравнении регрессии какого-либо существенного фактора, т. е. использование парной регрессии вместо множественной.

Наряду с ошибками спецификации могут иметь место ошибки выборки, которые имеют место в силу неоднородности данных в исходной статистической совокупности, что, как правило, бывает при изучении экономических процессов. Если совокупность неоднородна, то уравнение регрессии не имеет практического смысла. Для получения хорошего результата обычно исключают из совокупности единицы с аномальными значениями исследуемых признаков. И в этом случае результаты регрессии представляют собой выборочные характеристики.

Использование временной информации также представляет собой выборку из всего множества хронологических дат. Изменив временной интервал, можно получить другие результаты регрессии.

Наибольшую опасность в практическом использовании методов регрессии представляют ошибки измерения. Если ошибки спецификации можно уменьшить, изменяя форму модели (вид математической формулы), а ошибки выборки – увеличивая объем исходных данных, то ошибки измерения практически сводят на нет все усилия по количественной оценке связи между признаками.

Особенно велика роль ошибок измерения при исследовании на макроуровне. Так, в исследованиях спроса и потребления в качестве объясняющей переменной широко используется «доход на душу населения». Вместе с тем, статистическое измерение величины дохода сопряжено с рядом трудностей и не лишено возможных ошибок, например, в результате наличия скрытых доходов.

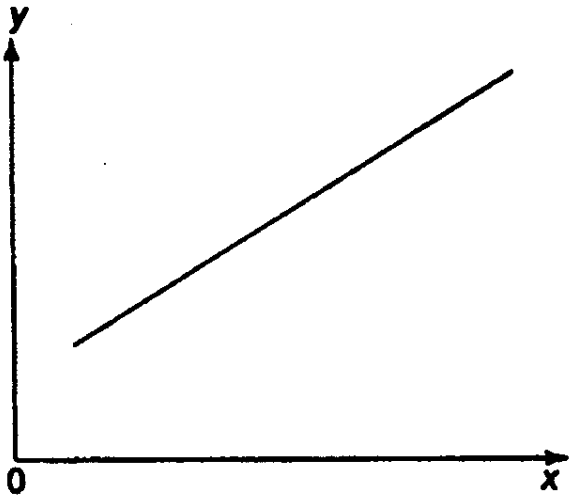
Предполагая, что ошибки измерения сведены к минимуму, основное внимание в эконометрических исследованиях уделяется ошибкам спецификации модели.

В парной регрессии выбор вида математической функции $\hat{y}_x = f(x)$ может быть осуществлен тремя методами:

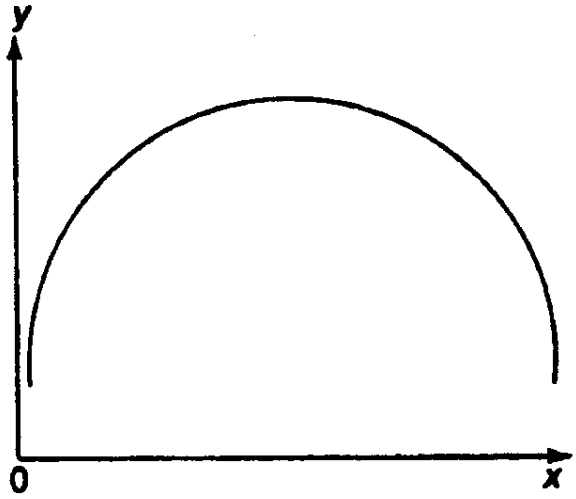
1) графическим;
2) аналитическим, т.е. исходя из теории изучаемой взаимосвязи;

3) экспериментальным.

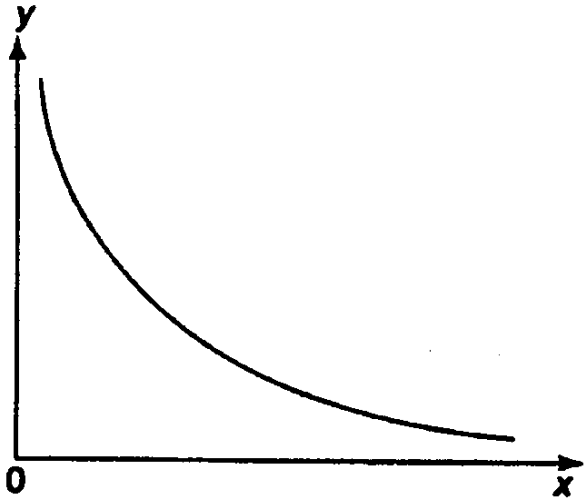
При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Он основан на поле корреляции. Основные типы кривых, используемые при количественной оценке связей, представлены на рис. 1.1:



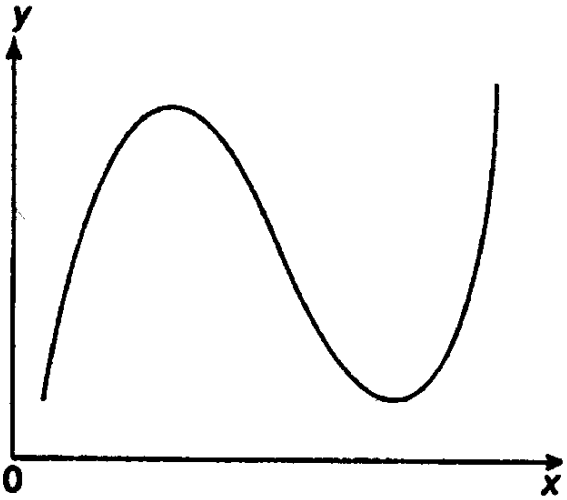
$$\hat{y}_x = a + b \cdot x$$



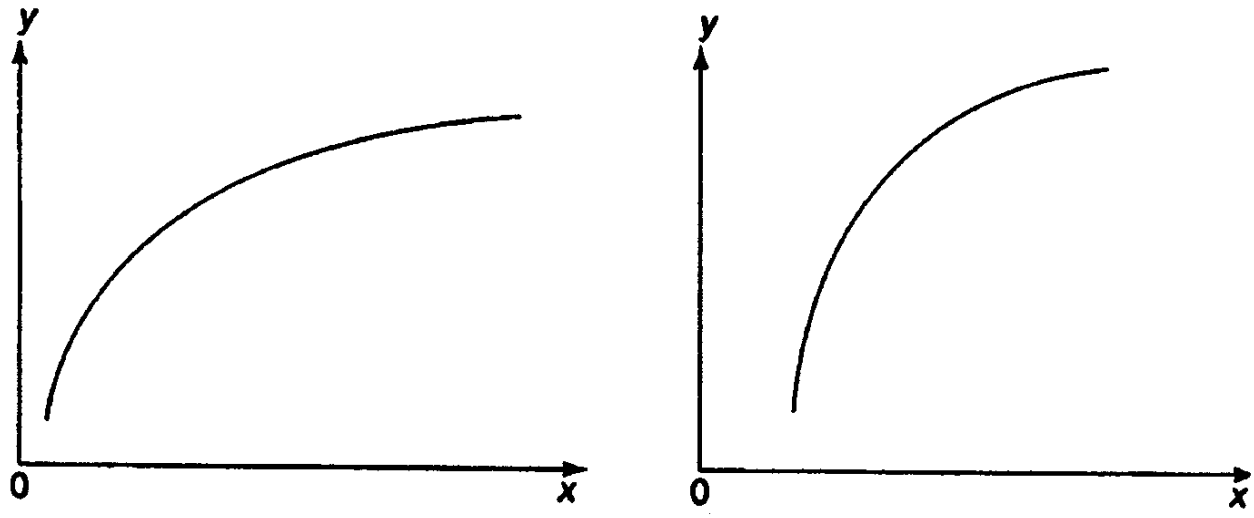
$$\hat{y}_x = a + b \cdot x + c \cdot x^2$$



$$\hat{y}_x = a + b/x$$



$$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$



$\hat{y}_x = a \cdot x^b$ $\hat{y}_x = a \cdot b^x$

Рис. 1.1. Основные типы кривых, используемые при количественной оценке связей между двумя переменными.

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 $\hat{y}_x = f(x)$, то фактические значения результативного признака совпадают с теоретическими $y = \hat{y}_x$, 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,

имеют место отклонения фактических данных от теоретических $\left(y-\hat{y}_x\right)$. Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

$$\sigma^2_{\text{ост}}=\frac{1}{n}\sum\left(y-\hat{y}_x\right)^2.$$

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. Значит, если мы выбираем параболу второй степени $\hat{y}_x=a+b\cdot x+c\cdot x^2$, то требуется объем информации уже не менее 14 наблюдений.

1.1. Линейная модель парной регрессии и корреляции

Рассмотрим простейшую модель парной регрессии – линейную регрессию. Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Линейная регрессия сводится к нахождению уравнения вида

$$\hat{y}_x=a+b\cdot x \text{ или } y=a+b\cdot x+\varepsilon. \qquad (1.1)$$

Уравнение вида $\hat{y}_x=a+b\cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Построение линейной регрессии сводится к оценке ее параметров – a и b. 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 $\hat{y}_x$ минимальна:

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_{x_i}\right)^2=\sum_{i=1}^{n}\varepsilon_i^2 \;\rightarrow\; \min. \qquad (1.2)$$

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

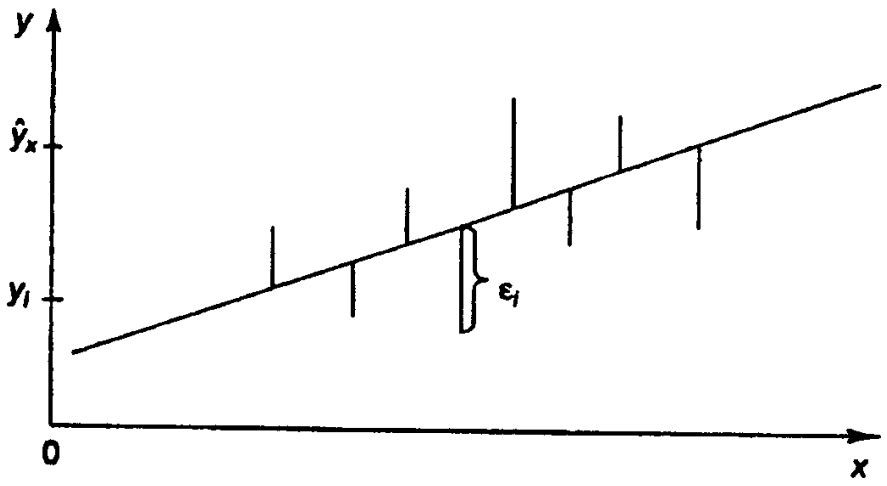


Рис. 1.2. Линия регрессии с минимальной дисперсией остатков.

Как известно из курса математического анализа, чтобы найти минимум функции (1.2), надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю. Обозначим $\sum_i \varepsilon_i^2$ через $S(a,\ b)$, тогда:

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

$$\begin{cases}\dfrac{\partial S}{\partial a}=-2\sum\left(y-a-b\cdot x\right)=0;\\[2ex] \dfrac{\partial S}{\partial b}=-2\sum x\left(y-a-b\cdot x\right)=0.\end{cases} \qquad (1.3)$$

После несложных преобразований, получим следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} a \cdot n + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum x \cdot y. \end{cases} \quad (1.4)$$

Решая систему уравнений (1.4), найдем искомые оценки параметров a и b. Можно воспользоваться следующими готовыми формулами, которые следуют непосредственно из решения системы (1.4):

$$a = \overline{y} - b \cdot \overline{x}, \qquad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}, \quad (1.5)$$

где $\operatorname{cov}(x, y) = \overline{y \cdot x} - \overline{y} \cdot \overline{x}$ – ковариация признаков x и y, $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x и

$$\overline{x} = \frac{1}{n}\sum x, \ \overline{y} = \frac{1}{n}\sum y, \ \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x, \ \overline{x^2} = \frac{1}{n}\sum x^2.$$

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности[3].

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.

[3] Более подробно смотри Приложение А.

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.

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 следующим формулам:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y}. \qquad (1.6)$$

Линейный коэффициент корреляции находится в пределах: $-1 \le r_{xy} \le 1$. Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную 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. Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = 1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}, \qquad (1.7)$$

где $\sigma_{\text{ост}}^2 = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2$, $\sigma_y^2 = \frac{1}{n}\sum\left(y - \bar{y}\right)^2 = \overline{y^2} - \bar{y}^2$.

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

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.

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

$$\overline{A}=\frac{1}{n}\sum\left|\frac{y-\hat{y}_x}{y}\right|\cdot 100\%. \qquad (1.8)$$

Средняя ошибка аппроксимации не должна превышать 8–10%.

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 $\overline{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

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

где $\sum(y-\overline{y})^2$ – общая сумма квадратов отклонений; $\sum\left(\hat{y}_x-\overline{y}\right)^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма

квадратов отклонений); $\sum\left(y-\hat{y}_x\right)^2$ – 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).

Таблица 1.1

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

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:

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}. \tag{1.9}$$

Фактическое значение F-критерия Фишера (1.9) сравнивается с табличным значением $F_{\text{табл}}\left(\alpha;\,k_1;\,k_2\right)$ 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

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}=\frac{\sum\left(\hat{y}_x-\overline{y}\right)^2}{\sum\left(y-\hat{y}_x\right)^2}\cdot\left(n-2\right). \tag{1.10}$$

Величина F-критерия связана с коэффициентом детерминации r_{xy}^2, и ее можно рассчитать по следующей формуле:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2). \qquad (1.11)$$

В парной линейной регрессии 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:

$$m_b = \sqrt{\frac{S_{\text{ост}}^2}{\sum (x - \overline{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}}, \qquad (1.12)$$

где $S_{\text{ост}}^2 = \dfrac{\sum \left(y - \hat{y}_x\right)^2}{n-2}$ – 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 его величина сравнивается с его стандартной ошибкой, т.е. определяется фактическое значение t-критерия Стьюдента: $t_b = \dfrac{b}{m_b}$ которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы $(n-2)$. Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\text{табл}} \cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора x ($b > 0$), уменьшение результативного признака при увеличении признака-фактора ($b < 0$) или его независимость от независимой

переменной ($b = 0$) (см. рис. 1.3), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-1{,}5 \le b \le 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

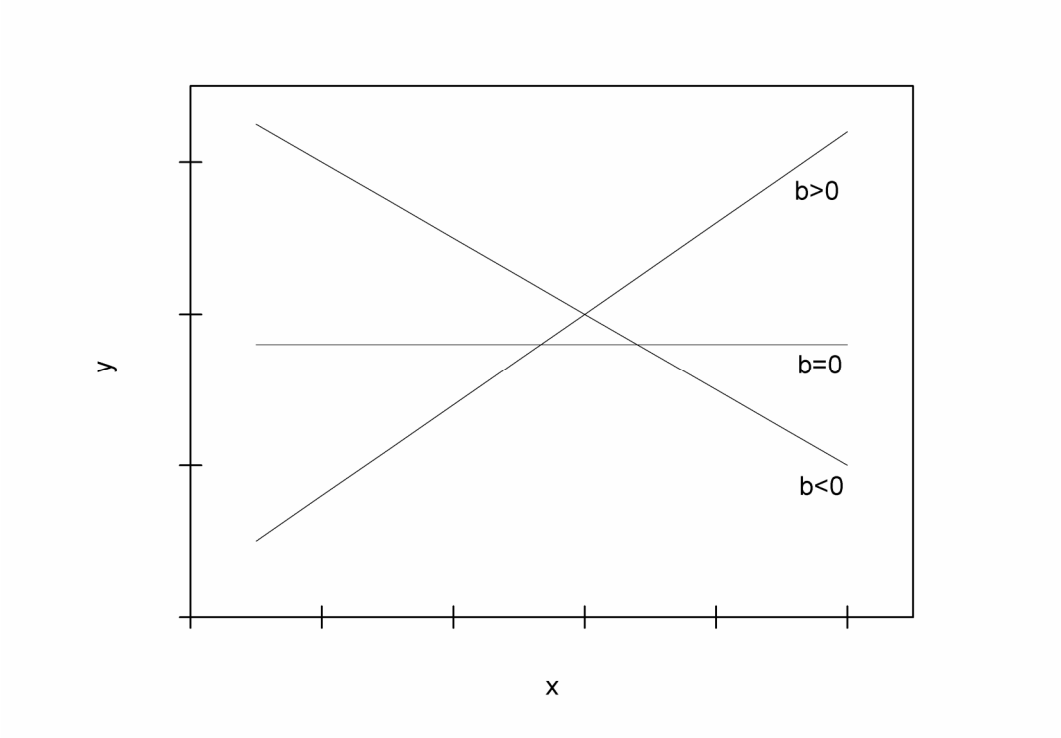


Рис. 1.3. Наклон линии регрессии в зависимости от значения параметра b.

Стандартная ошибка параметра a определяется по формуле:

$$m_a = \sqrt{S_{\text{ост}}^2 \cdot \frac{\sum x^2}{n \cdot \sum \left(x - \overline{x}\right)^2}} = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n}. \qquad (1.13)$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии. Вычисляется t-критерий: $t_a = \dfrac{a}{m_a}$, его величина сравнивается с табличным значением при $n - 2$ степенях свободы.

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r:

$$m_r = \sqrt{\frac{1-r^2}{n-2}}. \qquad (1.14)$$

Фактическое значение t-критерия Стьюдента определяется как $t_r = \frac{r}{m_r}$.

Существует связь между t-критерием Стьюдента и F-критерием Фишера:

$$t_b = t_r = \sqrt{F}. \qquad (1.15)$$

В прогнозных расчетах по уравнению регрессии определяется предсказываемое $\hat{y}_p$ значение как точечный прогноз $\hat{y}_x$ при $x_p = x_k$, т.е. путем подстановки в уравнение регрессии $\hat{y}_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно не реален. Поэтому он дополняется расчетом стандартной ошибки $\hat{y}_p$, т.е. $m_{\hat{y}_p}$, и соответственно интервальной оценкой прогнозного значения $\hat{y}_p$:

$$\hat{y}_p - \Delta_{\hat{y}_p} \le \hat{y}_p \le \hat{y}_p + \Delta_{\hat{y}_p},$$

где $\Delta_{\hat{y}_p} = m_{\hat{y}_p} \cdot t_{\text{табл}}$, а $m_{\hat{y}_p}$ – средняя ошибка прогнозируемого индивидуального значения:

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{n \cdot \sigma_x^2}}. \qquad (1.16)$$

Рассмотрим **пример**. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи.

Таблица 1.2

Расходы на продукты питания, y, тыс. руб.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Доходы семьи, x, тыс. руб.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим поле корреляции.

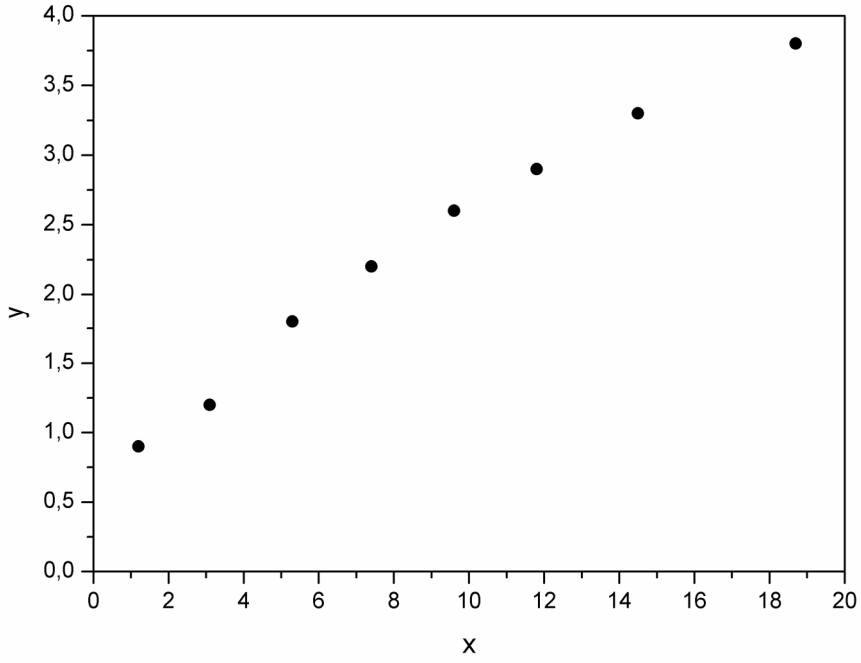


Рис. 1.4.

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

Для удобства дальнейших вычислений составим таблицу.

Таблица 1.3

	x	y	$x \cdot y$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$\left(y - \hat{y}_x\right)^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1,2	0,9	1,08	1,44	0,81	1,038	–0,138	0,0190	15,33
2	3,1	1,2	3,72	9,61	1,44	1,357	–0,157	0,0246	13,08
3	5,3	1,8	9,54	28,09	3,24	1,726	0,074	0,0055	4,11
4	7,4	2,2	16,28	54,76	4,84	2,079	0,121	0,0146	5,50
5	9,6	2,6	24,96	92,16	6,76	2,449	0,151	0,0228	5,81
6	11,8	2,9	34,22	139,24	8,41	2,818	0,082	0,0067	2,83
7	14,5	3,3	47,85	210,25	10,89	3,272	0,028	0,0008	0,85
8	18,7	3,8	71,06	349,69	14,44	3,978	–0,178	0,0317	4,68
Итого	71,6	18,7	208,71	885,24	50,83	18,717	–0,017	0,1257	52,19
Среднее значение	8,95	2,34	26,09	110,66	6,35	2,34	–	0,0157	6,52
σ	5,53	0,935	–	–	–	–	–	–	–
σ^2	30,56	0,874	–	–	–	–	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии $\hat{y}_x = a + b \cdot x$. Для этого воспользуемся формулами (1.5):

$$b = \frac{\operatorname{cov}(x,\ y)}{\sigma_x^2} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168;$$

$$a = \overline{y} - b \cdot \overline{x} = 2,34 - 0,168 \cdot 8,95 = 0,836.$$

Получили уравнение: $\hat{y}_x = 0,836 + 0,168 \cdot x$. Т.е. с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

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

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,168 \cdot \frac{5,53}{0,935} = 0,994.$$

Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками.

Коэффициент детерминации $r_{xy}^2 = 0,987$ (примерно тот же результат получим, если воспользуемся формулой (1.7)) показывает, что уравнением регрессии объясняется 98,7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,3%.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера. Сосчитаем фактическое значение F-критерия:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2) = \frac{0,987}{1 - 0,987} \cdot 6 = 455,54.$$

Табличное значение ($k_1 = 1,\ k_2 = n - 2 = 6,\ \alpha = 0,05$): $F_{\text{табл}} = 5,99$. Так как $F_{\text{факт}} > F_{\text{табл}}$, то признается статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента и доверительные интервалы

каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции

$$\left(S^2_{\text{ост}} = \frac{\sum \left(y - \hat{y}_x \right)^2}{n-2} = \frac{0,1257}{8-2} = 0,021 \right):$$

$$m_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{0,021}}{5,53 \cdot \sqrt{8}} = 0,0093,$$

$$m_a = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n} = \frac{\sqrt{0,021 \cdot 885,24}}{5,53 \cdot 8} = 0,0975,$$

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0,987}{6}} = 0,0465.$$

Фактические значения t-статистик: $t_b = \frac{0,168}{0,0093} = 18,065$,

$t_a = \frac{0,836}{0,0975} = 8,574$, $t_r = \frac{0,994}{0,0465} = 21,376$. Табличное значение t-критерия Стьюдента при $\alpha = 0,05$ и числе степеней свободы $\nu = n - 2 = 6$ есть $t_{\text{табл}} = 2,447$. Так как $t_b > t_{\text{табл}}$, $t_a > t_{\text{табл}}$ и $t_r > t_{\text{табл}}$, то признаем статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитаем доверительные интервалы для параметров регрессии a и b: $a \pm t \cdot m_a$ и $b \pm t \cdot m_b$. Получим, что $a \in [0,597;\ 1,075]$ и $b \in [0,145;\ 0,191]$.

Средняя ошибка аппроксимации (находим с помощью столбца 10 таблицы 1.3; $A_i = \left| \frac{y_i - \hat{y}_{x_i}}{y_i} \right| \cdot 100\%$) $\overline{A} = 6,52\%$ говорит о хорошем качестве уравнения регрессии, т.е. свидетельствует о хорошем подборе модели к исходным данным.

И, наконец, найдем прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \overline{x} = 1{,}1 \cdot 8{,}95 = 9{,}845$, т.е. найдем расходы на питание, если доходы семьи составят 9,85 тыс. руб.

$$\hat{y}_p = 0{,}836 + 0{,}168 \cdot 9{,}845 = 2{,}490 \text{ (тыс. руб.)}$$

Значит, если доходы семьи составят 9,845 тыс. руб., то расходы на питание будут 2,490 тыс. руб.

Найдем доверительный интервал прогноза. Ошибка прогноза

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{n \cdot \sigma_x^2}} = \sqrt{0{,}021 \cdot \left(1 + \frac{1}{8} + \frac{\left(9{,}845 - 8{,}95\right)^2}{8 \cdot 30{,}56}\right)} = 0{,}154,$$

а доверительный интервал ($\hat{y}_p - \Delta_{\hat{y}_p} \le \hat{y}_p \le \hat{y}_p + \Delta_{\hat{y}_p}$):

$$2{,}113 < \hat{y}_p < 2{,}867.$$

Т.е. прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:

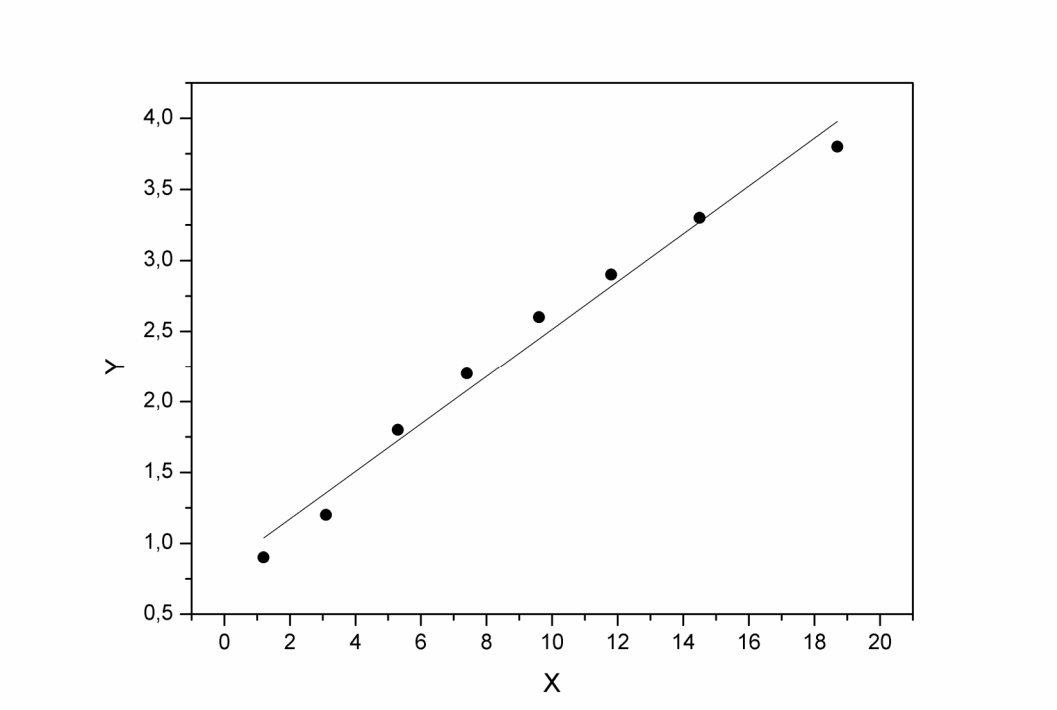


Рис. 1.5.

1.2. Нелинейные модели парной регрессии и корреляции

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Различают два класса нелинейных регрессий:

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

– полиномы различных степеней – $\hat{y}_x = a + b \cdot x + c \cdot x^2$,

$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$;

– равносторонняя гипербола – $\hat{y}_x = a + b/x$;

– полулогарифмическая функция – $\hat{y}_x = a + b \cdot \ln x$.

2. Регрессии, нелинейные по оцениваемым параметрам, например

– степенная – $\hat{y}_x = a \cdot x^b$;

– показательная – $\hat{y}_x = a \cdot b^x$;

– экспоненциальная – $\hat{y}_x = \mathrm{e}^{a+b \cdot x}$.

Регрессии нелинейные по включенным переменным приводятся к линейному виду простой заменой переменных, а дальнейшая оценка параметров производится с помощью метода наименьших квадратов. Рассмотрим некоторые функции.

Парабола второй степени $\hat{y}_x = a + b \cdot x + c \cdot x^2$ приводится к линейному виду с помощью замены: $x = x_1$, $x^2 = x_2$. В результате приходим к двухфакторному уравнению $\hat{y}_x = a + b \cdot x_1 + c \cdot x_2$, оценка параметров которого при помощи МНК, как будет показано в параграфе 2.2 приводит к системе следующих нормальных уравнений:

$$\begin{cases} a \cdot n + b \cdot \sum x_1 + c \cdot \sum x_2 = \sum y; \\ a \cdot \sum x_1 + b \cdot \sum x_1^2 + c \cdot \sum x_1 \cdot x_2 = \sum x_1 \cdot y; \\ a \cdot \sum x_2 + b \cdot \sum x_1 \cdot x_2 + c \cdot \sum x_2^2 = \sum x_2 \cdot y. \end{cases}$$

А после обратной замены переменных получим

$$\begin{cases} a \cdot n + b \cdot \sum x + c \cdot \sum x^2 = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 + c \cdot \sum x^3 = \sum x \cdot y; \\ a \cdot \sum x^2 + b \cdot \sum x^3 + c \cdot \sum x^4 = \sum x^2 \cdot y. \end{cases} \quad (1.17)$$

Парабола второй степени обычно применяется в случаях, когда для определенного интервала значений фактора меняется характер связи рассматриваемых признаков: прямая связь меняется на обратную или обратная на прямую.

Равносторонняя гипербола $\hat{y}_x = a + b/x$ 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 (например, кривая А.В. Филлипса), расходов на непродовольственные товары от доходов или общей суммы расходов (например, кривые Э. Энгеля) и в других случаях. Гипербола приводится к линейному уравнению простой заменой: $z = 1/x$. Система линейных уравнений при применении МНК будет выглядеть следующим образом:

$$\begin{cases} a \cdot n + b \cdot \sum \frac{1}{x} = \sum y; \\ a \cdot \sum \frac{1}{x} + b \cdot \sum \frac{1}{x^2} = \sum \frac{1}{x} \cdot y. \end{cases} \quad (1.18)$$

Аналогичным образом приводятся к линейному виду зависимости $\hat{y}_x = a + b \cdot \ln x$, $\hat{y}_x = a + b \cdot \sqrt{x}$ и другие.

Несколько иначе обстоит дело с регрессиями нелинейными по оцениваемым параметрам, которые делятся на два типа: нелинейные модели внутренне линейные (приводятся к линейному виду с помощью соответствующих преобразований, например, логарифмированием) и нелинейные модели внутренне нелинейные (к линейному виду не приводятся).

К внутренне линейным моделям относятся, например, степенная функция – $\hat{y}_x = a \cdot x^b$, показательная – $\hat{y}_x = a \cdot b^x$, экспоненциальная – $\hat{y}_x = \mathrm{e}^{a+b \cdot x}$, логистическая – $\hat{y}_x = \dfrac{a}{1+b \cdot \mathrm{e}^{-c \cdot x}}$, обратная – $\hat{y}_x = \dfrac{1}{a+b \cdot x}$.

К внутренне нелинейным моделям можно, например, отнести следующие модели: $\hat{y}_x = a + b \cdot x^c$, $\hat{y}_x = a \cdot \left(1 - \dfrac{1}{1-x^b}\right)$.

Среди нелинейных моделей наиболее часто используется степенная функция $y = a \cdot x^b \cdot \varepsilon$, которая приводится к линейному виду логарифмированием:

$$\ln y = \ln\left(a \cdot x^b \cdot \varepsilon\right);$$

$$\ln y = \ln a + b \cdot \ln x + \ln \varepsilon;$$

$$Y = A + b \cdot X + \mathrm{E},$$

где $Y = \ln y,\ X = \ln x,\ A = \ln a,\ \mathrm{E} = \ln \varepsilon$. Т.е. МНК мы применяем для преобразованных данных:

$$\begin{cases} A \cdot n + b \cdot \sum X = \sum Y, \\ A \cdot \sum X + b \cdot \sum X^2 = \sum X \cdot Y, \end{cases}$$

а затем потенцированием находим искомое уравнение.

Широкое использование степенной функции связано с тем, что параметр b в ней имеет четкое экономическое истолкование – он является коэффициентом эластичности. (Коэффициент эластичности показывает, на

сколько процентов измениться в среднем результат, если фактор изменится на 1%.) Формула для расчета коэффициента эластичности имеет вид:

$$Э = f'(x) \cdot \frac{x}{y}. \qquad (1.19)$$

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

$$\overline{Э} = f'(\overline{x}) \cdot \frac{\overline{x}}{\overline{y}}. \qquad (1.20)$$

Приведем формулы для расчета средних коэффициентов эластичности для наиболее часто используемых типов уравнений регрессии:

Таблица 1.5

Вид функции, y	Первая производная, y'	Средний коэффициент эластичности, $\overline{Э}$
1	**2**	**3**
$y = a + b \cdot x + \varepsilon$	b	$\frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$
$y = a + b \cdot x + c \cdot x^2 + \varepsilon$	$b + 2c \cdot x$	$\frac{(b + 2c \cdot \overline{x}) \cdot \overline{x}}{a + b \cdot \overline{x} + c \cdot \overline{x}^2}$
$y = a + \frac{b}{x} + \varepsilon$	$-\frac{b}{x^2}$	$-\frac{b}{a \cdot \overline{x} + b}$
$y = a \cdot x^b \cdot \varepsilon$	$a \cdot b \cdot x^{b-1}$	b
$y = a \cdot b^x \cdot \varepsilon$	$a \cdot \ln b \cdot b^x$	$\overline{x} \cdot \ln b$
$y = a + b \cdot \ln x + \varepsilon$	$\frac{b}{x}$	$\frac{b}{a + b \cdot \ln \overline{x}}$
$y = \frac{a}{1 + b \cdot e^{-c \cdot x + \varepsilon}}$	$\frac{a \cdot b \cdot c \cdot e^{-c \cdot x}}{(1 + b \cdot e^{-c \cdot x})^2}$	$\frac{b \cdot c \cdot \overline{x}}{b + e^{c \cdot \overline{x}}}$
$y = \frac{1}{a + b \cdot x + \varepsilon}$	$-\frac{b}{(a + b \cdot x)^2}$	$-\frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$

Возможны случаи, когда расчет коэффициента эластичности не имеет смысла. Это происходит тогда, когда для рассматриваемых признаков бессмысленно определение изменения в процентах.

