

# ЭКОНОМЕТРИКА

Тема лекции:

*Парная регрессия*

Преподаватель:

к.т.н., доцент Павел Александрович Прохоренков

## Модель парной регрессии

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

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

где  $X$  – неслучайная величина,

$Y$  и  $\varepsilon$  - случайные величины.

Величина  $Y$  называется **объясняемой** (зависимой) переменной, а  $X$  - **объясняющей** (независимой) переменной. Постоянные  $\alpha$  и  $\beta$  - параметры

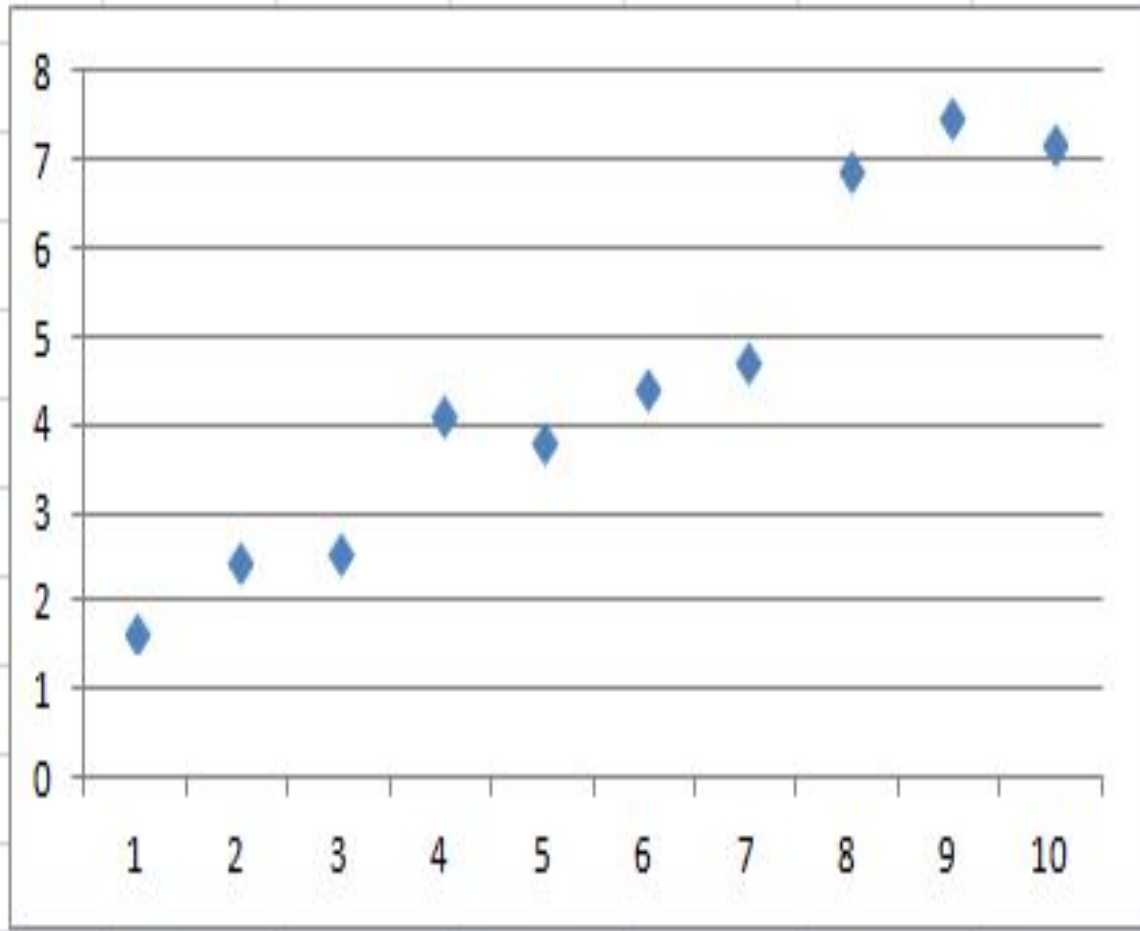
Наличие случайного члена  $\varepsilon$  (ошибка регрессии) связано с воздействием на зависимую переменную других неучтенных в уравнении факторов.

