

# Метод наименьших квадратов.

В математической статистике методы получения наилучшего приближения к исходным данным в виде аппроксимирующей функции получили название регрессионного анализа.

Основными задачами регрессионного анализа являются установление зависимости между переменными и оценка (прогноз) значений зависимой переменной.

В экономических исследованиях часто заданному значению одной переменной может соответствовать множество значений другой переменной.

Другими словами, каждому значению одной переменной соответствует условное распределение другой переменной.

Зависимость, при которой каждому значению одной переменной соответствует условное математическое ожидание другой называется **регрессионной**:

$$M(Y | X) = f(X)$$

# Метод наименьших квадратов.

## ■ Уравнение парной регрессии.

$$y_t = a_0 + a_1 x_t + u_t \quad (7.1)$$

### Постановка задачи.

**Дано:** выборка наблюдений за поведением переменных  $y_t$  и  $x_t$ .

Найти: 1. Оценки значений параметров  $a_0$  и  $a_1$ .

2. Оценки точности  $\sigma(a_0)$  и  $\sigma(a_1)$ .

3. Оценка рассеяния случайного возмущения  $\sigma_u$ .

4. Оценку точности прогнозирования  $\sigma(y(x_0))$ .

### Принятые обозначения:

| Выборка:        | Система уравнений наблюдений. |
|-----------------|-------------------------------|
| $y_1 \quad x_1$ | $y_1 = a_0 + a_1 x_1 + u_1$   |
| $y_2 \quad x_2$ | $y_2 = a_0 + a_1 x_2 + u_2$   |
| .....           | .....                         |
| $y_n \quad x_n$ | $y_n = a_0 + a_1 x_n + u_n$   |

*Вектор  $Y$*

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

*Матрица  $X$*

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

# Метод наименьших квадратов

Идея метода.

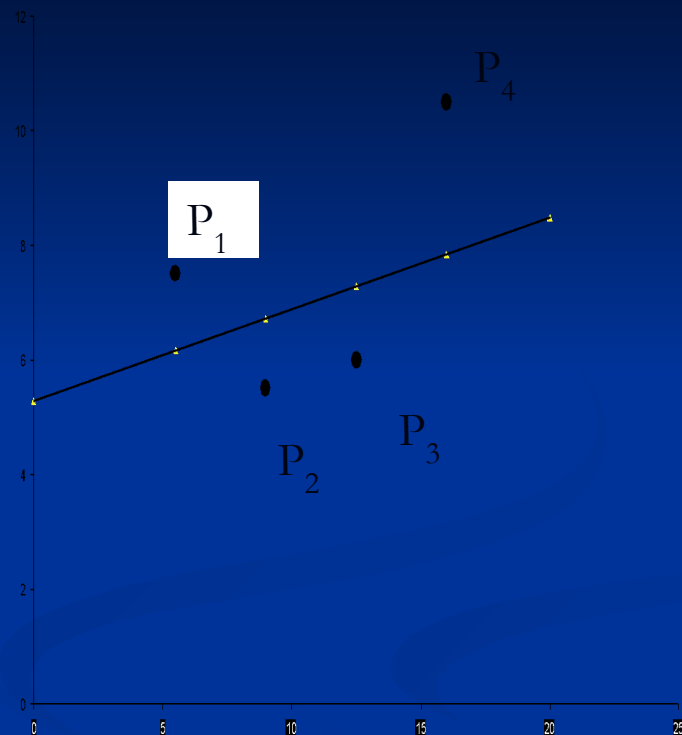
Пусть имеем выборку из 4-х точек ( $n=4$ ):