Уравнение нелинейной регрессии, так же, как и в случае линейной зависимости, дополняется показателем тесноты связи. В данном случае это индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}, \qquad (1.21)$$

где $\sigma^2_y = \frac{1}{n}\sum\left(y - \overline{y}\right)^2$ – общая дисперсия результативного признака y, $\sigma^2_{\text{ост}} = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2$ – остаточная дисперсия.

Величина данного показателя находится в пределах: $0 \le \rho_{xy} \le 1$. Чем ближе значение индекса корреляции к единице, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

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

$$\rho^2_{xy} = 1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y} = \frac{\sigma^2_{\text{объясн}}}{\sigma^2_y}, \qquad (1.22)$$

т.е. имеет тот же смысл, что и в линейной регрессии; $\sigma^2_{\text{объясн}} = \frac{1}{n}\sum\left(\hat{y}_x - \overline{y}\right)^2$.

Индекс детерминации ρ^2_{xy} можно сравнивать с коэффициентом детерминации r^2_{xy} для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем величина r^2_{xy} меньше ρ^2_{xy}. А близость этих показателей указывает на то, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейную функцию.

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

$$F = \frac{\rho_{xy}^2}{1-\rho_{xy}^2} \cdot \frac{n-m-1}{m}, \qquad (1.23)$$

где ρ_{xy}^2 – индекс детерминации, n – число наблюдений, m – число параметров при переменной x. Фактическое значение F-критерия (1.23) 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).

О качестве нелинейного уравнения регрессии можно также судить и по средней ошибке аппроксимации, которая, так же как и в линейном случае, вычисляется по формуле (1.8).

Рассмотрим **пример** из параграфа 1.1, предположив, что связь между признаками носит нелинейный характер, и найдем параметры следующих нелинейных уравнений: $y = a + b \cdot \ln x + \varepsilon$, $y = a + b \cdot \sqrt{x} + \varepsilon$ и $y = a \cdot x^b \cdot \varepsilon$.

Для нахождения параметров регрессии $\hat{y}_x = a + b \cdot \ln x$ делаем замену $z = \ln x$ и составляем вспомогательную таблицу ($\varepsilon = y - \hat{y}_x$).

Таблица 1.5

	x	z	y	$z \cdot y$	z^2	y^2	$\hat{y}_x$	ε	ε^2	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	1,2	0,182	0,9	0,164	0,033	0,81	0,499	0,401	0,1610	44,58
2	3,1	1,131	1,2	1,358	1,280	1,44	1,508	-0,308	0,0947	25,64
3	5,3	1,668	1,8	3,002	2,781	3,24	2,078	-0,278	0,0772	15,43
4	7,4	2,001	2,2	4,403	4,006	4,84	2,433	-0,233	0,0541	10,57
5	9,6	2,262	2,6	5,881	5,116	6,76	2,709	-0,109	0,0119	4,20
6	11,8	2,468	2,9	7,157	6,092	8,41	2,929	-0,029	0,0008	0,99
7	14,5	2,674	3,3	8,825	7,151	10,89	3,148	0,152	0,0232	4,62
8	18,7	2,929	3,8	11,128	8,576	14,44	3,418	0,382	0,1459	10,05
Итого	71,6	15,315	18,7	41,918	35,035	50,83	18,720	-0,020	0,5688	116,08
Среднее значение	8,95	1,914	2,34	5,240	4,379	6,35	–	–	0,0711	14,51
σ	–	0,846	0,935	–	–	–	–	–	–	–
σ^2	–	0,716	0,874	–	–	–	–	–	–	–

Найдем уравнение регрессии:

$$b = \frac{\operatorname{cov}(z,\ y)}{\sigma_z^2} = \frac{5,240 - 1,914 \cdot 2,34}{0,716} = 1,063,$$

$$a = \overline{y} - b \cdot \overline{z} = 2,34 - 1,063 \cdot 1,914 = 0,305.$$

Т.е. получаем следующее уравнение регрессии: $\hat{y}_x = 0,305 + 1,063 \cdot \ln x$. Теперь заполняем столбцы 8-11 нашей таблицы.

Индекс корреляции находим по формуле (1.21):

$$\rho_{xy} = \sqrt{1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{0,0711}{0,874}} = 0,958,$$

а индекс детерминации $\rho_{xy}^2 = 0,918$, который показывает, что 91,8% вариации результативного признака объясняется вариацией признака-фактора, а 8,2% приходится на долю прочих факторов.

Средняя ошибка аппроксимации: $\overline{A} = 14,51\%$, что недопустимо велико.

F -критерий Фишера:

$$F = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \cdot \frac{n - m - 1}{m} = \frac{0,919}{1 - 0,919} \cdot \frac{8 - 1 - 1}{1} = 68,07,$$

значительно превышает табличное $F_{\text{табл}} = 5,99$.

Изобразим на графике исходные данные и линию регрессии:

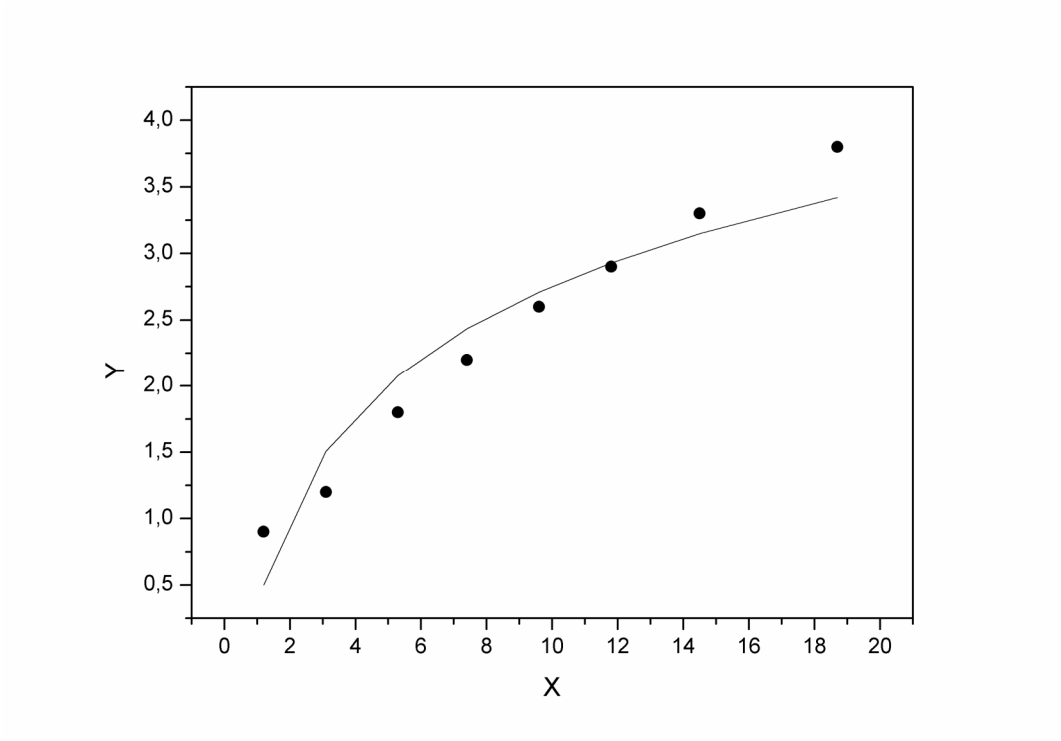


Рис. 1.6.

Для нахождения параметров регрессии $\hat{y}_x = a + b \cdot \sqrt{x}$ делаем замену $z = \sqrt{x}$ и составляем вспомогательную таблицу ($\varepsilon = y - \hat{y}_x$).

Таблица 1.6

	x	z	y	$z \cdot y$	z^2	y^2	$\hat{y}_x$	ε	ε^2	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	1,2	1,10	0,9	0,99	1,2	0,81	0,734	0,166	0,0276	18,46
2	3,1	1,76	1,2	2,11	3,1	1,44	1,353	-0,153	0,0235	12,77
3	5,3	2,30	1,8	4,14	5,3	3,24	1,857	-0,057	0,0033	3,19
4	7,4	2,72	2,2	5,98	7,4	4,84	2,247	-0,047	0,0022	2,12
5	9,6	3,10	2,6	8,06	9,6	6,76	2,599	0,001	0,0000	0,05
6	11,8	3,44	2,9	9,96	11,8	8,41	2,912	-0,012	0,0001	0,42
7	14,5	3,81	3,3	12,57	14,5	10,89	3,259	0,041	0,0017	1,20
8	18,7	4,32	3,8	16,43	18,7	14,44	3,740	0,060	0,0036	1,58
Итого	71,6	22,54	18,7	60,24	71,6	50,83	18,700	-0,001	0,0619	39,82
Среднее значение	8,95	2,82	2,34	7,53	8,95	6,35	–	–	0,0077	4,98
σ	–	1,00	0,935	–	–	–	–	–	–	–
σ^2	–	1,00	0,874	–	–	–	–	–	–	–

Найдем уравнение регрессии:

$$b = \frac{\operatorname{cov}(z,\, y)}{\sigma_z^2} = \frac{7,53 - 2,82 \cdot 2,34}{1,00} = 0,931,$$

$$a = \overline{y} - b \cdot \overline{z} = 2,34 - 0,931 \cdot 2,82 = -0,286.$$

Т.е. получаем следующее уравнение регрессии: $\hat{y}_x = -0,286 + 0,931 \cdot \sqrt{x}$.

Теперь заполняем столбцы 8-11 нашей таблицы.

Индекс корреляции находим по формуле (1.21):

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{0,0077}{0,874}} = 0,996,$$

а индекс детерминации $\rho^2 = 0,991$, который показывает, что 99,1% вариации результативного признака объясняется вариацией признака-фактора, а 0,9% приходится на долю прочих факторов.

Средняя ошибка аппроксимации: $\overline{A} = 0,0498 \cdot 100\% = 4,98\%$ показывает, что линия регрессии хорошо приближает исходные данные.

F -критерий Фишера:

$$F = \frac{\rho^2_{xy}}{1 - \rho^2_{xy}} \cdot \frac{n - m - 1}{m} = \frac{0,991}{1 - 0,991} \cdot \frac{8 - 1 - 1}{1} = 660,67,$$

значительно превышает табличное $F_{\text{табл}} = 5,99$.

Изобразим на графике исходные данные и линию регрессии:

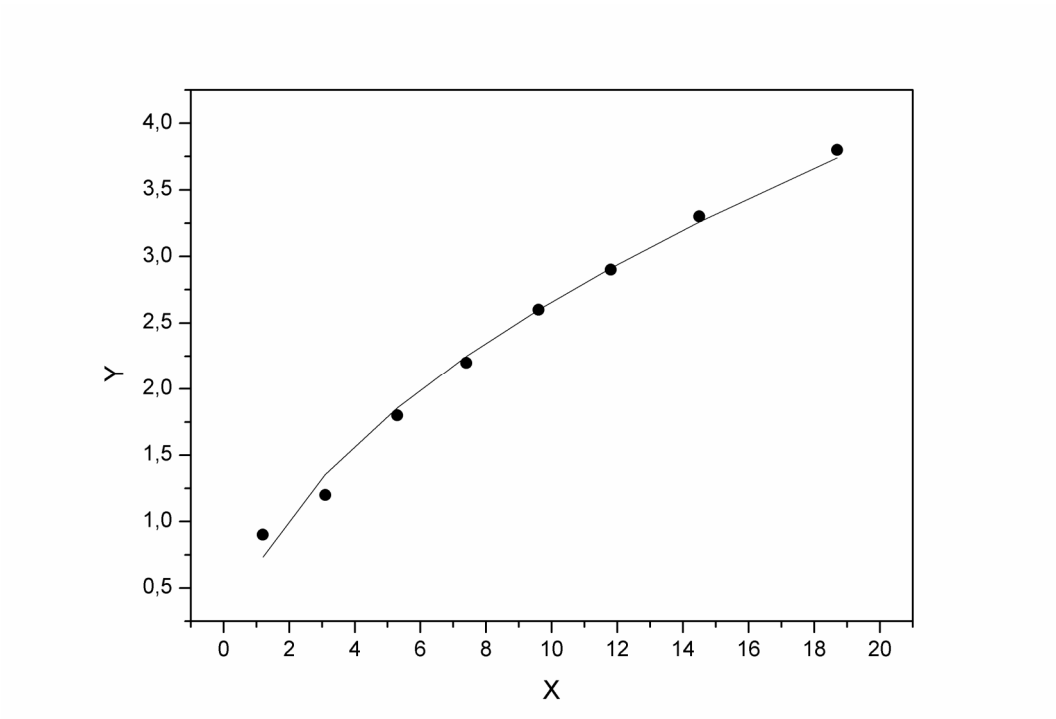


Рис. 1.7

Для нахождения параметров регрессии $y = a \cdot x^b \cdot \varepsilon$ необходимо провести ее линеаризацию, как было показано выше:

$$Y = A + b \cdot X + \mathrm{E},$$

где $Y = \ln y,\ X = \ln x,\ A = \ln a,\ \mathrm{E} = \ln \varepsilon$.

Составляем вспомогательную таблицу для преобразованных данных:

Таблица 1.7

	X	Y	$X \cdot Y$	X^2	Y^2	$\hat{y}_x$	ε	ε^2	A_i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0,182	-0,105	-0,019	0,033	0,011	0,8149	0,0851	0,0072	9,46
2	1,131	0,182	0,206	1,280	0,033	1,3747	-0,1747	0,0305	14,56
3	1,668	0,588	0,980	2,781	0,345	1,8473	-0,0473	0,0022	2,63
4	2,001	0,788	1,578	4,006	0,622	2,2203	-0,0203	0,0004	0,92
5	2,262	0,956	2,161	5,116	0,913	2,5627	0,0373	0,0014	1,43
6	2,468	1,065	2,628	6,092	1,134	2,8713	0,0287	0,0008	0,99
7	2,674	1,194	3,193	7,151	1,425	3,2165	0,0835	0,0070	2,53
8	2,929	1,335	3,910	8,576	1,782	3,7004	0,0996	0,0099	2,62
Итого	15,315	6,002	14,637	35,035	6,266	18,608	0,0919	0,0595	35,14
Среднее значение	1,914	0,750	1,830	4,379	0,783	–	–	0,0074	4,39
σ	0,846	0,470	–	–	–	–	–	–	–
σ^2	0,716	0,221	–	–	–	–	–	–	–

Найдем уравнение регрессии:

$$b = \frac{\mathrm{cov}(X,\ Y)}{\sigma_X^2} = \frac{1,830 - 1,914 \cdot 0,750}{0,716} = 0,551,$$

$$A = \overline{Y} - b \cdot \overline{X} = 0,750 - 0,551 \cdot 1,914 = -0,305.$$

Т.е. получаем следующее уравнение регрессии: $\hat{Y}_x = -0,305 + 0,551 \cdot X$. После потенцирования находим искомое уравнение регрессии:

$$\hat{y}_x = 0,737 \cdot x^{0,551}.$$

Теперь заполняем столбцы 7-10 нашей таблицы.

Индекс корреляции находим по формуле (1.21):

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma_y^2}} = \sqrt{1 - \frac{0,0074}{0,221}} = 0,983,$$

а индекс детерминации $\rho^2 = 0,967$, который показывает, что 96,7% вариации результативного признака объясняется вариацией признака-фактора, а 3,3% приходится на долю прочих факторов.

Средняя ошибка аппроксимации: $\overline{A} = 4,39\%$ показывает, что линия регрессии хорошо приближает исходные данные.

F -критерий Фишера:

$$F = \frac{\rho_{xy}^2}{1-\rho_{xy}^2} \cdot \frac{n-m-1}{m} = \frac{0,967}{1-0,967} \cdot \frac{8-1-1}{1} = 175,82,$$

значительно превышает табличное $F_{\text{табл}} = 5,99$.

Изобразим на графике исходные данные и линию регрессии:

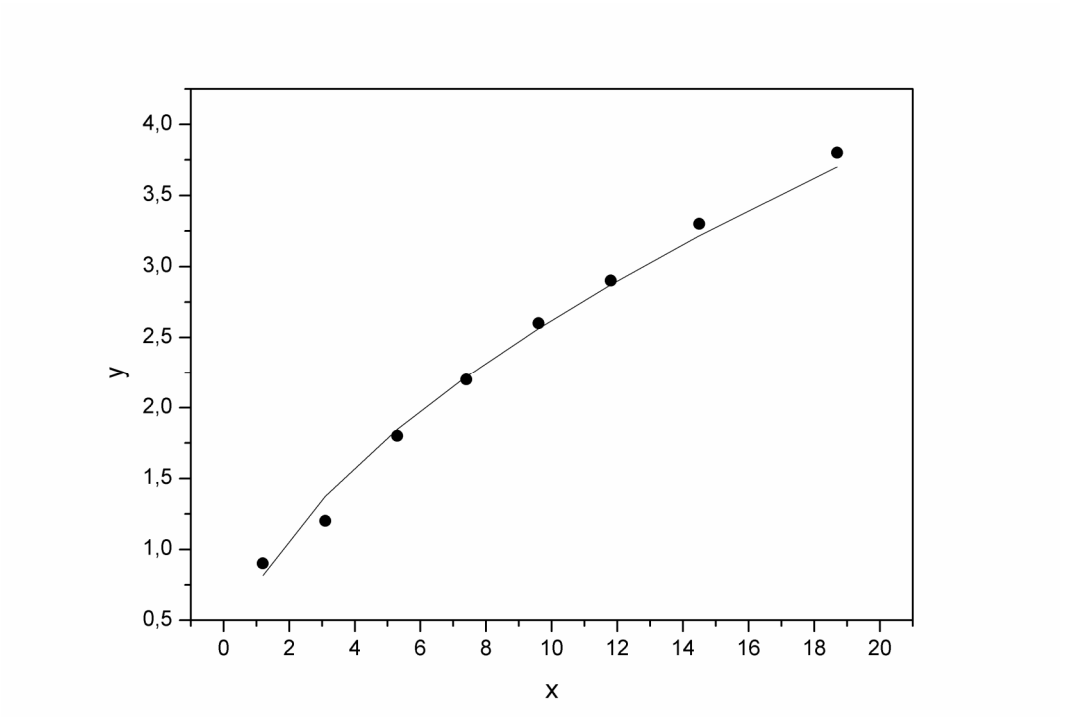


Рис. 1.8.

Сравним построенные модели по индексу детерминации и средней ошибке аппроксимации:

Таблица 1.8

Модель	Индекс детерминации, R^2 (r_{xy}^2, ρ_{xy}^2)	Средняя ошибка аппроксимации, $\overline{A}$, %
Линейная модель, $\hat{y}_x = a + b \cdot x$	0,987	6,52
Полулогарифмическая модель, $\hat{y}_x = a + b \cdot \ln x$	0,918	14,51
Модель с квадратным корнем, $\hat{y}_x = a + b \cdot \sqrt{x}$	0,991	4,98
Степенная модель, $y = a \cdot x^b \cdot \varepsilon$	0,967	4,39

Наиболее хорошо исходные данные аппроксимирует модель с квадратным корнем. Но в данном случае, так как индексы детерминации линейной модели и модели с квадратным корнем отличаются всего на 0,004, то вполне можно обойтись более простой линейной функцией.

2. Множественная регрессия и корреляция

Парная регрессия может дать хороший результат при моделировании, если влиянием других факторов, воздействующих на объект исследования, можно пренебречь. Если же этим влиянием пренебречь нельзя, то в этом случае следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии

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

где y – зависимая переменная (результативный признак), x_i – независимые, или объясняющие, переменные (признаки-факторы).

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и целом ряде других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

2.1. Спецификация модели. Отбор факторов при построении уравнения множественной регрессии

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Он включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

Включение в уравнение множественной регрессии того или иного набора факторов связано прежде всего с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям.

1. Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.

2. Факторы не должны быть интеркоррелированы и тем более находиться в точной функциональной связи.

Включение в модель факторов с высокой интеркорреляцией, может привести к нежелательным последствиям – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии.

Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. 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.

При дополнительном включении в регрессию $m+1$ фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться:

$$R^2_{m+1} \geq R^2_m \quad \text{и} \qquad S^2_{m+1} \leq S^2_m.$$

Если же этого не происходит и данные показатели практически не отличаются друг от друга, то 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.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Коэффициенты интеркорреляции (т.е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. 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. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

Пусть, например, при изучении зависимости $y = \hat{f}(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей:

Таблица 2.1

	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. 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 y $\left(r_{yx_2} = 0,7 < r_{yx_1} = 0,8\right)$, но зато значительно слабее межфакторная корреляция $r_{x_2x_3} = 0,2 < r_{x_1x_3} = 0,5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность 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.

Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

1. Затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы; параметры линейной регрессии теряют экономический смысл.

2. 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.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ $\left(i \neq j\right)$ были бы равны нулю. Так, для уравнения, включающего три объясняющих переменных

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный единице:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны единице, то определитель такой матрицы равен нулю:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И, наоборот, чем ближе к единице определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Существует ряд подходов преодоления сильной межфакторной корреляции. Самый простой путь устранения мультиколлинеарности состоит в исключении из модели одного или нескольких факторов. Другой подход

связан с преобразованием факторов, при котором уменьшается корреляция между ними.

Одним из путей учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Так, если $y = f(x_1, x_2, x_3)$, то возможно построение следующего совмещенного уравнения:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + \varepsilon .$$

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, 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.

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

1. Метод исключения – отсев факторов из полного его набора.
2. Метод включения – дополнительное введение фактора.
3. Шаговый регрессионный анализ – исключение ранее введенного фактора.

При отборе факторов также рекомендуется пользоваться следующим правилом: число включаемых факторов обычно в 6–7 раз меньше объема

совокупности, по которой строится регрессия. Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F-критерий меньше табличного значения.

2.2. Метод наименьших квадратов (МНК). Свойства оценок на основе МНК

Возможны разные виды уравнений множественной регрессии: линейные и нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция. В линейной множественной регрессии $\hat{y}_x = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m$ параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Рассмотрим линейную модель множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m + \varepsilon . \qquad (2.1)$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных $\hat{y}$ минимальна:

$$\sum_i \left(y_i - \hat{y}_{x_i} \right)^2 \rightarrow \min . \qquad (2.2)$$

Как известно из курса математического анализа, для того чтобы найти экстремум функции нескольких переменных, надо вычислить частные производные первого порядка по каждому из параметров и приравнять их к нулю.

Итак. Имеем функцию $m+1$ аргумента:

$$S(a,\ b_1,\ b_2,\ ...,\ b_m) = \sum \left(y - a - b_1 x_1 - b_2 x_2 - ... - b_m x_m\right)^2 .$$

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum \left(y - a - b_1 x_1 - b_2 x_2 - ... - b_m x_m\right) = 0; \\ \dfrac{\partial S}{\partial b_1} = -2\sum x_1 \left(y - a - b_1 x_1 - b_2 x_2 - ... - b_m x_m\right) = 0; \\ ..... \\ \dfrac{\partial S}{\partial b_m} = -2\sum x_m \left(y - a - b_1 x_1 - b_2 x_2 - ... - b_m x_m\right) = 0. \end{cases}$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (2.1):

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 + ... + b_m \sum x_m = \sum y, \\ a\sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + ... + b_m \sum x_1 x_m = \sum y x_1, \\ ..... \\ a\sum x_m + b_1 \sum x_1 x_m + b_2 \sum x_2 x_m + ... + b_m \sum x_m^2 = \sum y x_m. \end{cases} \quad (2.3)$$

Для двухфакторной модели данная система будет иметь вид:

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

Метод наименьших квадратов применим и к уравнению множественной регрессии в стандартизированном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_m t_{x_m} + \varepsilon, \quad (2.4)$$

где $t_y,\ t_{x_1},\ ...,\ t_{x_m}$ – стандартизированные переменные: $t_y = \dfrac{y - \overline{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i - \overline{x_i}}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\overline{t_y} = \overline{t_{x_i}} = 0$, а

среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; β_i – стандартизированные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии показывают, на сколько единиц изменится в среднем результат, если соответствующий фактор x_i изменится на одну единицу 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. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + \ldots + \beta_m, \end{cases} \quad (2.5)$$

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

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}. \quad (2.6)$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (2.4) к уравнению регрессии в натуральном масштабе переменных (2.1), при этом параметр a определяется как $a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \ldots - b_m\overline{x}_m$.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

На основе линейного уравнения множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m + \varepsilon \qquad (2.7)$$

могут быть найдены частные уравнения регрессии:

$$\begin{cases} y_{x_1 \cdot x_2, x_3, ..., x_m} = \hat{f}(x_1), \\ y_{x_2 \cdot x_1, x_3, ..., x_m} = \hat{f}(x_2), \\ ..... \\ y_{x_m \cdot x_1, x_2, ..., x_{m-1}} = \hat{f}(x_m), \end{cases} \qquad (2.8)$$

т.е. уравнения регрессии, которые связывают результативный признак с соответствующим фактором x_i при закреплении остальных факторов на среднем уровне. В развернутом виде систему (2.8) можно переписать в виде:

$$\begin{cases} y_{x_1 \cdot x_2, x_3, ..., x_m} = a + b_1 x_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + ... + b_m \bar{x}_m + \varepsilon, \\ y_{x_2 \cdot x_1, x_3, ..., x_m} = a + b_1 \bar{x}_1 + b_2 x_2 + b_3 \bar{x}_3 + ... + b_m \bar{x}_m + \varepsilon, \\ ..... \\ y_{x_m \cdot x_1, x_2, ..., x_{m-1}} = a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + ... + b_m x_m + \varepsilon. \end{cases}$$

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

$$\begin{cases} y_{x_1 \cdot x_2, x_3, ..., x_m} = A_1 + b_1 x_1, \\ y_{x_2 \cdot x_1, x_3, ..., x_m} = A_2 + b_2 x_2, \\ ..... \\ y_{x_m \cdot x_1, x_2, ..., x_{m-1}} = A_m + b_m x_m, \end{cases} \qquad (2.9)$$

где

$$\begin{cases} A_1 = a + b_2\overline{x}_2 + b_3\overline{x}_3 + ... + b_m\overline{x}_m, \\ A_2 = a + b_1\overline{x}_1 + b_3\overline{x}_3 + ... + b_m\overline{x}_m, \\ ..... \\ A_m = a + b_1\overline{x}_1 + b_2\overline{x}_2 + b_3\overline{x}_3 + ... + b_{m-1}x_{m-1}. \end{cases}$$

В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, ибо другие факторы закреплены на неизменном уровне. Эффекты влияния других факторов присоединены в них к свободному члену уравнения множественной регрессии. Это позволяет на основе частных уравнений регрессии определять частные коэффициенты эластичности:

$$Э_{y_{x_i}} = b_i \cdot \frac{x_i}{\hat{y}_{x_i \cdot x_1, x_2, ... x_{i-1}, x_{i+1}, ..., x_m}}, \qquad (2.10)$$

где b_i – коэффициент регрессии для фактора x_i в уравнении множественной регрессии, $\hat{y}_{x_i \cdot x_1, x_2, ... x_{i-1}, x_{i+1}, ..., x_m}$ – частное уравнение регрессии.

Наряду с частными коэффициентами эластичности могут быть найдены средние по совокупности показатели эластичности:

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_{x_i}}, \qquad (2.11)$$

которые показывают на сколько процентов в среднем изменится результат, при изменении соответствующего фактора на 1%. Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Рассмотрим **пример**[4] (для сокращения объема вычислений ограничимся только десятью наблюдениями). Пусть имеются следующие данные (условные) о сменной добыче угля на одного рабочего y (т), мощности пласта x_1 (м) и уровне механизации работ x_2 (%), характеризующие процесс добычи угля в 10 шахтах.

[4] Данные примера взяты из [5]

Таблица 2.2

№	1	2	3	4	5	6	7	8	9	10
x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

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 таблицу ($\varepsilon = y - \hat{y}_x$):

Таблица 2.3

№	x_1	x_2	y	x_1^2	x_2^2	y^2	$x_1 \cdot x_2$	$x_1 \cdot y$	$x_2 \cdot y$	$\hat{y}_x$	ε^2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	8	5	5	64	25	25	40	40	25	5,13	0,016
2	11	8	10	121	64	100	88	110	80	8,79	1,464
3	12	8	10	144	64	100	96	120	80	9,64	0,127
4	9	5	7	81	25	49	45	63	35	5,98	1,038
5	8	7	5	64	49	25	56	40	35	5,86	0,741
6	8	8	6	64	64	36	64	48	48	6,23	0,052
7	9	6	6	81	36	36	54	54	36	6,35	0,121
8	9	4	5	81	16	25	36	45	20	5,61	0,377
9	8	5	6	64	25	36	40	48	30	5,13	0,762
10	12	7	8	144	49	64	84	96	56	9,28	1,631
Сумма	94	63	68	908	417	496	603	664	445	68	6,329
Среднее значение	9,4	6,3	6,8	90,8	41,7	49,6	60,3	66,4	44,5	–	–
σ^2	2,44	2,01	3,36	–	–	–	–	–	–	–	–
σ	1,56	1,42	1,83	–	–	–	–	–	–	–	–

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

$$\begin{cases} 10a + 94b_1 + 63b_2 = 68, \\ 94a + 908b_1 + 603b_2 = 664, \\ 63a + 603b_1 + 417b_2 = 445. \end{cases}$$

Откуда получаем, что $a = -3,54$, $b_1 = 0,854$, $b_2 = 0,367$. Т.е. получили следующее уравнение множественной регрессии:

$$\hat{y}_x = -3,54 + 0,854 \cdot x_1 + 0,367 \cdot x_2 .$$

Оно показывает, что при увеличении только мощности пласта x_1 (при неизменном x_2) на 1 м добыча угля на одного рабочего y увеличится в среднем на 0,854 т, а при увеличении только уровня механизации работ x_2 (при неизменном x_1) на 1% – в среднем на 0,367 т.

Найдем уравнение множественной регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon,$$

при этом стандартизованные коэффициенты регрессии будут

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,854 \cdot \frac{1,56}{1,83} = 0,728,$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0,367 \cdot \frac{1,42}{1,83} = 0,285 .$$

Т.е. уравнение будет выглядеть следующим образом:

$$\hat{t}_y = 0,728 \cdot t_{x_1} + 0,285 \cdot t_{x_2} .$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что мощность пласта оказывает большее влияние на сменную добычу угля, чем уровень механизации работ.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности (2.11):

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_{x_i}} .$$

Вычисляем:

$$\overline{Э}_1 = 0,854 \cdot \frac{9,4}{6,8} = 1,18, \qquad \overline{Э}_2 = 0,367 \cdot \frac{6,3}{6,8} = 0,34 .$$

Т.е. увеличение только мощности пласта (от своего среднего значения) или только уровня механизации работ на 1% увеличивает в среднем сменную

добычу угля на 1,18% или 0,34% соответственно. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

2.3. Проверка существенности факторов и показатели качества регрессии

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата – показателя детерминации.

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком или, иначе, оценивает тесноту совместного влияния факторов на результат.

Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}, \qquad (2.12)$$

где σ^2_y – общая дисперсия результативного признака; $\sigma^2_{\text{ост}}$ – остаточная дисперсия.

Границы изменения индекса множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx_1x_2\ldots x_m} \geq r_{yx_i(\max)} \quad \left(i=\overline{1,m}\right).$$

При правильном включении факторов в регрессионную модель величина индекса множественной корреляции будет существенно отличаться от индекса корреляции парной зависимости. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны,

то индекс множественной корреляции может практически совпадать с индексом парной корреляции (различия в третьем, четвертом знаках). Отсюда ясно, что сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Расчет индекса множественной корреляции предполагает определение уравнения множественной регрессии и на его основе остаточной дисперсии:

$$\sigma^2_{\text{ост}} = \frac{1}{n}\sum\left(y - \hat{y}_{x_1 x_2 \ldots x_m}\right)^2. \qquad (2.13)$$

Можно пользоваться следующей формулой индекса множественной детерминации:

$$R^2_{yx_1x_2\ldots x_m} = 1 - \frac{\sum\left(y - \hat{y}_{x_1x_2\ldots x_m}\right)^2}{\sum\left(y - \overline{y}\right)^2}. \qquad (2.14)$$

При линейной зависимости признаков формула индекса множественной корреляции может быть представлена следующим выражением:

$$R_{yx_1x_2\ldots x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}, \qquad (2.15)$$

где β_i – стандартизованные коэффициенты регрессии; r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Формула индекса множественной корреляции для линейной регрессии получила название *линейного коэффициента множественной корреляции*, или, что то же самое, *совокупного коэффициента корреляции*.

Возможно также при линейной зависимости определение совокупного коэффициента корреляции через матрицу парных коэффициентов корреляции:

$$R_{yx_1x_2,\ldots,x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}, \qquad (2.16)$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_p} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_p} \\ r_{yx_2} & r_{x_2x_1} & 1 & \dots & r_{x_2x_p} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_p} & r_{x_px_1} & r_{x_px_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_p} \\ \dots & \dots & \dots & \dots \\ r_{x_px_1} & r_{x_px_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Как видим, величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от межфакторной корреляции. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции.

В рассмотренных показателях множественной корреляции (индекс и коэффициент) используется остаточная дисперсия, которая имеет систематическую ошибку в сторону преуменьшения, тем более значительную, чем больше параметров определяется в уравнении регрессии при заданном объеме наблюдений n. Если число параметров при x_i равно m и приближается к объему наблюдений, то остаточная дисперсия будет близка к нулю и коэффициент (индекс) корреляции приблизится к единице даже при слабой связи факторов с результатом. Для того чтобы не допустить возможного преувеличения тесноты связи, используется скорректированный индекс (коэффициент) множественной корреляции.