Задача моделирования: на основе выборочного наблюдения оценивается выборочное уравнение регрессии (или линия регрессии):

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

где коэффициенты ( $a$  и  $b$ ) - оценки параметров ( $\alpha$  и  $\beta$ ).

X	Y
1	1,688
2	2,487
3	2,551
4	4,125
5	3,837
6	4,451
7	4,713
8	6,906
9	7,483
10	7,181



## Метод наименьших квадратов

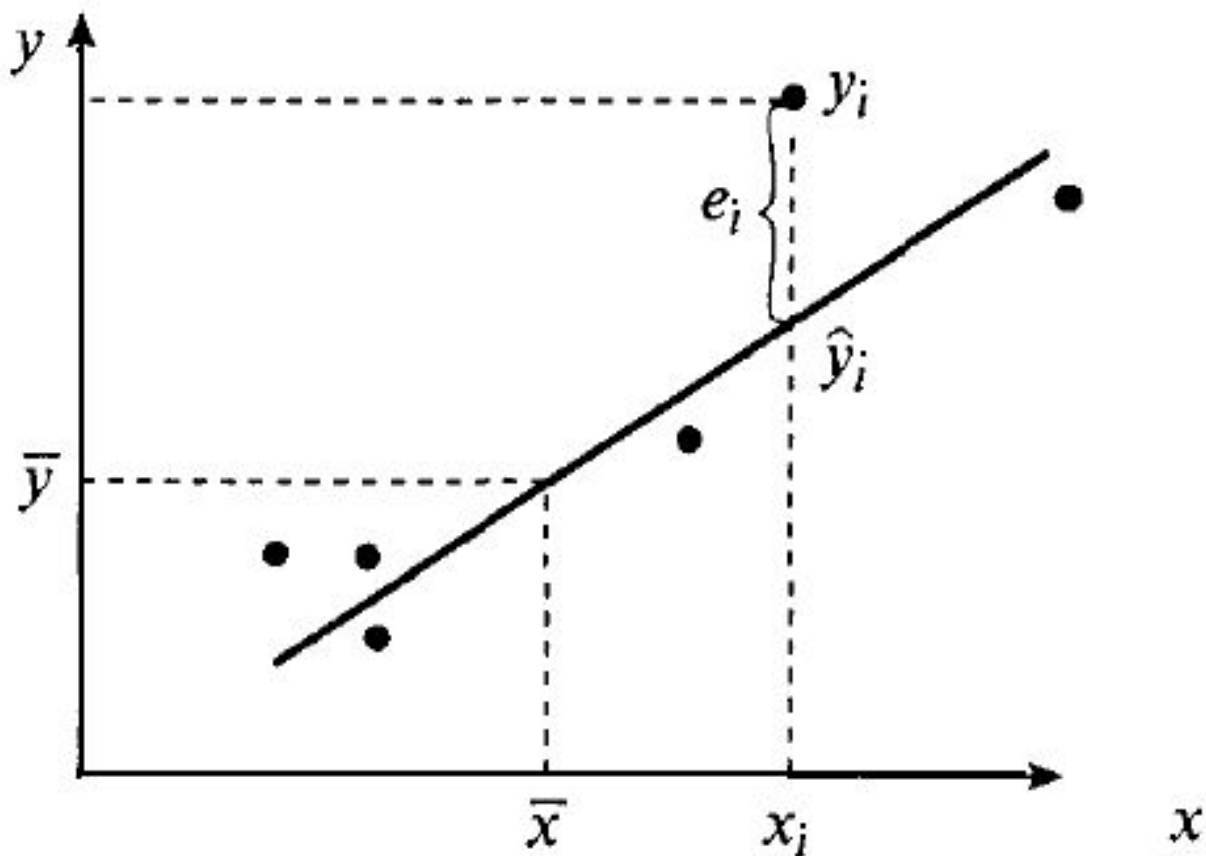
Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений  $(x_i, y_i), i = 1, \dots, n$  уравнением (2.2).

