Н. В. Артамонов

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

УДК 330.43
ББК 65.05
А86

Артамонов Н. В.
Введение в эконометрику
Электронное издание
М.: МЦНМО, 2019
251 с.
ISBN 978-5-4439-3381-8

Учебник знакомит читателя с базовыми понятиями и методами современной эконометрики, которая является неотъемлемой частью современного экономического образования. В первых двух главах подробно излагаются линейные и нелинейные регрессионные модели, их статистические свойства и возможности применения в экономике. В третьей главе рассматриваются возможные отклонения от стандартных предположений линейной модели регрессии, встречающиеся при моделировании экономических ситуаций и при анализе экономических данных. Обсуждаются корректировки регрессионной модели для описания таких ситуаций. Последняя глава посвящена регрессионным моделям временных рядов. Учебник основан на лекциях по курсу «Эконометрика-1», читаемых автором в МГИМО (У) МИД России на факультете Международных экономических отношений.

Книга предназначена студентам (бакалавриата и магистратуры), аспирантам и преподавателям, специалистам и исследователям, работающим в области прикладной экономики и финансов.

В настоящее издание добавлено новое приложение о линейной регрессии в языке R.

Подготовлено на основе книги: *Н. В. Артамонов*. Введение в эконометрику. — 3-е изд., испр. и доп. — М.: МЦНМО, 2019. — ISBN 978-5-4439-1381-0.

Учебное издание для вузов

12+

Издательство Московского центра
непрерывного математического образования
119002, Москва, Большой Власьевский пер., 11. Тел. (499) 241-08-04
http://www.mccme.ru

ISBN 978-5-4439-3381-8

Оглавление

Глава 1
Парная регрессия

§ 1.1. Парный коэффициент корреляции

1.1.1. Коэффициент корреляции

Пусть (X, Y) — двумерная *нормально распределенная* случайная величина. Тогда «степень зависимости» случайных величин X и Y характеризуется парным коэффициентом корреляции

$$\rho = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}} = \frac{\mathsf{E}(XY) - \mathsf{E}X \cdot \mathsf{E}Y}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}}.$$

Из определения коэффициента корреляции следует, что

1) $-1 \leqslant \rho \leqslant 1$;
2) коэффициент корреляции не меняется при линейных преобразованиях величин, т. е.

$$\operatorname{corr}(X, Y) = \operatorname{corr}(a_0 + a_1 X, b_0 + b_1 Y), \quad a_1, b_1 \neq 0.$$

Коэффициент корреляции принимает крайние значения ± 1 в том и только том случае, когда между случайными величинами X и Y существует *линейная функциональная* зависимость, т. е.

$$\rho = \pm 1 \iff Y = \beta_0 + \beta_1 X, \quad \beta_1 \neq 0,$$

причем

$$\beta_1 = \rho \sqrt{\frac{\operatorname{Var}(Y)}{\operatorname{Var}(X)}},$$

т. е. знак коэффициента β_1 совпадает со знаком коэффициента корреляции.

В общем случае коэффициент корреляции возникает при решении следующей экстремальной задачи: подобрать линейную функцию $l(x) = \beta_0 + \beta_1 x$ так, чтобы случайная величина $l(X)$ меньше всего отклонялась от Y в среднеквадратичном, т. е.

$$\mathsf{E}(Y - \beta_0 - \beta_1 X)^2 \underset{\beta_0, \beta_1}{\longrightarrow} \min.$$

Решение этой задачи задается равенствами

$$\beta_1^* = \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)} = \rho\sqrt{\frac{\operatorname{Var}(Y)}{\operatorname{Var}(X)}}, \quad \beta_0^* = \mathsf{E}Y - \beta_1^* \cdot \mathsf{E}X,$$

и наименьшее среднеквадратичное отклонение равно

$$\mathsf{E}(Y - \beta_0^* - \beta_1^* X)^2 = (1-\rho^2)\operatorname{Var}(Y).$$

Кроме того, для всех $x \in \mathbb{R}$ верно равенство

$$\mathsf{E}(Y \mid X = x) = \beta_0^* + \beta_1^* x,$$

т. е. наилучший прогноз случайной величины Y при условии, что известно значение случайной величины $X = x$, равен $\widehat{Y} = \beta_0^* + \beta_1^* x$. Рассмотрим три случая:

