

# **Эконометрика**

## **Тема 3**

### ***Тема 3. Модель парной линейной регрессии.***

- 1) Функциональная, статистическая и корреляционная зависимости.**
- 2) Линейная парная регрессия и коэффициент корреляции.**
- 3) Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели. Теорема Гаусса—Маркова.**
- 4) Интервальная оценка функции регрессии и ее параметров.**
- 5) Оценка значимости уравнения регрессии. Коэффициент детерминации.**
- 6) Геометрическая интерпретация регрессии и коэффициента детерминации.**
- 7) Коэффициент ранговой корреляции Спирмена.**

# 1. Функциональная, статистическая и корреляционная зависимости.

## Виды зависимостей

**Функциональная**  
- каждому значению одной переменной соответствует вполне **определенное значение** другой

**Статистическая (стохастическая, вероятностная)** - каждому значению одной переменной соответствует **определенное (условное) распределение** другой переменной

**Корреляционная** - каждому значению одной переменной соответствует **определенное условное математическое ожидание (среднее значение)** другой – иначе это функциональная зависимость между значениями одной переменной и условным математическим ожиданием другой

Корреляционная зависимость может быть представлена в виде:

$$M_x(Y) = \varphi(x) \quad , \quad \text{где } \varphi(x) \neq \text{const.}$$

В регрессионном анализе рассматриваются **односторонняя зависимость случайной переменной  $Y$  от одной (или нескольких) неслучайной независимой переменной  $X$**  ( $Y$  подвержена случайному разбросу за счет действия ряда **неконтролируемых факторов**)

# 1. Функциональная, статистическая и корреляционная зависимости.

**При этом:**

- 1)  $M_x(Y) = \varphi(x)$  - модельное уравнение регрессии (или просто **уравнение регрессии**);
- 2)  $\varphi(x)$  - модельная функция регрессии (или просто **функция регрессии**);
- 3) график функции  $\varphi(x)$  - модельная линия регрессии (или просто **линия регрессии**);
- 4)  $\hat{y} = \hat{\varphi}(x, b_0, b_1, \dots, b_p)$  где  $\hat{y}$  — условная (групповая) средняя переменной  $Y$  при фиксированном значении переменной  $X = x$ ,  $b_0, b_1, \dots, b_p$  — параметры кривой.

- **оценка** (приближенное выражение, аппроксимация) **по выборке функции регрессии** — используется на практике, т.к. точный закон условного распределения  $Y$ , как правило, **неизвестен**.

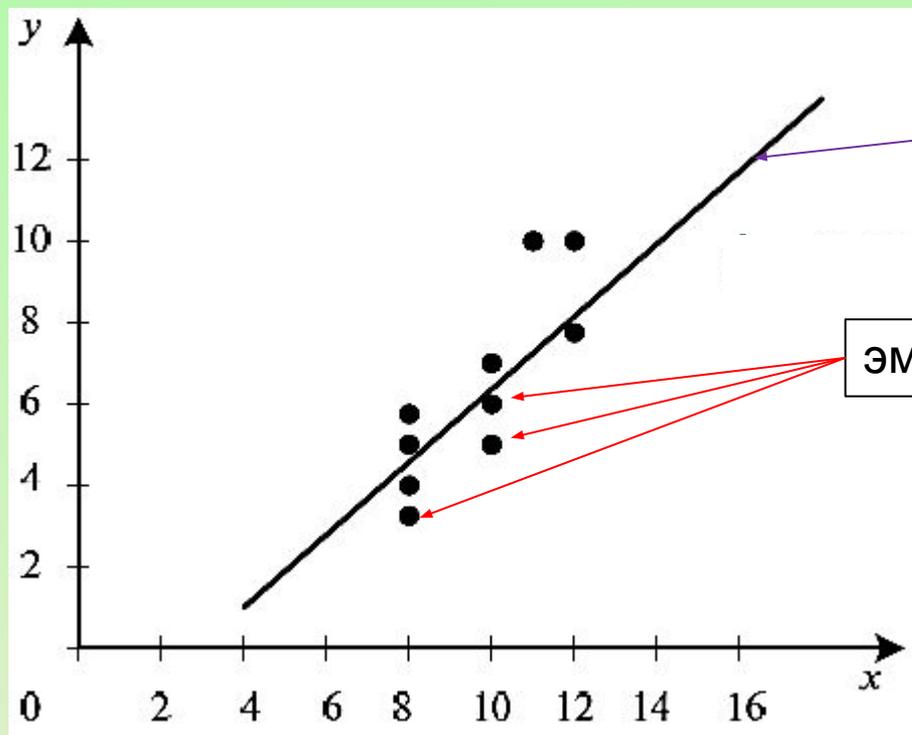
Уравнение 4) — это **выборочное уравнение регрессии** или **выборочная линия (кривая) регрессии**. При правильно определенной аппроксимирующей функции  $\hat{\varphi}(x, b_0, b_1, \dots, b_p)$  при  $n \rightarrow \infty$  она будет сходиться по вероятности к функции регрессии  $\varphi(x)$ .