На рисунке приведены диаграмма рассеяния и линия регрессии.

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

Найдем остаток  $e_i$  в  $i$ -ом наблюдении как разность между фактическим и расчетным значениями зависимой переменной, т.е.

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





Неизвестные значения (**a** и **b**) определяются методом **наименьших квадратов** (МНК). Суть МНК заключается в минимизации суммы квадратов остатков:

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

В данном выражении  $(x_i, y_i)$  - известные значения наблюдения (какие-то числа), (**a**, **b**) – неизвестные, которые надо найти.

Запишем необходимые условия экстремума:

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

После преобразований получим систему нормальных уравнений относительно  $(\mathbf{a}, \mathbf{b})$ : (справка: система уравнений является нормальной относительно переменных  $\mathbf{a}, \mathbf{b}$ , если это система алгебраических линейных уравнений относительно этих переменных):

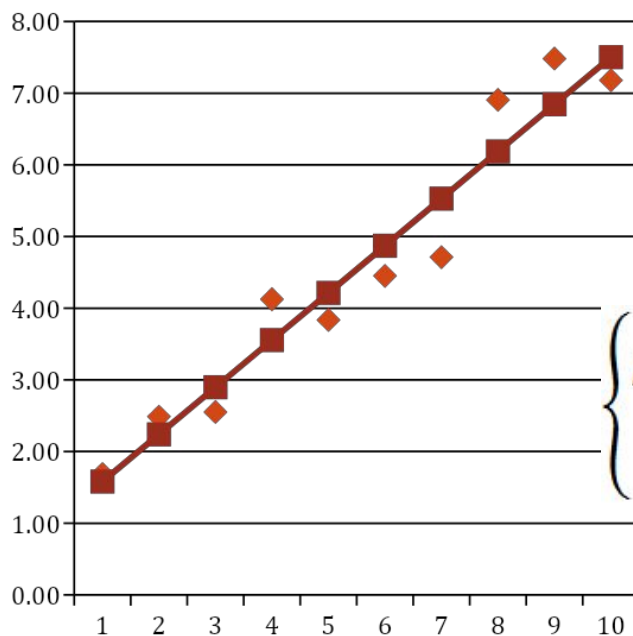
$$\begin{cases} \mathbf{a} + \mathbf{b}\bar{x} = \bar{y}, \\ \mathbf{a}\bar{x} + \mathbf{b}\bar{x}^2 = \overline{xy}. \end{cases}$$

Решение системы:

$$\begin{cases} \mathbf{b} = \frac{\text{cov}(x,y)}{\text{var}(x,y)} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2}, \\ \mathbf{a} = \bar{y} - \mathbf{b}\bar{x}. \end{cases} \quad (2.3)$$



10	X	Y	y1	XY	Xкв	Yкв	e
	1	1,69	1,58	1,69	1	2,85	0,11
	2	2,49	2,24	4,97	4	6,19	0,25
	3	2,55	2,90	7,65	9	6,51	-0,35
	4	4,13	3,56	16,50	16	17,02	0,57
	5	3,84	4,21	19,19	25	14,72	-0,38
	6	4,45	4,87	26,71	36	19,81	-0,42
	7	4,71	5,53	32,99	49	22,21	-0,82
	8	6,91	6,19	55,25	64	47,69	0,72
	9	7,48	6,85	67,35	81	56,00	0,64
	10	7,18	7,50	71,81	100	51,57	-0,32
сумма	55	45,422	45,422	304,102	385	244,559	
средние	5,5	4,5422	4,5422	30,4102	38,5	24,4559	



$$\hat{y} = a + bx$$

$$\left\{ \begin{aligned} b &= \frac{cov(x, y)}{var(x, y)} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = 0,658 \\ a &= \bar{y} - b\bar{x} = 0,923 \end{aligned} \right.$$

Можно показать, что

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

где  $r$  – коэффициент корреляции между  $X$  и  $Y$ ,

$S_x, S_y$  - их стандартные отклонения. Таким образом, если

коэффициент корреляции уже рассчитан, можно найти коэффициенты (a

b) парной регрессии.