1) $\rho > 0$; тогда $\beta_1^* > 0$ и при увеличении x ожидаемое (среднее) значение $\mathsf{E}(Y \mid X = x)$ случайной величины Y также увеличивается; в этом случае говорят о *прямой линейной зависимости* между величинами;
2) $\rho < 0$; тогда $\beta_1^* < 0$ и при увеличении x ожидаемое (среднее) значение $\mathsf{E}(Y \mid X = x)$ случайной величины y уменьшается; в этом случае говорят об *обратной линейной зависимости* между величинами;
3) $\rho = 0$; тогда $\beta_1^* = 0$, $\mathsf{E}(Y \mid X = x) = \beta_0^*$ и знание значения случайной величины X не улучшает прогноз Y.

Важное значение коэффициента корреляции обусловлено следующей теоремой.

Теорема. *Пусть* (X, Y) *— двумерная нормально распределенная случайная величина. Тогда случайные величины* X *и* Y *независимы в том и только том случае, когда* $\operatorname{corr}(X, Y) = 0$.

Таким образом, парный коэффициент корреляции можно рассматривать как *меру зависимости* двух случайных величин (факторов), имеющих совместное нормальное распределение, причем

- $\rho = 0 \iff$ величины независимы;
- $\rho = \pm 1 \iff$ между величинами имеется линейная функциональная зависимость: $y = \beta_0^* + \beta_1^* x$.

1.1.2. Выборочный коэффициент корреляции

Пусть $(x_i, y_i)_{i=1}^n$ — выборка из двумерной нормально распределенной случайной величины, n — объем выборки.

Напомним, что выборочные (неисправленные) оценки дисперсий случайных величин X и Y определяются как

$$\widehat{\mathrm{Var}}(X) = \widehat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\bar{x^2}) - (\bar{x})^2,$$

$$\widehat{\mathrm{Var}}(Y) = \widehat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = (\bar{y^2}) - (\bar{y})^2,$$

где

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

Напомним также, что $\widehat{\mathrm{Var}}(X)$ и $\widehat{\mathrm{Var}}(Y)$ — состоятельные, но смещенные оценки дисперсий $\mathrm{Var}(X)$ и $\mathrm{Var}(Y)$ соответственно.

Выборочный коэффициент ковариации определяется как [1]

$$\widehat{\mathrm{cov}}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y},$$

а *выборочный коэффициент корреляции* определяется равенством [2]

$$r = \widehat{\mathrm{corr}}(X, Y) = \frac{\widehat{\mathrm{cov}}(X, Y)}{\sqrt{\widehat{\mathrm{Var}}(X)\cdot\widehat{\mathrm{Var}}(Y)}}, \quad -1 \leqslant r \leqslant 1.$$

Выборочные коэффициенты ковариации и корреляции являются состоятельными оценками коэффициентов ковариации и корреляции в генеральной совокупности. Выборочный коэффициент корреляции может рассматриваться как выборочная «мера линейной зависимости» между случайными величинами.

Проверка значимости коэффициента корреляции. Проверка значимости подразумевает проверку статистической гипотезы

$$H_0 : \rho = 0$$

[1] В MS Excel функция КОВАР(·, ·).

[2] В MS Excel функция КОРРЕЛ(·, ·).

против двусторонней альтернативы

$$H_1 : \rho \neq 0.$$

Другими словами, проверяется статистическая гипотеза о том, что в генеральной совокупности случайные величины (факторы) X и Y *не коррелируют*. Так как двумерная случайная величина (X, Y) по предположению имеет совместное нормальное распределение, некоррелируемость означает независимость факторов. Проверка гипотезы о независимости факторов основана на следующем результате: при справедливости нулевой гипотезы t-статистика

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{H_0}{\sim} t_{n-2}$$

имеет распределение Стьюдента с $n-2$ степенями свободы. Таким образом, получаем следующий статистический критерий проверки нулевой гипотезы:

> при заданном уровне значимости α гипотеза H_0 отвергается в пользу альтернативы H_1 при $|t| > t_{кр}$,

где $t_{\text{кр}} = t(\alpha; n-2)$ есть *двустороннее* критическое значение распределения Стьюдента t_{n-2}. Напомним, что двустороннее критическое значение определяется как решение уравнения

$$\mathrm{P}(|t_{n-2}| > t_{\text{кр}}) = \alpha.$$

При $|t| < t_{\text{кр}}$ говорят, что данные *согласуются* с нулевой гипотезой или *не противоречат ей*, *и H_0 не отвергается*.