Скорректированный индекс множественной корреляции содержит поправку на число степеней свободы, а именно остаточная сумма квадратов $\sum\left(y-\hat{y}_{x_1x_2\ldots x_m}\right)^2$ делится на число степеней свободы остаточной вариации $\left(n-m-1\right)$, а общая сумма квадратов отклонений $\sum\left(y-\overline{y}\right)^2$ на число степеней свободы в целом по совокупности $\left(n-1\right)$.

Формула скорректированного индекса множественной детерминации имеет вид:

$$\widehat{R}^2=1-\frac{\sum\left(y-\hat{y}\right)^2\Big/\left(n-m-1\right)}{\sum\left(y-\overline{y}\right)\Big/\left(n-1\right)}, \qquad (2.17)$$

где m – число параметров при переменных x; n – число наблюдений.

Поскольку $\dfrac{\sum\left(y-\hat{y}_{x_1x_2\ldots x_m}\right)^2}{\sum\left(y-\overline{y}\right)^2}=1-R^2$, то величину скорректированного индекса детерминации можно представить в виде:

$$\widehat{R}^2=1-\left(1-R^2\right)\cdot\frac{n-1}{n-m-1}. \qquad (2.17а)$$

Чем больше величина m, тем сильнее различия $\widehat{R}^2$ и R^2.

Как было показано выше, ранжирование факторов, участвующих во множественной линейной регрессии, может быть проведено через стандартизованные коэффициенты регрессии (β-коэффициенты). Эта же цель может быть достигнута с помощью частных коэффициентов корреляции (для линейных связей). Кроме того, частные показатели корреляции широко используются при решении проблемы отбора факторов: целесообразность включения того или иного фактора в модель можно доказать величиной показателя частной корреляции.

Частные коэффициенты корреляции характеризуют 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

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

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

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 \ldots x_i \ldots x_m}}{1 - R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}}}, \qquad (2.18)$$

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

При двух факторах формула (2.18) примет вид:

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_2}}}; \qquad r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_1}}}. \qquad (2.18а)$$

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:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right) \cdot \left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}\right)}}. \quad (2.19)$$

При двух факторах данная формула примет вид:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}; \; r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}. \quad (2.19а)$$

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции первого порядка. Так, по уравнению $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$ возможно исчисление трех частных коэффициентов корреляции второго порядка:

$$r_{yx_1 \cdot x_2 x_3}, \; r_{yx_2 \cdot x_1 x_3}, \; r_{yx_3 \cdot x_1 x_2},$$

каждый из которых определяется по рекуррентной формуле. Например, при $i = 1$ имеем формулу для расчета $r_{yx_1 \cdot x_2 x_3}$:

$$r_{yx_1 \cdot x_2 x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_3 \cdot x_2} \cdot r_{x_1 x_3 \cdot x_2}}{\sqrt{\left(1 - r^2_{yx_3 \cdot x_2}\right)\left(1 - r^2_{x_1 x_3 \cdot x_2}\right)}}. \quad (2.20)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде. Если из стандартизованного уравнения регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \beta_3 t_{x_3} + \varepsilon$ следует, что $\beta_1 > \beta_2 > \beta_3$, т.е. по силе влияния на результат порядок факторов таков: x_1, x_2, x_3, то этот же порядок факторов определяется и по соотношению частных коэффициентов корреляции, $r_{yx_1 \cdot x_2 x_3} > r_{yx_2 \cdot x_1 x_3} > r_{yx_3 \cdot x_1 x_2}$.

В эконометрике частные коэффициенты корреляции обычно не имеют самостоятельного значения. Их используют на стадии формирования модели. Так, строя многофакторную модель, на первом шаге определяется уравнение регрессии с полным набором факторов и рассчитывается матрица частных коэффициентов корреляции. На втором шаге отбирается фактор с наименьшей и несущественной по t-критерию Стьюдента величиной показателя частной корреляции. Исключив его из модели, строится новое уравнение регрессии. Процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции существенно отличаются от нуля. Если исключен несущественный фактор, то множественные коэффициенты детерминации на двух смежных шагах построения регрессионной модели почти не отличаются друг от друга, $R^2_{m+1} \approx R^2_m$, где m – число факторов.

Из приведенных выше формул частных коэффициентов корреляции видна связь этих показателей с совокупным коэффициентом корреляции. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить совокупный коэффициент корреляции по формуле:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\left(1-r^2_{yx_1}\right)\cdot\left(1-r^2_{yx_2\cdot x_1}\right)\cdot\left(1-r^2_{yx_3\cdot x_1x_2}\right)\cdot\ldots\cdot\left(1-r^2_{yx_m\cdot x_1x_2\ldots x_{m-1}}\right)}. \quad (2.21)$$

В частности, для двухфакторного уравнения формула (2.21) принимает вид:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\left(1-r^2_{yx_1}\right)\cdot\left(1-r^2_{yx_2\cdot x_1}\right)}. \quad (2.21)$$

При полной зависимости результативного признака от исследуемых факторов коэффициент совокупного их влияния равен единице. Из единицы вычитается доля остаточной вариации результативного признака $\left(1-r^2\right)$, обусловленная последовательно включенными в анализ факторами. В результате подкоренное выражение характеризует совокупное действие всех исследуемых факторов.

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью F-критерия Фишера:

$$F = \frac{S_{\text{факт}}}{S_{\text{ост}}} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}, \qquad (2.22)$$

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

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, т.е. F_{x_i}.

Частный F-критерий построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. В общем виде для фактора x_i частный F-критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_m} - R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}}{1 - R^2_{yx_1 \ldots x_i \ldots x_m}} \cdot \frac{n-m-1}{1}, \qquad (2.23)$$

где $R^2_{yx_1 \ldots x_i \ldots x_m}$ – 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).

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы: 1 и $n-m-1$. Если фактическое значение F_{x_i} превышает $F_{\text{табл}}(\alpha, k_1, k_2)$, 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.

Для двухфакторного уравнения частные F-критерии имеют вид:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3), \quad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3). \quad (2.23а)$$

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.

Частный F-критерий оценивает значимость коэффициентов чистой регрессии. Зная величину F_{x_i}, можно определить и t-критерий для коэффициента регрессии при i-м факторе, t_{b_i}, а именно:

$$t_{b_i} = \sqrt{F_{x_i}}. \quad (2.24)$$

Оценка значимости коэффициентов чистой регрессии по t-критерию Стьюдента может быть проведена и без расчета частных F-критериев. В этом случае, как и в парной регрессии, для каждого фактора используется формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}}, \qquad (2.25)$$

где b_i – коэффициент чистой регрессии при факторе x_i, m_{b_i} – средняя квадратическая (стандартная) ошибка коэффициента регрессии b_i.

Для уравнения множественной регрессии $\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$ средняя квадратическая ошибка коэффициента регрессии может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y \sqrt{1 - R^2_{yx_1 \ldots x_m}}}{\sigma_{x_i} \sqrt{1 - R^2_{x_i x_1 \ldots x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}}, \qquad (2.26)$$

где σ_y – среднее квадратическое отклонение для признака y, σ_{x_i} – среднее квадратическое отклонение для признака x_i, $R^2_{yx_1 \ldots x_m}$ – коэффициент детерминации для уравнения множественной регрессии, $R^2_{x_i x_1 \ldots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии; $n - m - 1$ – число степеней свободы для остаточной суммы квадратов отклонений.

Как видим, чтобы воспользоваться данной формулой, необходимы матрица межфакторной корреляции и расчет по ней соответствующих коэффициентов детерминации $R^2_{x_i x_1 \ldots x_m}$. Так, для уравнения $\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$ оценка значимости коэффициентов регрессии b_1, b_2, b_3 предполагает расчет трех межфакторных коэффициентов детерминации: $R^2_{x_1 \cdot x_2 x_3}$, $R^2_{x_2 \cdot x_1 x_3}$, $R^2_{x_3 \cdot x_1 x_2}$.

Взаимосвязь показателей частного коэффициента корреляции, частного F-критерия и t-критерия Стьюдента для коэффициентов чистой регрессии может использоваться в процедуре отбора факторов. Отсев факторов при построении уравнения регрессии методом исключения практически можно

осуществлять не только по частным коэффициентам корреляции, исключая на каждом шаге фактор с наименьшим незначимым значением частного коэффициента корреляции, но и по величинам t_{b_i} и F_{x_i}. Частный F-критерий широко используется и при построении модели методом включения переменных и шаговым регрессионным методом.

Пример. Оценим качество уравнения, полученного в предыдущем параграфе. Сначала найдем значения парных коэффициентов корреляции:

$$r_{yx_1} = \frac{\overline{y \cdot x_1} - \overline{y} \cdot \overline{x}_1}{\sigma_y \cdot \sigma_{x_1}} = \frac{66,4 - 6,8 \cdot 9,4}{1,83 \cdot 1,56} = 0,869;$$

$$r_{yx_2} = \frac{\overline{y \cdot x_2} - \overline{y} \cdot \overline{x}_2}{\sigma_y \cdot \sigma_{x_2}} = \frac{44,5 - 6,8 \cdot 6,3}{1,83 \cdot 1,42} = 0,639;$$

$$r_{x_1x_2} = \frac{\overline{x_1 \cdot x_2} - \overline{x}_1 \cdot \overline{x}_2}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{60,3 - 9,4 \cdot 6,3}{1,56 \cdot 1,42} = 0,488.$$

Значения парных коэффициентов корреляции указывают на достаточно тесную связь сменной добычи угля на одного рабочего y с мощностью пласта x_1 и на умеренную связь с уровнем механизации работ x_2. В то же время межфакторная связь $r_{x_1x_2}$ не очень сильная ($r_{x_1x_2} = 0,49 < 0,7$), что говорит о том, что оба фактора являются информативными, т.е. и x_1, и x_2 необходимо включить в модель.

Теперь рассчитаем совокупный коэффициент корреляции $R_{yx_1x_2}$. Для этого сначала найдем определитель матрицы парных коэффициентов корреляции:

$$\Delta r = \begin{vmatrix} 1 & 0,87 & 0,64 \\ 0,87 & 1 & 0,49 \\ 0,64 & 0,49 & 1 \end{vmatrix} = 0,139064,$$

и определитель матрицы межфакторной корреляции:

$$\Delta r_{11} = \begin{vmatrix} 1 & 0,49 \\ 0,49 & 1 \end{vmatrix} = 0,7599.$$

Тогда коэффициент множественной корреляции по формуле (2.16):

$$R_{yx_1x_2} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}} = \sqrt{1 - \frac{0,139064}{0,7599}} = 0,904.$$

Т.е. можно сказать, что 81,7% (коэффициент детерминации $R^2_{yx_1x_2} = 0,817$) вариации результата объясняется вариацией представленных в уравнении признаков, что указывает на весьма тесную связь признаков с результатом.

Примерно тот же результат (различия связаны с ошибками округлений) для коэффициента множественной регрессии получим, если воспользуемся формулами (2.12) и (2.15):

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{0,6329}{3,36}} = 0,901;$$

$$R_{yx_1x_2} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = \sqrt{0,728 \cdot 0,87 + 0,285 \cdot 0,64} = 0,903.$$

Скорректированный коэффициент множественной детерминации

$$\widehat{R} = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-m-1} = 1 - \left(1 - 0,817\right) \cdot \frac{10-1}{10-2-1} = 0,765$$

указывает на умеренную связь между результатом и признаками. Это связано с малым количеством наблюдений.

Теперь найдем частные коэффициенты корреляции по формулам (2.18а) и (2.19а):

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}} = \sqrt{1 - \frac{1 - 0,817}{1 - 0,408}} = 0,831;$$

$$r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}} = \sqrt{1 - \frac{1 - 0,817}{1 - 0,755}} = 0,503.$$

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0,869 - 0,639 \cdot 0,488}{\sqrt{\left(1 - 0,489^2\right)\left(1 - 0,639^2\right)}} = 0,830;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0,639 - 0,869 \cdot 0,488}{\sqrt{\left(1 - 0,488^2\right)\left(1 - 0,869^2\right)}} = 0,498.$$

Т.е. можно сделать вывод, что фактор x_1 оказывает более сильное влияние на результат, чем признак x_2.

Оценим надежность уравнения регрессии в целом и показателя связи с помощью F-критерия Фишера. Фактическое значение F-критерия (2.22)

$$F_{\text{факт}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0,817}{1 - 0,817} \cdot \frac{10 - 2 - 1}{2} = 15,63.$$

Табличное значение F-критерия при пятипроцентном уровне значимости ($\alpha = 0,05$, $k_1 = 2$, $k_2 = 10 - 2 - 1 = 7$): $F_{\text{табл}} = 4,74$. Так как $F_{\text{факт}} = 15,63 > F_{\text{табл}} = 4,10$, то уравнение признается статистически значимым.

Оценим целесообразность включения фактора x_1 после фактора x_2 и x_2 после x_1 с помощью частного F-критерия Фишера (2.23а):

$$F_{x_1} = \frac{R_{yx_1x_2}^2 - r_{yx_2}^2}{1 - R_{yx_1x_2}^2} \cdot (n - 3) = \frac{0,817 - 0,408}{1 - 0,817} \cdot 7 = 15,65;$$

$$F_{x_2} = \frac{R_{yx_1x_2}^2 - r_{yx_1}^2}{1 - R_{yx_1x_2}^2} \cdot (n - 3) = \frac{0,817 - 0,755}{1 - 0,817} \cdot 7 = 2,37.$$

Табличное значение частного F-критерия при пятипроцентном уровне значимости ($\alpha = 0,05$, $k_1 = 1$, $k_2 = 10 - 2 - 1 = 7$): $F_{\text{табл}} = 5,59$. Так как $F_{x_1} = 15,65 > F_{\text{табл}} = 5,59$, а $F_{x_2} = 2,37 < F_{\text{табл}} = 5,59$, то включение фактора x_1 в модель статистически оправдано и коэффициент чистой

регрессии b_1 статистически значим, а дополнительное включение фактора x_2, после того, как уже введен фактор x_1, нецелесообразно.

Уравнение регрессии, включающее только один значимый аргумент x_2:

$$\hat{y} = -2,754 + 1,016x_1.$$

2.4. Линейные регрессионные модели с гетероскедастичными остатками

При оценке параметров уравнения регрессии применяется метод наименьших квадратов (МНК). При этом делаются определенные предпосылки относительно случайной составляющей ε. В модели

$$y = a + b_1x_1 + b_2x_2 + ... + b_mx_m + \varepsilon$$

случайная составляющая ε представляет собой ненаблюдаемую величину. После того как произведена оценка параметров модели, рассчитывая разности фактических и теоретических значений результативного признака y, можно определить оценки случайной составляющей $y - \hat{y}_x$. Поскольку они не являются реальными случайными остатками, их можно считать некоторой выборочной реализацией неизвестного остатка заданного уравнения, т.е. ε_i.

При изменении спецификации модели, добавлении в нее новых наблюдений выборочные оценки остатков ε_i могут меняться. Поэтому в задачу регрессионного анализа входит не только построение самой модели, но и исследование случайных отклонений ε_i, т.е. остаточных величин.

При использовании критериев Фишера и Стьюдента делаются предположения относительно поведения остатков ε_i – остатки представляют собой независимые случайные величины и их среднее значение равно 0; они

имеют одинаковую (постоянную) дисперсию и подчиняются нормальному распределению.

Статистические проверки параметров регрессии, показателей корреляции основаны на непроверяемых предпосылках распределения случайной составляющей ε_i. Они носят лишь предварительный характер. После построения уравнения регрессии проводится проверка наличия у оценок ε_i (случайных остатков) тех свойств, которые предполагались. Связано это с тем, что оценки параметров регрессии должны отвечать определенным критериям. Они должны быть несмещенными, состоятельными и эффективными. Эти свойства оценок, полученных по МНК, имеют чрезвычайно важное практическое значение в использовании результатов регрессии и корреляции.

Несмещенность оценки означает, что математическое ожидание остатков равно нулю. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным исследованиям.

Оценки считаются *эффективными*, если они характеризуются наименьшей дисперсией. В практических исследованиях это означает возможность перехода от точечного оценивания к интервальному.

Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки. Большой практический интерес представляют те результаты регрессии, для которых доверительный интервал ожидаемого значения параметра регрессии b_i имеет предел значений вероятности, равный единице. Иными словами, вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.

Указанные критерии оценок (несмещенность, состоятельность и эффективность) обязательно учитываются при разных способах оценивания. Метод наименьших квадратов строит оценки регрессии на основе минимизации суммы квадратов остатков. Поэтому очень важно исследовать поведение остаточных величин регрессии ε_i. Условия, необходимые для

получения несмещенных, состоятельных и эффективных оценок, представляют собой предпосылки МНК, соблюдение которых желательно для получения достоверных результатов регрессии.

Исследования остатков ε_i предполагают проверку наличия следующих пяти предпосылок МНК:

1) случайный характер остатков;

2) нулевая средняя величина остатков, не зависящая от x_i;

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.

Прежде всего, проверяется случайный характер остатков ε_i – первая предпосылка МНК. С этой целью стоится график зависимости остатков ε_i от теоретических значений результативного признака (рис. 2.1). Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и МНК оправдан, теоретические значения $\hat{y}_x$ хорошо аппроксимируют фактические значения y.

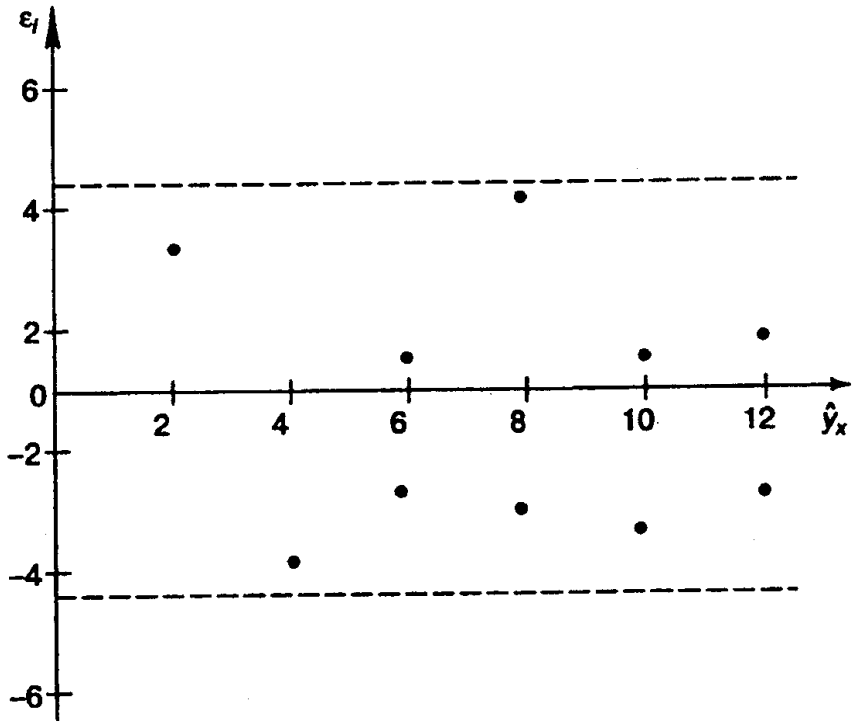


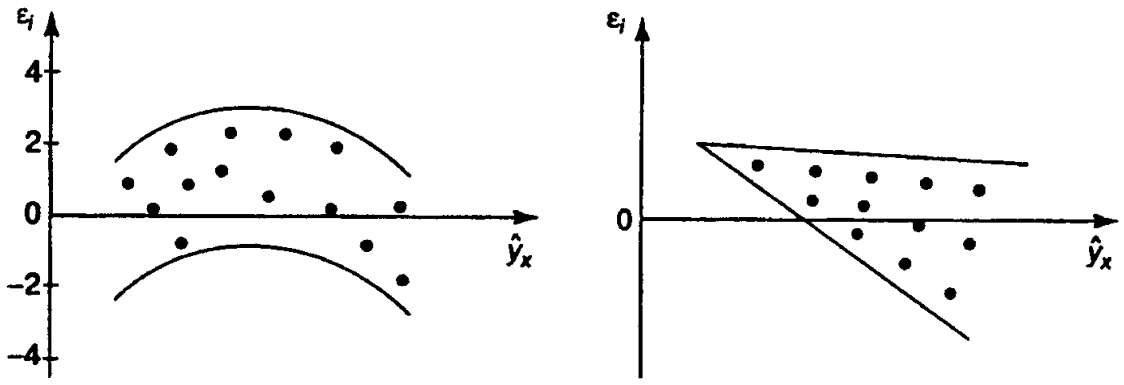
Рис. 2.1. Зависимость случайных остатков ε_i от теоретических значений $\hat{y}_x$.

Возможны следующие случаи, если ε_i зависит от $\hat{y}_x$ то:

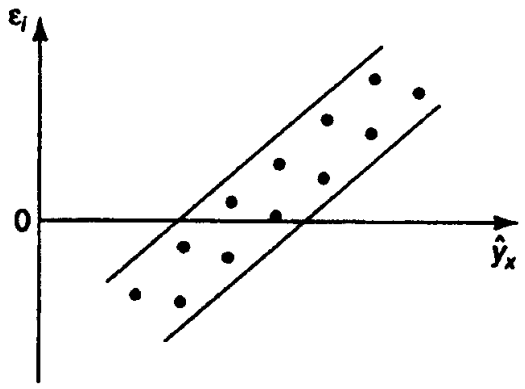
1) остатки ε_i не случайны (рис. 2.2а);

2) остатки ε_i не имеют постоянной дисперсии (рис. 2.2б);

3) остатки ε_i носят систематический характер (рис. 2.2в).



а **б**



в

Рис. 2.2. Зависимость случайных остатков ε_i от теоретических значений $\hat{y}_x$.

В этих случаях необходимо либо применять другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки ε_i не будут случайными величинами.

Вторая предпосылка МНК относительно нулевой средней величины остатков означает, что $\sum\left(y-\hat{y}_x\right)=0$. Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных.

Вместе с тем, несмещенность оценок коэффициентов регрессии, полученных МНК, зависит от независимости случайных остатков и величин x, что также исследуется в рамках соблюдения второй предпосылки МНК. С этой целью наряду с изложенным графиком зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$ строится график зависимости случайных остатков ε_i от факторов, включенных в регрессию x_j (рис. 2.3).

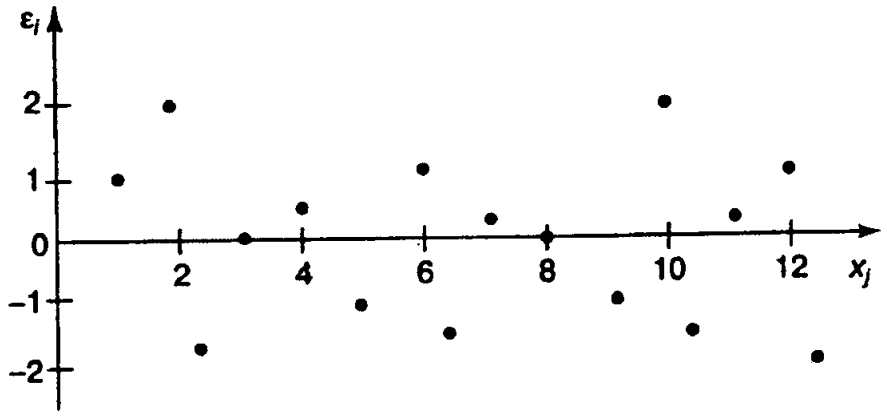


Рис. 2.3. Зависимость величины остатков от величины фактора x_j.

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 неадекватна. Причины неадекватности могут быть разные. Возможно, что нарушена третья предпосылка МНК и дисперсия остатков не постоянна для каждого значения фактора x_j. Может быть неправильна спецификация модели и в нее необходимо ввести дополнительные члены от x_j, например x_j^2. Скопление точек в определенных участках значений фактора x_j говорит о наличии систематической погрешности модели.

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.

Совершенно необходимым для получения по МНК состоятельных оценок параметров регрессии является соблюдение третьей и четвертой предпосылок.

В соответствии с третьей предпосылкой МНК требуется, чтобы дисперсия остатков была *гомоскедастичной*. Это значит, что для каждого

значения фактора x_j остатки ε_i имеют одинаковую дисперсию. Если это условие применения МНК не соблюдается, то имеет место *гетероскедастичность*. Наличие гетероскедастичности можно наглядно видеть из поля корреляции (рис. 2.4).

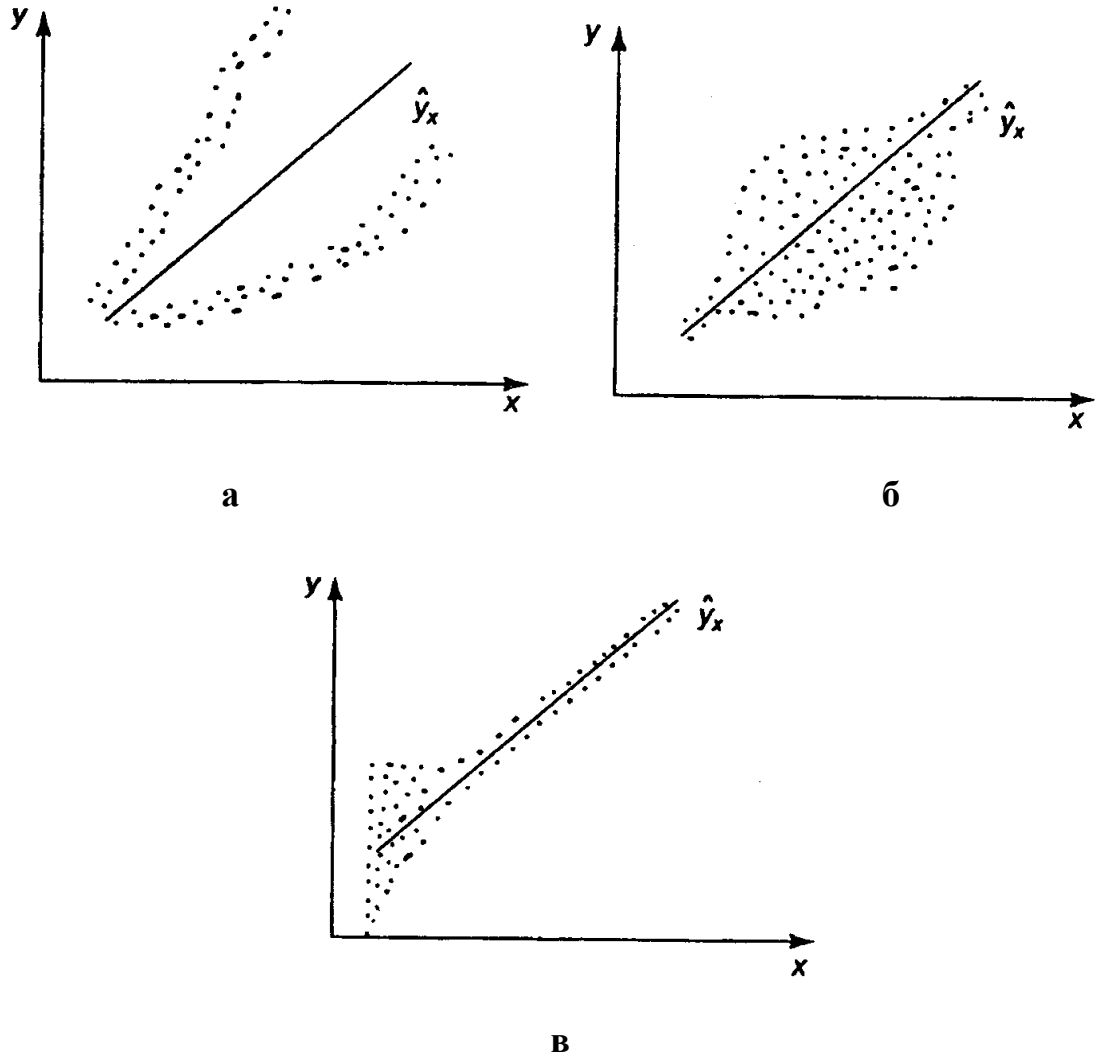


Рис. 2.4. Примеры гетероскедастичности.

На рис. 2.4 изображено: а – дисперсия остатков растет по мере увеличения x; б – дисперсия остатков достигает максимальной величины при средних значениях переменной x и уменьшается при минимальных и максимальных значениях x; в – максимальная дисперсия остатков при малых значениях x и дисперсия остатков однородна по мере увеличения значений x.

Наличие гомоскедастичности или гетероскедастичности можно видеть и по рассмотренному выше графику зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$. Так, для рис. 2.4а зависимость остатков от $\hat{y}_x$ представлена на рис. 2.5.

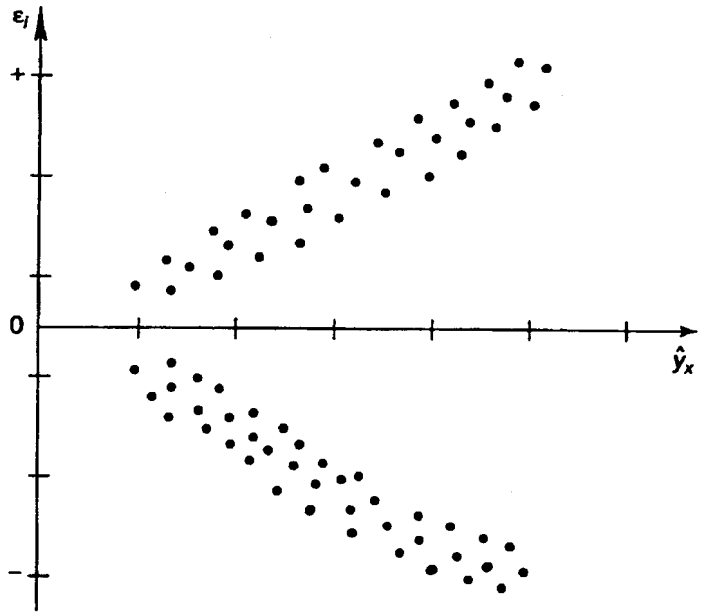


Рис. 2.5. Гетероскедастичность: большая дисперсия ε_i для больших значений $\hat{y}_x$.

Соответственно для зависимости, изображенной на полях корреляции рис. 2.4б и 2.4в гетероскедастичность остатков представлена на рис. 2.6 и 2.7.

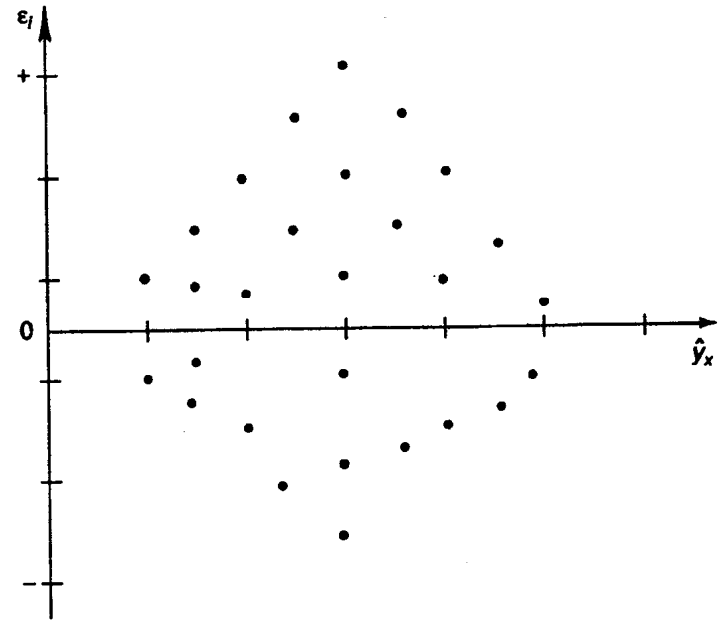


Рис. 2.6. Гетероскедастичность, соответствующая полю корреляции на рис. 2.4б.

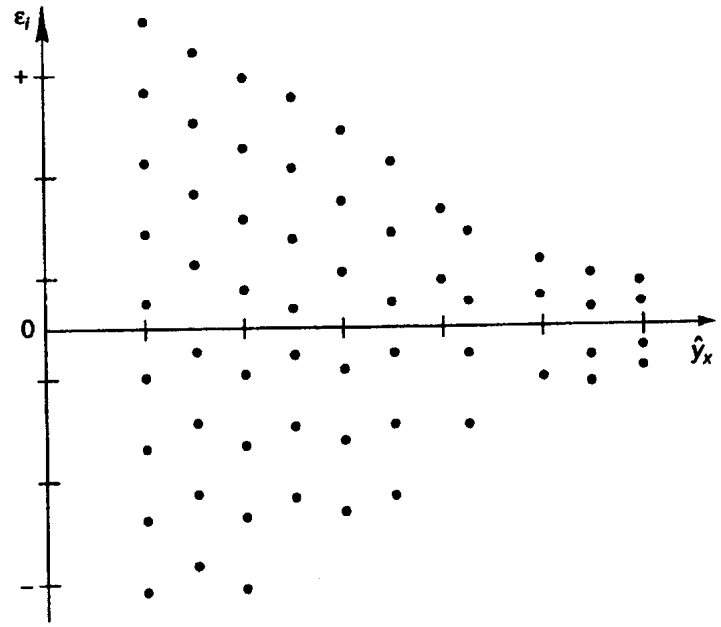


Рис. 2.7. Гетероскедастичность, соответствующая полю корреляции на рис. 2.4в.

Для множественной регрессии данный вид графиков является наиболее приемлемым визуальным способом изучения гомо- и гетероскедастичности.

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[5]. Коэффициент корреляции между ε_i и ε_j, где ε_i – остатки текущих наблюдений, ε_j – остатки предыдущих наблюдений (например, $j = i - 1$), может быть определен как

$$r_{\varepsilon_i \varepsilon_j} = \frac{\operatorname{cov}\left(\varepsilon_i,\ \varepsilon_j\right)}{\sigma_{\varepsilon_i} \cdot \sigma_{\varepsilon_j}},$$

т.е. по обычной формуле линейного коэффициента корреляции. Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированы и функция плотности вероятности $F\left(\varepsilon\right)$ зависит от j-й

[5] Подробнее об автокорреляции см. в разделе 4.

