

### Линейная регрессия

1. Вывести непосредственно формулу для МНК-оценки параметра  $\theta$  в регрессионной модели вида  $Y_i = \theta X_i + \varepsilon_i, i = 1, \dots, n$ .
2. Имеются данные о себестоимости  $Y$  (в у.е.) экземпляра книги в зависимости от тиража  $X$  (в тыс. экземпляров). Данные представлены в таблице:

тираж	1	2	3	4	5
себестоимость	6	5	4	4	3

Предполагается, что справедлива модель вида  $Y_i = a + bX_i + \varepsilon_i, i = 1, \dots, n$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ . Постройте:

- 1) МНК-оценки параметров  $a$  и  $b$  (с использованием и без использования матричной техники вычислений);
- 2) оценку дисперсии  $\sigma_\varepsilon^2$ ;
- 3) проверьте гипотезу  $H_0: b = 0$ ;

Постройте точечную и интервальную (уровня надёжности 0.95) оценку для себестоимости, если тираж  $X=6$ .

3. Имеются данные (из «Основы химии» Д.И. Менделеева) о количестве ( $Y$ ) азотнатриевой соли  $\text{NaNO}_3$ , которое можно растворить в 100 граммах воды в зависимости от температуры  $t$

$t$	0	4	10	15	21	29	36	51	68
$Y$	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Оцените по МНК параметры регрессионной модели  $Y_i = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 + \varepsilon_i$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ . Проверьте гипотезу о том, что параметр  $\theta_2 = 0$ . Следует ли для описания данного явления перейти к более простой регрессионной модели? Оцените коэффициент детерминации для «короткой» и «длинной» модели.

### Домашнее задание

1. В таблице указана динамика веса поросят

Возраст $X$ (недели)	0	1	2	3	4	5	6
Вес $Y$ (кг)	1,2	2,5	3,9	5,2	6,4	7,7	9,2

Предполагается, что справедлива модель вида  $Y_i = a + bX_i + \varepsilon_i$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ . Постройте:

- 1) МНК-оценки параметров  $a$  и  $b$ ;
- 2) оценку дисперсии  $\sigma_\varepsilon^2$ ;
- 3) проверьте гипотезу  $H_0: b = 0$ ;
- 4) постройте точечную и интервальную (уровня надёжности 0.95) веса в точках  $X_0=3$  и  $X_0=6$ ;
- 5) сделайте прогноз для значения  $Y$  в точке  $X=8$ .

2. Бюджетное обследование пяти случайным образом выбранных семей дало следующие результаты (в тыс. у.е.):

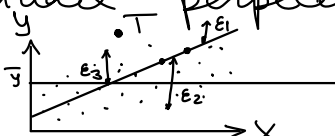
Семья	Накопления (Y)	Доход (X)	Имущество (Z)
1	3,0	40	60
2	6,0	55	36
3	5,0	45	36
4	3,5	30	15
5	1,5	30	90

Предполагается, что справедлива модель вида  $Y_i = a + b_1 X_i + b_2 Z + \varepsilon_i$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ .

- 1) Оцените параметры модели;
- 2) Спрогнозируйте накопления семьи, имеющей доход 40 тыс у.е. и имущество 25 тыс у.е.;
- 3) Предположим, что доход семьи вырос на 10 тыс у.е. в то время, как стоимость имущества не изменилась. Оцените, как вырастут накопления семьи;
- 4) Оцените, как вырастут накопления семьи, если её доход увеличился на 5 тыс, а имущество на 15 тысяч;
- 5) Проверьте гипотезы (на уровне значимости 0.05) о том, что а)  $b_1$  и  $b_2$  равны нулю (т.е. модель является тривиальной), б)  $b_1 = 0$  (т.е. величина дохода не существенна), в)  $b_2 = 0$  (т.е. стоимость имущества не существенна), г)  $b_1 = 0,8$  (такое значение было вычислено согласно данным по другой стране)

парная линейная регрессионная модель:  

$$y = \beta_0 + \beta_1 X + \varepsilon$$



Метод наименьших квадратов

$$Q = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min$$

многомерная линейная регрессия модель:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nn} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Свойства МНК оценок:

1) несмещенная  $E[\hat{\beta}] = \beta$

2) коб. м-чс оценки имеют вид  $K_{\hat{\beta}} = \sigma_{\varepsilon}^2 (X^T X)^{-1}$

3)  $n \rightarrow \infty: \hat{\beta} \sim N(\beta; \sigma_{\varepsilon}^2 (X^T X)^{-1})$

4)  $E \varepsilon = 0; K_{\varepsilon} = \sigma^2 I \Rightarrow BLUE$

лучшая  
линейная  
несмещенная  
оценка

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

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = ESS$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = TSS$$

residual sqs  
 $TSS = RSS + ESS$  — estimate sqs  
total sum of squares

коэф. детерминации

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2_{adj} = (1 - R^2) \frac{n-1}{n-p-1} \in [0, 1] \quad n\text{-наблюд} \quad p\text{-кон-во параметров}$$

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-p-1} \frac{RSS}{TSS}$$

$$\hat{\sigma}_{\varepsilon} = \sqrt{\frac{RSS}{n-p-1}}$$

1. Вывести непосредственно формулу для МНК-оценки параметра  $\theta$  в регрессионной модели вида  $Y_i = \theta X_i + \varepsilon_i, i = 1, \dots, n$ .

$$y_i = \theta x_i + \varepsilon_i$$

$$\varepsilon_i = (y_i - \theta x_i)^2$$

$$Q(x, y) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \theta x_i)^2 \rightarrow \min_{\theta \in \mathbb{R}}$$

$$\frac{\partial Q}{\partial \theta} = \sum_{i=1}^n -2x_i(y_i - \theta x_i) = -2 \sum_{i=1}^n x_i(y_i - \theta x_i) = 0$$

$$\sum_{i=1}^n (x_i y_i) = \theta \sum_{i=1}^n x_i^2 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^n (x_i y_i)}{\sum_{i=1}^n x_i^2}$$

2. Имеются данные о себестоимости  $Y$  (в у.е.) экземпляра книги в зависимости от тиража  $X$  (в тыс. экземпляров). Данные представлены в таблице:

тираж	1	2	3	4	5
себестоимость	6	5	4	4	3

Предполагается, что справедлива модель вида  $Y_i = a + bX_i + \varepsilon_i, i = 1, \dots, n$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ . Постройте:

- МНК-оценки параметров  $a$  и  $b$  (с использованием и без использования матричной техники вычислений);
- оценку дисперсии  $\sigma_\varepsilon^2$ ;
- проверьте гипотезу  $H_0: b = 0$ ;

Постройте точечную и интервальную (уровня надёжности 0.95) оценку для себестоимости, если тираж  $X=6$ .

Формулы МНК (без матричной техники)

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, \hat{a} = \bar{Y} - b\bar{X}$$

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{Y} = \frac{6+5+4+4+3}{5} = 4,4$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (1-3)(6-4,4) + (2-3)(5-4,4) + (3-3)(4-4,4) + (4-3)(4-4,4) + (5-3)(3-4,4) =$$

$$= (-2) \cdot 1,6 + (-1) \cdot 0,6 + 0 \cdot (-0,4) + 1 \cdot (-0,4) + 2 \cdot (-0,4) = -3,2 - 0,6 + 0 - 0,4 - 0,8 = -5$$

$$\sum (X_i - \bar{X})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$b = \frac{-5}{10} = -0,5 \quad \hat{a} = 4,4 - (-0,5) \cdot 3 = 4,4 + 1,5 = 5,9$$

$$\hat{a} = 5,9 \quad \hat{b} = -0,5$$

большая формула

$$y_i = a + bx_i + \varepsilon_i, i = 1, \dots, n, \varepsilon_i \sim N(0, \sigma^2 I)$$

$$\varepsilon_i = y_i - a - bx_i$$

$$МНК: \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min$$

$$Q = \sum_{i=1}^n (y_i - a - bx_i)^2 \rightarrow \min_{a, b}$$

$$\begin{cases} \frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \rightarrow \sum_{i=1}^n y_i - na - b \sum_{i=1}^n x_i = 0 \\ \frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0 \rightarrow \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

$$na = \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \quad \hat{a} = \frac{\sum_{i=1}^n y_i}{n} - b \frac{\sum_{i=1}^n x_i}{n} \quad \hat{a} = \bar{y} - b\bar{x}$$