Пример. На основе $n = 62$ выборочных данных, был рассчитан выборочный коэффициент корреляции $r = 0{,}68$ между дневными логарифмическими доходностями[1] биржевых индексов NASDAQ и FTSE. Проверим значимость коэффициента корреляции, т. е. проверим статистическую гипотезу H_0 о *независимости* доходностей обоих биржевых индексов (в предположении их *нормальной распределенности!*). Вычислим значение t-статистики:

$$t = \frac{0{,}68 \cdot \sqrt{62-2}}{\sqrt{1-0{,}68^2}} \approx 7{,}1838.$$

[1] Логарифмическая доходность рассчитывается как $h_t = \ln(S_t/S_{t-1})$.

Критическое значение распределения Стьюдента при уровне значимости $\alpha = 5\%$ равно $t_{\text{кр}} = t(5\%; 62-2) \approx 2{,}003$. Так как $|t| > t_{\text{кр}}$, гипотеза H_0 о независимости доходностей *отвергается*, коэффициент корреляции значим.

Доверительный интервал для коэффициента корреляции. Задача о построении доверительного интервала для коэффициента корреляции связана с той проблемой, что в общем случае (при $\rho \neq 0$) t-статистика имеет неизвестное распределение. Однако Фишер заметил, что если взять *z-преобразование Фишера* [1] от выборочного коэффициента корреляции

$$z(r) = \frac{1}{2} \ln \frac{1+r}{1-r},$$

то эта статистика при больших объемах выборки (а фактически уже при $n > 6$) имеет распределение, близкое к нормальному:

$$z(r) \approx \mathcal{N}\left(z(\rho), \frac{1}{n-3}\right).$$

Следовательно, при заданной доверительной вероятности γ приближенный (асимптотический) доверительный интервал для z-преобразования Фишера коэффициента корреляции определяется как

$$\mathrm{P}\left(z(r) - \frac{z_\gamma}{\sqrt{n-3}} < z(\rho) < z(r) + \frac{z_\gamma}{\sqrt{n-3}}\right) \approx \gamma. \tag{1.1}$$

Здесь z_γ есть *двустороннее* критическое значение стандартного нормального распределения при уровне значимости $1-\gamma$ и находится как решение уравнения

$$\Phi(z_\gamma) = \frac{1+\gamma}{2},$$

где $\Phi(x)$ — функция стандартного нормального распределения.

Доверительный интервал для коэффициента корреляции получается применением к интервалу (1.1) *обратного преобразования Фишера* [2]

$$z^{-1}(x) = \operatorname{th}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

[1] В MS Excel функция ФИШЕР(·).

[2] В MS Excel функция ФИШЕРОБР(·).

Таким образом, асимптотический доверительный интервал для коэффициента корреляции имеет вид

$$\mathrm{P}\left(z^{-1}\left(z(r)-\frac{z_\gamma}{\sqrt{n-3}}\right)<\rho<z^{-1}\left(z(r)+\frac{z_\gamma}{\sqrt{n-3}}\right)\right)\approx\gamma.$$

Замечание. Зная доверительный интервал для коэффициента корреляции, можно проверить его значимость, т. е. статистическую гипотезу $H_0\colon \rho=0$ при уровне значимости $\alpha=1-\gamma$. Нулевая гипотеза отвергается тогда и только тогда, когда нуль не принадлежит доверительному интервалу.