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

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

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

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

## Анализ вариации зависимой переменной

Цель регрессионного анализа состоит в объяснении поведения зависимой переменной  $y$ .

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

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

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



Разброс значений зависимой переменной характеризуется выборочной дисперсией  $\mathit{var}(\mathbf{y})$ , которую можно представить как:

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

Можно доказать, что  $\mathit{cov}(\hat{\mathbf{y}}, \mathbf{e}) = \mathbf{0}$ , тогда

$$\mathit{var}(\mathbf{y}) = \mathit{var}(\hat{\mathbf{y}}) + \mathit{var}(\mathbf{e}). \quad (2.5)$$

Таким образом, дисперсия  $\mathit{var}(\mathbf{y})$  разложена на 2 части:

- $\mathit{var}(\hat{\mathbf{y}})$  – часть объясненная регрессионным уравнением,
- $\mathit{var}(\mathbf{e})$  – необъясненная часть.



Коэффициентом детерминации  $R^2$  называется отношение:

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

$R^2$  характеризует долю вариации зависимой переменной, объясненную с помощью уравнения регрессии.

Отношение  $\frac{\text{var}(e)}{\text{var}(y)}$  представляет собой долю необъясненной регрессии.

Если  $R^2 = 1$ , то регрессия точно описывает выборку:

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

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

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

$$\text{var}(y) = \text{var}(e), \text{var}(\hat{y}) = 0, \hat{y}_i = \bar{y}, i = 1, \dots, n,$$

т.е. переменная  $X$  не улучшает качества предсказания  $y$  по сравнению с

горизонтальной прямой  $\hat{y}_i = \bar{y}$ .

Чем ближе к единице  $R^2$ , тем  $\hat{y}$  более точно аппроксимирует  $y$ .

Замечание. Вычисление  $R^2$  корректно, если константа  $a$  включена в

уравнение регрессии.

## F – тест на качество оценивания

Для определения статистической значимости коэффициента детерминации  $R^2$  проверяется гипотеза:  $H_0: F \equiv 0$  для  $F$  – статистики:

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

Величина  $F$  имеет распределение Фишера с  $v_1 - 1, v_2 = n - 2$ .

Проверку значимости  $R^2$  можно выполнить двумя способами.

1. Критическое значение  $F_{\text{крит}}$  при заданных  $\alpha, v_1, v_2$

определяется по таблице  $F$ -распределения Фишера или в Excel с помощью функции

$$F_{\text{крит}} = \text{FRASPOBR}(\alpha, v_1, v_2).$$

Из сравнения наблюдаемого значения  $F$  с критическим, получаем:

- Если  $F < F_{\text{крит}}$ , то  $H_0$  принимается, т.е.  $R^2$  незначим.
- Если  $F > F_{\text{крит}}$ , то  $H_0$  отвергается, т.е.  $R^2$  значим.

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

Значимость  $F = \text{FPACII}(F, \nu_1, \nu_2)$ .



Из сравнения значимости  $F$  с заданным стандартным уровнем значимости, получаем:

- Если значимость  $F$  больше стандартного уровня, то  $R^2$  незначим.
- Если значимость  $F$  меньше стандартного уровня, то  $R^2$  значим.

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



## Средняя ошибка аппроксимации

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

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

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

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

# Свойства коэффициентов регрессии и проверка гипотез

*Случайные составляющие коэффициентов регрессии*

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

- Неслучайную  $(\alpha + \beta X)$ ;
- Случайную  $(\varepsilon)$ .