$$P_1 = (x_1, y_1)$$

$$P_2 = (x_2, y_2)$$

$$P_3 = (x_3, y_3)$$

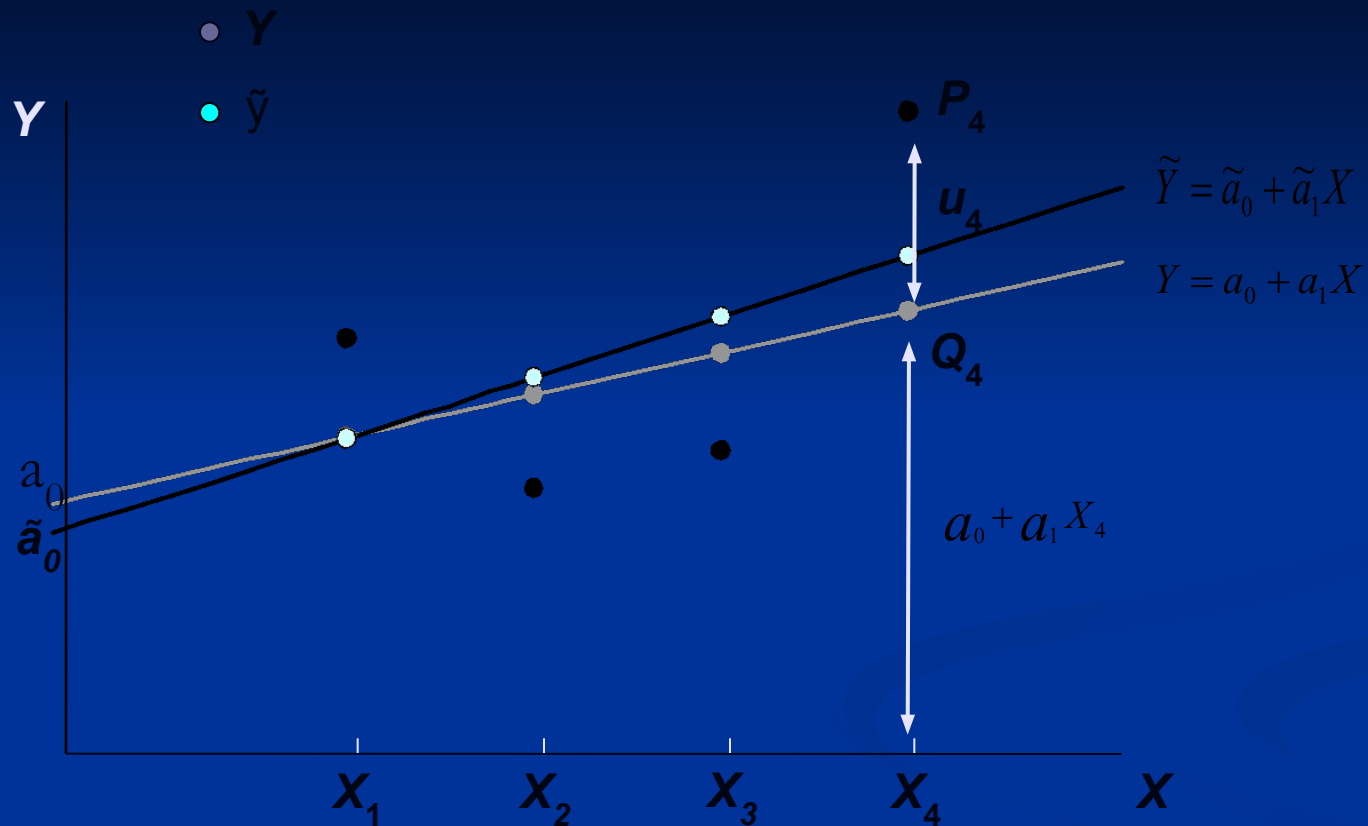
$$P_4 = (x_4, y_4)$$



На практике мы имеем возможность наблюдать только исходные точки. Предполагаем, что существует теоретическая прямая, которая наилучшим образом проходит через них.

Задача: оценить с некоторой точностью, как может проходить эта прямая.

# Метод наименьших квадратов



Любое значение  $Y$  можно представить в виде суммы неслучайной величины  $a_0 + a_1 x$  и случайной величины  $u$ .

Идея метода заключается в том, чтобы найти такие значения параметров, которые обеспечат минимум суммы квадратов случайных отклонений.

# Реализация метода наименьших квадратов

Итак, оценки параметров модели парной регрессии согласно МНК будем искать из условия:

$$S = \sum u_i^2 = \sum (y_i - \tilde{a}_0 + \tilde{a}_1 x_i)^2 = \min$$

Условиями минимума функции являются равенство нулю первых производных и положительность вторых производных по  $\tilde{a}_0$  и  $\tilde{a}_1$ .

$$\frac{\partial S}{\partial a_0} = \sum 2(y_i - \tilde{a}_0 - \tilde{a}_1 x_i)(-1) = 0 \quad (7.2)$$

$$\frac{\partial S}{\partial a_1} = \sum 2(y_i - \tilde{a}_0 - \tilde{a}_1 x_i)(-x_i) = 0$$

при этом:

$$\frac{\partial^2 S}{\partial \tilde{a}_0^2} = \sum 1 > 0$$

$$\frac{\partial^2 S}{\partial \tilde{a}_1^2} = \sum x_i^2 > 0$$

Система уравнений (7.2) называется системой нормальных уравнений для вычисления оценок параметров модели (7.1)

# Реализация метода наименьших квадратов

Упростим систему нормальных уравнений (7.2)

$$\begin{aligned} n\tilde{a}_0 + \tilde{a}_1 \sum x_i &= \sum y_i \\ \sum x_i \tilde{a}_0 + \tilde{a}_1 \sum x_i^2 &= \sum x_i y_i \end{aligned} \quad (7.3)$$