Пример. На основе $n=100$ выборочных данных, был рассчитан выборочный коэффициент корреляции $r=0{,}68$ между дневными логарифмическими доходностями биржевых индексов NASDAQ и FTSE. Построим доверительный интервал для коэффициента корреляции с доверительной вероятностью $\gamma=0{,}95$. Применим z-преобразование Фишера $z=z(0{,}68)\approx 0{,}8291$. Критическое значение z_γ определяется как решение уравнения

$$\Phi(z_\gamma)=\frac{1+0{,}95}{2}=0{,}975,$$

откуда получаем $z_\gamma=1{,}96$. Доверительный интервал для $z(\rho)$ равен

$$\left(0{,}8291-\frac{1{,}96}{\sqrt{100-3}};\ 0{,}8291+\frac{1{,}96}{\sqrt{100-3}}\right)=(0{,}6301;\ 1{,}0281).$$

Применив обратное преобразование Фишера, получаем доверительный интервал для коэффициента корреляции

$$\mathrm{P}(0{,}5581<\rho<0{,}7732)=0{,}95$$

($0{,}5581=z^{-1}(0{,}6301)$ и $0{,}7732=z^{-1}(1{,}0281)$).

Проверим значимость коэффициента корреляции, т. е. проверим нулевую гипотезу о *независимости* доходностей обоих биржевых индексов (в предположении их *нормальной распределенности*!). Так как нуль не принадлежит доверительному интервалу, нулевая гипотеза отвергается при уровне значимости $\alpha=1-0{,}95=0{,}05$.

§ 1.2. Подгонка прямой. Метод наименьших квадратов

Рассмотрим следующую вспомогательную задачу. Пусть на координатной плоскости заданы n точек с координатами $(x_i, y_i)_{i=1}^n$. Требуется найти прямую, «меньше всего отклоняющуюся от заданных точек». Так как прямая задается уравнением

$$y = f(x) = \beta_0 + \beta_1 x,$$

зависящим от двух параметров β_0 и β_1, необходимо по заданным значениям $\{x_i\}$ и $\{y_i\}$ найти значения этих параметров «оптимальной» прямой. Основной вопрос: что понимать под «наименьшим отклонением прямой от точек» и, более общим образом, как определить «меру отклонения прямой от точек»? Приведем несколько возможных подходов к определению меры μ отклонения прямой от заданных точек:

1) сумма модулей отклонений в каждой точке x_i:

$$\mu = \sum_{i=1}^{n} |y_i - f(x_i)| = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|;$$

2) сумма квадратов отклонений в каждой точке x_i:

$$\mu = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2;$$

3) сумма отклонений в каждой точке x_i с заданной весовой функцией $\omega(\cdot) > 0$:

$$\mu = \sum_{i=1}^{n} \omega\big(y_i - f(x_i)\big) = \sum_{i=1}^{n} \omega\big(y_i - (\beta_0 + \beta_1 x_i)\big).$$

С вероятностной точки зрения в случае нормального распределения выборочных данных «наилучшими вероятностными и статистическими свойствами» обладают оценки параметров прямой, полученные минимизацией суммы квадратов отклонений (второй случай). Этот метод получения оценок параметров оптимальной прямой называется *методом наименьших квадратов* (сокращенно МНК) или *Ordinary Least Squares* (сокращенно OLS), а полученные оценки параметров называются МНК- или OLS-оценками.

Итак, в качестве меры отклонения прямой от заданных на плоскости точек $(x_i, y_i)_1^n$ возьмем сумму квадратов отклонений в каждой точке [1]:

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2.$$

Тогда параметры прямой, для которой эта мера отклонения минимальна, находятся как решение экстремальной задачи без ограничений:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \to \min.$$

Согласно необходимым условиям существования экстремума параметры оптимальной прямой находятся как решение системы уравнений

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = \displaystyle\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-1) = 0, \\ \dfrac{\partial S}{\partial \beta_1} = \displaystyle\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i) = 0. \end{cases}$$

После простых преобразований приходим к системе линейных уравнений

$$\begin{cases} n\beta_0 + \beta_1 \displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ \beta_0 \displaystyle\sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \end{cases} \tag{1.2}$$

называемой системой *нормальных уравнений*. Найдем явные формулы для решения этой системы. Для удобства разделим каждое уравнение в системе (1.2) на n:

$$\begin{cases} \beta_0 + \beta_1 \bar{x} = \bar{y}, \\ \beta_0 \bar{x} + \beta_1 \bar{x}^2 = \overline{xy}. \end{cases}$$

Выразим β_0 из первого уравнения:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

[1] Очевидно, $S(\beta_0, \beta_1)$ есть многочлен второго порядка от параметров β_0 и β_1.