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

$$\begin{cases} \mathbf{b} = \boldsymbol{\beta} + \frac{\text{cov}(x, \boldsymbol{\varepsilon})}{\text{var}(x)} \\ \mathbf{a} = \boldsymbol{\alpha} + \left[ \frac{1}{n} \sum \boldsymbol{\varepsilon}_i - \frac{\bar{x} \text{cov}(x, \boldsymbol{\varepsilon})}{\text{var}(x)} \right] \end{cases} \quad (3.1)$$

Здесь коэффициенты  $(\mathbf{a}, \mathbf{b})$  разложены на две составляющие: неслучайную, равную истинным значениям  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ , и случайную, зависящую от  $\boldsymbol{\varepsilon}$ .

На практике нельзя разложить коэффициенты регрессии на составляющие, так как значения  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  или фактические значения  $\boldsymbol{\varepsilon}$  в выборке неизвестны.

## Предпосылки регрессионного анализа. Условия Гаусса - Маркова

Линейная регрессионная модель с двумя переменными имеет вид:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

где  $Y$  - объясняемая переменная,

$X$  - объясняющая переменная,

$\varepsilon$  - случайный член.

Для того, чтобы регрессионный анализ, основанный на МНК, давал наилучшие, из всех возможных, результаты, должны выполняться определенные условия, которые называются условиями Гаусса – Маркова:



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

$$M(\varepsilon_i) = 0, \quad i = 1, \dots, n.$$

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

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

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

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

4. Объясняющая переменная  $X_i$  есть величина неслучайная.



При выполнении условий Гаусса – Маркова модель называется классической нормальной регрессионной моделью. Также обычно предполагают, что случайный член распределен нормально, т.е.

$$\varepsilon_i \sim N(0; \sigma^2).$$

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

Рассмотрим, что означают условия Гаусса – Маркова.

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

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

Под дисперсией  $\sigma^2$  имеется в виду возможное поведение случайного члена до того, как сделана выборка. Величина  $\sigma^2$  неизвестна и оценка  $\hat{\sigma}^2$  - одна из задач регрессионного анализа.

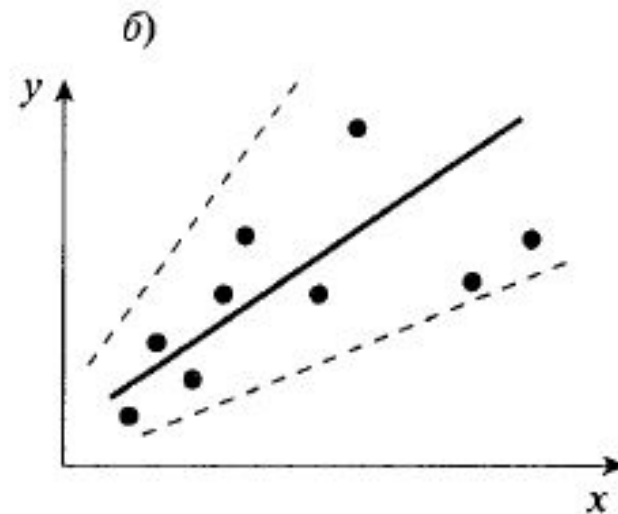
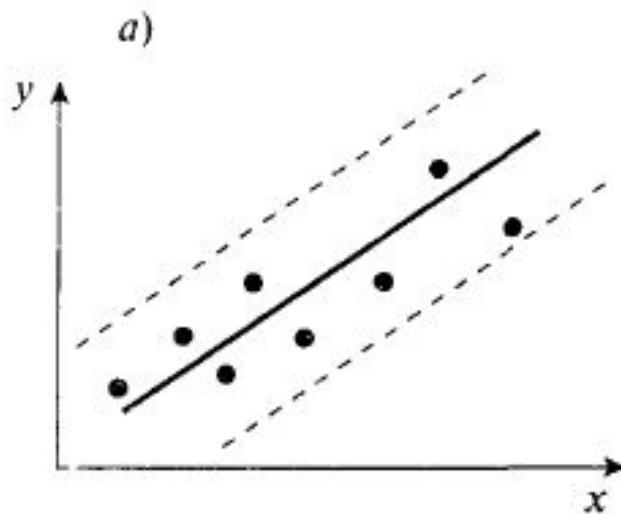
Условие независимости дисперсии случайного члена от номера наблюдения называется **гомоскедастичностью** (этот термин означает «одинаковый разброс»).