## 2. Линейная парная регрессия и коэффициент корреляции.

Рассмотрим зависимость между сменной добычей угля на одного рабочего  $Y(m)$  и мощностью пласта  $X(m)$  для  $n=10$  шахт:

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	8	11	12	9	8	8	9	9	8	12
$y_i$	5	10	10	7	5	6	6	5	6	8

Изобразим полученную зависимость *графически*:



Предполагаемая *линейная корреляционная (регрессионная) зависимость* между переменными  $X$  и  $Y$

эмпирические точки *поля корреляции*

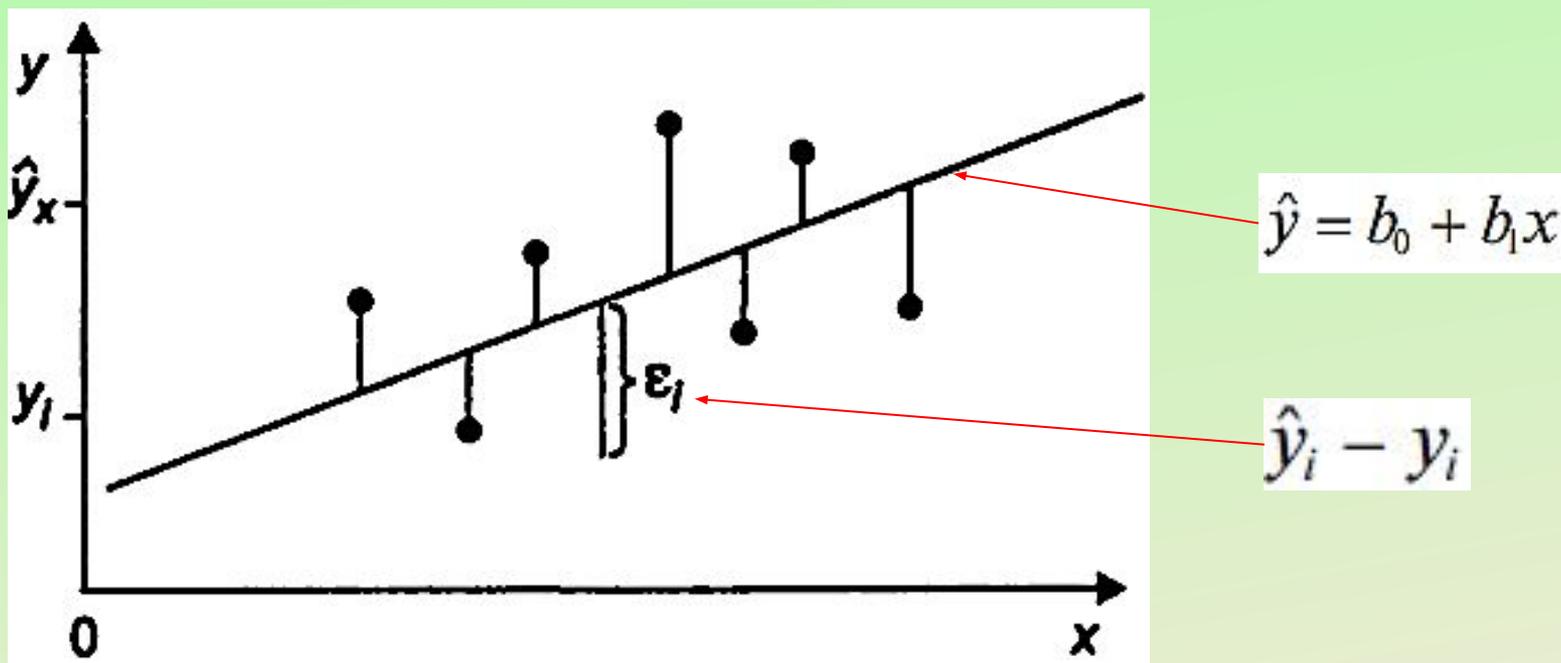
уравнение регрессии будем искать в виде:

$$\hat{y} = b_0 + b_1x$$

## 2. Линейная парная регрессия и коэффициент корреляции.

Согласно **методу наименьших квадратов** неизвестные параметры  $b_0$  и  $b_1$  выбираются таким образом, чтобы сумма квадратов отклонений эмпирических значений  $y_i$  от значений  $\hat{y}_i$ , найденных по уравнению регрессии, была **минимальной**:

$$S = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (b_0 + b_1 x_i - y_i)^2 \rightarrow \min$$



## 2. Линейная парная регрессия и коэффициент корреляции.

Определение коэффициентов уравнения парной линейной регрессии по методу наименьших квадратов:

$$S = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (b_0 + b_1 x_i - y_i)^2 \rightarrow \min$$

$$\begin{cases} \frac{\partial S}{\partial b_0} = 2 \sum_{i=1}^n (b_0 + b_1 x_i - y_i) = 0; \\ \frac{\partial S}{\partial b_1} = 2 \sum_{i=1}^n (b_0 + b_1 x_i - y_i) x_i = 0, \end{cases}$$

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i; \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}$$

*/n обе части уравнения*

(частные производные функции  $S(b_0, b_1)$ )

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y}; \\ b_0 \bar{x} + b_1 \bar{x}^2 = \overline{xy}, \end{cases} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\overline{xy} = \frac{\sum_{i=1}^n x_i y_i}{n} \quad \overline{x^2} = \frac{\sum_{i=1}^n x_i^2}{n}$$

(система нормальных уравнений)

$$b_0 = \bar{y} - b_1 \bar{x}$$

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

(выборочная ковариация)

$$\hat{y} = \bar{y} - b_1 \bar{x} + b_1 x$$

$$b_1 = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\text{Cov}(X, Y)}{s_x^2}$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2$$

(выборочная дисперсия X)

$$\hat{y} - \bar{y} = b_1 (x - \bar{x})$$

## 2. Линейная парная регрессия и коэффициент корреляции.

Для рассматриваемого примера (слайд 5):