и подставим во второе уравнение:

$$(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1(\bar{x^2}) = \overline{xy}.$$

После преобразования получаем (формально), что решение системы имеет вид

$$\widehat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{(\bar{x^2}) - (\bar{x})^2} = \frac{\widehat{\mathrm{cov}}(x, y)}{\widehat{\mathrm{Var}}(x)} = \widehat{\mathrm{corr}}(x, y)\sqrt{\frac{\widehat{\mathrm{Var}}(y)}{\widehat{\mathrm{Var}}(x)}} = \widehat{\mathrm{corr}}(x, y)\frac{\widehat{\sigma}_y}{\widehat{\sigma}_x}$$

и

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

где $\widehat{\sigma}_x = \sqrt{\widehat{\mathrm{Var}}(x)}$ и $\widehat{\sigma}_y = \sqrt{\widehat{\mathrm{Var}}(y)}$ — выборочные стандартные отклонения величин x и y соответственно.

Несложно показать, что функция $S(\beta_0, \beta_1)$ выпукла. Следовательно, решение системы нормальных уравнений (1.2) будет глобальным минимумом функции $S(\beta_0, \beta_1)$. Таким образом, оптимальная прямая задается уравнением

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

Замечание 1. Из первого уравнения системы (1.2) следует, что

$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x},$$

т. е. оптимальная прямая проходит через точку с координатами $(\bar{x}, \bar{y})$.

Замечание 2. Несложно заметить, что система нормальных уравнений (1.2) имеет единственное решение тогда и только тогда, когда $\widehat{\mathrm{Var}}(x) \neq 0$, т. е. когда не все значения x_i совпадают.

Замечание 3. Метод наименьших квадратов может быть применен для нахождения параметров любой функции, меньше всего отклоняющейся от заданных точек. Эта задача корректно разрешима в случае, когда неизвестные *параметры входят в функцию линейно*. Тогда система нормальных уравнений будет *системой линейных уравнений* и в общем случае будет иметь единственное решение.

§ 1.3. Парная линейная модель регрессии

Перейдем теперь к задаче количественного описания зависимости между двумя экономическими факторами y и x (например, y — уровень зарплаты индивидуума, а x — уровень образования (в годах)).

Естественно ожидать, что значение фактора y не всегда однозначно определяется значением фактора x. Так, уровень зарплаты зависит не только от уровня образования, но и от множества других факторов (стажа работы, возраста, индивидуальных способностей, места работы и пр.). Кроме того, учесть *все* факторы, влияющие на y помимо x, просто не представляется возможным в силу недостаточного количества информации или невозможности ее получения (например, как оценить или измерить индивидуальные способности индивидуума, несомненно влияющие на уровень зарплаты?). Также для одного значения фактора x могут наблюдаться различные значения фактора y.

Обычно для описания ситуаций с недостаточной информацией используют различные вероятностные математические модели. Рассмотрим подробно модель зависимости между факторами, описываемую уравнением

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.3}$$

где y_i и ε_i суть случайные величины, а x_i — *неслучайная* (детерминированная) величина, i — номер наблюдения. Фактор y называется *зависимой переменной* (dependent variable), а фактор x называется *регрессором* или *объясняющей переменной* (explanatory variable). Параметр β_1 называется параметром *наклона прямой* (slope), а β_0 — *константой*, *свободным членом* или *параметром сдвига* (intercept).

Уравнение (1.3) называется *уравнением регрессии* или *регрессионным уравнением*, а случайные величины ε_i называются *ошибками* регрессии. Ошибки регрессии удобно представлять себе как «неучтенные факторы», влияющие на y помимо фактора x. Таким образом, уравнение (1.3) отражает наши представления о характере зависимости между факторами.

Относительно ошибок регрессии будем предполагать выполнение следующих условий, называемых иногда условиями Гаусса—Маркова:

1) $\mathrm{E}\varepsilon_i = 0$, $i = 1, \ldots, n$ (ошибки регрессии несистематические);
2) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ не зависит от i;
3) $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$ (некоррелируемость ошибок для разных наблюдений);
4) $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$ (нормальная распределенность ошибок регрессии).