Зависимость дисперсии случайного члена от номера наблюдения называется гетероскедастичностью.

Таким образом:

- $D(\varepsilon_i) = \sigma^2, i = 1, \dots, n$  - гомоскедастичность
- $D(\varepsilon_i) = \sigma_i^2, i = 1, \dots, n$  - гетероскедастичность.

Характерные диаграммы рассеяния для случая гомоскедастичности (а) и случай гетероскедастичности (б) показаны на рисунке:





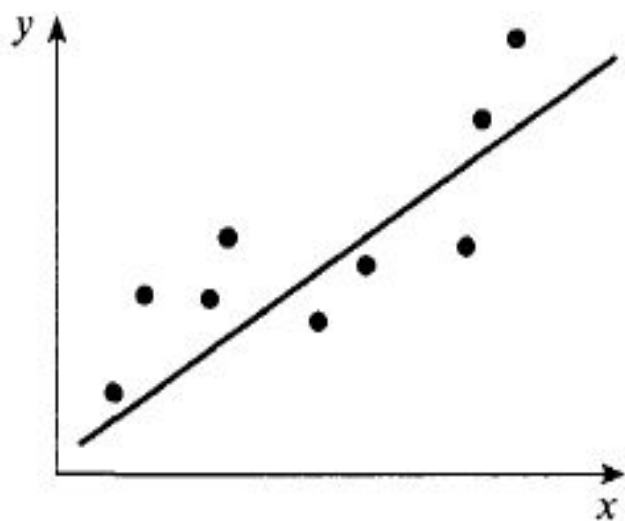
Если условие гомоскедастичности не выполняется, то оценки коэффициентов регрессии будут *неэффективными*, хотя и *несмещенными*.

Существуют специальные методы позволяющие определить наличие гетероскедастичности и устранить ее.

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

Типичный вид данных при автокорреляции остатков показан на рисунке:





(Можно представить, что точки выборки разбросаны вокруг линии регрессии по какому-то закону).

Если условие независимости остатков не выполняется, то оценки коэффициентов регрессии, полученные по методу МНК, будут *неэффективными*, хотя и *несмещенными*.

Существуют специальные методы позволяющие определить наличие автокорреляции и устранить ее.

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

Нарушение четвертого условия может быть связано с ошибками измерения объясняющих переменных или с использованием лаговых переменных. (Для справки: лаговые переменные – это переменные относящиеся к предыдущим моментам времени).

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

## Теорема Гаусса - Маркова

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

1) Несмещенность, т.е.  $M(a) = \alpha$ ,  $M(b) = \beta$ .

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

2) Эффективность, т.е. имеет наименьшую дисперсию в классе всех линейных несмещенных оценок, равную:

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

3) Состоятельность, т.е.

$$\lim_{n \rightarrow \infty} D(a) = 0, \quad \lim_{n \rightarrow \infty} D(b) = 0.$$

(Это означает, что при достаточно большом  $n$  оценки  $(a, b)$  близки к  $(\alpha, \beta)$ ).



## Расчет стандартных ошибок коэффициентов регрессии

Полученные теоретические дисперсии  $D(a)$  и  $D(b)$  зависят от дисперсии  $\sigma^2$  случайного члена.