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.

При несоблюдении основных предпосылок МНК приходится корректировать модель, изменяя ее спецификацию, добавлять (исключать) некоторые факторы, преобразовывать исходные данные для того, чтобы получить оценки коэффициентов регрессии, которые обладают свойством несмещенности, имеют меньшее значение дисперсии остатков и обеспечивают в связи с этим более эффективную статистическую проверку значимости параметров регрессии.

2.5. Обобщенный метод наименьших квадратов (ОМНК)

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).

Обобщенный метод наименьших квадратов применяется к преобразованным данным и позволяет получать оценки, которые обладают не только свойством несмещенности, но и имеют меньшие выборочные дисперсии. Остановимся на использовании ОМНК для корректировки гетероскедастичности.

Как и раньше, будем предполагать, что среднее значение остаточных величин равно нулю. А вот дисперсия их не остается неизменной для разных значений фактора, а пропорциональна величине K_i, т.е.

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

где $\sigma_{\varepsilon_i}^2$ – дисперсия ошибки при конкретном i-м значении фактора; σ^2 – постоянная дисперсия ошибки при соблюдении предпосылки о гомоскедастичности остатков; K_i – коэффициент пропорциональности, меняющийся с изменением величины фактора, что и обусловливает неоднородность дисперсии.

При этом предполагается, что σ^2 неизвестна, а в отношении величин K_i выдвигаются определенные гипотезы, характеризующие структуру гетероскедастичности.

В общем виде для уравнения $y_i = a + bx_i + \varepsilon_i$ при $\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i$ модель примет вид: $y_i = a + bx_i + \sqrt{K_i}\varepsilon_i$. В ней остаточные величины гетероскедастичны. Предполагая в них отсутствие автокорреляции, можно перейти к уравнению с гомоскедастичными остатками, поделив все переменные, зафиксированные в ходе i-го наблюдения, на $\sqrt{K_i}$. Тогда дисперсия остатков будет величиной постоянной, т. е. $\sigma_{\varepsilon_i}^2 = \sigma^2$.

Иными словами, от регрессии y по x мы перейдем к регрессии на новых переменных: $y/\sqrt{K}$ и $x/\sqrt{K}$. Уравнение регрессии примет вид:

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

а исходные данные для данного уравнения будут иметь вид:

$$y = \begin{vmatrix} \frac{y_1}{\sqrt{K_1}} \\ \frac{y_2}{\sqrt{K_2}} \\ \cdots\cdots \\ \frac{y_n}{\sqrt{K_n}} \end{vmatrix}, \qquad x = \begin{vmatrix} \frac{x_1}{\sqrt{K_1}} \\ \frac{x_2}{\sqrt{K_2}} \\ \cdots\cdots \\ \frac{x_n}{\sqrt{K_n}} \end{vmatrix}.$$

По отношению к обычной регрессии уравнение с новыми, преобразованными переменными представляет собой взвешенную регрессию, в которой переменные y и x взяты с весами $1/\sqrt{K}$.

Оценка параметров нового уравнения с преобразованными переменными приводит к взвешенному методу наименьших квадратов, для которого необходимо минимизировать сумму квадратов отклонений вида

$$S(a, b) = \sum_{i=1}^{n} \frac{1}{K_i} (y_i - a - bx_i)^2.$$

Соответственно получим следующую систему нормальных уравнений:

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

Если преобразованные переменные x и y взять в отклонениях от средних уровней, то коэффициент регрессии b можно определить как

$$b = \frac{\sum \frac{1}{K} \cdot x \cdot y}{\sum \frac{1}{K} \cdot x^2}.$$

При обычном применении метода наименьших квадратов к уравнению линейной регрессии для переменных в отклонениях от средних уровней коэффициент регрессии b определяется по формуле:

$$b = \frac{\sum x \cdot y}{\sum x^2}.$$

Как видим, при использовании обобщенного МНК с целью корректировки гетероскедастичности коэффициент регрессии b представляет собой взвешенную величину по отношению к обычному МНК с весом $1/K$.

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

$$y = a + b_1 x_1 + b_2 x_2 + \varepsilon,$$

для которой дисперсия остаточных величин оказалась пропорциональна K_i^2. K_i представляет собой коэффициент пропорциональности, принимающий различные значения для соответствующих i значений факторов x_1 и x_2. Ввиду того, что

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

рассматриваемая модель примет вид

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + K_i \varepsilon_i,$$

где ошибки гетероскедастичны.

Для того чтобы получить уравнение, где остатки ε_i гомоскедастичны, перейдем к новым преобразованным переменным, разделив все члены исходного уравнения на коэффициент пропорциональности K. Уравнение с преобразованными переменными составит

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

Это уравнение не содержит свободного члена. Вместе с тем, найдя переменные в новом преобразованном виде и применяя обычный МНК к ним, получим иную спецификацию модели:

$$\frac{y_i}{K_i} = A + b_1 \frac{x_{1i}}{K_i} + b_2 \frac{x_{2i}}{K_i} + \varepsilon_i .$$

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

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

предположить, что $e = \varepsilon \cdot x_1$, т.е. $K = x_1$ и $\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot x_1$, то обобщенный МНК предполагает оценку параметров следующего трансформированного уравнения:

$$\frac{y}{x_1} = b_1 + b_2 \frac{x_2}{x_1} + \ldots + b_m \frac{x_m}{x_1} + \varepsilon .$$

Применение в этом случае обобщенного МНК приводит к тому, что наблюдения с меньшими значениями преобразованных переменных x/K имеют при определении параметров регрессии относительно больший вес, чем с первоначальными переменными. 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. Пусть y – издержки производства, x_1 – объем продукции, x_2 – основные производственные фонды, x_3 – численность работников, тогда уравнение

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$$

является моделью издержек производства с объемными факторами. Предполагая, что $\sigma_{\varepsilon_i}^2$ пропорциональна квадрату численности работников x_3, 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

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

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 фовдовооруженности труда; и с изменением фондовооруженности труда на единицу при неизменном уровне производительности труда.

Если предположить, что в модели с первоначальными переменными дисперсия остатков пропорциональна квадрату объема продукции, $\sigma^2_{\varepsilon_i} = \sigma^2 \cdot x_1^2$, можно перейти к уравнению регрессии вида

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

В нем новые переменные: y/x_1 – затраты на единицу (или на 1 руб. продукции), x_2/x_1 – фондоемкость продукции, x_3/x_1 – трудоемкость продукции.

Гипотеза о пропорциональности остатков величине фактора может иметь реальное основание: при обработке недостаточно однородной совокупности, включающей как крупные, так и мелкие предприятия, большим объемным значениям фактора может соответствовать большая дисперсия результативного признака и большая дисперсия остаточных величин.

При наличии одной объясняющей переменной гипотеза $\sigma_{\varepsilon_i}^2 = \sigma^2 x^2$ трансформирует линейное уравнение

$$y = a + bx + e$$

в уравнение

$$\frac{y}{x} = b + \frac{a}{x} + \varepsilon,$$

в котором параметры a и b поменялись местами, константа стала коэффициентом наклона линии регрессии, а коэффициент регрессии – свободным членом.

Пример. Рассматривая зависимость сбережений y от дохода x, по первоначальным данным было получено уравнение регрессии

$$y = -1{,}081 + 0{,}1178 \cdot x.$$

Применяя обобщенный МНК к данной модели в предположении, что ошибки пропорциональны доходу, было получено уравнение для преобразованных данных:

$$\frac{y}{x} = 0{,}1026 - 0{,}8538 \cdot \frac{1}{x}.$$

Коэффициент регрессии первого уравнения сравнивают со свободным членом второго уравнения, т.е. 0,1178 и 0,1026 – оценки параметра b зависимости сбережений от дохода.

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 обобщенного МНК. 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 обобщенного МНК позволяет получить оценки параметров модели, обладающие меньшей дисперсией.

2.6. Регрессионные модели с переменной структурой (фиктивные переменные)

До сих пор в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель фактор, имеющий два или более качественных уровней. Это могут быть разного рода атрибутивные признаки, такие, например, как профессия, пол, образование, климатические условия, принадлежность к определенному региону. Чтобы ввести такие переменные 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*.

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

$$y = a + bx + \varepsilon,$$

где y – количество потребляемого кофе; x – цена.

Аналогичные уравнения могут быть найдены отдельно для лиц мужского пола: $y_1 = a_1 + b_1 x_1 + \varepsilon_1$ и женского пола: $y_2 = a_2 + b_2 x_2 + \varepsilon_2$.

Различия в потреблении кофе проявятся в различии средних $\overline{y}_1$ и $\overline{y}_2$. Вместе с тем сила влияния x на y может быть одинаковой, т.е. $b \approx b_1 \approx b_2$. В этом случае возможно построение общего уравнения регрессии с включением в него фактора «пол» в виде фиктивной переменной. Объединяя

уравнения y_1 и y_2 и, вводя фиктивные переменные, можно прийти к следующему выражению:

$$y = a_1 z_1 + a_2 z_2 + bx + \varepsilon,$$

где z_1 и z_2 – фиктивные переменные, принимающие значения:

$$z_1 = \begin{cases} 1 - \text{мужской пол}, \\ 0 - \text{женский пол}; \end{cases} \qquad z_2 = \begin{cases} 0 - \text{мужской пол}, \\ 1 - \text{женский пол}. \end{cases}$$

В общем уравнении регрессии зависимая переменная y рассматривается как функция не только цены x но и пола (z_1, z_2). Переменная z рассматривается как дихотомическая переменная, принимающая всего два значения: 1 и 0. При этом когда $z_1 = 1$, то $z_2 = 0$, и наоборот.

Для лиц мужского пола, когда $z_1 = 1$ и $z_2 = 0$, объединенное уравнение регрессии составит: $\hat{y} = a_1 + bx$, а для лиц женского пола, когда $z_1 = 0$ и $z_2 = 1$: $\hat{y} = a_2 + bx$. Иными словами, различия в потреблении для лиц мужского и женского пола вызваны различиями свободных членов уравнения регрессии: $a_1 \neq a_2$. Параметр b является общим для всей совокупности лиц, как для мужчин, так и для женщин.

Однако при введении двух фиктивных переменных z_1 и z_2 в модель $y = a_1 z_1 + a_2 z_2 + bx + \varepsilon$ применение МНК для оценивания параметров a_1 и a_2 приведет к вырожденной матрице исходных данных, а следовательно, и к невозможности получения их оценок. Объясняется это тем, что при использовании МНК в данном уравнении появляется свободный член, т.е. уравнение примет вид

$$y = A + a_1 z_1 + a_2 z_2 + bx + \varepsilon.$$

Предполагая при параметре A независимую переменную, равную 1, имеем следующую матрицу исходных данных:

$$\begin{bmatrix} 1 & 1 & 0 & x_1 \\ 1 & 1 & 0 & x_2 \\ 1 & 0 & 1 & x_3 \\ 1 & 1 & 0 & x_4 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & x_n \end{bmatrix}.$$

В рассматриваемой матрице существует линейная зависимость между первым, вторым и третьим столбцами: первый равен сумме второго и третьего столбцов. Поэтому матрица исходных факторов вырождена. Выходом из создавшегося затруднения может явиться переход к уравнениям

$$y = A + A_1 z_1 + bx + \varepsilon$$

или

$$y = A + A_2 z_2 + bx + \varepsilon,$$

т.е. каждое уравнение включает только одну фиктивную переменную z_1 или z_2.

Предположим, что определено уравнение

$$y = A + A_1 z_1 + bx + \varepsilon,$$

где z_1 принимает значения 1 для мужчин и 0 для женщин.

Теоретические значения размера потребления кофе для мужчин будут получены из уравнения

$$\hat{y} = A + A_1 + bx.$$

Для женщин соответствующие значения получим из уравнения

$$\hat{y} = A + bx.$$

Сопоставляя эти результаты, видим, что различия в уровне потребления мужчин и женщин состоят в различии свободных членов данных уравнений: A – для женщин и $A + A_1$ – для мужчин.

Теперь качественный фактор принимает только два состояния, которым соответствуют значения 1 и 0. Если же число градаций

качественного признака-фактора превышает два, то в модель вводится несколько фиктивных переменных, число которых должно быть меньше числа качественных градаций. Только при соблюдении этого положения матрица исходных фиктивных переменных не будет линейно зависима и возможна оценка параметров модели.

Пример. Проанализируем зависимость цены двухкомнатной квартиры от ее полезной площади. При этом в модель могут быть введены фиктивные переменные, отражающие тип дома: «хрущевка», панельный, кирпичный.

При использовании трех категорий домов вводятся две фиктивные переменные: z_1 и z_2. Пусть переменная z_1 принимает значение 1 для панельного дома и 0 для всех остальных типов домов; переменная z_2 принимает значение 1 для кирпичных домов и 0 для остальных; тогда переменные z_1 и z_2 принимают значения 0 для домов типа «хрущевки».

Предположим, что уравнение регрессии с фиктивными переменными составило:

$$\hat{y} = 320 + 500x + 2200z_1 + 1600z_2.$$

Частные уравнения регрессии для отдельных типов домов, свидетельствуя о наиболее высоких ценах квартир в панельных домах, будут иметь следующий вид: «хрущевки» – $\hat{y} = 320 + 500x$; панельные – $\hat{y} = 2520 + 500x$; кирпичные – $\hat{y} = 1920 + 500x$.

Параметры при фиктивных переменных z_1 и z_2 представляют собой разность между средним уровнем результативного признака для соответствующей группы и базовой группы. В рассматриваемом примере за базу сравнения цены взяты дома «хрущевки», для которых $z_1 = z_2 = 0$. Параметр при z_1, равный 2200, означает, что при одной и той же полезной площади квартиры цена ее в панельных домах в среднем на 2200 долл. США выше, чем в «хрущевках». Соответственно параметр при z_2 показывает, что

в кирпичных домах цена выше в среднем на 1600 долл. при неизменной величине полезной площади по сравнению с указанным типом домов.

В отдельных случаях может оказаться необходимым введение двух и более групп фиктивных переменных, т.е. двух и более качественных факторов, каждый из которых может иметь несколько градаций. Например, при изучении потребления некоторого товара наряду с факторами, имеющими количественное выражение (цена, доход на одного члена семьи, цена на взаимозаменяемые товары и др.), учитываются и качественные факторы. С их помощью оцениваются различия в потреблении отдельных социальных групп населения, дифференциация в потреблении по полу, национальному составу и др. При построении такой модели из каждой группы фиктивных переменных следует исключить по одной переменной. Так, если модель будет включать три социальные группы, три возрастные категории и ряд экономических переменных, то она примет вид:

$$y = a + b_1 s_1 + b_2 s_2 + b_3 z_1 + b_4 z_2 + b_5 x_1 + b_6 x_2 + \ldots + b_{m+4} x_m + \varepsilon,$$

где y – потребление;

$$s_i = \begin{cases} 1 - \text{если наблюдения относятся к } i\text{-й социальной группе } (i = 1,\ 2), \\ 0 - \text{в остальных случаях;} \end{cases}$$

$$z_i = \begin{cases} 1 - \text{если наблюдения относятся к } j\text{-й возрастной группе } (j = 1,\ 2), \\ 0 - \text{в остальных случаях;} \end{cases}$$

$x_1,\ x_2,\ \ldots, x_m$ – экономические (количественные) переменные.

До сих пор мы рассматривали фиктивные переменные как факторы, которые используются в регрессионной модели наряду с количественными переменными. Вместе с тем возможна регрессия только на фиктивных переменных. Например, изучается дифференциация заработной платы рабочих высокой квалификации по регионам страны. Модель заработной платы может иметь вид:

$$\hat{y} = a + b_1 z_1 + b_2 z_2 + \ldots + b_m z_m,$$

где y – средняя заработная плата рабочих высокой квалификации по отдельным предприятиям;

$$z_1 = \begin{cases} 1 - \text{если предприятие находится в Северо-Западном районе;} \\ 0 - \text{если предприятие находится в остальных районах;} \end{cases}$$

$$z_2 = \begin{cases} 1 - \text{если предприятие находится в Волго-Вятском районе;} \\ 0 - \text{если предприятие находится в остальных районах;} \end{cases}$$

……..

$$z_m = \begin{cases} 1 - \text{если преприятие находится в Дальневосточном районе;} \\ 0 - \text{если предприятие находится в остальных районах.} \end{cases}$$

Поскольку последний район, указанный в модели, обозначен z_m, то в исследование включено $m+1$ район.

Мы рассмотрели модели с фиктивными переменными, в которых последние выступают факторами. Может возникнуть необходимость построить модель, в которой дихотомический признак, т.е. признак, который может принимать только два значения, играет роль результата. Подобного вида модели применяются, например, при обработке данных социологических опросов. В качестве зависимой переменной y рассматриваются ответы на вопросы, данные в альтернативной форме: «да» или «нет». Поэтому зависимая переменная имеет два значения: 1, когда имеет место ответ «да», и 0 – во всех остальных случаях. Модель такой зависимой переменной имеет вид:

$$y = a + b_1 x_1 + ... + b_m x_m + \varepsilon .$$

Модель является вероятностной линейной моделью. В ней y принимает значения 1 и 0, которым соответствуют вероятности p и $1-p$. Поэтому при решении модели находят оценку условной вероятности события y при фиксированных значениях x. Для оценки параметров линейно-вероятностной модели применяются методы Logit-, Probit- и Tobit-анализа. Такого рода модели используют при работе с неколичественными переменными. Как правило, это модели выбора из заданного набора

альтернатив. Зависимая переменная y представлена дискретными значениями (набор альтернатив), объясняющие переменные x_j – характеристики альтернатив (время, цена), z_j – характеристики индивидов (возраст, доход, уровень образования). Модель такого рода позволяет предсказать долю индивидов в генеральной совокупности, которые выбирают данную альтернативу.

Среди моделей с фиктивными переменными наибольшими прогностическими возможностями обладают модели, в которых зависимая переменная y рассматривается как функция ряда экономических факторов x_i и фиктивных переменных z_j. Последние обычно отражают различия в формировании результативного признака по отдельным группам единиц совокупности, т.е. в результате неоднородной структуры пространственного или временного характера.

3. Системы эконометрических уравнений

При использовании отдельных уравнений регрессии, например для экономических расчетов, в большинстве случаев предполагается, что аргументы (факторы) можно изменять независимо друг от друга. Однако это предположение является очень грубым: практически изменение одной переменной, как правило, не может происходить при абсолютной неизменности других. Ее изменение повлечет за собой изменения во всей системе взаимосвязанных признаков. Следовательно, отдельно взятое уравнение множественной регрессии не может характеризовать истинные влияния отдельных признаков на вариацию результирующей переменной. Именно поэтому в последние десятилетия в экономических исследованиях важное место заняла проблема описания структуры связей между переменными системой так называемых одновременных уравнений, называемых также структурными уравнениями.

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

Возможна *система независимых уравнений*, когда каждая зависимая переменная y рассматривается как функция одного и того же набора факторов x:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n + \varepsilon_m. \end{cases} \qquad (3.1)$$

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 регрессии. Так как фактические значения зависимой

переменной отличаются от теоретических на величину случайной ошибки, то в каждом уравнении присутствует величина случайной ошибки ε_i.

Если зависимая переменная y одного уравнения выступает в виде фактора x в другом уравнении, то исследователь может строить модель в виде *системы рекурсивных уравнений:*

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + \varepsilon_2, \\ ..... \\ y_m = b_{m1}y_1 + ... + b_{m,m-1}y_{m-1} + a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n + \varepsilon_m. \end{cases} \quad (3.2)$$

В данной системе зависимая переменная y включает в каждое последующее уравнение в качестве факторов все зависимые переменные предшествующих уравнений наряду с набором собственно факторов x. Каждое уравнение этой системы может рассматриваться самостоятельно, и его параметры определяются методом наименьших квадратов (МНК).

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:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + ... + b_{1m}y_m + a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + ... + b_{2m}y_m + a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + ... + b_{3m}y_m + a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n + \varepsilon_2, \\ ..... \\ y_m = b_{m1}y_1 + b_{m2}y_2 + ... + b_{m,m-1}y_{m-1} + a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n + \varepsilon_n. \end{cases} \quad (3.3)$$

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.

3.1. Структурная и приведенная формы модели

Система совместных, одновременных уравнений (или структурная форма модели) обычно содержит эндогенные и экзогенные переменные.

Эндогенные переменные – это зависимые переменные, число которых равно числу уравнений в системе и которые 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 b_{ik} и экзогенных переменных – коэффициенты

a_{ij}, которые называются *структурными коэффициентами* модели. Все переменные в модели выражены в отклонениях от среднего уровня, т.е. под x подразумевается $x-\overline{x}$, а под y – соответственно $y-\overline{y}$. Поэтому свободный член в каждом уравнении системы (3.3) отсутствует.

Использование МНК для оценивания структурных коэффициентов модели дает, как принято считать в теории, смещенные и несостоятельные оценки. Поэтому обычно для определения структурных коэффициентов модели структурная форма модели преобразуется в *приведенную форму* модели.

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

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{1n}x_n + u_1, \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{2n}x_n + u_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = \delta_{m1}x_1 + \delta_{m2}x_2 + \ldots + \delta_{mn}x_n + u_m, \end{cases} \qquad (3.4)$$

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

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + \varepsilon_1, \\ y_2 = b_{21} y_1 + a_{22} x_2 + \varepsilon_2 \end{cases} \quad (3.5)$$

приведенная форма модели имеет вид

$$\begin{cases} y_1 = \delta_{11} x_1 + \delta_{12} x_2 + u_1, \\ y_2 = \delta_{21} x_1 + \delta_{22} x_2 + u_2. \end{cases} \quad (3.6)$$

Из первого уравнения (3.5) можно выразить y_2 следующим образом (ради упрощения опускаем случайную величину):

$$y_2 = \frac{y_1 - a_{11} x_1}{b_{12}}.$$

Подставляя во второе уравнение (3.5), имеем

$$\frac{y_1 - a_{11} x_1}{b_{12}} = b_{21} y_1 + a_{22} x_2,$$

откуда

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

Поступая аналогично со вторым уравнением системы (3.5), получим

$$y_2 = \frac{a_{11} b_{21}}{1 - b_{12} b_{21}} x_1 + \frac{a_{22}}{1 - b_{12} b_{21}} x_2,$$

т.е. система (3.5) принимает вид

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

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

$$\delta_{11} = \frac{a_{11}}{1 - b_{12}b_{21}}, \quad \delta_{12} = \frac{a_{22}b_{12}}{1 - b_{12}b_{21}},$$

$$\delta_{21} = \frac{a_{11}b_{21}}{1 - b_{12}b_{21}}, \quad \delta_{22} = \frac{a_{22}}{1 - b_{12}b_{21}}.$$

Следует заметить, что приведенная форма модели хотя и позволяет получить значения эндогенной переменной через значения экзогенных переменных, но аналитически она уступает структурной форме модели, так как в ней отсутствуют оценки взаимосвязи между эндогенными переменными.

3.2. Проблема идентификации

При переходе от приведенной формы модели к структурной эконометрист сталкивается с проблемой идентификации. Идентификация – это единственность соответствия между приведенной и структурной формами модели.

Структурная модель (3.3) в полном виде содержит $m \cdot (m + n - 1)$ параметров, а приведенная форма модели в полном виде содержит $m \cdot n$ параметров. Т.е. в полном виде структурная модель содержит большее число параметров, чем приведенная форма модели. Соответственно $m \cdot (m + n - 1)$ параметров структурной модели не могут быть однозначно определены из $m \cdot n$ параметров приведенной формы модели.

Чтобы получить единственно возможное решение для структурной модели, необходимо предположить, что некоторые из структурных коэффициентов модели ввиду слабой взаимосвязи признаков с эндогенной переменной из левой части системы равны нулю. Тем самым уменьшится число структурных коэффициентов модели. Уменьшение числа структурных коэффициентов модели возможно и другим путем: например, путем приравнивания некоторых коэффициентов друг к другу, т.е. путем предположений, что их воздействие на формируемую эндогенную

переменную одинаково. На структурные коэффициенты могут накладываться, например, ограничения вида $b_{ik} + a_{ij} = 0$.

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:

Таблица 4.1

$D+1=H$	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 не меньше, чем число эндогенных переменных в системе без одного.

Целесообразность проверки условия идентификации модели через определитель матрицы коэффициентов, отсутствующих в данном уравнении, но присутствующих в других, 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.

Рассмотрим **пример.** 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: $C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot C_{t-1} + \varepsilon_1$. 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 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.

	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, имеет вид

	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, имеет вид

	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, имеет вид

	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}	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:

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

3.3. Методы оценки параметров структурной формы модели

Коэффициенты структурной модели могут быть оценены разными способами в зависимости от вида системы одновременных уравнений. Наибольшее распространение в литературе получили следующие методы оценивания коэффициентов структурной модели:

1) косвенный метод наименьших квадратов;

2) двухшаговый метод наименьших квадратов;

3) трехшаговый метод наименьших квадратов;

4) метод максимального правдоподобия с полной информацией;

5) метод максимального правдоподобия при ограниченной информации.

Рассмотрим вкратце сущность каждого из этих методов.

Косвенный метод наименьших квадратов (КМНК) применяется в случае точно идентифицируемой структурной модели. 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.

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).

Основная идея ДМНК – на основе приведенной формы модели получить для сверхидентифицируемого уравнения теоретические значения эндогенных переменных, содержащихся в правой части уравнения.

Далее, подставив их вместо фактических значений, можно применить обычный МНК к структурной форме сверхидентифицируемого уравнения. Метод получил название двухшагового МНК, ибо дважды используется МНК: на первом шаге при определении приведенной формы модели и нахождении на ее основе оценок теоретических значений эндогенной переменной $\hat{y}_i = \delta_{i1}x_1 + \delta_{i2}x_2 + ... + \delta_{in}x_n$ и на втором шаге применительно к структурному сверхидентифицируемому уравнению при определении структурных коэффициентов модели по данным теоретических (расчетных) значений эндогенных переменных.

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 структурные коэффициенты по ним находятся из системы приведенных уравнений.

Для примера, рассмотренного в предыдущем параграфе, необходимо применить именно двухшаговый метод наименьших квадратов. Но можно сделать следующее замечание. Если из модели исключить тождество дохода, число эндогенных переменных модели снизится на единицу – переменная Y_t станет экзогенной. А число предопределенных переменных модели не изменится, т.к. из модели будет исключена эндогенная переменная G_t, но ее место займет переменная Y_t. В правых частях функции потребления и функции денежного рынка будут находиться только предопределенные переменные. Функция инвестиций постулирует зависимость эндогенной переменной I_t от эндогенной переменной r_t (которая зависит только от предопределенных переменных) и предопределенной переменной I_{t-1}. Таким образом, мы получим рекурсивную систему. Ее параметры можно оценивать обычным МНК, и нет необходимости исследования уравнения на идентификацию.

Косвенный и двухшаговый методы наименьших квадратов подробно описаны в литературе и рассматриваются как традиционные методы оценки коэффициентов структурной модели. Эти методы достаточно легко реализуемы.

Метод максимального правдоподобия рассматривается как наиболее общий метод оценивания, результаты которого при нормальном распределении признаков совпадают с МНК. Однако при большом числе уравнений системы этот метод приводит к достаточно сложным вычислительным процедурам. Поэтому в качестве модификации

используется метод максимального правдоподобия при ограниченной информации (метод наименьшего дисперсионного отношения), разработанный в 1949 г. Т.Андерсоном и Н.Рубиным.

В отличие от метода максимального правдоподобия в данном методе сняты ограничения на параметры, связанные с функционированием системы в целом. Это делает решение более простым, но трудоемкость вычислений остается достаточно высокой. Несмотря на его значительную популярность, к середине 60-х годов он был практически вытеснен двухшаговым методом наименьших квадратов (ДМНК) в связи с гораздо большей простотой последнего.

Дальнейшим развитием ДМНК является трехшаговый МНК (ТМНК), предложенный в 1962 г. А.Зельнером и Г.Тейлом. Этот метод оценивания пригоден для всех видов уравнений структурной модели. Однако при некоторых ограничениях на параметры более эффективным оказывается ДМНК.

4. Временные ряды

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 за несколько последовательных моментов или периодов времени. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

1) факторы, формирующие тенденцию ряда;

2) факторы, формирующие циклические колебания ряда;

3) случайные факторы.

Рассмотрим воздействие каждого фактора на временной ряд в отдельности.

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. На рис. 4.1 показан гипотетический временной ряд, содержащий возрастающую тенденцию.

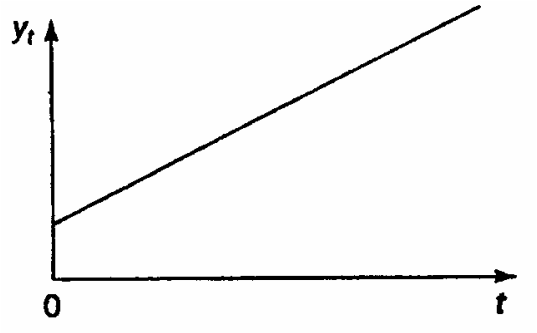


Рис. 4.1.

Также изучаемый показатель может быть подвержен циклическим колебаниям. Эти колебания могут носить сезонный характер, поскольку экономическая деятельность ряда отраслей экономики 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, связанные с общей динамикой конъюнктуры рынка. На рис. 4.2 представлен гипотетический временной ряд, содержащий только сезонную компоненту.

Рис. 4.2.

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. Пример ряда, содержащего только случайную компоненту, приведен на рис. 4.3.

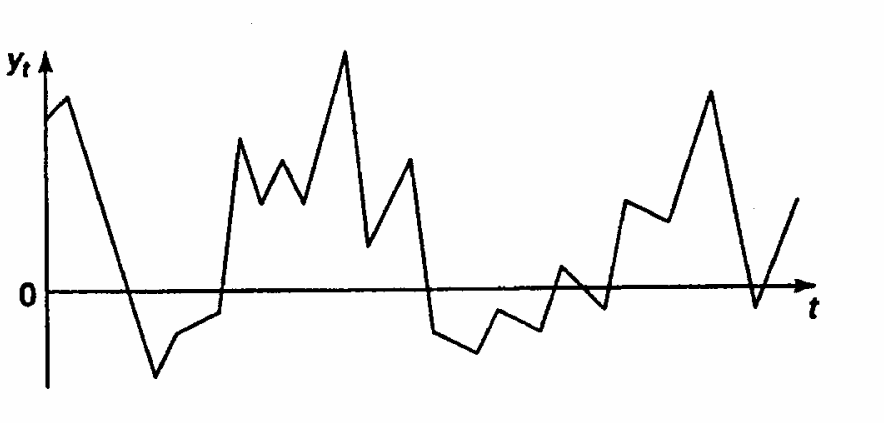


Рис. 4.3.

Очевидно, что реальные данные не следуют целиком и полностью из каких-либо описанных выше моделей. Чаще всего они содержат все три компоненты. Каждый их уровень формируется под воздействием тенденции, сезонных колебаний и случайной компоненты.

В большинстве случаев фактический уровень временного ряда можно представить как сумму или произведение трендовой, циклической и случайной компонент. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется *аддитивной моделью* временного ряда. Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется *мультипликативной моделью* временного ряда. 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.

4.1. 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.

Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени.

Формула для расчета коэффициента автокорреляции имеет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \overline{y}_1)(y_{t-1} - \overline{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \overline{y}_1)^2 \sum_{t=2}^{n}(y_{t-1} - \overline{y}_2)^2}}, \qquad (4.1)$$

где

$$\overline{y}_1 = \frac{1}{n-1}\sum_{t=2}^{n} y_t, \qquad \overline{y}_2 = \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}.$$

Эту величину называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда t и y_{t-1}.

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:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \overline{y}_3)(y_{t-2} - \overline{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \overline{y}_3)^2 \sum_{t=3}^{n}(y_{t-2} - \overline{y}_4)^2}}, \qquad (4.2)$$

где

$$\overline{y}_3 = \frac{1}{n-2}\sum_{t=3}^{n} y_t, \qquad \overline{y}_4 = \frac{1}{n-2}\sum_{t=3}^{n} y_{t-2}.$$

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют *лагом*. 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 $n/4$.

Свойства коэффициента автокорреляции.

1. Он 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 некоторых временных рядов, имеющих сильную нелинейную тенденцию (например, параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

2. 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, а следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, т.е. 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.

Рассмотрим **пример**. Пусть имеются некоторые условные данные об общем количестве правонарушений на таможне одного из субъектов РФ (например, Республики Татарстан).