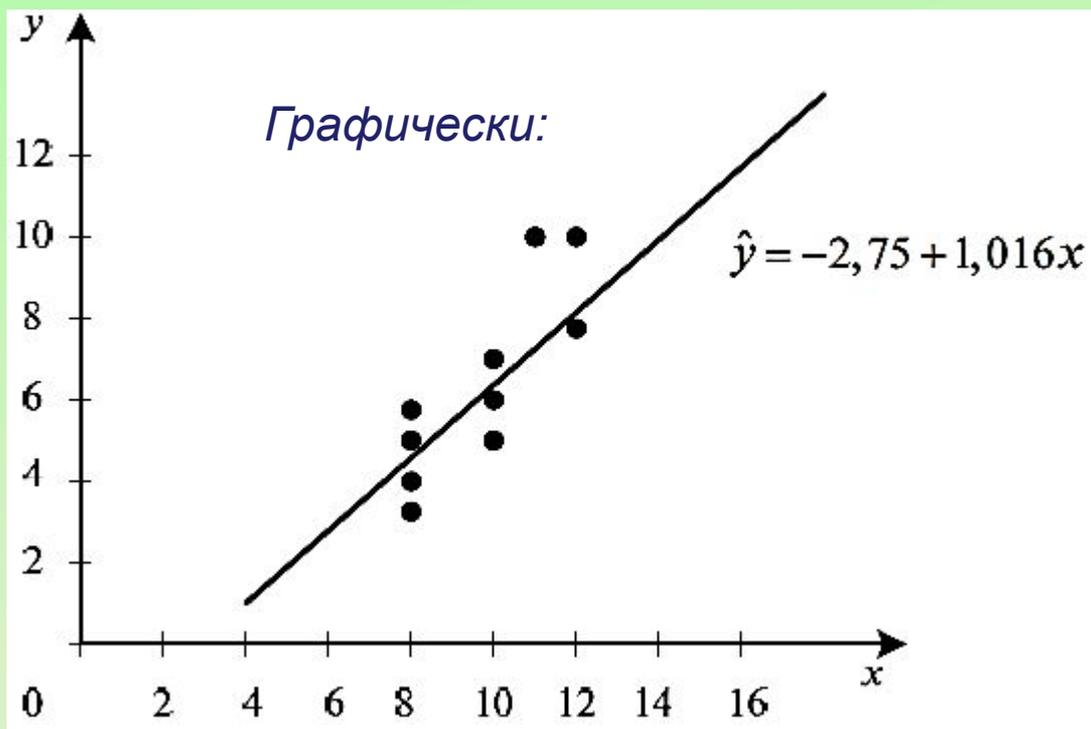
$$\bar{x} = 94/10 = 9,4 \text{ (м)}; \quad \bar{y} = 68/10 = 6,8 \text{ (т)};$$

$$s_x^2 = 908/10 - 9,4^2 = 2,44;$$

$$\text{Cov}(X, Y) = 664/10 - 9,4 \cdot 6,8 = 2,48; \quad b_1 = 2,48/2,44 = 1,016$$

Т.о., уравнение  
регрессии  $Y$  по  $X$ :

$$\hat{y} - 6,8 = 1,016(x - 9,4) \quad \text{или} \quad \hat{y} = 2,75 + 1,016x$$



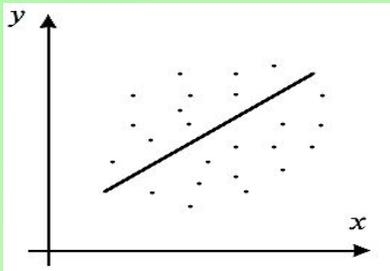
при увеличении мощности пласта  $X$  на 1 м добыча угля на одного рабочего  $Y$  увеличивается в среднем на 1,016 т (свободный член в уравнении регрессии не имеет реального смысла)

## 2. Линейная парная регрессия и коэффициент корреляции.

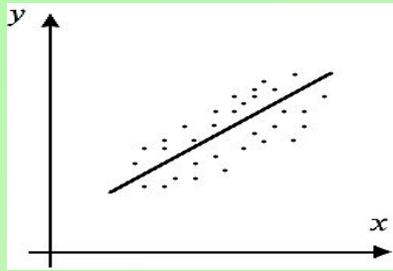
**Выборочный коэффициент корреляции** - показатель тесноты линейной связи переменных X и Y:

$$r = b_1 \frac{s_x}{s_y}, \quad s_x, s_y - \text{средние квадратные отклонения X и Y}$$