По данным выборки отклонения  $\varepsilon_i$ , а следовательно, и их дисперсии  $\sigma^2$  найти невозможно, поэтому их заменяют наблюдаемыми остатками  $e_i$  и их выборочной дисперсией.

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

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

Несмещенной оценкой дисперсии  $\sigma^2$  является величина остаточной дисперсии, которую находят по формуле:

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

которая служит мерой разброса зависимой переменной вокруг линии регрессии.

Величина  $S$  называется стандартной ошибкой регрессии.

Заменяя в теоретических дисперсиях  $D(a)$  и  $D(b)$ , неизвестную  $\sigma^2$  на ее оценку  $S^2$ , получим оценки дисперсий коэффициентов уравнения регрессии:

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

Величины  $S_a$  и  $S_b$  называются стандартными ошибками коэффициентов регрессии.

### Статистические свойства МНК - оценок $(a, b)$

Пусть выполняется условие нормальности распределения случайного члена, т.е.  $\varepsilon_i \sim N(0; \sigma^2)$ . Тогда МНК – оценки коэффициентов регрессии также имеют нормальное распределение, т.е.

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

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

# Проверка гипотез, относящихся к коэффициентам регрессии ( $\alpha$ , $\beta$ )

*Проверка гипотезы  $H_0: \beta = \beta_0$*

Пусть в теоретической зависимости

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

Случайный член  $\varepsilon$  распределен нормально с неизвестной дисперсией  $\sigma^2$ .

Хотя величина  $\beta$  неизвестна, можно предполагать, что она равна заданной величине  $\beta_0$ .

Выдвигаются гипотезы

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



Задача состоит в проверке гипотезы  $H_0$  на основании выборочных данных.

Пусть по выборочным данным получена оценка  $b$ .

В качестве критерия проверки гипотезы  $H_0$  принимают случайную величину, равную

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

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

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

$t_{\text{крит}}$ .

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

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

Любое значение  $\beta$ , совместимое с оценкой  $b$ , удовлетворяет условию

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

Разрешив это неравенство относительно  $\beta$ , получаем

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

т.е. получили **доверительный интервал** для величины  $\beta$ .

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

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

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

## Проверка гипотезы $H_0: \beta = 0$

Пусть по выборке получена оценка коэффициента регрессии  $b$ .

Гипотеза  $H_0: \beta = 0$  используется для установления значимости коэффициента регрессии  $b$ .

Тогда величина  $t$ , используемая в качестве критерия проверки гипотезы  $H_0$ , равна

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

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



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

Значимость  $t = \text{СТЮДРАСП}(t; \nu; 2)$ .

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

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

## Пример

Регрессионная модель зависимости расходов на питание  $y$  от личного дохода  $x$  имеет вид:

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

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

Найти стандартные ошибки коэффициентов регрессии, оценить значимость коэффициента регрессии  $b$  при 5%-ном уровне значимости, построить доверительный интервал для коэффициента  $\beta$  при том же уровне значимости.

1. Расчет стандартных ошибок коэффициентов регрессии  $(a, b)$ .

Находим несмещенную оценку дисперсии  $\sigma^2$  (ее также называют остаточной дисперсией):

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

Тогда

$$S_a^2 = \frac{\overline{x^2} \cdot S^2}{n \cdot \text{var}(x)} = \frac{132 \cdot 3,3}{5 \cdot 32} = 2,7225$$

$$S_b^2 = \frac{S^2}{n \cdot \text{var}(x)} = \frac{3,3}{5 \cdot 32} = 0,020625$$

Тогда стандартные ошибки коэффициентов регрессии равны:

Тогда стандартные ошибки коэффициентов регрессии равны:

$$S_a = \sqrt{2,7225} = 1,65$$

$$S_b = \sqrt{0,020625} = 0,1436.$$

2. Найдем наблюдаемое значение критерия  $t$ :

$$t = \frac{b}{S_b} = \frac{0,775}{0,1436} = 5,3969.$$

Для вычисления значимости  $t$  воспользуемся Excel

Значимость  $t = \text{СТЮДРАСП}(t; v; 2)$ .