Для решения системы (7.3) выразим из первого уравнения  $\tilde{a}_0$ , подставим его во второе уравнение.

$$\tilde{a}_0 = \frac{1}{n} (\sum y_i - \tilde{a}_1 \sum x_i) \quad (7.4)$$

$$\sum x_i \frac{1}{n} (\sum y_i - \tilde{a}_1 \sum x_i) + \sum x_i^2 \tilde{a}_1 = \sum x_i y_i$$

Или

$$\sum x_i (\sum y_i - \tilde{a}_1 \sum x_i) + n \sum x_i^2 \tilde{a}_1 = n \sum x_i y_i$$

Откуда получаем:

$$\tilde{a}_1 (\sum x_i^2 - \sum x_i \sum x_i) = n \sum x_i y_i - \sum x_i \sum y_i$$

$$\tilde{a}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{(n \sum x_i^2 - \sum x_i \sum x_i)} \quad (7.5)$$

# Реализация метода наименьших квадратов

Вычислив с помощью (7.5) оценку  $\tilde{a}_1$ , с помощью выражения (7.4) получим значение оценки параметра  $\tilde{a}_0$ .

$$\begin{aligned}\text{Cov}(x, y) &= \frac{1}{(n-1)} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)} \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) = \\ &= \frac{1}{(n-1)} \left( \sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + \sum \bar{x} \bar{y} \right) = \frac{1}{(n-1)} \left( \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right) \\ \sigma^2(x) &= \frac{1}{(n-1)} \sum (x_i - \bar{x})^2 = \frac{1}{(n-1)} \sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2) = \frac{1}{(n-1)} \left( \sum x_i^2 - 2\bar{x} \sum x_i + \sum \bar{x}^2 \right) = \\ &= \frac{1}{(n-1)} \left( \sum x_i^2 - n\bar{x}^2 \right) = \frac{1}{(n-1)} \left( \sum x_i^2 - \frac{1}{n} \sum x_i \sum x_i \right)\end{aligned}$$

Тогда выражение (7.5) можно записать в виде:

$$a_1 = \frac{\text{Cov}(x, y)}{\sigma^2(x)} \quad (7.6)$$

# Реализация метода наименьших квадратов

**Вопрос.** Как связано полученное решение со случайными возмущениями?

$$\begin{aligned}\text{Cov}(x, y) &= \text{Cov}(x, a_0 + a_1 x + u) = \\ \text{Cov}(x, a_0) &+ \text{Cov}(x, a_1 x) + \text{Cov}(x, u) = \\ &= a_1 \sigma^2(x) + \text{Cov}(x, u)\end{aligned}\tag{7.7}$$

Подставляя (7.7) в (7.6) получим выражение:

$$\tilde{a}_1 = a_1 + \frac{\text{Cov}(x, u)}{\sigma^2(x)}$$

Условие несмещенности оценки параметра  $\tilde{a}_1$

$$M(\tilde{a}_1) = a_1 + \frac{1}{\sigma^2(x)} \text{Cov}(x, u)$$

# Характеристики точности уравнения парной регрессии

Вычислим дисперсии параметров уравнения регрессии и дисперсию прогнозирования эндогенной переменной.

1. Дисперсия параметра  $\tilde{a}_1$

$$\begin{aligned}\sigma^2(\tilde{a}_1) &= \sigma^2 \left( \frac{\text{Cov}(x, y)}{\sigma^2(x)} \right) = \sigma^2 \left( \frac{\frac{1}{(n-1)} \sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{(n-1)} \sum (x_i - \bar{x})^2} \right) = \\ &= \sigma^2 \left( \frac{\sum (x_i - \bar{x})(y_i - a_0 - a_1 x)}{\sum (x_i - \bar{x})^2} \right) = \sigma^2 \left( \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2} \right) = \\ \sigma^2(\tilde{a}_1) &= \sum \frac{(x_i - \bar{x})^2 \sigma^2(u_i)}{[\sum (x_i - \bar{x})^2]^2} = \sigma_u^2 \frac{1}{n \sigma^2(x)}\end{aligned}\tag{7.8}$$

# Характеристики точности уравнения парной регрессии

Дисперсия параметра  $\tilde{a}_0$

$$\sigma^2(\bar{y}) = \sigma^2\left(\frac{1}{n} \sum (y_i)\right) = \frac{1}{n^2} \sum \sigma^2(y_i) = \frac{\sigma_u^2}{n}$$

Дисперсия  $\sigma^2(\tilde{a}_1)$  известна (7.8), необходимо вычислить дисперсию  $y$ .