Таблица 4.1

Год	Квартал	t	Количество возбужденных дел, y_t
1999	I	1	375
	II	2	371
	III	3	869
	IV	4	1015
2000	I	5	357
	II	6	471
	III	7	992
	IV	8	1020
2001	I	9	390
	II	10	355
	III	11	992
	IV	12	905
2002	I	13	461
	II	14	454
	III	15	920
	IV	16	927

Построим поле корреляции:

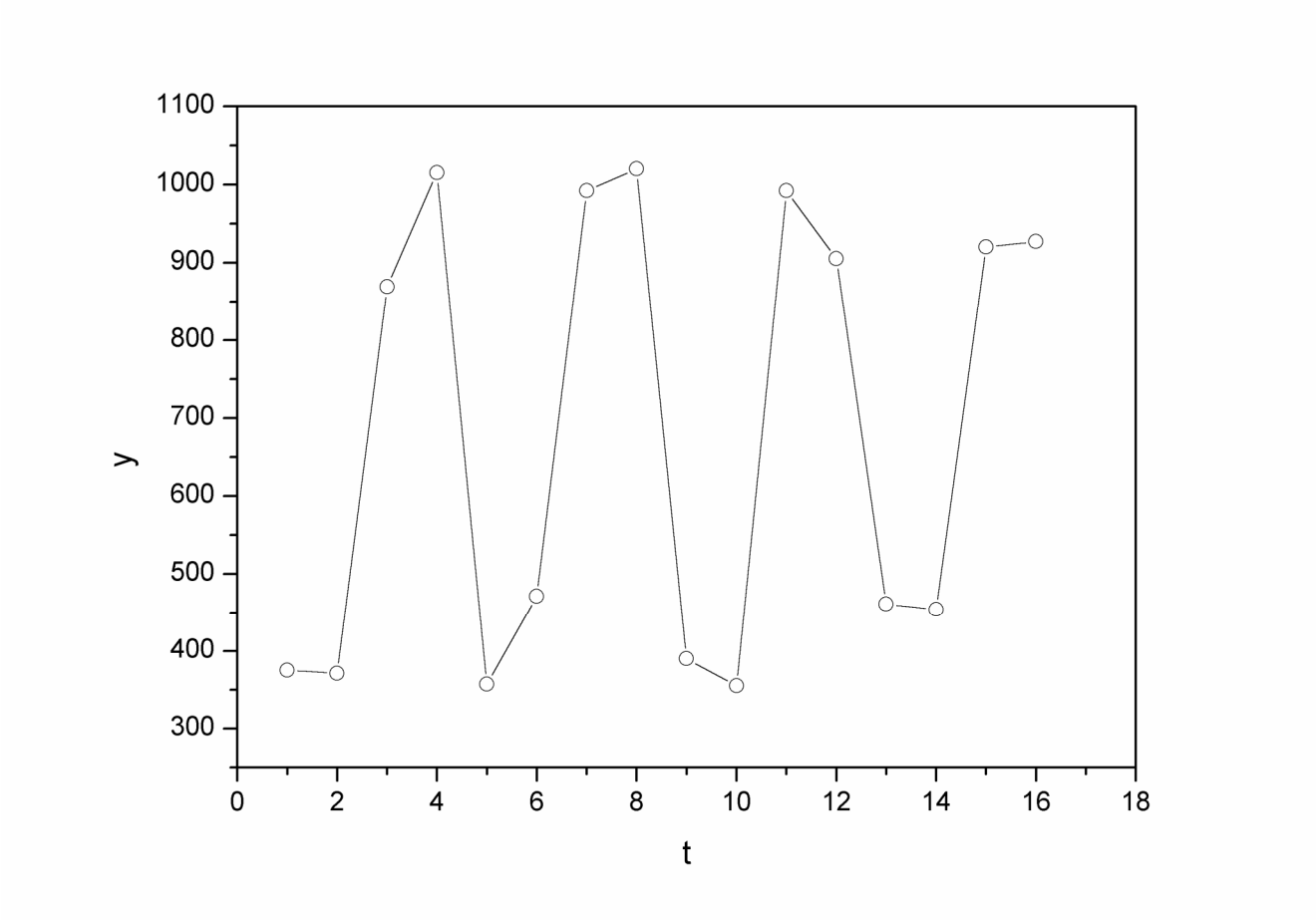


Рис. 4.4.

Уже исходя из графика видно, что значения y образуют пилообразную фигуру. Рассчитаем несколько последовательных коэффициентов автокорреляции. Для этого составляем первую вспомогательную таблицу.

Таблица 4.2

t	y_t	y_{t-1}	$y_t - \overline{y}_1$	$y_{t-1} - \overline{y}_2$	$(y_t - \overline{y}_1)\times$ $\times(y_{t-1} - \overline{y}_2)$	$(y_t - \overline{y}_1)^2$	$(y_{t-1} - \overline{y}_2)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	375	–	–	–	–	–	–
2	371	375	-328,33	-288,13	94601,72	107800,59	83018,90
3	869	371	169,67	-292,13	-49565,70	28787,91	85339,94
4	1015	869	315,67	205,87	64986,98	99647,55	42382,46
5	357	1015	-342,33	351,87	-120455,66	117189,83	123812,50
6	471	357	-228,33	-306,13	69898,66	52134,59	93715,58
7	992	471	292,67	-192,13	-56230,69	85655,73	36913,94
8	1020	992	320,67	328,87	105458,74	102829,25	108155,48
9	390	1020	-309,33	356,87	-110390,60	95685,05	127356,20
10	355	390	-344,33	-273,13	94046,85	118563,15	74600,00
11	992	355	292,67	-308,13	-90180,41	85655,73	94944,10
12	905	992	205,67	328,87	67638,69	42300,15	108155,48

1	2	3	4	5	6	7	8
13	461	905	-238,33	241,87	-57644,88	56801,19	58501,10
14	454	461	-245,33	-202,13	49588,55	60186,81	40856,54
15	920	454	220,67	-209,13	-46148,72	48695,25	43735,36
16	927	920	227,67	256,87	58481,59	51833,63	65982,20
Сумма	10499	9947	9,05	0,05	74085,16	1153766,39	1187469,73
Среднее значение	699,33	663,13	–	–	–	–	–

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 (4.1):

$$r_1 = \frac{74085,16}{\sqrt{1153756,39 \cdot 1187469,73}} = 0,063294.$$

Составляем вспомогательную таблицу для расчета коэффициента автокорреляции второго порядка.

Таблица 4.3

t	y_t	y_{t-2}	$y_t - \bar{y}_3$	$y_{t-2} - \bar{y}_4$	$(y_t - \bar{y}_3) \times \times (y_{t-2} - \bar{y}_4)$	$(y_t - \bar{y}_3)^2$	$(y_{t-2} - \bar{y}_4)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	375	–	–	–	–	–	–
2	371	–	–	–	–	–	–
3	869	375	145,57	-269,79	-39273,33	21190,62	72786,64
4	1015	371	291,57	-273,79	-79828,95	85013,06	74960,96
5	357	869	-366,43	224,21	-82157,27	134270,94	50270,12
6	471	1015	-252,43	370,21	-93452,11	63720,90	137055,44
7	992	357	268,57	-287,79	-77291,76	72129,84	82823,08
8	1020	471	296,57	-173,79	-51540,90	87953,76	30202,96
9	390	992	-333,43	347,21	-115770,23	111175,56	120554,78
10	355	1020	-368,43	375,21	-138238,62	135740,66	140782,54
11	992	390	268,57	-254,79	-68428,95	72129,84	64917,94
12	905	355	181,57	-289,79	-52617,17	32967,66	83978,24
13	461	992	-262,43	347,21	-91118,32	68869,50	120554,78
14	454	905	-269,43	260,21	-70108,38	72592,52	67709,24
15	920	461	196,57	-183,79	-36127,60	38639,76	33778,76
16	927	454	203,57	-190,79	-38839,12	41440,74	36400,82
Сумма	10128	9027	-0,02	-0,06	-1034792,71	1037835,43	1116776,36
Среднее значение	723,43	644,79	–	–	–	–	–

Следовательно

$$r_2 = \frac{-1034792,71}{\sqrt{1037835,43 \cdot 1116776,36}} = -0,961183.$$

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

Таблица 4.4

Лаг	Коэффициент автокорреляции уровней
1	0,063294
2	–0,961183
3	–0,036290
4	0,964735
5	0,050594
6	–0,976516
7	–0,069444
8	0,964629
9	0,162064
10	-0,972918
11	-0,065323
12	0,985761

Коррелограмма:

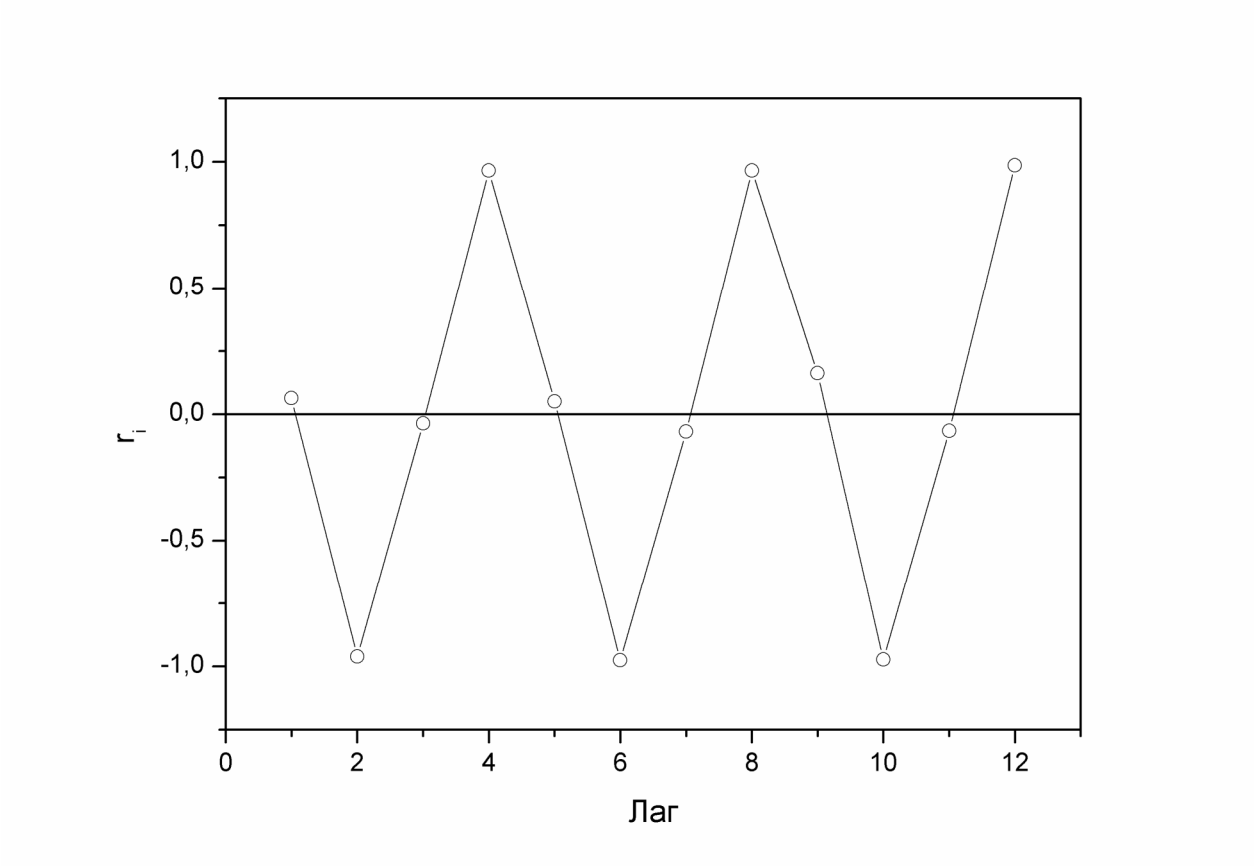


Рис. 4.5.

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

4. 2. 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*.

Поскольку зависимость от времени может принимать разные формы, для ее формализации можно использовать различные виды функций. Для построения трендов чаще всего применяются следующие функции:

линейный тренд: $\hat{y}_t = a + b \cdot t$;

гипербола: $\hat{y}_t = a + b/t$;

экспоненциальный тренд: $\hat{y}_t = \mathrm{e}^{a+b \cdot t}$ (или $\hat{y}_t = a \cdot b^t$);

степенная функция: $\hat{y}_t = a \cdot t^b$;

полиномы различных степеней: $\hat{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + ... + b_m \cdot t^m$.

Параметры каждого из перечисленных выше трендов можно определить обычным МНК, используя в качестве независимой переменной время $t = 1, 2, ..., n$, а в качестве зависимой переменной – фактические уровни временного ряда $\hat{y}_t$. Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

Существует несколько способов определения типа тенденции. К числу наиболее распространенных способов относятся качественный анализ изучаемого процесса, построение и визуальный анализ графика зависимости уровней ряда от времени. В этих же целях можно использовать и коэффициенты автокорреляции уровней ряда. Тип тенденции можно

определить путем сравнения коэффициентов автокорреляции первого порядка, рассчитанных по исходным и преобразованным уровням ряда. Если временной ряд имеет линейную тенденцию, то его соседние уровни $\hat{y}_t$ и $\hat{y}_{t-1}$ тесно коррелируют. В этом случае коэффициент автокорреляции первого порядка уровней 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 ряда. Чем сильнее выражена нелинейная тенденция в изучаемом временном ряде, тем в большей степени будут различаться значения указанных коэффициентов.

Выбор наилучшего уравнения в случае, когда ряд содержит нелинейную тенденцию, можно осуществить путем перебора основных форм тренда, расчета по каждому уравнению скорректированного коэффициента детерминации и средней ошибки аппроксимации. Этот метод легко реализуется при компьютерной обработке данных.

4.3. Моделирование сезонных колебаний

Простейший подход к моделированию сезонных колебаний – это расчет значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели временного ряда.

Общий вид аддитивной модели следующий:

$$Y = T + S + E. \qquad (4.3)$$

Эта модель предполагает, что каждый уровень временного ряда может быть представлен как сумма трендовой (T), сезонной (S) и случайной (E) компонент.

Общий вид мультипликативной модели выглядит так:

$$Y = T \cdot S \cdot E. \qquad (4.4)$$

Эта модель предполагает, что каждый уровень временного ряда может быть представлен как произведение трендовой (T), сезонной (S) и случайной (E) компонент.

Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты.

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 ($T + E$) в аддитивной или ($T \cdot E$) в мультипликативной 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. Построение аддитивной модели временного ряда. Обратимся к данным об объеме правонарушений на таможне за четыре года, представленным в табл. 4.1.

Было показано, что данный временной ряд содержит сезонные колебания периодичностью 4, т.к. количество правонарушений в первый-второй кварталы ниже, чем в третий-четвертый. 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. 4.5).

1.2. Разделив полученные суммы на 4, найдем скользящие средние (гр. 4 табл. 4.5). 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. 4.5).

Таблица 4.5

№ квартала, t	Количество правонарушений, y_t	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	**2**	**3**	**4**	**5**	**6**
1	375	–	–	–	–
2	371	2630	657,5	–	–
3	869	2612	653	655,25	213,75
4	1015	2712	678	665,5	349,5
5	357	2835	708,75	693,75	-336,75
6	471	2840	710	709,375	-238,375
7	992	2873	718,25	714,125	277,875
8	1020	2757	689,25	703,75	316,25
9	390	2757	689,25	689,25	-299,25
10	355	2642	660,5	674,875	-319,875
11	992	2713	678,25	669,375	322,625
12	905	2812	703	690,625	214,375
13	461	2740	685	694	-233
14	454	2762	690,5	687,75	-233,75
15	920	–	–	–	–
16	927	–	–	–	–

Шаг 2. Найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (гр. 6 табл. 4.5). Используем эти оценки для расчета значений сезонной компоненты S (табл. 4.6). 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.

Таблица 4.6

Показатели	Год	№ квартала, i			
		I	II	III	IV
	1999	–	–	213,75	349,5
	2000	-336,75	-238,375	277,875	316,25
	2001	-299,25	-319,875	322,625	214,375
	2002	-233	-233,75	–	–
Всего за i-й квартал		-869	-792	814,25	880,125
Средняя оценка сезонной компоненты для i-го квартала, $\overline{S}_i$		-289,667	-264	271,417	293,375
Скорректированная сезонная компонента, S_i		-292,448	-266,781	268,636	290,593

Для данной модели имеем:

$$-289,667 - 264 + 271,417 + 293,375 = 11,125.$$

Корректирующий коэффициент: $k = 11,125/4 = 2,781.$

Рассчитываем скорректированные значения сезонной компоненты ($S_i = \overline{S}_i - k$) и заносим полученные данные в таблицу 4.6.

Проверим равенство нулю суммы значений сезонной компоненты:

$$-292,448 - 266,781 + 268,636 + 290,593 = 0.$$

Шаг 3. Исключим влияние сезонной компоненты, вычитая ее значение из каждого уровня исходного временного ряда. Получим величины

$T + E = Y - S$ (гр. 4 табл. 4.7). Эти значения рассчитываются за каждый момент времени и содержат только тенденцию и случайную компоненту.

Таблица 4.7

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	375	-292,448	667,448	672,700	380,252	-5,252	27,584
2	371	-266,781	637,781	673,624	406,843	-35,843	1284,721
3	869	268,636	600,364	674,547	943,183	-74,183	5503,117
4	1015	290,593	724,407	675,470	966,063	48,937	2394,830
5	357	-292,448	649,448	676,394	383,946	-26,946	726,087
6	471	-266,781	737,781	677,317	410,536	60,464	3655,895
7	992	268,636	723,364	678,240	946,876	45,124	2036,175
8	1020	290,593	729,407	679,163	969,756	50,244	2524,460
9	390	-292,448	682,448	680,087	387,639	2,361	5,574
10	355	-266,781	621,781	681,010	414,229	-59,229	3508,074
11	992	268,636	723,364	681,933	950,569	41,431	1716,528
12	905	290,593	614,407	682,857	973,450	-68,450	4685,403
13	461	-292,448	753,448	683,780	391,332	69,668	4853,630
14	454	-266,781	720,781	684,703	417,922	36,078	1301,622
15	920	268,636	651,364	685,627	954,263	-34,263	1173,953
16	927	290,593	636,407	686,550	977,143	-50,143	2514,320

Шаг 4. Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $(T + E)$ с помощью линейного тренда. Результаты аналитического выравнивания следующие:

$$T = 671,777 + 0,9233 \cdot t.$$

Подставляя в это уравнение значения $t = 1, 2, ..., 16$, найдем уровни T для каждого момента времени (гр. 5 табл. 4.7).

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. 4.7).

На одном графике отложим фактические значения уровней временного ряда и теоретические, полученные по аддитивной модели.

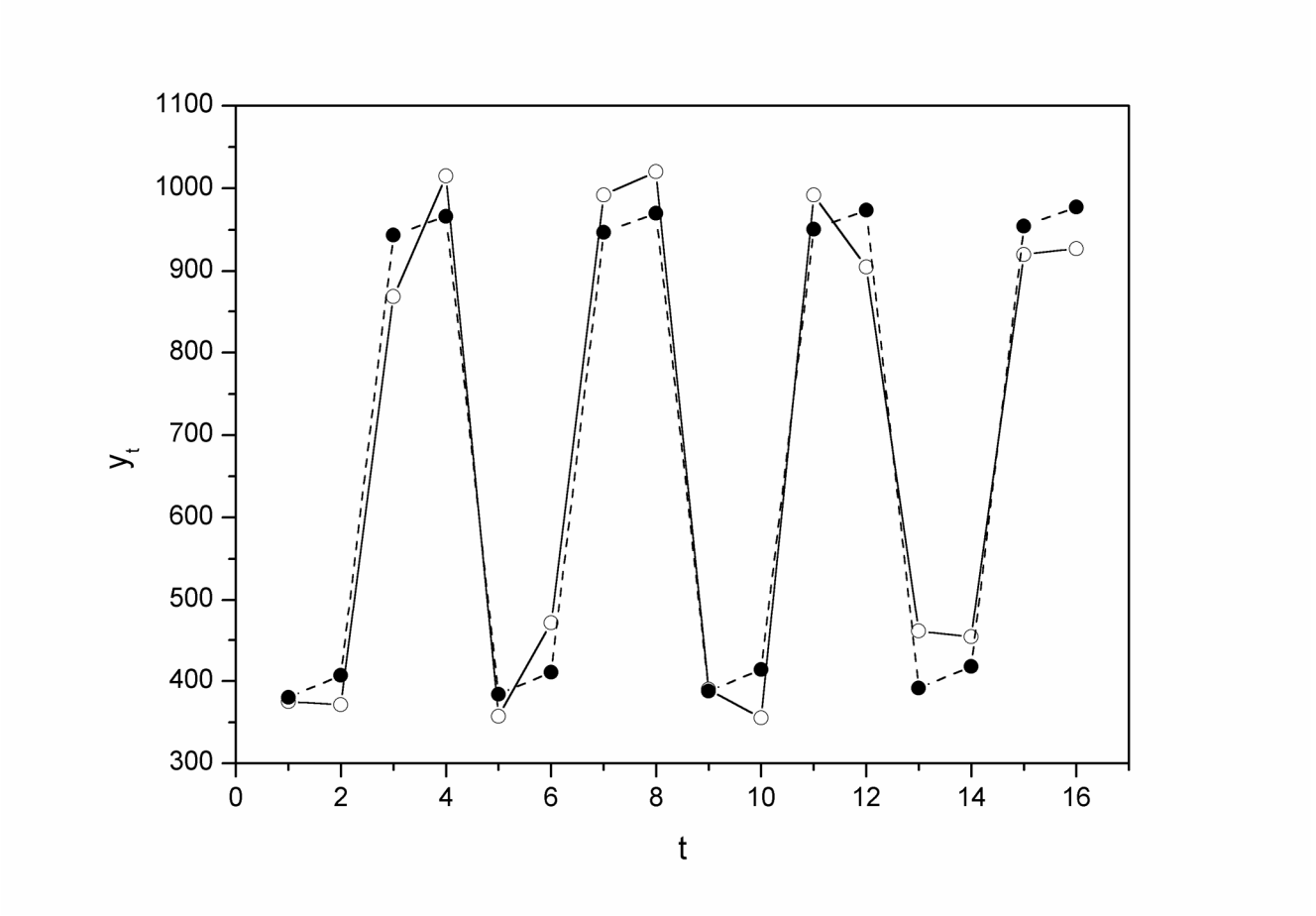


Рис. 4.6.

Для оценки качества построенной модели применим сумму квадратов полученных абсолютных ошибок.

$$R^2 = 1 - \frac{E^2}{\left(y_t - \overline{y}\right)^2} = 1 - \frac{37911{,}973}{1252743{,}75} = 0{,}970.$$

Следовательно, можно сказать, что аддитивная модель объясняет 97% общей вариации уровней временного ряда количества правонарушений по кварталам за 4 года.

Шаг 6. Прогнозирование по аддитивной модели. Предположим, что по нашему примеру необходимо дать прогноз об общем объеме правонарушений на I и II кварталы 2003 года. Прогнозное значение F_t уровня временного ряда в аддитивной модели есть сумма трендовой и сезонной компонент. Для определения трендовой компоненты воспользуемся уравнением тренда

$$T = 671,777 + 0,9233 \cdot t.$$

Получим

$$T_{17} = 671,777 + 0,9233 \cdot 17 = 687,473;$$

$$T_{18} = 671,777 + 0,9233 \cdot 18 = 688,396.$$

Значения сезонных компонент за соответствующие кварталы равны: $S_1 = -292,448$ и $S_2 = -266,781$. Таким образом,

$$F_{17} = T_{17} + S_1 = 687,473 - 292,448 \approx 395;$$

$$F_{18} = T_{18} + S_2 = 688,396 - 266,781 \approx 422.$$

Т.е. в первые два квартала 2003 г. следовало ожидать порядка 395 и 422 правонарушений соответственно.

Построение мультипликативной модели рассмотрим на данных предыдущего примера.

Шаг 1. Методика, применяемая на этом шаге, полностью совпадает с методикой построения аддитивной модели.

Таблица 4.8

№ квартала, t	Количество правонарушений, $\hat{y}_t$	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	**2**	**3**	**4**	**5**	**6**
1	375	–	–	–	–
2	371	2630	657,5	–	–
3	869	2612	653	655,25	1,3262
4	1015	2712	678	665,5	1,5252
5	357	2835	708,75	693,75	0,5146
6	471	2840	710	709,375	0,6640
7	992	2873	718,25	714,125	1,3891
8	1020	2757	689,25	703,75	1,4494
9	390	2757	689,25	689,25	0,5658
10	355	2642	660,5	674,875	0,5260
11	992	2713	678,25	669,375	1,4820
12	905	2812	703	690,625	1,3104
13	461	2740	685	694	0,6643
14	454	2762	690,5	687,75	0,6601
15	920	–	–	–	–
16	927	–	–	–	–

Шаг 2. Найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированные скользящие средние (гр. 6 табл. 4.8). Эти оценки используются для расчета сезонной компоненты S (табл. 4.9). 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 4.9

Показатели	Год	№ квартала, i			
		I	II	III	IV
	1999	–	–	1,3262	1,5252
	2000	0,5146	0,6640	1,3891	1,4494
	2001	0,5658	0,5260	1,4820	1,3104
	2002	0,6643	0,6601	–	–
Всего за i-й квартал		1,7447	1,8501	4,1973	4,2850
Средняя оценка сезонной компоненты для i-го квартала, $\overline{S}_i$		0,5816	0,6167	1,3991	1,4283
Скорректированная сезонная компонента, S_i		0,5779	0,6128	1,3901	1,4192

Имеем

$$0,5816 + 0,6167 + 1,3991 + 1,4283 = 4,0257.$$

Определяем корректирующий коэффициент:

$$k = 4 \big/ 4,0257 = 0,9936.$$

Скорректированные значения сезонной компоненты S_i получаются при умножении ее средней оценки $\overline{S}_i$ на корректирующий коэффициент k.

Проверяем условие равенство 4 суммы значений сезонной компоненты:

$$0,5779 + 0,6128 + 1,3901 + 1,4192 = 4.$$

Шаг 3. Разделим каждый уровень исходного ряда на соответствующие значения сезонной компоненты. В результате получим величины $T \cdot E = Y/S$ (гр. 4 табл. 4.10), которые содержат только тенденцию и случайную компоненту.

Таблица 4.10

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	375	0,5779	648,9012	654,9173	378,4767	0,9908
2	371	0,6128	605,4178	658,1982	403,3439	0,9198
3	869	1,3901	625,1349	661,4791	919,5221	0,9451
4	1015	1,4192	715,1917	664,7600	943,4274	1,0759
5	357	0,5779	617,7539	668,0409	386,0608	0,9247
6	471	0,6128	768,6031	671,3218	411,3860	1,1449
7	992	1,3901	713,6177	674,6027	937,7652	1,0578
8	1020	1,4192	718,7148	677,8836	962,0524	1,0602
9	390	0,5779	674,8572	681,1645	393,6450	0,9907
10	355	0,6128	579,3081	684,4454	419,4281	0,8464
11	992	1,3901	713,6177	687,7263	956,0083	1,0377
12	905	1,4192	637,6832	691,0072	980,6774	0,9228
13	461	0,5779	797,7159	694,2881	401,2291	1,1490
14	454	0,6128	740,8616	697,5690	427,4703	1,0621
15	920	1,3901	661,8229	700,8499	974,2515	0,9443
16	927	1,4192	653,1849	704,1308	999,3024	0,9277

Шаг 4. Определим компоненту T в мультипликативной модели. Для этого рассчитаем параметры линейного тренда, используя уровни $T \cdot E$. В результате получим уравнение тренда:

$$T = 651,6364 + 3,2809 \cdot t.$$

Подставляя в это уравнение значения $t = 1, 2, ..., 16$, найдем уровни T для каждого момента времени (гр. 5 табл. 4.10).

Шаг 5. Найдем уровни ряда, умножив значения T на соответствующие значения сезонной компоненты (гр. 6 табл. 4.10). На одном графике откладываем фактические значения уровней временного ряда и теоретические, полученные по мультипликативной модели.

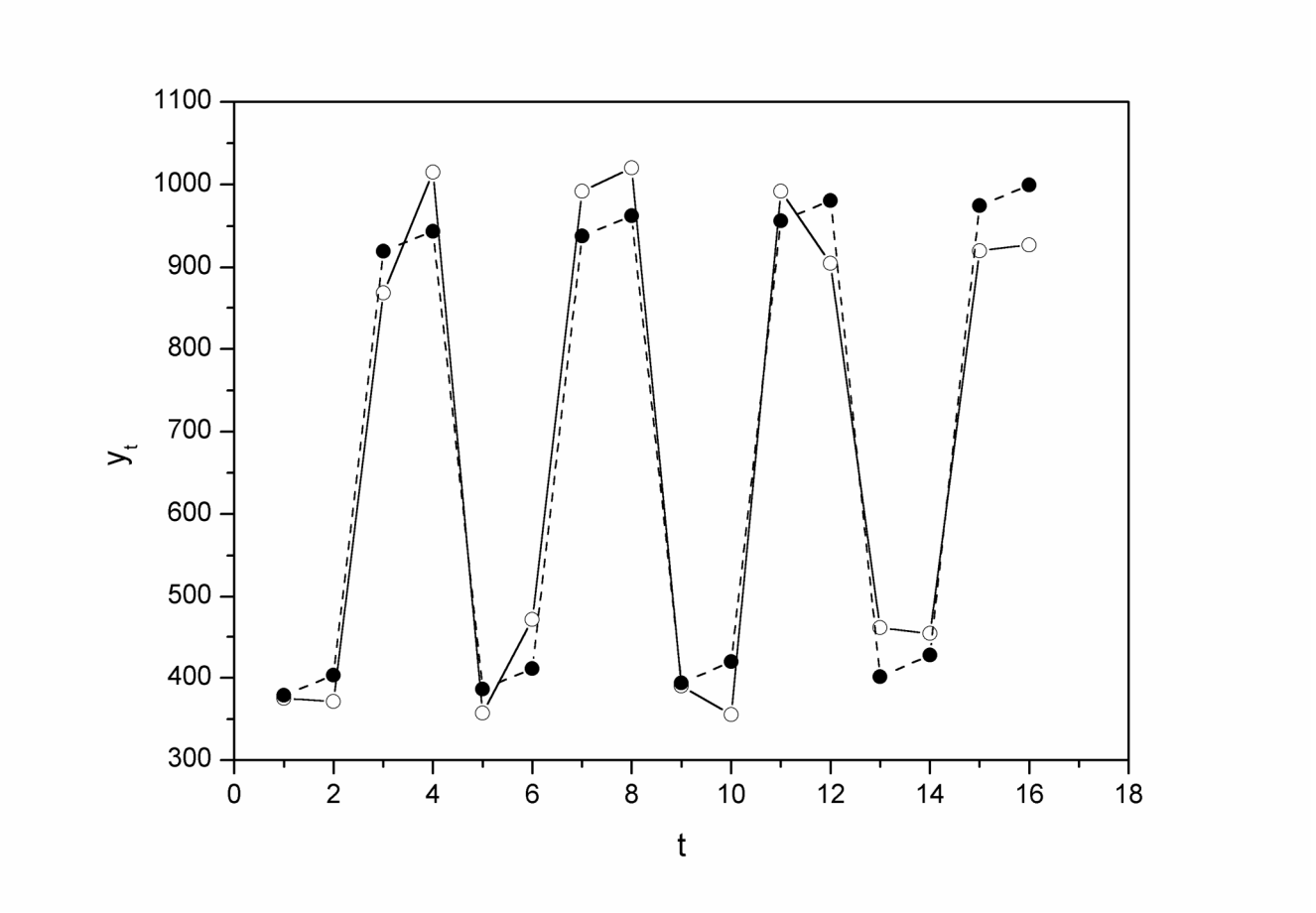


Рис. 4.7.

Расчет ошибки в мультипликативной модели производится по формуле:

$$E = Y / (T \cdot S).$$

Для сравнения мультипликативной модели и других моделей временного ряда можно, по аналогии с аддитивной моделью, использовать сумму квадратов абсолютных ошибок $(y_t - T \cdot S)^2$:

$$R^2 = 1 - \frac{(y_t - T \cdot S)^2}{(y_t - \bar{y})^2} = 1 - \frac{43065{,}02}{1252743{,}75} = 0{,}966.$$

Сравнивая показатели детерминации аддитивной и мультипликативной моделей, делаем вывод, что они примерно одинаково аппроксимируют исходные данные.

Шаг 6. Прогнозирование по мультипликативной модели. Если предположить, что по нашему примеру необходимо дать прогноз об общем объеме правонарушений на I и II кварталы 2003 года, прогнозное значение F_t уровня временного ряда в мультипликативной модели есть произведение трендовой и сезонной компонент. Для определения трендовой компоненты воспользуемся уравнением тренда

$$T = 651,6364 + 3,2809 \cdot t.$$

Получим

$$T_{17} = 651,6364 + 3,2809 \cdot 17 = 707,4117;$$

$$T_{18} = 651,6364 + 3,2809 \cdot 18 = 710,6926.$$

Значения сезонных компонент за соответствующие кварталы равны: $S_1 = 0,5779$ и $S_2 = 0,6128$. Таким образом

$$F_{17} = T_{17} \cdot S_1 = 707,4117 \cdot 0,5779 \approx 409;$$

$$F_{18} = T_{18} \cdot S_2 = 710,6926 \cdot 0,6128 \approx 436.$$

Т.е. в первые два квартала 2003 г. следовало ожидать порядка 409 и 436 правонарушений соответственно.

Таким образом, аддитивная и мультипликативная модели дают примерно одинаковый результат по прогнозу.

4.4. Автокорреляция в остатках. Критерий Дарбина-Уотсона

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

1. Она 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.

От истинной автокорреляции остатков следует отличать ситуации, когда причина автокорреляции заключается в неправильной спецификации функциональной формы модели. В этом случае следует изменить форму модели, а не использовать специальные методы расчета параметров уравнения регрессии при наличии автокорреляции в остатках.

Один из более распространенных методов определения автокорреляции в остатках – это расчет критерия Дарбина-Уотсона:

$$d = \frac{\sum_{t=2}^{n} \left(\varepsilon_t - \varepsilon_{t-1} \right)^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (4.5)$$

Т.е. величина d есть отношение суммы квадратов разностей последовательных значений остатков к остаточной сумме квадратов по модели регрессии.

Можно показать, что при больших значениях n существует следующее соотношение между критерием Дарбина-Уотсона d и коэффициентом автокорреляции остатков первого порядка r_1:

$$d = 2 \cdot \left(1 - r_1 \right). \qquad (4.6)$$

Таким образом, если в остатках существует полная положительная автокорреляция и $r_1 = 1$, то $d = 0$. Если в остатках полная отрицательная автокорреляция, то $r_1 = -1$ и, следовательно, $d = 4$. Если автокорреляция остатков отсутствует, то $r_1 = 0$ и $d = 2$. Т.е. $0 \le d \le 4$.

Алгоритм выявления автокорреляции остатков на основе критерия Дарбина-Уотсона следующий. 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. Далее по специальным таблицам (см. приложение Е) 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. По этим значениям числовой промежуток $[0;\ 4]$ разбивают на пять отрезков. Принятие или отклонение каждой из гипотез с вероятностью $1-\alpha$ осуществляется следующим образом:

$0 < d < d_L$ – есть положительная автокорреляция остатков, H_0 отклоняется, с вероятностью $P = 1-\alpha$ принимается H_1;

$d_L < d < d_U$ – зона неопределенности;

$d_U < d < 4 - d_U$ – нет оснований отклонять H_0, т.е. автокорреляция остатков отсутствует;

$4 - d_U < d < 4 - d_L$ – зона неопределенности;

$4 - d_L < d < 4$ – есть отрицательная автокорреляция остатков, H_0 отклоняется, с вероятностью $P = 1-\alpha$ принимается H_1^*.