Находим значимость  $t = 0,01247$ .

Вывод: т.к. значимость  $t$  меньше заданного стандартного уровня значимости  $(0,05)$ , то коэффициент  $b$  значим.

3. Найдем доверительный интервал для  $\beta$ .

Для этого требуется найти  $t_{\text{крит}}$ . Воспользуемся таблицами распределения Стьюдента или функцией Excel

$$t_{\text{крит}} = \text{СТЮДРАСПОБР}(\alpha, v)$$

Получаем:

$$t_{\text{крит}} = 3,1824.$$

Доверительный интервал для  $\beta$  имеет вид:

$$b - t_{\text{крит}} S_b < \beta < b + t_{\text{крит}} S_b$$

Подставляя значения, получаем:

$$0,775 - 3,18 \cdot 0,1436 < \beta < 0,775 + 3,18 \cdot 0,1436$$

Подсчитывая, получаем доверительный интервал для  $\beta$ :

$$0,318 < \beta < 1,23.$$

Стемп Параметры страницы Описание

СТЮДРАСП =СТЮДРАСП(C2;3;2)

A	B	C	D	E	F	G	H	I
	t	5,3969						
	значимость t	=СТЮДРАСП(C2;3;2)						

**Аргументы функции**

СТЮДРАСП

X	C2		= 5,3969
Степени_свободы	3		= 3
Хвосты	2		= 2

= 0,012468211

Возвращает t-распределение Стьюдента.

**Хвосты** число возвращаемых хвостов распределения (1 или 2).

Значение: 0,012468211

[Справка по этой функции](#)

# Взаимосвязь критериев

В парном регрессионном анализе эквивалентны  $t$  – критерий для

$H_0: \beta = 0$ ;  $t$  – критерий для  $H_0: \rho = 0$ ;  $F$  – критерий для  $R^2$ :

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

Связь между этими критериями выражается равенством:

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



Причем соотношение между критическими значениями критериев при любом уровне значимости имеет вид:

$$(t_b)_{\text{крит}} = (t_r)_{\text{крит}} = (\sqrt{F})_{\text{крит}},$$

и все эти критерии дают одинаковый результат.

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

# Прогнозирование в регрессионных моделях

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

Различают точечное и интервальное прогнозирование. При точечном прогнозировании оценка – некоторое конкретное число, при интервальном – интервал, в котором находится истинное значение зависимой переменной с заданным уровнем значимости.

Рассмотрим регрессионную модель

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

Действительное значение зависимой переменной при  $x = x_p$  (х предсказания):

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

где  $M(\varepsilon_p) = 0$ ,  $D(\varepsilon_p) = \sigma^2$ . Значения  $\alpha, \beta, \varepsilon_p$  неизвестны.

Предсказанным значением является оценка  $y_p$  (точечный прогноз):

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

Ошибка предсказания равна разности между предсказанным и действительным значениями:

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



Ошибка предсказания имеет нулевой математическое ожидание:

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

Дис

Ошибка предсказания имеет нулевой математическое ожидание:

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

Дисперсия прогноза равна:

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

Из формулы следует, что чем больше  $x_p$  отклоняется от выборочного

Из среднего  $\bar{x}$ , тем больше дисперсия ошибки предсказания, и чем больше выборочного

объем выборки  $n$ , тем меньше дисперсия предсказания.

среднего  $\bar{x}$ , тем больше дисперсия ошибки предсказания, и чем больше

объем выборки  $n$ , тем меньше дисперсия предсказания.



Заменяя в дисперсии прогноза  $\sigma^2$  на ее оценку  $S^2$  и извлекая квадратный корень, получим стандартную ошибку предсказания:

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

Доверительный интервал для действительного значения  $y_p$  определяется выражением

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

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

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