Из условия $\mathsf{E}\varepsilon_i = 0$ следует, что

$$\mathsf{E}y_i = \beta_0 + \beta_1 x_i,$$

т. е. среднее значение фактора y при заданном значении x_i равно $\beta_0 + \beta_1 x_i$ и не зависит от ошибок регрессии. Отсюда термин: несистематические ошибки.

Очевидно, $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i)$ (так как значения x_i детерминированны). Следовательно, условие постоянства дисперсий ошибок регрессии влечет за собой постоянство дисперсий случайных величин y_i. Следует напомнить, что дисперсию $\mathrm{Var}(y_i)$ можно рассматривать как «меру разброса» значений случайной y_i величины относительно своего среднего значения (математического ожидания) $\mathsf{E}y_i = \beta_0 + \beta_1 x_i$. Если смотреть на ошибки регрессии как на «неучтенные факторы», условие постоянства дисперсий можно описательно трактовать следующим образом: «степень влияния» невключенных в модель факторы в разных наблюдениях постоянна. Условие постоянства дисперсий ошибок называется *гомоскедастичностью* (homoskedasticity), и говорят, что ошибки модели регрессии гомоскедастичны или однородны. При нарушении условия постоянства дисперсий ошибок регрессии говорят, что ошибки *гетероскедастичны* или *неоднородны*.

Условие некоррелируемости (независимости в случае нормального распределения) ошибок для разных наблюдений можно трактовать как «локальность» их влияния: не включенные в модель факторы, которые моделируются ошибками регрессии, влияют только на «свое» наблюдение и не влияют на другие. В случае пространственных выборок (cross-sectional data) это условие обычно считается выполненным. Оно, как правило, нарушается в случае построения регрессионных моделей для временны́х рядов.

1.3.1. Теорема Гаусса—Маркова

Итак, мы предполагаем, что зависимость между факторами y и x описывается уравнением регрессии (1.3), но параметры уравнения β_0, β_1 и σ^2 нам неизвестны.

Основная задача — получить «наилучшие» оценки параметров регрессии на основе выборочных данных. Ограничимся рассмотрением только оценок параметров, линейных относительно y_i. Под «наилучшими» будем подразумевать несмещенные оценки с минималь-

ной дисперсией[1]. Такие оценки называются BLUE-*оценками* (BLUE = Best Linear Unbiased Estimators) или *эффективными оценками*.

Основным результатом является следующая теорема.

Теорема (Гаусс—Марков)**.** *Пусть для линейной модели парной регрессии*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

выполнены условия 1—3 *на ошибки регрессии* ε_i*. Тогда* OLS-*оценки* $\widehat{\beta}_0$ *и* $\widehat{\beta}_1$ *параметров* β_0 *и* β_1 *являются* BLUE-*оценками, т. е. среди несмещенных линейных* (*относительно* y_i) *оценок имеют наименьшую дисперсию.*

Доказательство. Докажем несмещенность OLS-оценок. Рассмотрим сначала оценку параметра β_1. Для нее имеем следующее выражение:

$$\widehat{\beta}_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\bar{x}^2 - (\bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n((\bar{x})^2 - \bar{x}^2)} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}.$$

Так как величины x_i неслучайны и $\mathsf{E}y_i = \beta_0 + \beta_1 x_i$ (условие 1 на ошибки регрессии), мы получаем

$$\mathsf{E}\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})\mathsf{E}y_i}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum(x_i - \bar{x})^2} =$$

$$= \beta_0 \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} + \beta_1 \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})x_i} = \beta_1.$$

При выводе мы воспользовались равенствами

$$\sum(x_i - \bar{x}) = 0, \quad \sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})x_i.$$

Далее, так как

$$\mathsf{E}(\bar{y}) = \mathsf{E}\Big(\frac{1}{n}\sum_{i=1}^{n} y_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x},$$

[1] Напомним, что оценка параметров вероятностной модели в математической статистике рассматривается как случайная величина.