Если фактическое значение критерия Дарбина-Уотсона попадает в зону неопределенности, то на практике предполагают существование автокорреляции остатков и отклоняют гипотезу H_0.

Пример. Проверим гипотезу о наличии автокорреляции в остатках для аддитивной модели нашего временного ряда. Исходные данные и промежуточные расчеты заносим в таблицу:

Таблица 4.11

t	y_t	$\varepsilon_t = E$	ε_{t-1}	$(\varepsilon_t - \varepsilon_{t-1})^2$	ε_t^2
1	**2**	**3**	**4**	**5**	**6**
1	375	-5,252	–	–	27,584
2	371	-35,843	-5,252	935,8093	1284,7
3	869	-74,183	-35,843	1469,956	5503,1
4	1015	48,937	-74,183	15158,53	2394,8
5	357	-26,946	48,937	5758,23	726,09
6	471	60,464	-26,946	7640,508	3655,9
7	992	45,124	60,464	235,3156	2036,2
8	1020	50,244	45,124	26,2144	2524,5
9	390	2,361	50,244	2292,782	5,574
10	355	-59,229	2,361	3793,328	3508,1
11	992	41,431	-59,229	10132,44	1716,5
12	905	-68,450	41,431	12073,83	4685,4
13	461	69,668	-68,45	19076,58	4853,6
14	454	36,078	69,668	1128,288	1301,6
15	920	-34,263	36,078	4947,856	1174
16	927	-50,143	-34,263	252,1744	2514,3
Сумма		-0,002	50,141	84921,85	37911,97

Фактическое значение критерия Дарбина-Уотсона для данной модели составляет:

$$d = \frac{84921,85}{37911,97} = 2,24.$$

Сформулируем гипотезы: H_0 – в остатках нет автокорреляции; H_1 – в остатках есть положительная автокорреляция; H_1^* – в остатках есть отрицательная автокорреляция. Зададим уровень значимости $\alpha = 0,05$. По таблице значений критерия Дарбина-Уотсона определим для числа наблюдений $n = 16$ и числа независимых параметров модели $k = 1$ (мы рассматриваем только зависимость от времени t) критические значения $d_L = 1,10$ и $d_U = 1,37$. Фактическое значение d-критерия Дарбина-Уотсона попадает в интервал $d_U < d < 4 - d_U$ (1,37<2,24<2,63). Следовательно, нет основания отклонять гипотезу H_0 об отсутствии автокорреляции в остатках.

Существует несколько ограничений на применение критерия Дарбина-Уотсона.

1. Он неприменим к моделям, включающим в качестве независимых переменных лаговые значения результативного признака.

2. Методика расчета и использования критерия Дарбина-Уотсона направлена только на выявление автокорреляции остатков первого порядка.

3. Критерий Дарбина-Уотсона дает достоверные результаты только для больших выборок.

Приложение А[6]

Случайные переменные

Дискретная случайная переменная

Ваше интуитивное понимание вероятности почти наверняка соответствует задачам этой книги, и поэтому мы опустим традиционный раздел чистой теории вероятностей, хотя он мог бы быть весьма увлекательным. Многие люди непосредственно сталкивались с вероятностями, участвуя в лотереях и азартных играх, и их заинтересованность в том, чем они занимались, часто приводила к удивительно высокой практической компетентности, обычно при полном отсутствии формальной подготовки.

Мы начнем непосредственно с дискретных случайных переменных. Случайная переменная – это любая переменная, значение которой не может быть точно предсказано. Дискретной называется случайная величина, имеющая определенный набор возможных значений. Пример – сумма выпавших очков при бросании двух игральных костей. Пример случайной величины, не являющейся дискретной, – температура в комнате. Она может принять любое из непрерывного диапазона значений и является примером непрерывной случайной величины. К рассмотрению таких величин в этом приложении мы перейдем позже.

Продолжая разговор о примере с двумя игральными костями, предположим, что одна из них зеленая, а другая – красная. Если их бросить, то возможны 36 элементарных исходов эксперимента, поскольку на зеленой кости может выпасть любое число от 1 до 6 и то же самое – на красной. Случайная переменная, определенная как их сумма, которую мы обозначим через x, может принимать только одно из 11 числовых значений — от 2 до

[6] За основу приложения А взят учебник [4].

12. Взаимосвязь между исходами эксперимента и значениями случайной величины в данном случае показана в табл. А.1.

Таблица А.1

Красная	Зеленая					
	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Предположив, что кости «правильные», мы можем воспользоваться табл. А.1 для определения вероятности каждого значения x. Поскольку на костях имеется 36 различных комбинаций, каждый исход имеет вероятность 1/36. Лишь одна из возможных комбинаций {зеленая=1, красная=1} дает сумму, равную 2, так что вероятность $X = 2$ равна 1/36. Чтобы получить сумму $x = 7$, нам потребуются сочетания {зеленая=1, красная=6}, либо {зеленая=2, красная=5}, либо {зеленая=3, красная=4}, либо {зеленая=4, красная=3}, либо {зеленая=5, красная=2}, либо {зеленая=6, красная=1}. В данном случае нас устроят 6 возможных исходов, и поэтому вероятность получения 7 равна 6/36. Все эти вероятности приведены в табл. А.2. Если все их сложить, то получится ровно 1. Это будет так, поскольку с вероятностью 100% рассматриваемая сумма примет одно из значений от 2 до 12.

Таблица А.2

Значения x	2	3	4	5	6	7	8	9	10	11	12
Вероятность	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Совокупность всех возможных значений случайной переменной описывается генеральной совокупностью, из которой извлекаются эти значения. В нашем случае генеральная совокупность – это набор чисел от 2 до 12.

Математическое ожидание дискретной случайной величины

Математическое ожидание дискретной случайной величины – это взвешенное среднее всех ее возможных значений, причем в качестве весового коэффициента берется вероятность соответствующего исхода. Вы можете рассчитать его, перемножив все возможные значения случайной величины на их вероятности и просуммировав полученные произведения. Математически если случайная величина обозначена как x, то ее математическое ожидание обозначается как $M(x)$ или m_x.

Предположим, что x может принимать n конкретных значений $(x_1, x_2, ..., x_n)$ и что вероятность получения x_i равна p_i. Тогда

$$M(x) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{i=1}^{n} x_i p_i . \qquad (A.1)$$

В случае с двумя костями величинами от x_1 до x_n были числа от 2 до 12. Математическое ожидание рассчитывается так:

$$M(x) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ... + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

Прежде чем пойти дальше, рассмотрим еще более простой пример случайной переменной – число очков, выпадающее при бросании лишь одной игральной кости.

В данном случае возможны шесть исходов: $x_1 = 1$, $x_2 = 2$, …, $x_6 = 6$. Каждый исход имеет вероятность 1/6, поэтому здесь

$$M(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3,5. \qquad (A.2)$$

В данном случае математическим ожиданием случайной переменной является число, которое само по себе не может быть получено при бросании кости.

Математическое ожидание случайной величины часто называют ее *средним по генеральной совокупности*. Для случайной величины x это значение часто обозначается как μ.

Математические ожидания функций дискретных случайных переменных

Пусть $g(x)$ – некоторая функция от x. Тогда $M(g(x))$ – математическое ожидание $g(x)$ записывается как

$$M(g(x)) = \sum g(x_i)\, p_i, \qquad \text{(A.3)}$$

где суммирование производится по всем возможным значениям x. В табл. А.3 показана последовательность практического расчета *математического ожидания функции* от x.

Таблица А.3

x	Вероятность	Функция от x	Функция, взвешенная по вероятности
1	**2**	**3**	**4**
x_1	p_1	$g(x_1)$	$g(x_1)\, p_1$
x_2	p_2	$g(x_2)$	$g(x_2)\, p_2$
…	…	…	…
x_n	p_n	$g(x_n)$	$g(x_n)\, p_n$
Всего			$M(g(X)) = \sum g(x_i)\, p_i$

Предположим, что x может принимать n различных значений от x_1 до x_n с соответствующими вероятностями от p_1 до p_n. В первой колонке записываются все возможные значения x. Во второй – записываются соответствующие вероятности. В третьей колонке рассчитываются значения функции для соответствующих величин x. В четвертой колонке перемножаются числа из колонок 2 и 3. Ответ приводится в суммирующей строке колонки 4.

Рассчитаем математическое ожидание величины x^2. Для этого рассмотрим пример с числами, выпадающими при бросании одной кости. Использовав схему, приведенную в табл. А.3, заполним табл. А.4.

Таблица А.4

x_i	p_i	x_i^2	$x_i^2 p_i$
1	**2**	**3**	**4**
1	1/6	1	0,167
2	1/6	4	0,667
3	1/6	9	1,500
4	1/6	16	2,667
5	1/6	25	4,167
6	1/6	36	6,000
Всего			15,167

В четвертой ее колонке даны шесть значений x^2, взвешенных по соответствующим вероятностям, которые в данном примере все равняются 1/6. По определению, величина $M\left(x^2\right)$ равна $\sum x_i^2 p_i$, она приведена как сумма в четвертой колонке и равна 15,167.

Математическое ожидание x, как уже было показано, равно 3,5, и 3,5 в квадрате равно 12,25. Таким образом, величина $M\left(x^2\right)$ не равна μ^2, и, следовательно, нужно аккуратно проводить различия между $M\left(x^2\right)$ и $\left\{M\left(x\right)\right\}^2$.

Правила расчета математического ожидания

Существуют три правила, которые часто используются. Эти правила практически самоочевидны, и они одинаково применимы для дискретных и непрерывных случайных переменных.

Правило 1. Математическое ожидание суммы нескольких переменных равно сумме их математических ожиданий. Например, если имеются три случайные переменные x, y и z, то

$$M\left(x+y+z\right)=M\left(x\right)+M\left(y\right)+M\left(z\right). \qquad \text{(A.4)}$$

Правило 2. Если случайная переменная умножается на константу, то ее математическое ожидание умножается на ту же константу. Если x – случайная переменная и a – константа, то

$$M\left(a\cdot x\right)=a\cdot M\left(x\right). \qquad \text{(A.5)}$$

Правило 3. Математическое ожидание константы есть она сама. Например, если a – константа, то

$$M\left(a\right)=a. \qquad \text{(A.6)}$$

Следствие из трех правил:

$$M\left(a+b\cdot x\right)=a+b\cdot M\left(x\right).$$

Независимость случайных переменных

Две случайные переменные x и y называются независимыми, если

$$M\left(f\left(x\right)\cdot g\left(y\right)\right)=M\left(f\left(x\right)\right)\cdot M\left(g\left(y\right)\right) \qquad \text{(A.7)}$$

для любых функций $f\left(x\right)$ и $g\left(y\right)$. Из независимости следует как важный частный случай, что $M\left(x\cdot y\right)=M\left(x\right)\cdot M\left(y\right)$.

Теоретическая дисперсия дискретной случайной переменной

Теоретическая дисперсия является мерой разброса для вероятностного распределения. Она определяется как математическое ожидание квадрата разности между величиной x и ее средним, т.е. величины $\left(x-\mu\right)^2$, где μ – математическое ожидание x. Дисперсия обычно обозначается как σ_x^2 или $D\left(x\right)$, и если ясно, о какой переменной идет речь, то нижний индекс может быть опущен:

$$\sigma_x^2\equiv D\left(x\right)=M\left(\left(x-\mu\right)^2\right)=\sum_{i=1}^{n}\left(x_i-\mu\right)^2\,p_i\,. \qquad \text{(A.8)}$$

Из σ_x^2 можно получить σ_x – *среднее квадратическое отклонение* – столь же распространенную меру разброса для распределения вероятностей; среднее квадратическое отклонение случайной переменной есть квадратный корень из ее дисперсии.

Мы проиллюстрируем расчет дисперсии на примере с одной игральной костью. Поскольку $\mu = M(x)$, то $(x-\mu)^2$ в этом случае равно $(x-3{,}5)^2$. Мы рассчитаем математическое ожидание величины $(x-3{,}5)^2$, используя схему, представленную в табл. А.5. Дополнительный столбец $(x-\mu)$ представляет определенный этап расчета $(x-\mu)^2$. Суммируя последний столбец в табл. I.5, получим значение дисперсии σ_x^2, равное 2,92. Следовательно, стандартное отклонение (σ_x) равно $\sqrt{2{,}92}$, то есть 1,71.

Таблица А.5

x_i	p_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(x_i-\mu)^2 p_i$
1	**2**	**3**	**4**	**5**
1	1/6	–2,5	6,25	1,042
2	1/6	–1,5	2,25	0,375
3	1/6	–0,5	0,25	0,042
4	1/6	0,5	0,25	0,042
5	1/6	1,5	2,25	0,375
6	1/6	2,5	6,25	1,042
Всего				2,92

Одним из важных приложений правил расчета математического ожидания является формула расчета теоретической дисперсии случайной переменной, которая может быть записана как

$$\sigma_x^2 = M(x^2) - \mu^2. \qquad \text{(А.9)}$$

Это выражение иногда оказывается более удобным, чем первоначальное определение. Доказательство предоставляется читателю в качестве упражнения.

Вероятность в непрерывном случае

С дискретными случайными переменными очень легко обращаться, поскольку они по определению принимают значения из некоторого конечного набора. Каждое из этих значений связано с определенной вероятностью, характеризующей его «вес». Если эти «веса» известны, то не составит труда рассчитать *теоретическое среднее* (математическое ожидание) и дисперсию.

Вы можете представить указанные «веса» как определенные количества «пластичной массы», равные вероятностям соответствующих значений. Сумма вероятностей и, следовательно, суммарный «вес» этой «массы» равен единице. Это показано на рис. А.1 для примера, где величина есть сумма очков, выпавших при бросании двух игральных костей. Величина x принимает значения от 2 до 12, и для всех этих значений показано количество соответствующей «массы».

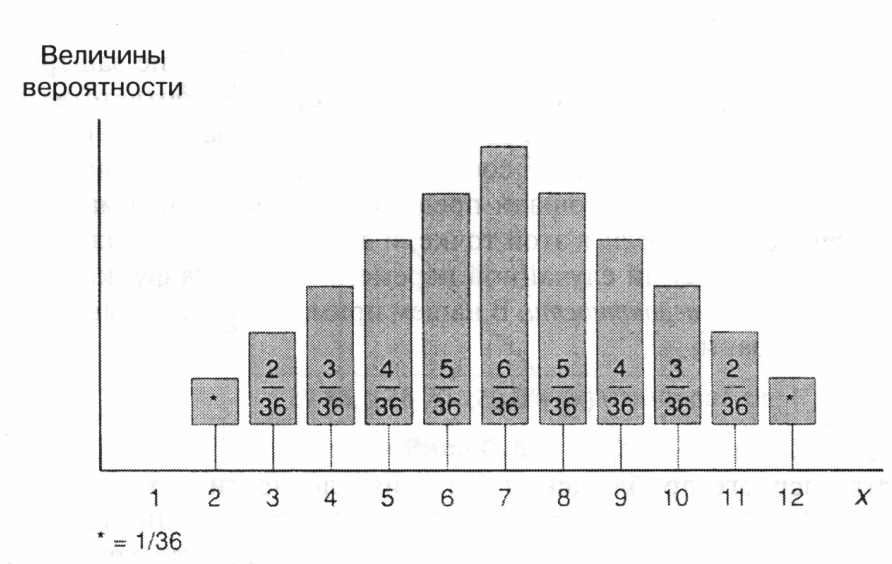


Рис. А.1.

К сожалению, анализ часто проводится для непрерывных случайных величин, которые могут принимать бесконечное число значений. Поскольку невозможно представить себе «пластичную массу», разделенную на бесконечное число частей, используем далее другой подход.

Проиллюстрируем наши рассуждения на примере температуры в комнате. Для определенности предположим, что она меняется в пределах от 55 до 75° по Фаренгейту, и вначале допустим, что все значения в этом диапазоне равновероятны.

Поскольку число различных значений, принимаемых показателем температуры, бесконечно, здесь бессмысленно пытаться разделить «пластичную массу» на малые части. Вместо этого можно «размазать» ее по всему диапазону. Поскольку все температуры от 55 до 75° F равновероятны, она должна быть «размазана» равномерно, как это показано на рис. А.2.

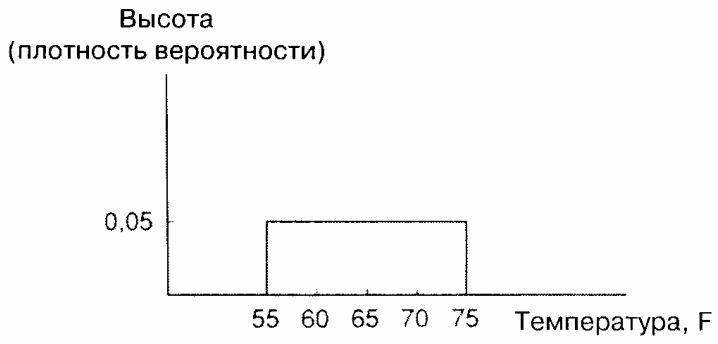


Рис. А.2.

В этом примере, как и во всех остальных, мы будем полагать, что «пластичная масса размазана» на единичной площади. Это связано с тем, что совокупная вероятность всегда равняется единице. В данном случае наша «масса» покрыла прямоугольник, и поскольку основание этого прямоугольника равно 20, его высота h определяется из соотношения:

$$20 \cdot h = 1, \qquad \text{(A.10)}$$

так как произведение основания и высоты равно площади. Следовательно, высота равна 0,05, как это показано на рисунке.

Найдя высоту прямоугольника, мы можем ответить на вопросы типа: с какой вероятностью температура будет находиться в диапазоне от 65 до 70°F? Ответ определяется величиной «замазанной» площади (или, говоря более формально, *совокупной вероятностью*), лежащей в диапазоне от 65 до 70°F, представленной заштрихованной фигурой на рис. А.3. Основание заштрихованного прямоугольника равно 5, его высота равна 0,05 и,

соответственно, площадь – 0,25. Искомая вероятность равна 1/4, что в любом случае очевидно, поскольку промежуток от 65 до 70°F составляет 1/4 всего диапазона.

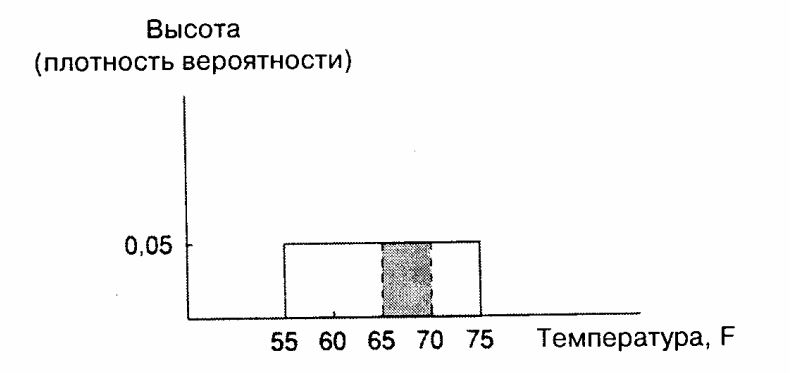


Рис. А.3.

Высота заштрихованной площади представляет то, что формально называется *плотностью вероятности* в этой точке, и если эта высота может быть записана как функция значений случайной переменной, то эта функция называется *функцией плотности вероятности*. В нашем примере она записывается как $f(x)$, где x – температура, и

$$f(x) = 0,05; \qquad 55 \le x \le 75. \tag{A.11}$$

В качестве первого приближения функция плотности вероятности показывает вероятность нахождения случайной переменной внутри единичного интервала вокруг данной точки. В нашем примере эта функция всюду равна 0,05, откуда вытекает, что температура находится, например, между 60 и 61°F с вероятностью 0,05.

В нашем случае график функции плотности вероятности горизонтален, и ее указанная интерпретация точна, однако в общем случае эта функция непрерывно меняется, и ее интерпретация дает лишь приближение. Далее мы рассмотрим пример, когда эта функция непостоянна, поскольку не все температуры равновероятны. Предположим, что центральное отопление работает таким образом, что температура никогда не падает ниже 65°F, а в жаркие дни температура превосходит этот уровень, не превышая, как и ранее, 75°F. Мы будем считать, что плотность вероятности максимальна при

температуре 65°F и далее она равномерно убывает до нуля при 75°F (рис. А.4).

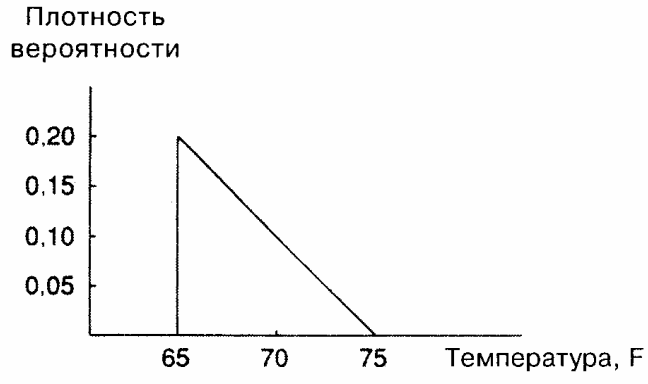


Рис. А.4.

Общая «замазанная» площадь, как всегда, равна единице, поскольку совокупная вероятность равна единице. Площадь треугольника равна половине произведения основания на высоту, поэтому получаем:

$$\frac{1}{2}\cdot 10\cdot h = 1, \qquad \text{(A.12)}$$

и высота при 65°F равна 0,20.

Предположим вновь, что мы хотим знать вероятность нахождения температуры в промежутке между 65 и 70°F. Она представлена заштрихованной площадью на рис. А.5, и если вы немного помните геометрию, то сможете проверить, что она равна 0,75. Если вы предпочитаете процентное измерение, то это означает, что с вероятностью 75% температура попадет в диапазон 65-70°F и только с вероятностью 25% – в диапазон 70-75°F.

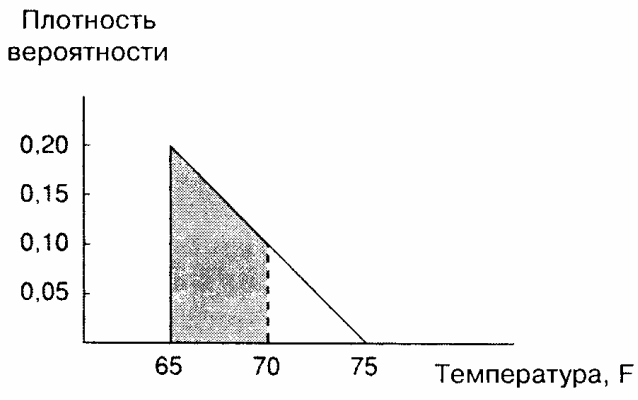


Рис. А.5.

В данном случае функция плотности вероятности записывается как $f(x)$, где

$$f(x) = 1{,}5 - 0{,}02x; \qquad 65 \le x \le 75. \tag{A.13}$$

Прежде чем продолжить изложение, упомянем о хорошей и плохой новостях. «Плохая новость» – это то, что если вы хотите рассчитать вероятности для более сложных функций с криволинейными графиками, то элементарная геометрия становится неприменимой. Вообще говоря, вы должны воспользоваться интегральным исчислением или специальными таблицами (если последние существуют). Интегральное исчисление используется также и при определении математического ожидания и дисперсии непрерывной случайной величины.

«Хорошая новость» – в том, что специальные таблицы существуют для всех функций, которые будут интересовать нас на практике. Кроме того, математическое ожидание и дисперсия имеют практически тот же смысл для непрерывных случайных величин, что и для дискретных, для них верны те же самые правила.

Постоянная и случайная составляющие случайной переменной

Часто вместо рассмотрения случайной величины как единого целого можно и удобно разбить ее на постоянную и чисто случайную составляющие, где постоянная составляющая всегда есть ее математическое ожидание. Если x – случайная переменная и μ – ее математическое ожидание, то декомпозиция случайной величины записывается следующим образом:

$$x = \mu + \varepsilon, \tag{A.14}$$

где ε – чисто случайная составляющая.

Конечно, можно было бы посмотреть на это по-другому и сказать, что случайная составляющая ε определяется как разность между x и μ

$$\varepsilon = x - \mu. \tag{A.15}$$

Из определения следует, что математическое ожидание величины ε равно нулю:

$$M(\varepsilon) = M(x - \mu) = M(x) - M(\mu) = \mu - \mu = 0.$$

Поскольку весь разброс значений x обусловлен ε, неудивительно, что теоретическая дисперсия x равна теоретической дисперсии ε. Последнее нетрудно доказать. По определению,

$$\sigma_x^2 = M\left((x - \mu)^2\right) = M\left(\varepsilon^2\right)$$

и

$$\sigma_\varepsilon^2 = M\left((\varepsilon - M(\varepsilon))^2\right) = M\left((\varepsilon - 0)^2\right) = M\left(\varepsilon^2\right).$$

Таким образом, σ^2 может быть эквивалентно определена как дисперсия x или ε.

Обобщая, можно утверждать, что если x – случайная переменная, определенная по формуле (А.14), где μ – заданное число и ε – случайный член с $M(\varepsilon) = 0$ и $\sigma_\varepsilon^2 = \sigma^2$, то математическое ожидание величины x равно μ, а дисперсия – σ^2.

Способы оценивания и оценки

До сих пор мы предполагали, что имеется точная информация о рассматриваемой случайной переменной, в частности – об ее распределении вероятностей (в случае дискретной переменной) или о функции плотности распределения (в случае непрерывной переменной). С помощью этой информации можно рассчитать теоретическое математическое ожидание, дисперсию и любые другие характеристики, в которых мы можем быть заинтересованы.

Однако на практике, за исключением искусственно простых случайных величин (таких, как число выпавших очков при бросании игральной кости), мы не знаем точного вероятностного распределения или плотности

распределения вероятностей. Это означает, что неизвестны также и теоретическое математическое ожидание, и дисперсия. Мы, тем не менее, можем нуждаться в оценках этих или других теоретических характеристик генеральной совокупности.

Процедура оценивания всегда одинакова. Берется выборка из n наблюдений, и с помощью подходящей формулы рассчитывается оценка нужной характеристики. Нужно следить за терминами, делая важное различие между способом или формулой оценивания и рассчитанным по ней для данной выборки числом, являющимся значением оценки. *Способ оценивания* – это общее правило, или формула, в то время как *значение оценки* – это конкретное число, которое меняется от выборки к выборке.

В табл. А.6 приведены формулы оценивания для двух важнейших характеристик генеральной совокупности. *Выборочное среднее* $\overline{x}$ обычно дает оценку для математического ожидания, а формула s^2 – оценку дисперсии генеральной совокупности.

Таблица А.6

Характеристики генеральной совокупности	Формулы оценивания
Среднее, μ	$\overline{x} = \frac{1}{n}\sum x_i$
Дисперсия, σ^2	$s^2 = \frac{1}{n-1}\sum \left(x_i - \overline{x}\right)^2$

Отметим, что это обычные формулы оценки математического ожидания и дисперсии генеральной совокупности, однако не единственные. Возможно, вы настолько привыкли использовать $\overline{x}$ в качестве оценки для μ, что даже не задумывались об альтернативах. Конечно, не все формулы оценки, которые можно представить, одинаково хороши. Причина, по которой в действительности используется $\overline{x}$, в том, что эта оценка в наилучшей степени соответствует двум очень важным критериям – несмещенности и эффективности. Эти критерии будут рассмотрены ниже.

Оценки как случайные величины

Получаемая *оценка* представляет частный случай случайной переменной. Причина здесь в том, что сочетание значений x в выборке случайно, поскольку x – случайная переменная и, следовательно, случайной величиной является и функция набора ее значений. Возьмем, например, $\overline{x}$ – оценку математического ожидания:

$$\overline{x} = \frac{1}{n}\left(x_1 + x_1 + ... + x_n\right).$$

Выше мы показали, что величина x в i-м наблюдении может быть разложена на две составляющие: постоянную часть μ и чисто случайную составляющую ε_i:

$$x_i = \mu + \varepsilon_i. \qquad \text{(A.17)}$$

Следовательно,

$$\overline{x} = \mu + \overline{\varepsilon}, \qquad \text{(A.18)}$$

где $\overline{\varepsilon}$ – выборочное среднее величин ε_i.

Отсюда можно видеть, что $\overline{x}$, подобно x, имеет как фиксированную, так и чисто случайную составляющие. Ее фиксированная составляющая – μ, то есть математическое ожидание x, а ее случайная составляющая – $\overline{\varepsilon}$, то есть среднее значение чисто случайной составляющей в выборке.

Функции плотности вероятности для x и $\overline{x}$ показаны на одинаковых графиках (рис. А.6). Как показано на рисунке, величина x считается нормально распределенной. Можно видеть, что распределения, как x, так и $\overline{x}$, симметричны относительно μ – теоретического среднего. Разница между ними в том, что распределение $\overline{x}$ уже и выше. Величина $\overline{x}$, вероятно, должна быть ближе к μ, чем значение единичного наблюдения x, поскольку ее случайная составляющая $\overline{\varepsilon}$ есть среднее от чисто случайных составляющих ε_1, ε_2, ..., ε_n в выборке, которые, по-видимому, «гасят» друг

друга при расчете среднего. Далее теоретическая дисперсия величины $\overline{\varepsilon}$ составляет лишь часть теоретической дисперсии ε.

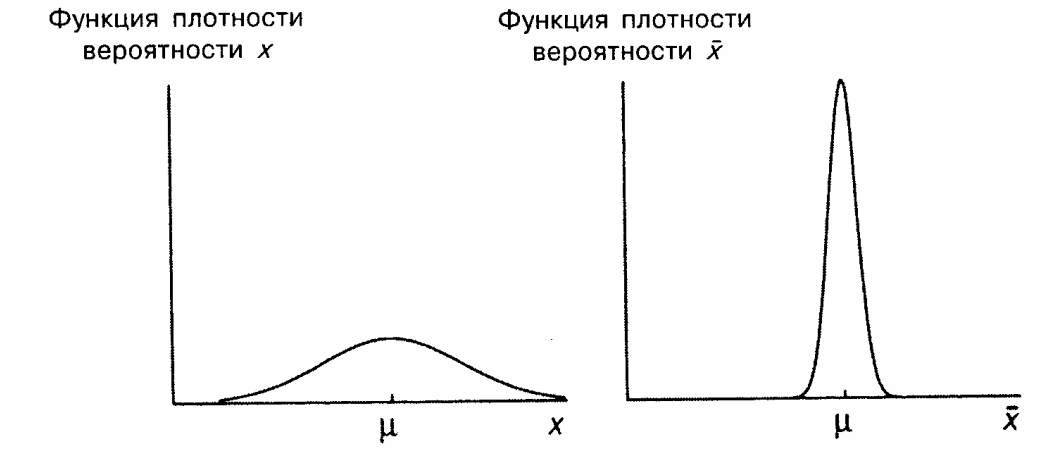


Рис. А.6.

Величина s^2 – оценка теоретической дисперсии x – также является случайной переменной. Вычитая (А.18) из (А.17), имеем:

$$x_i - \overline{x} = \varepsilon_i - \overline{\varepsilon}\,.$$

Следовательно,

$$s^2 = \frac{1}{n-1}\sum\left(x_i - \overline{x}\right)^2 = \frac{1}{n-1}\sum\left(\varepsilon_i - \overline{\varepsilon}\right)^2.$$

Таким образом, s^2 зависит от (и только от) чисто случайной составляющей наблюдений x в выборке. Поскольку эти составляющие меняются от выборки к выборке, также от выборки к выборке меняется и величина оценки s^2.

Несмещенность

Поскольку оценки являются случайными переменными, их значения лишь по случайному совпадению могут в точности равняться характеристикам генеральной совокупности. Обычно будет присутствовать определенная ошибка, которая может быть большой или малой, положительной или отрицательной, в зависимости от чисто случайных составляющих величин x в выборке.

Хотя это и неизбежно, на интуитивном уровне желательно, тем не менее, чтобы оценка в среднем за достаточно длительный период была аккуратной. Выражаясь формально, мы хотели бы, чтобы математическое ожидание оценки равнялось бы соответствующей характеристике генеральной совокупности. Если это так, то оценка называется *несмещенной*. Если это не так, то оценка называется *смещенной*, и разница между ее математическим ожиданием и соответствующей теоретической характеристикой генеральной совокупности называется *смещением*.

Начнем с выборочного среднего. Является ли оно несмещенной оценкой теоретического среднего? Равны ли $M\left(x\right)$ и μ? Да, это так, что непосредственно вытекает из (А.18).

Величина x включает две составляющие – μ и $\overline{\varepsilon}$. Значение $\overline{\varepsilon}$ равно средней чисто случайных составляющих величин x в выборке, и, поскольку математическое ожидание такой составляющей в каждом наблюдении равно нулю, математическое ожидание $\overline{\varepsilon}$ равно нулю. Следовательно,

$$M\left(\overline{x}\right)=M\left(\mu+\overline{\varepsilon}\right)=M\left(\mu\right)+M\left(\overline{\varepsilon}\right)=\mu+0=\mu. \qquad \text{(А.19)}$$

Тем не менее полученная оценка – не единственно возможная несмещенная оценка μ. Предположим для простоты, что у нас есть выборка всего из двух наблюдений – x_1 и x_2. Любое взвешенное среднее наблюдений x_1 и x_2 было бы несмещенной оценкой, если сумма весов равна единице. Чтобы показать это, предположим, что мы построили обобщенную формулу оценки:

$$Z=\lambda_1 x_1+\lambda_2 x_2. \qquad \text{(А.20)}$$

Математическое ожидание Z равно:

$$M\left(Z\right)=M\left(\lambda_1 x_1+\lambda_2 x_2\right)=\lambda_1 M\left(x_1\right)+\lambda_2 M\left(x_2\right)=\left(\lambda_1+\lambda_2\right)\mu. \qquad \text{(А.21)}$$

Если сумма λ_1 и λ_2 равна единице, то мы имеем $M\left(Z\right)=\mu$ и Z является несмещенной оценкой μ.

Таким образом, в принципе число несмещенных оценок бесконечно. Как выбрать одну из них? Почему в действительности мы всегда используем выборочное среднее с $\lambda_1 = \lambda_2 = 0,5$? Возможно, вы полагаете, что было бы несправедливым давать разным наблюдениям различные веса или что подобной асимметрии следует избегать в принципе. Мы, однако, не заботимся здесь о справедливости или о симметрии как таковой. Дальше мы увидим, что имеется и более осязаемая причина.