корреляционная связь слабее:



корреляционная связь сильнее:



$r$  - непосредственная **оценка** генерального коэффициента корреляции  $\rho$  между X и Y **лишь в случае двумерного нормального закона распределения** случайных величин X и Y

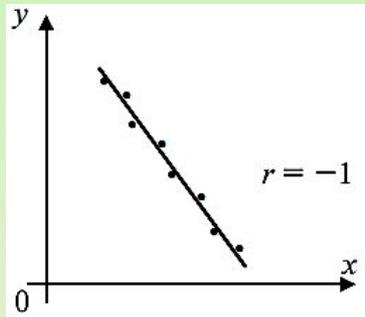
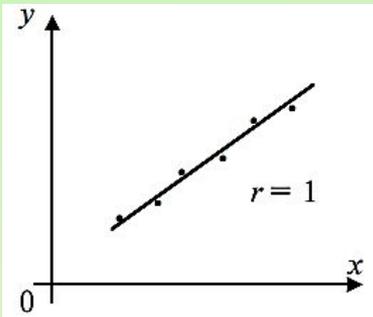
Для примера слайда 5:

$$r = \frac{10 \cdot 664 - 94 \cdot 68}{\sqrt{10 \cdot 908 - 94^2} \sqrt{10 \cdot 496 - 68^2}} = 0,866$$

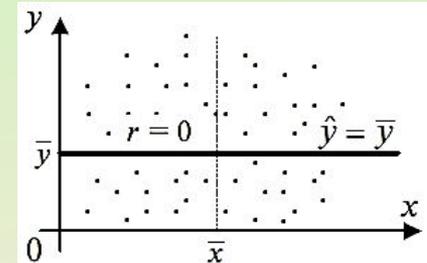
**Свойства выборочного коэффициента корреляции:**

1)  $-1 \leq r \leq 1$  - диапазон значений  $r$ ;

2) Чем ближе  $|r|$  к единице, тем **теснее связь**. При  $r = \pm 1$  корреляционная связь - **линейная функциональная зависимость** (все наблюдаемые значения располагаются на прямой линии):



3) При  $r = 0$  линейная корреляционная связь **отсутствует** (линия регрессии параллельна оси OX):



### 3. Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели. Теорема Гаусса—Маркова.

**Парная регрессионная модель:**  $Y = \varphi(X) + \varepsilon$

где  $\varepsilon$  - случайная переменная (случайный член), характеризующая отклонение от функции регрессии - возмущение, ошибка.

**Для линейного регрессионного анализа:**  $M_x(Y) = \beta_0 + \beta_1 x$

Пусть для оценки параметров линейной функции регрессии есть выборка, содержащая  $n$  пар значений переменных  $(x_i, y_i)$ , где  $i=1, 2, \dots, n$ . Тогда **линейная парная регрессионная модель:**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (*)$$

#### Основные предпосылки регрессионного анализа:

- 1) В модели (\*) возмущение  $\varepsilon_i$  (или  $y_i$ ) – величина **случайная**, а  $x_i$  - величина **неслучайная**;
- 2) Математическое ожидание возмущения равно нулю:  $M(\varepsilon_i) = 0$ ;
- 3) Дисперсия возмущения (или  $y_i$ ) постоянна для любого  $i$ :  $D(\varepsilon_i) = \sigma^2$  или  $D(y_i) = \sigma^2$ ;
- 4) Возмущения  $\varepsilon_i$  и  $\varepsilon_j$  (или переменные  $y_i$  и  $y_j$ ) не коррелированы:  $M(\varepsilon_i \varepsilon_j) = 0$  ( $i \neq j$ );
- 5) Возмущение  $\varepsilon_i$  (или зависимая переменная  $y_i$ ) есть **нормально распределенная** случайная величина – это требование необходимо для оценки **точности уравнения регрессии и его параметров**