для оценки константы $\widehat{\beta}_0$ в уравнении регрессии получаем

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \;\Rightarrow\; \mathsf{E}(\widehat{\beta}_0) = \mathsf{E}(\bar{y}) - \mathsf{E}(\widehat{\beta}_1 \bar{x}) = \\ = \beta_0 + \beta_1 \bar{x} - \bar{x} \cdot \mathsf{E}(\widehat{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.$$

Итак, $\widehat{\beta}_0$ и $\widehat{\beta}_1$ — несмещенные (unbiased) оценки параметров β_0 и β_1 уравнения регрессии.

Вычислим теперь дисперсии оценок $\widehat{\beta}_0$ и $\widehat{\beta}_1$. Для этого воспользуемся тем фактом, что из условий 2 и 3 на ошибки регрессии следует, что $\mathrm{Var}(y_i) = \sigma^2$ и $\mathrm{cov}(y_i, y_j) = 0$ при $i \neq j$. Следовательно, используя свойства дисперсии, для оценки $\widehat{\beta}_1$ получаем

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}\left(\frac{\sum\limits_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{\mathrm{Var}\left(\sum\limits_{i=1}^{n} (x_i - \bar{x}) y_i \right)}{\left(\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} =$$

$$= \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \,\mathrm{Var}(y_i)}{\left(\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2}{\left(\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \sigma^2 \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \frac{\sigma^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}.$$

Для нахождения дисперсии оценки $\widehat{\beta}_0$ сначала перепишем ее в виде

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \frac{1}{n} y_i - \bar{x} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right) y_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right) y_i.$$

Следовательно,

$$\mathrm{Var}(\widehat{\beta}_0) = \mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right) y_i \right) =$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 \mathrm{Var}(y_i) =$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - 2 \frac{\bar{x}(x_i - \bar{x})}{n \sum\limits_{i=1}^{n} (x_i - \bar{x})^2} + \frac{(\bar{x})^2 (x_i - \bar{x})^2}{\left(\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \right) =$$

$$= \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} - \frac{2\bar{x} \sum_{i=1}^{n} (x_i - \bar{x})}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{(\bar{x})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \right) =$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma^2 \cdot \bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Покажем теперь, что любая другая линейная несмещенная оценка имеет бóльшую дисперсию. Пусть $\widetilde{\beta}_1 = \sum c_i y_i$ — произвольная линейная (по y_i) несмещенная оценка параметра наклона β_1. Представим ее коэффициенты c_i как $c_i = \omega_i + \theta_i$, где

$$\widehat{\beta}_1 = \sum \omega_i y_i, \quad \omega_i = (x_i - \bar{x}) / \sum (x_i - \bar{x})^2.$$

Так как $\mathsf{E}\widetilde{\beta}_1 = \mathsf{E}\widehat{\beta}_1 = \beta_1$, мы получаем

$$0 = \mathsf{E}\widetilde{\beta}_1 - \mathsf{E}\widehat{\beta}_1 = \mathsf{E}(\widetilde{\beta}_1 - \widehat{\beta}_1) = \mathsf{E}\left(\sum \theta_i y_i\right) = \sum \theta_i \mathsf{E} y_i =$$
$$= \sum \theta_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum \theta_i + \beta_1 \sum \theta_i x_i.$$

Так как это равенство должно быть выполнено для произвольных значений β_0 и β_1, получаем, что

$$\sum \theta_i = 0, \quad \sum \theta_i x_i = 0.$$

Далее,

$$\operatorname{Var}(\widetilde{\beta}_1) = \operatorname{Var}\left(\sum c_i y_i\right) = \sum c_i^2 \operatorname{Var}(y_i) =$$
$$= \sigma^2 \sum_i (\omega_i + \theta_i)^2 = \sigma^2 \left(\sum \omega_i^2 + 2 \sum \omega_i \theta_i + \sum \theta_i^2\right).$$

По условию $\omega_i = (x_i - \bar{x}) / \left(\sum (x_i - \bar{x})^2\right)$, поэтому

$$\sum \omega_i \theta_i = \sum \frac{(x_i \theta_i - \bar{x} \theta_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i \theta_i - \bar{x} \sum \theta_i}{\sum (x_i - \bar{x})^2} = 0.$$

Так как $\operatorname{Var}(\widehat{\beta}_1) = \sigma^2 \sum \omega_i^2$, окончательно получаем

$$\operatorname{Var}(\widetilde{\beta}_1) = \sigma^2 \left(\sum \omega^2 + \sum \theta_i^2\right) = \sigma^2 \sum \omega_i^2 + \sigma^2 \sum \theta_i^2 =$$
$$= \operatorname{Var}(\widehat{\beta}_1) + \sigma^2 \sum \theta_i^2 \geqslant \operatorname{Var}(\widehat{\beta}_1).$$

Таким образом, $\operatorname{Var}(\widetilde{\beta}_1) \geqslant \operatorname{Var}(\widehat{\beta}_1)$.