До сих пор мы рассматривали только оценки теоретического среднего. Выше утверждалось, что величина s^2, определяемая в соответствии с табл. А.6, является оценкой теоретической дисперсии σ^2. Можно показать, что математическое ожидание s^2 равно σ^2, и эта величина является несмещенной оценкой теоретической дисперсии, если наблюдения в выборке независимы друг от друга. Доказательство этого математически несложно, но трудоемко, и поэтому мы его опускаем.

Эффективность

Несмещенность – желательное свойство оценок, но это не единственное такое свойство. Еще одна важная их сторона – это надежность. Конечно, немаловажно, чтобы оценка была точной в среднем за длительный период, но, как однажды заметил Дж. М. Кейнс, «в долгосрочном периоде мы все умрем». Мы хотели бы, чтобы наша оценка с максимально возможной вероятностью давала бы близкое значение к теоретической характеристике, что означает желание получить функцию плотности вероятности, как можно более «сжатую» вокруг истинного значения. Один из способов выразить это требование – сказать, что мы хотели бы получить сколь возможно малую дисперсию.

Предположим, что мы имеем две оценки теоретического среднего, рассчитанные на основе одной и той же информации, что обе они являются несмещенными и что их функции плотности вероятности показаны на рис.

A.7. Поскольку функция плотности вероятности для оценки B более «сжата», чем для оценки A, с ее помощью мы скорее получим более точное значение. Формально говоря, эта оценка более эффективна.

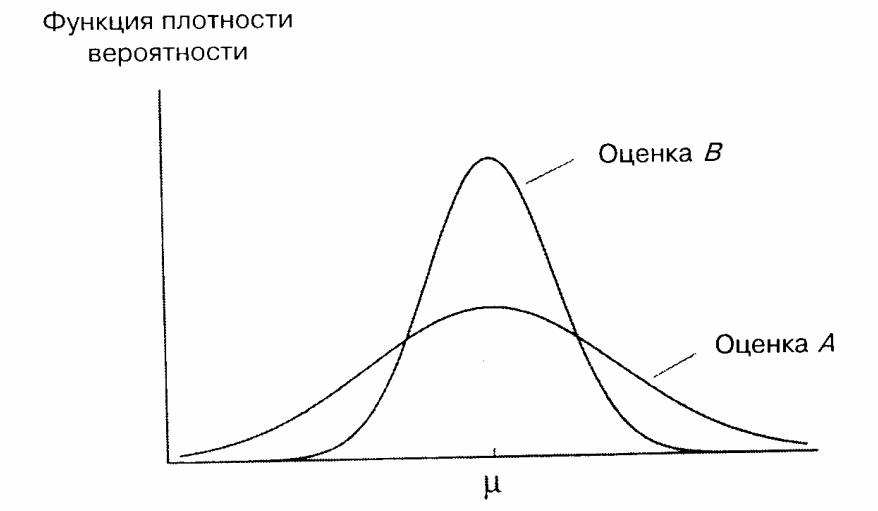


Рис. A.7.

Важно заметить, что мы использовали здесь слово «скорее». Даже хотя оценка B более эффективна, это не означает, что она всегда дает более точное значение. При определенном стечении обстоятельств значение оценки A может быть ближе к истине. Однако вероятность того, что оценка A окажется более точной, чем B, составляет менее 50%.

Это напоминает вопрос о том, пользоваться ли ремнями безопасности при управлении автомобилем. Множество обзоров в разных странах показало, что значительно менее вероятно погибнуть или получить увечья в дорожном происшествии, если воспользоваться ремнями безопасности. В то же время не раз отмечались странные случаи, когда не сделавший этого индивид чудесным образом уцелел, но погиб бы, будучи пристегнут ремнями. Упомянутые обзоры не отрицают этого. В них лишь делается вывод, что преимущество на стороне тех, кто пользуется ремнями безопасности. Подобным же преимуществом обладает и эффективная оценка. (Неприятный комментарий: в тех странах, где пользование ремнями безопасности сделано обязательным, сократилось предложение для трансплантации почек людей, ставших жертвами аварий.)

Мы говорили о желании получить оценку как можно с меньшей дисперсией, и эффективная оценка – это та, у которой дисперсия минимальна. Сейчас мы рассмотрим дисперсию обобщенной оценки теоретического среднего и покажем, что она минимальна в том случае, когда оба наблюдения имеют равные веса.

Если наблюдения x_1 и x_2 независимы, теоретическая дисперсия обобщенной оценки равна:

$$D(Z) = D(\lambda_1 x_1 + \lambda_2 x_2) = (\lambda_1^2 + \lambda_2^2)\sigma^2. \qquad \text{(A.21)}$$

Мы уже выяснили, что для несмещенности оценки необходимо равенство единице суммы λ_1 и λ_2. Следовательно, для несмещенных оценок $\lambda_2 = (1-\lambda_1)$ и

$$\lambda_1^2 + \lambda_2^2 = \lambda_1^2 + (1-\lambda_1)^2 = 2\lambda_1^2 - 2\lambda_1 + 1. \qquad \text{(A.22)}$$

Поскольку мы хотим выбрать λ_1 так, чтобы минимизировать дисперсию, нам нужно минимизировать при этом $(2\lambda_1^2 - 2\lambda_1 + 1)$. Эту задачу можно решить графически или с помощью дифференциального исчисления. В любом случае минимум достигается при $\lambda_1 = 0{,}5$. Следовательно, λ_2 также равно 0,5.

Итак, мы показали, что выборочное среднее имеет наименьшую дисперсию среди оценок рассматриваемого типа. Это означает, что оно имеет наиболее «сжатое» вероятностное распределение вокруг истинного среднего и, следовательно (в вероятностном смысле), наиболее точно. Строго говоря, выборочное среднее – это наиболее эффективная оценка среди всех несмещенных оценок. Конечно, мы показали это только для случая с двумя наблюдениями, но сделанные выводы верны для выборок любого размера, если наблюдения не зависят друг от друга.

Два заключительных замечания: во-первых, эффективность оценок можно сравнивать лишь тогда, когда они используют одну и ту же

информацию, например один и тот же набор наблюдений нескольких случайных переменных. Если одна из оценок использует в 10 раз больше информации, чем другая, то она вполне может иметь меньшую дисперсию, но было бы неправильно считать ее более эффективной. Во-вторых, мы ограничиваем понятие эффективности сравнением распределений несмещенных оценок. Существуют определения эффективности, обобщающие это понятие на случай возможного сравнения смещенных оценок, но в этом пособии мы придерживаемся данного простого определения.

Противоречия между несмещенностью и минимальной дисперсией

В данном обзоре мы уже выяснили, что для оценки желательны несмещенность и наименьшая возможная дисперсия. Эти критерии совершенно различны, и иногда они могут противоречить друг другу. Может случиться так, что имеются две оценки теоретической характеристики, одна из которых является несмещенной (A на рис. А.8), другая же смещена, но имеет меньшую дисперсию (B).

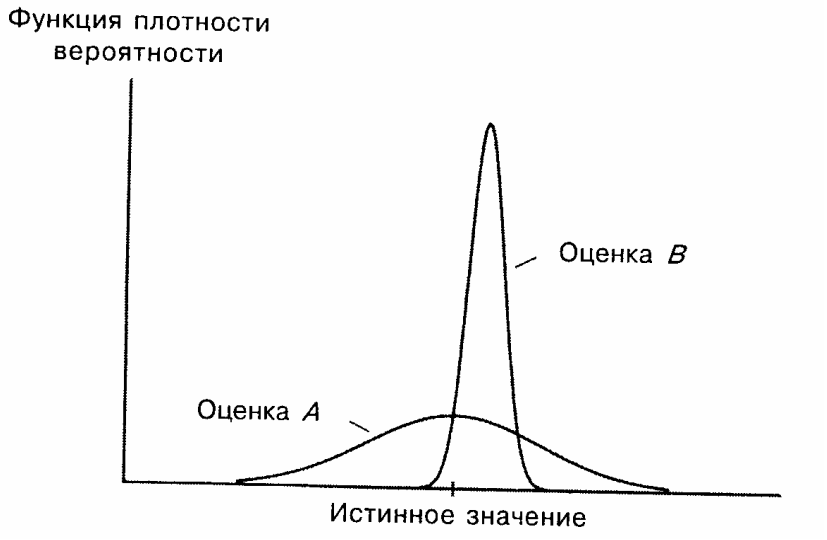


Рис. А.8.

Оценка A хороша своей несмещенностью, но преимуществом оценки B является то, что ее значения практически всегда близки к истинному значению. Какую из них вы бы выбрали?

Данный выбор зависит от обстоятельств. Если возможные ошибки вас не очень тревожат при условии, что за длительный период они «погасят» друг друга, то, по-видимому, вы выберете A. С другой стороны, если для вас приемлемы малые ошибки, но неприемлемы большие, то вам следует выбрать B.

Формально говоря, выбор определяется функцией потерь, стоимостью сделанной ошибки как функцией ее размера. Обычно выбирают оценку, дающую наименьшее ожидание потерь, и делается это путем взвешивания функции потерь по функции плотности вероятности. (Если вы не любите риск, то можете также пожелать учесть дисперсию потерь.)

Влияние увеличения размера выборки на точность оценок

Будем по-прежнему предполагать, что мы исследуем случайную переменную x с неизвестным математическим ожиданием μ и теоретической дисперсией σ^2 и что для оценивания μ используется $\bar{x}$. Каким образом точность оценки x зависит от числа наблюдений n?

Ответ неудивителен: при увеличении n оценка $\bar{x}$, вообще говоря, становится более точной. В единичном эксперименте большая по размеру выборка необязательно даст более точную оценку, чем меньшая выборка, – всегда может присутствовать элемент везения, – но общая тенденция должна быть именно такой. Поскольку дисперсия $\bar{x}$ выражается формулой σ^2/n (доказательство этого факта мы опускаем), она тем меньше, чем больше размер выборки, и, значит, тем сильнее «сжата» функция плотности вероятности для $\bar{x}$.

Это показано на рис. А.9. Мы предполагаем, что x нормально распределена со средним 25 и стандартным отклонением 50. Если размер выборки равен 25, то стандартное отклонение величины $\bar{x}$, равное $\sigma/\sqrt{n}$, составит: $50/\sqrt{25}=10$. Если размер выборки равен 100, то это стандартное отклонение равно 5. На рис. А.9 показаны соответствующие функции

плотности вероятности. Вторая ($n=100$) выше первой в окрестности μ, что говорит о более высокой вероятности получения с ее помощью аккуратной оценки. За пределами этой окрестности вторая функция всюду ниже первой.

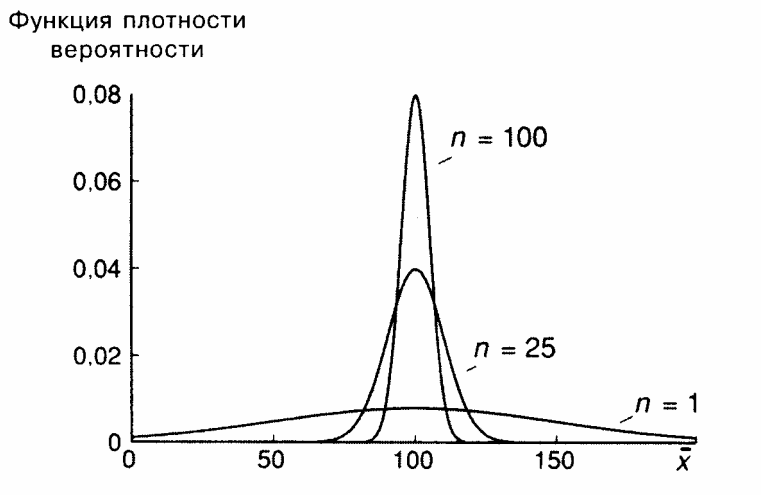


Рис. А.9.

Чем больше размер выборки, тем уже и выше будет график функции плотности вероятности для $\overline{x}$. Если n становится действительно большим, то график функции плотности вероятности будет неотличим от вертикальной прямой, соответствующей $\overline{x}=\mu$. Для такой выборки случайная составляющая x становится действительно очень малой, и поэтому $\overline{x}$ обязательно будет очень близкой к μ. Это вытекает из того факта, что стандартное отклонение $\overline{x}$, равное $\sigma/\sqrt{n}$, становится очень малым при больших n.

В пределе, при стремлении n к бесконечности, $\sigma/\sqrt{n}$ стремится к нулю и $\overline{x}$ стремится в точности к μ.

Состоятельность

Вообще говоря, если предел оценки по вероятности равен истинному значению характеристики генеральной совокупности, то эта оценка называется *состоятельной*. Иначе говоря, состоятельной называется такая

оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений.

В большинстве конкретных случаев несмещенная оценка является и состоятельной. Для этого можно построить контрпримеры, но они, как правило, будут носить искусственный характер.

Иногда бывает, что оценка, смещенная на малых выборках, является состоятельной (иногда состоятельной может быть даже оценка, не имеющая на малых выборках конечного математического ожидания). На рис. А.10 показано, как при различных размерах выборки может выглядеть распределение вероятностей. Тот факт, что при увеличении размера выборки распределение становится симметричным вокруг истинного значения, указывает на асимптотическую несмещенность. То, что в конечном счете оно превращается в единственную точку истинного значения, говорит о состоятельности оценки.

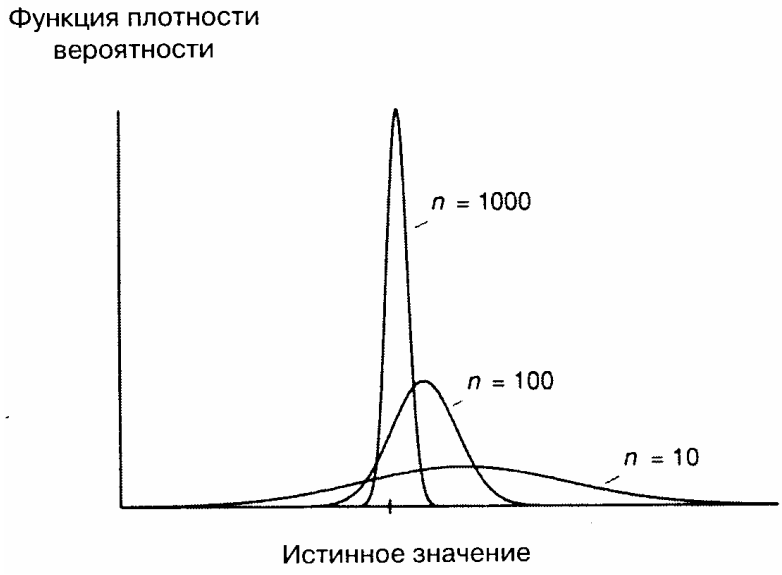


Рис. А.10.

Оценки, типа показанных на рис. А.10, весьма важны в регрессионном анализе. Иногда невозможно найти оценку, несмещенную на малых выборках. Если при этом вы можете найти хотя бы состоятельную оценку, это может быть лучше, чем не иметь никакой оценки, особенно если вы можете предположить направление смещения на малых выборках.

Нужно, однако, иметь в виду, что состоятельная оценка в принципе может на малых выборках работать хуже, чем несостоятельная (например, иметь большую среднеквадратичную ошибку), и поэтому требуется осторожность. Подобно тому, как вы можете предпочесть смещенную оценку несмещенной, если ее дисперсия меньше, вы можете предпочесть состоятельную, но смещенную оценку несмещенной или несостоятельную оценку им обеим (также в случае меньшей дисперсии).

Приложение В

Тестовые задания

Парная регрессия и корреляция

1. Наиболее наглядным видом выбора уравнения парной регрессии является:

а) аналитический;

б) графический;

в) экспериментальный (табличный).

2. Рассчитывать параметры парной линейной регрессии можно, если у нас есть:

а) не менее 5 наблюдений;

б) не менее 7 наблюдений;

в) не менее 10 наблюдений.

3. Суть метода наименьших квадратов состоит в:

а) минимизации суммы остаточных величин;

б) минимизации дисперсии результативного признака;

в) минимизации суммы квадратов остаточных величин.

4. Коэффициент линейного парного уравнения регрессии:

а) показывает среднее изменение результата с изменением фактора на одну единицу;

б) оценивает статистическую значимость уравнения регрессии;

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

5. На основании наблюдений за 50 семьями построено уравнение регрессии $\hat{y} = 284{,}56 + 0{,}672x$, ***где*** y – ***потребление,*** x – ***доход. Соответствуют ли знаки и значения коэффициентов регрессии теоретическим представлениям?***

а) да;

б) нет;

в) ничего определенного сказать нельзя.

6. Суть коэффициента детерминации r_{xy}^2 состоит в следующем:

а) оценивает качество модели из относительных отклонений по каждому наблюдению;

б) характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака;

в) характеризует долю дисперсии y, вызванную влиянием не учтенных в модели факторов.

7. Качество модели из относительных отклонений по каждому наблюдению оценивает:

а) коэффициент детерминации r_{xy}^2;

б) F-критерий Фишера;

в) средняя ошибка аппроксимации $\overline{A}$.

8. Значимость уравнения регрессии в целом оценивает:

а) F-критерий Фишера;

б) t-критерий Стьюдента;

в) коэффициент детерминации r_{xy}^2.

9. Классический метод к оцениванию параметров регрессии основан на:

а) методе наименьших квадратов:

б) методе максимального правдоподобия:

в) шаговом регрессионном анализе.

10. Остаточная сумма квадратов равна нулю:

а) когда правильно подобрана регрессионная модель;

б) когда между признаками существует точная функциональная связь;

в) никогда.

11. Объясненная (факторная) сумма квадратов отклонений в линейной парной модели имеет число степеней свободы, равное:

а) $n-1$;

б) 1;

в) $n-2$.

12. Остаточная сумма квадратов отклонений в линейной парной модели имеет число степеней свободы, равное:

а) $n-1$;

б) 1;

в) $n-2$.

13. Общая сумма квадратов отклонений в линейной парной модели имеет число степеней свободы, равное:

а) $n-1$;

б) 1;

в) $n-2$.

14. Для оценки значимости коэффициентов регрессии рассчитывают:

а) F-критерий Фишера;

б) t-критерий Стьюдента;

в) коэффициент детерминации r_{xy}^2.

15. Какое уравнение регрессии нельзя свести к линейному виду:

а) $\hat{y}_x = a + b \cdot \ln x$;

б) $\hat{y}_x = a \cdot x^b$:

в) $\hat{y}_x = a + b \cdot x^c$.

16. Какое из уравнений является степенным:

а) $\hat{y}_x = a + b \cdot \ln x$;

б) $\hat{y}_x = a \cdot x^b$:

в) $\hat{y}_x = a + b \cdot x^c$.

17. Параметр b в степенной модели является:

а) коэффициентом детерминации;

б) коэффициентом эластичности;

в) коэффициентом корреляции.

18. Коэффициент корреляции r_{xy} может принимать значения:

а) от –1 до 1;

б) от 0 до 1;

в) любые.

19. Для функции $y = a + \frac{b}{x} + \varepsilon$ средний коэффициент эластичности имеет вид:

а) $\overline{Э} = \frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$;

б) $\overline{Э} = -\frac{b}{a \cdot \overline{x} + b}$;

в) $\overline{Э} = -\frac{b \cdot \overline{x}}{a + b \cdot \overline{x}}$.

20. Какое из следующих уравнений нелинейно по оцениваемым параметрам:

а) $y = a + b \cdot x + \varepsilon$;

б) $y = a + b \cdot \ln x + \varepsilon$;

в) $y = a \cdot x^b \cdot \varepsilon$.

Множественная регрессия и корреляция

1. Добавление в уравнение множественной регрессии новой объясняющей переменной:

а) уменьшает значение коэффициента детерминации;

б) увеличивает значение коэффициента детерминации;

в) не оказывает никакого влияние на коэффициент детерминации.

2. Скорректированный коэффициент детерминации:

а) меньше обычного коэффициента детерминации;

б) больше обычного коэффициента детерминации;

в) меньше или равен обычному коэффициенту детерминации;

3. С увеличением числа объясняющих переменных скорректированный коэффициент детерминации:

а) увеличивается;

б) уменьшается;

в) не изменяется.

4. Число степеней свободы для остаточной суммы квадратов в линейной модели множественной регрессии равно:

а) $n-1$;

б) m;

в) $n-m-1$.

5. Число степеней свободы для общей суммы квадратов в линейной модели множественной регрессии равно:

а) $n-1$;

б) m;

в) $n-m-1$.

6. Число степеней свободы для факторной суммы квадратов в линейной модели множественной регрессии равно:

а) $n-1$;

б) m;

в) $n-m-1$.

7. Множественный коэффициент корреляции $R_{yx_1x_2}=0,9$. ***Определите, какой процент дисперсии зависимой переменной*** y ***объясняется влиянием факторов*** x_1 ***и*** x_2***:***

а) 90%;

б) 81%;

в) 19%.

8. Для построения модели линейной множественной регрессии вида $\hat{y} = a + b_1 x_1 + b_2 x_2$ ***необходимое количество наблюдений должно быть не менее:***

а) 2;

б) 7;

в) 14.

9. Стандартизованные коэффициенты регрессии β_i***:***

а) позволяют ранжировать факторы по силе их влияния на результат;

б) оценивают статистическую значимость факторов;

в) являются коэффициентами эластичности.

10. Частные коэффициенты корреляции:

а) характеризуют тесноту связи рассматриваемого набора факторов с исследуемым признаком;

б) содержат поправку на число степеней свободы и не допускают преувеличения тесноты связи;

в) характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании других факторов, включенных в уравнение регрессии.

11. Частный F***-критерий:***

а) оценивает значимость уравнения регрессии в целом;

б) служит мерой для оценки включения фактора в модель;

в) ранжирует факторы по силе их влияния на результат.

12. Несмещенность оценки параметра регрессии, полученной по МНК, означает:

а) что она характеризуется наименьшей дисперсией;

б) что математическое ожидание остатков равно нулю;

в) увеличение ее точности с увеличением объема выборки.

13. Эффективность оценки параметра регрессии, полученной по МНК, означает:

а) что она характеризуется наименьшей дисперсией;

б) что математическое ожидание остатков равно нулю;

в) увеличение ее точности с увеличением объема выборки.

14. Состоятельность оценки параметра регрессии, полученной по МНК, означает:

а) что она характеризуется наименьшей дисперсией;

б) что математическое ожидание остатков равно нулю;

в) увеличение ее точности с увеличением объема выборки.

15. Укажите истинное утверждение:

а) скорректированный и обычный коэффициенты множественной детерминации совпадают только в тех случаях, когда обычный коэффициент множественной детерминации равен нулю;

б) стандартные ошибки коэффициентов регрессии определяются значениями всех параметров регрессии;

в) при наличии гетероскедастичности оценки параметров регрессии становятся смещенными.

16. При наличии гетероскедастичности следует применять:

а) обычный МНК;

б) обобщенный МНК;

в) метод максимального правдоподобия.

17. Фиктивные переменные – это:

а) атрибутивные признаки (например, как профессия, пол, образование), которым придали цифровые метки;

б) экономические переменные, принимающие количественные значения в некотором интервале;

в) значения зависимой переменной за предшествующий период времени.

18. Если качественный фактор имеет три градации, то необходимое число фиктивных переменных:

а) 4;

б) 3;

в) 2.

Системы эконометрических уравнений

1. Наибольшее распространение в эконометрических исследованиях получили:

а) системы независимых уравнений;

б) системы рекурсивных уравнений;

в) системы взаимозависимых уравнений.

2. Эндогенные переменные – это:

а) предопределенные переменные, влияющие на зависимые переменные, но не зависящие от них, обозначаются через x.;

б) зависимые переменные, число которых равно числу уравнений в системе и которые обозначаются через y;

в) значения зависимых переменных за предшествующий период времени.

3. Экзогенные переменные – это:

а) предопределенные переменные, влияющие на зависимые переменные, но не зависящие от них, обозначаются через x;

б) зависимые переменные, число которых равно числу уравнений в системе и которые обозначаются через y;

в) значения зависимых переменных за предшествующий период времени.

4. Лаговые переменные – это:

а) предопределенные переменные, влияющие на зависимые переменные, но не зависящие от них, обозначаются через x.;

б) зависимые переменные, число которых равно числу уравнений в системе и которые обозначаются через y;

в) значения зависимых переменных за предшествующий период времени.

5. Для определения параметров структурную форму модели необходимо преобразовать в:

а) приведенную форму модели;

б) рекурсивную форму модели;

в) независимую форму модели.

6. Модель идентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

7. Модель неидентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

8. Модель сверхидентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

9. Уравнение идентифицируемо, если:

а) $D+1<H$;

б) $D+1=H$;

в) $D+1>H$.

10. Уравнение неидентифицируемо, если:

а) $D+1<H$;

б) $D+1=H$;

в) $D+1>H$.

11. Уравнение сверхидентифицируемо, если:

а) $D+1<H$;

б) $D+1=H$;

в) $D+1>H$.

12. Для определения параметров точно идентифицируемой модели:

а) применяется двушаговый МНК;

б) применяется косвенный МНК;

б) ни один из существующих методов применить нельзя.

13. Для определения параметров сверхидентифицируемой модели:

а) применяется двушаговый МНК;

б) применяется косвенный МНК;

б) ни один из существующих методов применить нельзя.

14. Для определения параметров неидентифицируемой модели:

а) применяется двушаговый МНК;

б) применяется косвенный МНК;

б) ни один из существующих методов применить нельзя.

Временные ряды

1. Аддитивная модель временного ряда имеет вид:

а) $Y=T\cdot S\cdot E$;

б) $Y=T+S+E$;

в) $Y=T\cdot S+E$.

2. Мультипликативная модель временного ряда имеет вид:

а) $Y = T \cdot S \cdot E$;

б) $Y = T + S + E$;

в) $Y = T \cdot S + E$.

3. Коэффициент автокорреляции:

а) характеризует тесноту линейной связи текущего и предыдущего уровней ряда;

б) характеризует тесноту нелинейной связи текущего и предыдущего уровней ряда;

в) характеризует наличие или отсутствие тенденции.

4. Аддитивная модель временного ряда строится, если:

а) значения сезонной компоненты предполагаются постоянными для различных циклов;

б) амплитуда сезонных колебаний возрастает или уменьшается;

в) отсутствует тенденция.

5. Мультипликативная модель временного ряда строится, если:

а) значения сезонной компоненты предполагаются постоянными для различных циклов;

б) амплитуда сезонных колебаний возрастает или уменьшается;

в) отсутствует тенденция.

6. На основе поквартальных данных построена аддитивная модель временного ряда. Скорректированные значения сезонной компоненты за первые три квартала равны: 7 – I квартал, 9 – II квартал и –11 – III квартал. Значение сезонной компоненты за IV квартал есть:

а) 5;

б) –4;

в) –5.

7. На основе поквартальных данных построена мультипликативная модель временного ряда. Скорректированные значения сезонной компоненты за первые три квартала равны: 0,8 – I

квартал, 1,2 – II квартал и 1,3 – III квартал. Значение сезонной компоненты за IV квартал есть:

а) 0,7;

б) 1,7;

в) 0,9.

8. Критерий Дарбина-Уотсона применяется для:

а) определения автокорреляции в остатках;

б) определения наличия сезонных колебаний;

в) для оценки существенности построенной модели.

Приложение С

Вопросы к экзамену

1. Определение эконометрики. Эконометрический метод и этапы эконометрического исследования.
2. Парная регрессия. Способы задания уравнения парной регрессии.
3. Линейная модель парной регрессии. Смысл и оценка параметров.
4. Оценка существенности уравнения в целом и отдельных его параметров (F-критерий Фишера и t-критерий Стьюдента).
5. Прогноз по линейному уравнению регрессии. Средняя ошибка аппроксимации.
6. Нелинейная регрессия. Классы нелинейных регрессий.
7. Регрессии нелинейные относительно включенных в анализ объясняющих переменных.
8. Регрессии нелинейные по оцениваемым параметрам.
9. Коэффициенты эластичности для разных видов регрессионных моделей.
10. Корреляция и F-критерий Фишера для нелинейной регрессии.
11. Отбор факторов при построении уравнения множественной регрессии.
12. Оценка параметров уравнения множественной регрессии.
13. Множественная корреляция.
14. Частные коэффициенты корреляции.
15. F-критерий Фишера и частный F-критерий Фишера для уравнения множественной регрессии.
16. t-критерий Стьюдента для уравнения множественной регрессии.
17. Фиктивные переменные во множественной регрессии.
18. Предпосылки МНК: гомоскедастичность и гетероскедастичность.
19. Предпосылки МНК: автокорреляция остатков.
20. Обобщенный МНК.

21. Общие понятия о системах эконометрических уравнений.

22. Структурная и приведенная формы модели.

23. Проблема идентификации. Необходимое условие идентифицируемости.

24. Проблема идентификации. Достаточное условие идентифицируемости.

25. Методы оценки параметров структурной формы модели.

26. Основные элементы временного ряда.

27. Автокорреляция уровней временного ряда и выявление его структуры.

28. Моделирование сезонных колебаний: аддитивная модель временного ряда.

29. Моделирование сезонных колебаний: мультипликативная модель временного ряда.

30. Критерий Дарбина-Уотсона.

Приложение D

Варианты индивидуальных заданий

D.1. Парная регрессия и корреляция

Пример. По территориям региона приводятся данные за 199X г.

Таблица D.1

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Требуется:

1. Построить линейное уравнение парной регрессии y от x.

2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость параметров регрессии и корреляции с помощью F-критерия Фишера и t-критерия Стьюдента.

4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 107% от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике построить исходные данные и теоретическую прямую.

Решение

1. Для расчета параметров уравнения линейной регрессии строим расчетную таблицу D.2.

Таблица D.2

	x	y	yx	x^2	y^2	$\hat{y}_x$	$y-\hat{y}_x$	A_i
1	78	133	10374	6084	17689	149	-16	12,0
2	82	148	12136	6724	21904	152	-4	2,7
3	87	134	11658	7569	17956	157	-23	17,2
4	79	154	12166	6241	23716	150	4	2,6
5	89	162	14418	7921	26244	159	3	1,9
6	106	195	20670	11236	38025	174	21	10,8
7	67	139	9313	4489	19321	139	0	0,0
8	88	158	13904	7744	24964	158	0	0,0
9	73	152	11096	5329	23104	144	8	5,3
10	87	162	14094	7569	26244	157	5	3,1
11	76	159	12084	5776	25281	147	12	7,5
12	115	173	19895	13225	29929	183	-10	5,8
Итого	1027	1869	161808	89907	294377	1869	0	68,9
Среднее значение	85,6	155,8	13484,0	7492,3	24531,4	–	–	5,7
σ	12,84	16,05	–	–	–	–	–	–
σ^2	164,94	257,76	–	–	–	–	–	–

$$b=\frac{\overline{y\cdot x}-\overline{y}\cdot\overline{x}}{\overline{x^2}-\overline{x}^2}=\frac{13484-155,8\cdot 85,6}{7492,3-85,6^2}=\frac{147,52}{164,94}=0,89;$$

$$a=\overline{y}-b\cdot\overline{x}=155,8-0,89\cdot 85,6=79,62.$$

Получено уравнение регрессии: $y=79,62+0,89\cdot x$.

С увеличением среднедушевого прожиточного минимума на 1 руб. среднедневная заработная плата возрастает в среднем на 0,89 руб.

2. Тесноту линейной связи оценит коэффициент корреляции:

$$r_{xy}=b\cdot\frac{\sigma_x}{\sigma_y}=0,89\cdot\frac{12,84}{16,05}=0,712;\qquad r_{xy}^2=0,51.$$

Это означает, что 51% вариации заработной платы (y) объясняется вариацией фактора x – среднедушевого прожиточного минимума.

Качество модели определяет средняя ошибка аппроксимации:

$$\overline{A} = \frac{1}{n}\sum A_i = \frac{68,9}{12} = 5,74\%.$$

Качество построенной модели оценивается как хорошее, так как $\overline{A}$ не превышает 8-10%.

1. Оценку значимости уравнения регрессии в целом проведем с помощью F-критерия Фишера. Фактическое значение F-критерия:

$$F_{\text{факт}} = \frac{r_{xy}^2}{1-r_{xy}^2}\cdot(n-2) = \frac{0,51}{1-0,51}\cdot 10 = 10,41.$$

Табличное значение критерия при пятипроцентном уровне значимости и степенях свободы $k_1 = 1$ и $k_2 = 12-2 = 10$ составляет $F_{\text{табл}} = 4,96$. Так как $F_{\text{факт}} = 10,41 > F_{\text{табл}} = 4,96$, то уравнение регрессии признается статистически значимым.

Оценку статистической значимости параметров регрессии проведем с помощью t-статистики Стьюдента и путем расчета доверительного интервала каждого из показателей.

Табличное значение t-критерия для числа степеней свободы $df = n-2 = 12-2 = 10$ и $\alpha = 0,05$ составит $t_{\text{табл}} = 2,23$.

Определим случайные ошибки m_a, m_b, $m_{r_{xy}}$:

$$m_a = S_{\text{ост}}\cdot\frac{\sqrt{\sum x^2}}{n\cdot\sigma_x} = 12,6\cdot\frac{\sqrt{89907}}{12\cdot 12,84} = 24,5;$$

$$m_b = \frac{S_{\text{ост}}}{\sigma_x\cdot\sqrt{n}} = \frac{12,6}{12,95\cdot\sqrt{12}} = 0,281;$$

$$m_{r_{xy}} = \sqrt{\frac{1-r_{xy}^2}{n-2}} = \sqrt{\frac{1-0,51}{12-2}} = 0,219.$$

Тогда

$$t_a = \frac{a}{m_a} = \frac{79,616}{24,6} = 3,2;$$

$$t_b = \frac{b}{m_b} = \frac{0,89}{0,281} = 3,2;$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0,712}{0,219} = 3,3.$$

Фактические значения t-статистики превосходят табличное значение:

$$t_a = 3,2 > t_{\text{табл}} = 2,3; \quad t_b = 3,3 > t_{\text{табл}} = 2,3; \quad t_{r_{xy}} = 3,3 > t_{\text{табл}} = 2,3,$$

поэтому параметры a, b и r_{xy} не случайно отличаются от нуля, а статистически значимы.