Модель (\*) - **классическая нормальная линейная регрессионная модель**  
(Classical Normal Linear Regression model)

### 3. Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели. Теорема Гаусса—Маркова.

- Оценка параметров парной регрессионной модели
  - 1) Метод наименьших квадратов (ММК)
  - 2) Метод максимального правдоподобия

#### Оценка по ММК:

- 1) Оценка модели (\*) слайда 10 – уравнение  $\hat{y} = b_0 + b_1 x$  ( $b_0$  и  $b_1$  определяются на основе ММК);
- 2) Воздействие неучтенных случайных факторов и ошибок наблюдений в модели (\*) слайда 10 определяется с помощью дисперсии возмущений (ошибок) или **остаточной дисперсии**  $\sigma^2$ .

Несмещенная оценка этой дисперсии - **выборочная остаточная дисперсия**  $s^2$ :

$$s^2 = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

где  $\hat{y}_i$  — групповая средняя, найденная по уравнению  $s^2$  регрессии;  
 $e_i = \hat{y}_i - y_i$  — выборочная оценка возмущения  $\varepsilon_i$  или остаток регрессии.

- 3) **Теорема Гаусса—Маркова.** Если регрессионная модель (\*) удовлетворяет предпосылкам 1)-4), то оценки  $b_0$ ,  $b_1$  имеют **наименьшую дисперсию** в классе всех линейных несмещенных оценок (Best Linear Unbiased Estimator, или **BLUE**). Т.о., оценки  $b_0$  и  $b_1$  в определенном смысле являются **наиболее эффективными линейными оценками** параметров  $\beta_0$  и  $\beta_1$ .

### 3. Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели. Теорема Гаусса—Маркова.

#### Оценка по методу максимального правдоподобия:

- 1) Для его применения *должен быть известен вид закона распределения вероятностей имеющихся выборочных данных*;
- 2) Если выполняется предпосылка 5) модели (\*) слайда 10 (*нормальная классическая регрессионная модель*) и  $y_i$  – независимые нормально распределенные случайные величины с мат. ожиданием  $M(y_i) = \beta_0 + \beta_1 x_i$  и постоянной дисперсией  $\sigma^2$  то *плотность* нормально распределенной случайной величины  $y_i$ :

$$\varphi_N(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

- 3) *Функция правдоподобия*, выражающая плотность вероятности *совместного* появления результатов выборки:

$$L(y_i x_i; \dots; y_n, x_n; \beta_0, \beta_1; \sigma^2) = \prod_{i=1}^n \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

- 4) В качестве оценок параметров  $\beta_0$ ,  $\beta_1$  и  $\sigma^2$  принимаются такие значения  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ ,  $\hat{\sigma}^2$ , которые *максимизируют* функцию правдоподобия L;
- 5) Функция L достигает максимума при *минимизации функции*  $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ , поэтому оценки ММК параметров уравнения (\*) *совпадают* с оценками метода максимального правдоподобия.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \frac{\sum_{i=1}^n e_i^2}{n}$$

- оценка остаточной дисперсии модели (\*), являющаяся *смещенной* в отличие оценки ММК

## ***4. Интервальная оценка функции регрессии и ее параметров.***

## ***4. Интервальная оценка функции регрессии и ее параметров.***

### ***Вопросы изученные в Теме 3:***

- 1) Функциональная, статистическая и корреляционная зависимости.
- 2) Линейная парная регрессия и коэффициент корреляции.
- 3) Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели. Теорема Гаусса—Маркова.
- 4) Интервальная оценка функции регрессии и ее параметров.
- 5) Оценка значимости уравнения регрессии. Коэффициент детерминации.
- 6) Геометрическая интерпретация регрессии и коэффициента детерминации.
- 7) Коэффициент ранговой корреляции Спирмена.