$$\begin{aligned} \sigma^2(\tilde{a}_0) &= \sigma^2(\bar{y} - \tilde{a}_1 \bar{x}) = \sigma^2(\bar{y}) + \bar{x}^2 \sigma^2(\tilde{a}_1) - 2\bar{x} \text{Cov}(\bar{y}, \tilde{a}_1) \\ \text{Cov}(\bar{y}, \tilde{a}_1) &= 0 \end{aligned} \quad (7.9)$$

В результате получаем:

$$\sigma^2(\tilde{a}_0)^2 = \sigma_u^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n \sigma^2(x)} \right) \quad (7.10)$$

# Характеристики точности уравнения парной регрессии

Дисперсия прогноза эндогенной переменной.

$$\begin{aligned}\sigma^2(\tilde{y}) &= \sigma^2(\tilde{a}_0 + \tilde{a}_1 x + u) = \\ &= \sigma^2(\tilde{a}_0) + x^2 \sigma^2(\tilde{a}_1) + \sigma_u^2 + \\ &+ 2x \text{Cov}(\tilde{a}_0, \tilde{a}_1) + 2\text{Cov}(\tilde{a}_0, u) + 2x \text{Cov}(\tilde{a}_1, u)\end{aligned}\quad (7.11)$$

Ковариации между случайными возмущениями и оценками параметров равны нулю, т.к. эти переменные независимые.

$$\text{Cov}(\tilde{a}_0, \tilde{a}_1) = \text{Cov}(\bar{y} - \tilde{a}_1 \bar{x}, \tilde{a}_1) = \text{Cov}(\bar{y}, \tilde{a}_1) + \bar{x} \text{Cov}(\tilde{a}_1, \bar{a}_1) = \bar{x} \sigma^2(\tilde{a}_1) \quad (7.12)$$

Подставляя в (7.11) (7.10), (7,8) и (7,12), получаем:

$$\sigma^2(\tilde{y}) = \sigma_u^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{n \sigma^2(x)} \right) \quad (7.13)$$

# Пример применения МНК

| X    | Y     | $\tilde{Y}$ | U     | $U^2$ | $\sigma(Y)$ |
|------|-------|-------------|-------|-------|-------------|
| 1,0  | 1,91  | 2,17        | -0,26 | 0,07  | 1,20        |
| 2,0  | 2,76  | 2,72        | 0,04  | 0,00  | 1,19        |
| 3,0  | 2,67  | 3,26        | -0,59 | 0,35  | 1,17        |
| 4,0  | 4,03  | 3,80        | 0,23  | 0,05  | 1,16        |
| 5,0  | 4,12  | 4,34        | -0,22 | 0,05  | 1,15        |
| 6,0  | 2,81  | 4,88        | -2,07 | 4,30  | 1,15        |
| 7,0  | 6,53  | 5,42        | 1,11  | 1,22  | 1,14        |
| 8,0  | 6,24  | 5,97        | 0,27  | 0,07  | 1,14        |
| 9,0  | 9,03  | 6,51        | 2,52  | 6,36  | 1,13        |
| 10,0 | 6,87  | 7,05        | -0,18 | 0,03  | 1,13        |
| 11,0 | 9,09  | 7,59        | 1,50  | 2,24  | 1,13        |
| 12,0 | 7,08  | 8,13        | -1,05 | 1,11  | 1,13        |
| 13,0 | 7,79  | 8,68        | -0,89 | 0,78  | 1,14        |
| 14,0 | 8,75  | 9,22        | -0,47 | 0,22  | 1,14        |
| 15,0 | 11,19 | 9,76        | 1,43  | 2,05  | 1,15        |
| 16,0 | 10,15 | 10,30       | -0,15 | 0,02  | 1,15        |
| 17,0 | 10,52 | 10,84       | -0,32 | 0,10  | 1,16        |
| 18,0 | 10,89 | 11,38       | -0,49 | 0,24  | 1,17        |
| 19,0 | 10,59 | 11,93       | -1,34 | 1,78  | 1,19        |
| 20,0 | 13,40 | 12,47       | 0,93  | 0,87  | 1,20        |

X-стаж работы сотрудника;

Y- часовая оплата труда.

**Модель:**  $Y_t = a_0 + a_1 X_t + U_t$

$$\begin{aligned} \sum x_i &= 210; \quad \sum y_i = 146.42; \\ \sum x_i^2 &= 2870; \quad \sum x_i y_i = 1897.66 \end{aligned}$$

$$\tilde{a}_1 = \frac{20 \cdot 1897.66 - 210 \cdot 146.42}{20 \cdot 2870 - 44100} = 1.63$$

$$\tilde{a}_0 = \frac{1}{20} (155.42 - 1.63 \cdot 210) = 0.54$$

$$\tilde{\sigma}_u^2 = \frac{21.93}{20 - 2} = 1.10$$

# Пример применения МНК

Графическое отображение результатов