Аналогично можно показать, что для произвольной несмещенной оценки $\widetilde{\beta}_0$ параметра β_0 всегда выполняется неравенство

$$\mathrm{Var}(\widetilde{\beta}_0) \geqslant \mathrm{Var}(\widehat{\beta}_0).$$

Теорема доказана. □

Замечание 1. Из доказательства видно, что для несмещенности OLS-оценок достаточно *только* условия 1 на ошибки регрессии.

Замечание 2. Можно показать, что

$$\mathrm{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Замечание 3. Из теоремы Гаусса—Маркова следует, что среди линейных по y несмещенных оценок параметров β_0 и β_1 наилучшими (т. е. с минимальной дисперсией) будут OLS-оценки. Однако могут существовать и нелинейные оценки параметров β_0 и β_1 с дисперсией, меньшей, чем у OLS-оценок.

Найдем теперь оценку третьего параметра уравнения регрессии — дисперсии ошибок σ^2. Обозначим через

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

прогноз фактора y при заданном значении x_i. Значения $\widehat{y}_i$ также называются *подогнанными* (fitted value) или *предсказанными* значениями зависимой переменной.

Определение. *Остатки* (residual) модели регрессии определяются равенством $e_i = y_i - \widehat{y}_i$.

Важно различать в модели регрессии ошибки ε_i и остатки e_i. Остатки также являются случайными величинами, но, в отличие от ошибок (имеющих теоретический характер), они наблюдаемы. Кроме того, для остатков всегда выполнено соотношение $\sum_{i=1}^{n} e_i = 0$, следующее из первого уравнения системы (1.2), т. е. остатки *всегда зависимы*, в отличие от ошибок регрессии ε_i. Но, тем не менее, можно считать, что остатки в некотором смысле «моделируют» ошибки регрессии и «наследуют» их свойства. На этом основаны методы исследования отклонений выборочных данных от предположений теоремы Гаусса—Маркова.

Введем следующее обозначение:

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Величина RSS называется *остаточной суммой квадратов* (residual sum of squares) в модели регрессии. Можно показать, что

$$\mathrm{E}(\mathrm{RSS}) = (n-2)\sigma^2.$$

Следовательно, статистика

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

является несмещенной оценкой дисперсии ошибок регрессии. Выборочная *стандартная ошибка* регрессии SER (Standard Error of Regression) определяется как

$$\mathrm{SER} = s = \sqrt{s^2} = \sqrt{\frac{\mathrm{RSS}}{n-2}}.$$

1.3.2. Статистические свойства OLS-оценок коэффициентов

При доказательстве теоремы Гаусса—Маркова мы нашли дисперсии оценок параметров регрессии $\widehat{\beta}_0$ и $\widehat{\beta}_1$. В выражениях для дисперсий участвует дисперсия ошибок σ^2, значение которой в большинстве прикладных задач неизвестно. Поэтому в прикладных вычислениях используют *оценки* дисперсий величин $\widehat{\beta}_0$ и $\widehat{\beta}_1$:

$$\widehat{\mathrm{Var}}(\widehat{\beta}_0) = \frac{s^2 \cdot \bar{x}^2}{\sum (x_i - \bar{x})^2},$$

$$\widehat{\mathrm{Var}}(\widehat{\beta}_1) = \frac{s^2}{\sum (x_i - \bar{x})^2},$$

$$\widehat{\mathrm{cov}}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{s^2 \cdot \bar{x}}{\sum (x_i - \bar{x})^2},$$

получаемые формальной заменой неизвестного параметра σ^2 в выражениях для дисперсии и ковариации оценок коэффициентов на его несмещенную оценку s^2. Стандартные ошибки оценок коэффи-