Рассчитаем доверительные интервалы для параметров регрессии a и b. Для этого определим предельную ошибку для каждого показателя:

$$\Delta_a = t_{\text{табл}} \cdot m_a = 2,23 \cdot 24,5 = 54,64;$$

$$\Delta_b = t_{\text{табл}} \cdot m_b = 2,23 \cdot 0,281 = 0,62.$$

Доверительные интервалы

$$\gamma_a = a \pm \Delta_a = 79,62 \pm 54,64;$$

$$\gamma_{a_{\min}} = 79,62 - 54,64 = 24,98;$$

$$\gamma_{a_{\max}} = 79,62 + 54,64 = 134,26;$$

$$\gamma_b = b \pm \Delta_b = 0,89 \pm 0,62;$$

$$\gamma_{b_{\min}} = 0,89 - 0,62 = 0,27;$$

$$\gamma_{b_{\max}} = 0,89 + 0,62 = 1,51.$$

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью $p = 1 - \alpha = 0,95$ параметры a и b, находясь в указанных границах, не принимают нулевых значений, т.е. не являются статистически незначимыми и существенно отличны от нуля.

1. Полученные оценки уравнения регрессии позволяют использовать его для прогноза. Если прогнозное значение прожиточного минимума составит: $x_p = \overline{x} \cdot 1,07 = 85,6 \cdot 1,07 = 91,6$ руб., тогда прогнозное значение заработной платы составит: $\overline{y}_p = 79,62 + 0,89 \cdot 91,6 = 161,14$ руб.

1. Ошибка прогноза составит:

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1+\frac{1}{n}+\frac{\left(x_p-\overline{x}\right)^2}{\sum\left(x-\overline{x}\right)^2}} = 12,6 \cdot \sqrt{1+\frac{1}{12}+\frac{\left(91,6-85,6\right)^2}{12 \cdot 12,84^2}} = 13,22.$$

Предельная ошибка прогноза, которая в 95% случаев не будет превышена, составит:

$$\Delta_{\hat{y}_p} = t_{\text{табл}} \cdot m_{\hat{y}_p} = 2,23 \cdot 13,22 = 29,48.$$

Доверительный интервал прогноза:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm \Delta_{\hat{y}_p} = 161,14 \pm 29,48;$$

$$\gamma_{\hat{y}_{p\min}} = 161,14 - 29,48 = 131,66 \text{ руб.};$$

$$\gamma_{\hat{y}_{p\max}} = 161,14 + 29,48 = 190,62 \text{ руб.}$$

Выполненный прогноз среднемесячной заработной платы является надежным ($p = 1-\alpha = 1-0,05 = 0,95$) и находится в пределах от 131,66 руб. до 190,62 руб.

1. В заключение решения задачи построим на одном графике исходные данные и теоретическую прямую (рис. D.1):

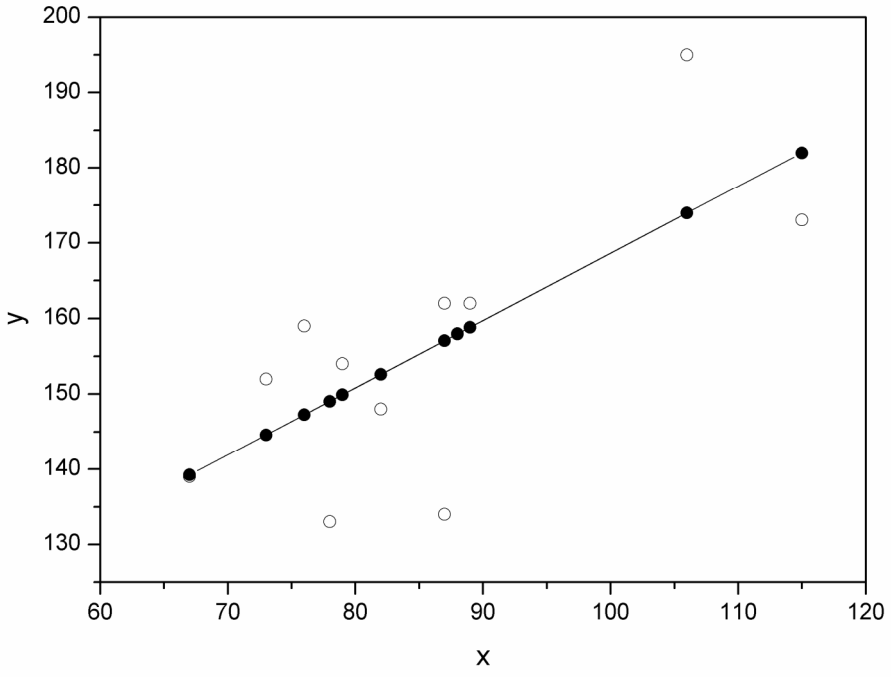


Рис. D.1.

Варианты индивидуальных заданий

Задача 1. По территориям региона приводятся данные за 199X г. (см. таблицу своего варианта).

Требуется:

1. Построить линейное уравнение парной регрессии y от x.

2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость параметров регрессии и корреляции с помощью F-критерия Фишера и t-критерия Стьюдента.

4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 107% от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике построить исходные данные и теоретическую прямую.

Вариант 1

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	81	124
2	77	131
3	85	146
4	79	139
5	93	143
6	100	159
7	72	135
8	90	152
9	71	127
10	89	154
11	82	127
12	111	162

Вариант 2

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	74	122
2	81	134
3	90	136
4	79	125
5	89	120
6	87	127
7	77	125
8	93	148
9	70	122
10	93	157
11	87	144
12	121	165

Вариант 3

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	77	123
2	85	152
3	79	140
4	93	142
5	89	157
6	81	181
7	79	133
8	97	163
9	73	134
10	95	155
11	84	132
12	108	165

Вариант 4

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	83	137
2	88	142
3	75	128
4	89	140
5	85	133
6	79	153
7	81	142
8	97	154
9	79	132
10	90	150
11	84	132
12	112	166

Вариант 5

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	79	134
2	91	154
3	77	128
4	87	138
5	84	133
6	76	144
7	84	160
8	94	149
9	79	125
10	98	163
11	81	120
12	115	162

Вариант 6

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	92	147
2	78	133
3	79	128
4	88	152
5	87	138
6	75	122
7	81	145
8	96	141
9	80	127
10	102	151
11	83	129
12	94	147

Вариант 7

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	75	133
2	78	125
3	81	129
4	93	153
5	86	140
6	77	135
7	83	141
8	94	152
9	88	133
10	99	156
11	80	124
12	112	156

Вариант 8

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	69	124
2	83	133
3	92	146
4	97	153
5	88	138
6	93	159
7	74	145
8	79	152
9	105	168
10	99	154
11	85	127
12	94	155

Вариант 9

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	78	133
2	94	139
3	85	141
4	73	127
5	91	154
6	88	142
7	73	122
8	82	135
9	99	142
10	113	168
11	69	124
12	83	130

Вариант 10

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	97	161
2	73	131
3	79	135
4	99	147
5	86	139
6	91	151
7	85	135
8	77	132
9	89	161
10	95	159
11	72	120
12	115	160

D.2. Множественная регрессия и корреляция

Пример. По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,0	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Решение

Для удобства проведения расчетов поместим результаты промежуточных расчетов в таблицу:

№	y	x_1	x_2	yx_1	yx_2	x_1x_2	x_1^2	x_2^2	y^2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7,0	3,9	10,0	27,3	70,0	39,0	15,21	100,0	49,0
2	7,0	3,9	14,0	27,3	98,0	54,6	15,21	196,0	49,0
3	7,0	3,7	15,0	25,9	105,0	55,5	13,69	225,0	49,0
4	7,0	4,0	16,0	28,0	112,0	64,0	16,0	256,0	49,0
5	7,0	3,8	17,0	26,6	119,0	64,6	14,44	289,0	49,0
6	7,0	4,8	19,0	33,6	133,0	91,2	23,04	361,0	49,0
7	8,0	5,4	19,0	43,2	152,0	102,6	29,16	361,0	64,0
8	8,0	4,4	20,0	35,2	160,0	88,0	19,36	400,0	64,0
9	8,0	5,3	20,0	42,4	160,0	106,0	28,09	400,0	64,0
10	10,0	6,8	20,0	68,0	200,0	136,0	46,24	400,0	100,0
11	9,0	6,0	21,0	54,0	189,0	126,0	36,0	441,0	81,0
12	11,0	6,4	22,0	70,4	242,0	140,8	40,96	484,0	121,0
13	9,0	6,8	22,0	61,2	198,0	149,6	46,24	484,0	81,0

1	2	3	4	5	6	7	8	9	10
14	11,0	7,2	25,0	79,2	275,0	180,0	51,84	625,0	121,0
15	12,0	8,0	28,0	96,0	336,0	224,0	64,0	784,0	144,0
16	12,0	8,2	29,0	98,4	348,0	237,8	67,24	841,0	144,0
17	12,0	8,1	30,0	97,2	360,0	243,0	65,61	900,0	144,0
18	12,0	8,5	31,0	102,0	372,0	263,5	72,25	961,0	144,0
19	14,0	9,6	32,0	134,4	448,0	307,2	92,16	1024,0	196,0
20	14,0	9,0	36,0	126,0	504,0	324,0	81,0	1296,0	196,0
Сумма	192	123,8	446	1276,3	4581	2997,4	837,74	10828,0	1958,0
Ср. знач.	9,6	6,19	22,3	63,815	229,05	149,87	41,887	541,4	97,9

Найдем средние квадратические отклонения признаков:

$$\sigma_y = \sqrt{\overline{y^2} - \overline{y}^2} = \sqrt{97,9 - 9,6^2} = 2,396;$$

$$\sigma_{x_1} = \sqrt{\overline{x_1^2} - \overline{x}_1^2} = \sqrt{41,887 - 6,19^2} = 1,890;$$

$$\sigma_{x_2} = \sqrt{\overline{x_2^2} - \overline{x}_2^2} = \sqrt{541,4 - 22,3^2} = 6,642.$$

1. Вычисление параметров линейного уравнения множественной регрессии.

Для нахождения параметров линейного уравнения множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2$$

необходимо решить следующую систему линейных уравнений относительно неизвестных параметров a, b_1, b_2:

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 = \sum y; \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum y x_1; \\ a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum y x_2 \end{cases}$$

либо воспользоваться готовыми формулами:

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}; \qquad b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2};$$

$$a = \overline{y} - b_1 \overline{x}_1 - b_2 \overline{x}_2.$$

Рассчитаем сначала парные коэффициенты корреляции:

$$r_{yx_1} = \frac{\operatorname{cov}(y,\, x_1)}{\sigma_y \cdot \sigma_{x_1}} = \frac{63{,}815 - 6{,}19 \cdot 9{,}6}{1{,}890 \cdot 2{,}396} = 0{,}970;$$

$$r_{yx_2} = \frac{\operatorname{cov}(y,\, x_2)}{\sigma_y \cdot \sigma_{x_2}} = \frac{229{,}05 - 22{,}3 \cdot 9{,}6}{6{,}642 \cdot 2{,}396} = 0{,}941;$$

$$r_{x_1x_2} = \frac{\operatorname{cov}(x_1,\, x_2)}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{149{,}87 - 6{,}19 \cdot 22{,}3}{1{,}890 \cdot 6{,}642} = 0{,}943.$$

Находим

$$b_1 = \frac{2{,}396}{1{,}890} \cdot \frac{0{,}970 - 0{,}941 \cdot 0{,}943}{1 - 0{,}943^2} = 0{,}946;$$

$$b_2 = \frac{2{,}396}{6{,}642} \cdot \frac{0{,}941 - 0{,}970 \cdot 0{,}943}{1 - 0{,}943^2} = 0{,}0856;$$

$$a = 9{,}6 - 0{,}946 \cdot 6{,}19 - 0{,}0856 \cdot 22{,}3 = 1{,}835.$$

Таким образом, получили следующее уравнение множественной регрессии:

$$\hat{y} = 1{,}835 + 0{,}946 \cdot x_1 + 0{,}0856 \cdot x_2.$$

Коэффициенты β_1 и β_2 стандартизованного уравнения регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon,$ находятся по формулам:

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0{,}946 \cdot \frac{1{,}890}{2{,}396} = 0{,}746;$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0{,}0856 \cdot \frac{6{,}642}{2{,}396} = 0{,}237.$$

Т.е. уравнение будет выглядеть следующим образом:

$$\hat{t}_y = 0{,}746 \cdot t_{x_1} + 0{,}237 \cdot t_{x_2}.$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что ввод в действие новых основных фондов оказывает большее влияние на выработку продукции, чем удельный вес рабочих высокой квалификации.

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

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_{x_i}}.$$

Вычисляем:

$$\overline{Э}_1 = 0,946 \cdot \frac{6,19}{9,6} = 0,61; \qquad \overline{Э}_2 = 0,0856 \cdot \frac{22,3}{9,6} = 0,20.$$

Т.е. увеличение только основных фондов (от своего среднего значения) или только удельного веса рабочих высокой квалификации на 1% увеличивает в среднем выработку продукции на 0,61% или 0,20% соответственно. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

1. Коэффициенты парной корреляции мы уже нашли:

$$r_{yx_1} = 0,970; \qquad r_{yx_2} = 0,941; \qquad r_{x_1x_2} = 0,943.$$

Они указывают на весьма сильную связь каждого фактора с результатом, а также высокую межфакторную зависимость (факторы x_1 и x_2 явно коллинеарны, т.к. $r_{x_1x_2} = 0,943 > 0,7$). При такой сильной межфакторной зависимости рекомендуется один из факторов исключить из рассмотрения.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии.

При двух факторах частные коэффициенты корреляции рассчитываются следующим образом:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0,970 - 0,941 \cdot 0,943}{\sqrt{\left(1 - 0,941^2\right) \cdot \left(1 - 0,943^2\right)}} = 0,734;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0,941 - 0,970 \cdot 0,943}{\sqrt{\left(1 - 0,970^2\right) \cdot \left(1 - 0,943^2\right)}} = 0,325.$$

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи. Именно по этой причине рекомендуется при наличии сильной коллинеарности (взаимосвязи) факторов исключать из исследования тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи.

Коэффициент множественной корреляции определить через матрицу парных коэффициентов корреляции:

$$R_{yx_1x_2} = \sqrt{1 - \frac{\Delta_r}{\Delta_{r_{11}}}},$$

где

$$\Delta_r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1x_2} \\ r_{yx_2} & r_{x_2x_1} & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta_{r_{11}} = \begin{vmatrix} 1 & r_{x_1x_2} \\ r_{x_2x_1} & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

$$\Delta_r = \begin{vmatrix} 1 & 0,970 & 0,941 \\ 0,970 & 1 & 0,943 \\ 0,941 & 0,943 & 1 \end{vmatrix} = 1 + 0,8607 + 0,8607 -$$

$$-0,8855 - 0,8892 - 0,9409 = 0,0058$$

$$\Delta_{r_{11}} = \begin{vmatrix} 1 & 0,943 \\ 0,943 & 1 \end{vmatrix} = 1 - 0,8892 = 0,1108.$$

Коэффициент множественной корреляции

$$R_{yx_1x_2} = \sqrt{1 - \frac{0,0058}{0,1108}} = 0,973.$$

Аналогичный результат получим при использовании других формул:

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{0,305}{5,74}} = 0,973;$$

$$R_{yx_1x_2} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = \sqrt{0,746 \cdot 0,970 + 0,237 \cdot 0,941} = 0,973;$$

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{yx_2 \cdot x_1}\right)} =$$

$$= \sqrt{1 - \left(1 - 0,970^2\right) \cdot \left(1 - 0,325^2\right)} = 0,973.$$

Коэффициент множественной корреляции показывает на весьма сильную связь всего набора факторов с результатом.

2. Нескорректированный коэффициент множественной детерминации $R^2_{yx_1x_2} = 0,947$ оценивает долю вариации результата за счет представленных в уравнении факторов в общей вариации результата. Здесь эта доля составляет $94,7\%$ и указывает на весьма высокую степень обусловленности вариации результата вариацией факторов, иными словами – на весьма тесную связь факторов с результатом.

Скорректированный коэффициент множественной детерминации

$$\hat{R}^2 = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-m-1)} = 1 - (1 - 0,947)\frac{20-1}{20-2-1} = 0,941$$

определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Он дает такую оценку тесноты связи, которая не зависит от числа факторов и поэтому может сравниваться по разным моделям с разным числом факторов. Оба коэффициента указывают на весьма высокую (более 94%) детерминированность результата y в модели факторами x_1 и x_2.

3. Оценку надежности уравнения регрессии в целом и показателя тесноты связи $R_{yx_1x_2}$ дает F-критерий Фишера:

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

В нашем случае фактическое значение F-критерия Фишера:

$$F_{\text{факт}} = \frac{0,973^2}{1-0,973^2} \cdot \frac{20-2-1}{2} = 151,88.$$

Получили, что $F_{\text{факт}} > F_{\text{табл}} = 3,49$ (при $n = 20$), т.е. вероятность случайно получить такое значение F-критерия не превышает допустимый уровень значимости 5%. Следовательно, полученное значение не случайно, оно сформировалось под влиянием существенных факторов, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи $R^2_{yx_1x_2}$.

4. С помощью частных F-критериев Фишера оценим целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1 при помощи формул:

$$F_{\text{част, } x_1} = \frac{R^2_{yx_1x_2} - R^2_{yx_2}}{1-R^2_{yx_1}} \cdot \frac{n-m-1}{m};$$

$$F_{\text{част, } x_2} = \frac{R^2_{yx_1x_2} - R^2_{yx_1}}{1-R^2_{yx_2}} \cdot \frac{n-m-1}{m}.$$

Найдем $R^2_{yx_1}$ и $R^2_{yx_2}$.

$$R^2_{yx_1} = r^2_{yx_1} = 0,970^2 = 0,941;$$

$$R^2_{yx_2} = r^2_{yx_2} = 0,941^2 = 0,885.$$

Имеем

$$F_{\text{част, } x_1} = \frac{0,947-0,885}{1-0,941} \cdot \frac{20-2-1}{2} = 8,9322;$$

$$F_{\text{част, } x_2} = \frac{0,947-0,941}{1-0,885} \cdot \frac{20-2-1}{2} = 0,4435.$$

Получили, что $F_{\text{част, } x_2} < F_{\text{табл}} = 3,49$. Следовательно, включение в модель фактора x_2 после того, как в модель включен фактор x_1 статистически нецелесообразно: прирост факторной дисперсии за счет дополнительного признака x_2 оказывается незначительным, несущественным; фактор x_2 включать в уравнение после фактора x_1 не следует.

Если поменять первоначальный порядок включения факторов в модель и рассмотреть вариант включения x_1 после x_2, то результат расчета частного F-критерия для x_1 будет иным. $F_{\text{част, } x_1} > F_{\text{табл}} = 3,49$, т.е. вероятность его случайного формирования меньше принятого стандарта $\alpha = 0,05\ (5\%)$. Следовательно, значение частного F-критерия для дополнительно включенного фактора x_1 не случайно, является статистически значимым, надежным, достоверным: прирост факторной дисперсии за счет дополнительного фактора x_1 является существенным. Фактор x_1 должен присутствовать в уравнении, в том числе в варианте, когда он дополнительно включается после фактора x_2.

5. Общий вывод состоит в том, что множественная модель с факторами x_1 и x_2 с $R^2_{yx_1x_2} = 0,947$ содержит неинформативный фактор x_2. Если исключить фактор x_2, то можно ограничиться уравнением парной регрессии:

$$\hat{y}_x = \alpha_0 + \alpha_1 x = 1,99 + 1,23 \cdot x, \qquad r^2_{yx} = 0,941.$$

Варианты индивидуальных заданий

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного

веса рабочих высокой квалификации в общей численности рабочих x_2 (%) (смотри таблицу своего варианта).

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Вариант 1

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6	3,6	9	11	9	6,3	21
2	6	3,6	12	12	11	6,4	22
3	6	3,9	14	13	11	7	24
4	7	4,1	17	14	12	7,5	25
5	7	3,9	18	15	12	7,9	28
6	7	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8	30
8	8	5,3	19	18	13	8,6	31
9	9	5,6	20	19	14	9,5	33
10	10	6,8	21	20	14	9	36

Вариант 2

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6	3,5	10	11	10	6,3	21
2	6	3,6	12	12	11	6,4	22
3	7	3,9	15	13	11	7	23
4	7	4,1	17	14	12	7,5	25
5	7	4,2	18	15	12	7,9	28
6	8	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8,4	31
8	9	5,3	20	18	14	8,6	31
9	9	5,6	20	19	14	9,5	35
10	10	6	21	20	15	10	36

Вариант 3

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,7	9	11	11	6,3	22
2	7	3,7	11	12	11	6,4	22
3	7	3,9	11	13	11	7,2	23
4	7	4,1	15	14	12	7,5	25
5	8	4,2	17	15	12	7,9	27
6	8	4,9	19	16	13	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,1	20	18	13	8,6	32
9	10	5,6	20	19	14	9,5	35
10	10	6,1	21	20	15	9,5	36

Вариант 4

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,5	9	11	10	6,3	22
2	7	3,6	10	12	10	6,5	22
3	7	3,9	12	13	11	7,2	24
4	7	4,1	17	14	12	7,5	25
5	8	4,2	18	15	12	7,9	27
6	8	4,5	19	16	13	8,2	30
7	9	5,3	19	17	13	8,4	31
8	9	5,5	20	18	14	8,6	33
9	10	5,6	21	19	14	9,5	35
10	10	6,1	21	20	15	9,6	36

Вариант 5

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,6	9	11	10	6,3	21
2	7	3,6	11	12	11	6,9	23
3	7	3,7	12	13	11	7,2	24
4	8	4,1	16	14	12	7,8	25
5	8	4,3	19	15	13	8,1	27
6	8	4,5	19	16	13	8,2	29
7	9	5,4	20	17	13	8,4	31
8	9	5,5	20	18	14	8,8	33
9	10	5,8	21	19	14	9,5	35
10	10	6,1	21	20	14	9,7	34

Вариант 6

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,5	9	11	10	6,3	21
2	7	3,6	10	12	10	6,8	22
3	7	3,8	14	13	11	7,2	24
4	7	4,2	15	14	12	7,9	25
5	8	4,3	18	15	12	8,1	26
6	8	4,7	19	16	13	8,3	29
7	9	5,4	19	17	13	8,4	31
8	9	5,6	20	18	13	8,8	32
9	10	5,9	20	19	14	9,6	35
10	10	6,1	21	20	14	9,7	36

Вариант 7

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	11	11	10	6,8	21
2	7	3,8	12	12	11	7,4	23
3	7	3,9	16	13	11	7,8	24
4	7	4,1	17	14	12	7,5	26
5	7	4,6	18	15	12	7,9	28
6	8	4,5	18	16	12	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,5	20	18	13	8,7	32
9	9	6,1	20	19	13	9,5	33
10	10	6,8	21	20	14	9,7	35

Вариант 8

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	9	11	11	7,1	22
2	7	4,1	14	12	11	7,5	23
3	7	4,3	16	13	12	7,8	25
4	7	4,1	17	14	12	7,6	27
5	8	4,6	17	15	12	7,9	29
6	8	4,7	18	16	13	8,1	30
7	9	5,3	20	17	13	8,5	32
8	9	5,5	20	18	14	8,7	32
9	11	6,9	21	19	14	9,6	33
10	10	6,8	21	20	15	9,8	36

Вариант 9

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,9	12	11	11	7,1	22
2	7	4,2	13	12	12	7,5	25
3	7	4,3	15	13	13	7,8	26
4	7	4,4	17	14	12	7,9	27
5	8	4,6	18	15	13	8,1	30
6	8	4,8	19	16	13	8,4	31
7	9	5,3	19	17	13	8,6	32
8	9	5,7	20	18	14	8,8	32
9	10	6,9	21	19	14	9,6	34
10	10	6,8	21	20	14	9,9	36

Вариант 10

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,6	12	11	10	7,2	23
2	7	4,1	14	12	11	7,6	25
3	7	4,3	16	13	12	7,8	26
4	7	4,4	17	14	11	7,9	28
5	7	4,5	18	15	12	8,2	30
6	8	4,8	19	16	12	8,4	31
7	8	5,3	20	17	12	8,6	32
8	8	5,6	20	18	13	8,8	32
9	9	6,7	21	19	13	9,2	33
10	10	6,9	22	20	14	9,6	34

D.3. Системы эконометрических уравнений

Пример решения типовой задачи смотри в разделе 3.

Варианты индивидуальных заданий

Даны системы эконометрических уравнений.

Требуется

1. Применив необходимое и достаточное условие идентификации, определите, идентифицируемо ли каждое из уравнений модели.

2. Определите метод оценки параметров модели.

3. Запишите в общем виде приведенную форму модели.

Вариант 1

Модель протекционизма Сальватора (упрощенная версия):

$$\begin{cases} M_t = a_1 + b_{12}N_t + b_{13}S_t + b_{14}E_{t-1} + b_{15}M_{t-1} + \varepsilon_1, \\ N_t = a_2 + b_{21}M_t + b_{23}S_t + b_{26}Y_t + \varepsilon_2, \\ S_t = a_3 + b_{31}M_t + b_{32}N_t + b_{36}X_t + \varepsilon_3. \end{cases}$$

где M – доля импорта в ВВП; N – общее число прошений об освобождении от таможенных пошлин; S – число удовлетворенных прошений об освобождении от таможенных пошлин; E – фиктивная переменная, равная 1 для тех лет, в которые курс доллара на международных валютных рынках был искусственно завышен, и 0 – для всех остальных лет; Y – реальный ВВП; X – реальный объем чистого экспорта; t – текущий период; $t-1$ – предыдущий период.

Вариант 2

Макроэкономическая модель (упрощенная версия модели Клейна):

$$\begin{cases} C_t = a_1 + b_{12}Y_t + b_{13}T_t + \varepsilon_1, \\ I_t = a_2 + b_{21}Y_t + b_{24}K_{t-1} + \varepsilon_2, \\ Y_t = C_t + I_t, \end{cases}$$

где C – потребление; I – инвестиции; Y – доход; T – налоги; K – запас капитала; t – текущий период; $t-1$ – предыдущий период.

Вариант 3

Макроэкономическая модель экономики США (одна из версий):

$$\begin{cases} C_t = a_1 + b_{11}Y_t + b_{12}C_{t-1} + \varepsilon_1, \\ I_t = a_2 + b_{21}Y_t + b_{23}r_t + \varepsilon_2, \\ r_t = a_3 + b_{31}Y_t + b_{34}M_t + b_{35}r_{t-1} + \varepsilon_3, \\ Y_t = C_t + I_t + G_t, \end{cases}$$

где C – потребление; Y – ВВП; I – инвестиции; r – процентная ставка; M – денежная масса; G – государственные расходы; t – текущий период; $t-1$ – предыдущий период.

Вариант 4

Модель Кейнса (одна из версий):

$$\begin{cases} C_t = a_1 + b_{11}Y_t + b_{12}Y_{t-1} + \varepsilon_1, \\ I_t = a_2 + b_{21}Y_t + \varepsilon_2, \\ Y_t = C_t + I_t + G_t, \end{cases}$$

где C – потребление; Y – ВВП; I – валовые инвестиции; G – государственные расходы; t – текущий период; $t-1$ – предыдущий период.

Вариант 5

Модель денежного и товарного рынков:

$$\begin{cases} R_t = a_1 + b_{12}Y_t + b_{14}M_t + \varepsilon_1, \\ Y_t = a_2 + b_{21}R_t + b_{23}I_t + b_{25}G_t + \varepsilon_2, \\ I_t = a_3 + b_{31}R_t + \varepsilon_3, \end{cases}$$

где R – процентные ставки; Y – реальный ВВП; M – денежная масса; I – внутренние инвестиции; G – реальные государственные расходы.

Вариант 6

Модифицированная модель Кейнса:

$$\begin{cases} C_t = a_1 + b_{11}Y_t + \varepsilon_1, \\ I_t = a_2 + b_{21}Y_t + b_{22}Y_{t-1} + \varepsilon_2, \\ Y_t = C_t + I_t + G_t, \end{cases}$$

где C – потребление; Y – доход; I – инвестиции; G – государственные расходы; t – текущий период; $t-1$ – предыдущий период.

Вариант 7

Макроэкономическая модель:

$$\begin{cases} C_t = a_1 + b_{11} D_t + \varepsilon_1, \\ I_t = a_2 + b_{22} Y_t + b_{23} Y_{t-1} + \varepsilon_2, \\ Y_t = D_t + T_t, \\ D_t = C_t + I_t + G_t, \end{cases}$$

где C – расходы на потребление; Y – чистый национальный продукт; D – чистый национальный доход; I – инвестиции; T – косвенные налоги; G – государственные расходы; t – текущий период; $t-1$ – предыдущий период.

Вариант 8

Гипотетическая модель экономики:

$$\begin{cases} C_t = a_1 + b_{11} Y_t + b_{12} J_t + \varepsilon_1, \\ J_t = a_2 + b_{21} Y_{t-1} + \varepsilon_2, \\ T_t = a_3 + b_{31} Y_t + \varepsilon_3, \\ Y_t = C_t + J_t + G_t, \end{cases}$$

где C – совокупное потребление в период t; Y – совокупный доход в период t; J – инвестиции в период t; T – налоги в период t; G – государственные доходы в период t.

Вариант 9

Модель денежного рынка:

$$\begin{cases} R_t = a_1 + b_{11} M_t + b_{12} Y_t + + \varepsilon_1, \\ Y_t = a_2 + b_{21} R_t + b_{22} I_t + \varepsilon_2, \\ I_t = a_3 + b_{33} R_t + \varepsilon_3, \end{cases}$$

где R – процентные ставки; Y – ВВП; M – денежная масса; I – внутренние инвестиции.

Вариант 10

Конъюнктурная модель имеет вид:

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

где C – расходы на потребление; Y – ВВП; I – инвестиции; r – процентная ставка; M – денежная масса; G – государственные расходы; t – текущий период; $t-1$ – предыдущий период.

D.4. Временные ряды

Пример решения типовой задачи смотри в разделе 4.

Варианты индивидуальных заданий

Имеются условные данные об объемах потребления электроэнергии (y_t) жителями региона за 16 кварталов.

Требуется:

1. Построить автокорреляционную функцию и сделать вывод о наличии сезонных колебаний.

2. Построить аддитивную модель временного ряда (для нечетных вариантов) или мультипликативную модель временного ряда (для четных вариантов).

3. Сделать прогноз на 2 квартала вперед.

Варианты 1, 2

t	y_t	t	y_t
1	5,8	9	7,9
2	4,5	10	5,5
3	5,1	11	6,3
4	9,1	12	10,8
5	7,0	13	9,0
6	5,0	14	6,5
7	6,0	15	7,0
8	10,1	16	11,1

Варианты 3, 4

t	y_t	t	y_t
1	5,5	9	8,0
2	4,6	10	5,6
3	5,0	11	6,4
4	9,2	12	10,9
5	7,1	13	9,1
6	5,1	14	6,4
7	5,9	15	7,2
8	10,0	16	11,0

Варианты 5, 6

t	y_t	t	y_t
1	5,3	9	8,2
2	4,7	10	5,5
3	5,2	11	6,5
4	9,1	12	11,0
5	7,0	13	8,9
6	5,0	14	6,5
7	6,0	15	7,3
8	10,1	16	11,2

Варианты 7, 8

t	y_t	t	y_t
1	5,5	9	8,3
2	4,8	10	5,4
3	5,1	11	6,4
4	9,0	12	10,9
5	7,1	13	9,0
6	4,9	14	6,6
7	6,1	15	7,5
8	10,0	16	11,2

Варианты 9, 10

t	y_t	t	y_t
1	5,6	9	8,2
2	4,7	10	5,6
3	5,2	11	6,4
4	9,1	12	10,8
5	7,0	13	9,1
6	5,1	14	6,7
7	6,0	15	7,5
8	10,2	16	11,3

Приложение Е

Математико-статистические таблицы

Е.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	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
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
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

1	2	3	4	5	6	7	8	9	10	11
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

Е.2. Критические значения t-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01 (двухсторонний)

Число степеней свободы d.f.	α			Число степеней свободы d.f.	α		
	00,10	0,05	0,01		00,10	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
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

Е.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
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

Литература

Основная:

1. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2002. – 344 с.

2. Практикум по эконометрике: Учебн. пособие / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2003. – 192 с.

3. Эконометрика: Учебно-методическое пособие / Шалабанов А.К., Роганов Д.А. – Казань: ТИСБИ, 2002. – 56 с.

4. Доугерти К. Введение в эконометрику: Пер. с англ. – М.: ИНФРА-М, 1999. – 402 с.

Дополнительная:

5. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов / Под ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2002. – 311 с.

6. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учебник. – М.: Дело, 2001. – 400 с.

7. Катышев П.К., Магнус Я.Р., Пересецкий А.А. Сборник задач к начальному курсу эконометрики. – М.: Дело, 2002. – 208 с.

8. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2-х т. – Т. 1. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М: ЮНИТИ-ДАНА, 2001. – 656 с.

9. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2-х т. – Т. 2. Айвазян С.А. Основы эконометрики. – М: ЮНИТИ-ДАНА, 2001. – 432 с.

10. Эконометрика: Учебник / Тихомиров Н.П., Дорохина Е.Ю. – М.: Издательство «Экзамен», 2003. – 512 с.

11. Сборник задач по эконометрике: Учебное пособие для студентов экономических вузов / Сост. Е.Ю. Дорохина, Л.Ф. Преснякова, Н.П. Тихомиров. – М.: Издательство «Экзамен», 2003. – 224 с.

12. Кулинич Е.И. Эконометрия. – М.: Финансы и статистика, 2001. – 304 с.

13. Эконометрика: Учебн. пособие для вузов / А.И. Орлов – М.: Издательство «Экзамен», 2002. – 576 с.

14. Мардас А.Н. Эконометрика. – СПб: Питер, 2001. – 144 с.

1. Гмурман В.Е. Теория вероятностей и математическая статистика: Учебн. пособие для вузов. – М.: Высш. шк., 2002. – 479 с.