Формулы МНК (с матричной техникой)

$$Y = X\beta + \varepsilon, где$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \quad Y = \begin{bmatrix} 6 \\ 5 \\ 4 \\ 4 \\ 3 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 1+1+1+1+1 & 1+2+3+4+5 \\ 1+2+3+4+5 & 1+4+9+16+25 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 6 \\ 5 \\ 4 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 6+5+4+4+3 \\ 6+10+12+16+15 \end{bmatrix} = \begin{bmatrix} 22 \\ 59 \end{bmatrix}$$

$$(X^T X)^{-1}$$

$$X^T X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$\det = 5 \cdot 55 - 15^2 = 275 - 225 = 50$$

$$2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i) = 0$$

$$2 \sum_{i=1}^n (y_i \bar{x} - \bar{y} \bar{x} - bx_i)(-x_i) = 0$$

$$\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0$$

$$\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - b(\sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n x_i^2) = 0$$

$$b = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{-\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (\bar{x} - x_i)} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\downarrow \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b}) = \sum_{i=1}^n (a_i - \bar{a})b_i - \sum_{i=1}^n (a_i - \bar{a})\bar{b} =$$

$$= \sum_{i=1}^n (a_i - \bar{a})b_i - \bar{b} \sum_{i=1}^n (a_i - \bar{a})$$

$$\sum_{i=1}^n (a_i - \bar{a}) = \sum_{i=1}^n a_i - \sum_{i=1}^n \bar{a} = \sum_{i=1}^n a_i - n\bar{a} =$$

$$= \sum_{i=1}^n a_i - \frac{n \sum_{i=1}^n a_i}{n} = 0$$

$$(X^T X)^{-1} = \frac{1}{\det} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \frac{1}{50} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \frac{1}{50} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} \begin{bmatrix} 22 \\ 59 \end{bmatrix} =$$

$$55 \cdot 22 + (-15) \cdot 59 = 1210 - 885 = 325$$

$$(-15) \cdot 22 + 5 \cdot 59 = -330 + 295 = -35$$

$$\hat{\beta} = \frac{1}{50} \begin{bmatrix} 325 \\ -35 \end{bmatrix} = \begin{bmatrix} 6,5 \\ -0,7 \end{bmatrix} \Rightarrow \hat{a} = 6,5 \quad \hat{b} = -0,7$$

2) Оценка дисперсии

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}$$

$$\hat{a} = 6,5 \quad \hat{b} = -0,7$$

$$\hat{y}_i = 6,5 - 0,7x_i \quad y_i = 6,5 - 0,7x_i + \varepsilon$$

$x_i$	$y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	квадрат
1	6	5,8	0,2	0,04
2	5	5,1	-0,1	0,01
3	4	4,4	-0,4	0,16
4	4	3,7	0,3	0,09
5	3	3	0	0

Сумма остатков:  $0,04 + 0,01 + 0,16 + 0,09 + 0 = 0,3$

$$\hat{\sigma}^2 = \frac{0,3}{5-1-1} = 0,1$$

3) Проверка гипотезы  $H_0: b = 0$

$H_0: b = 0$

$H_1: b \neq 0$

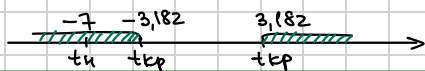
$$T = \frac{\hat{b} - b}{\hat{\sigma}} \sim t(n-p-1)$$

$$t_n = \frac{-0,7 - 0}{0,1} = -7$$

$$(X^T X)^{-1} = \frac{1}{50} \begin{pmatrix} 55 & -15 \\ -15 & 5 \end{pmatrix} = \begin{pmatrix} \frac{55}{50} & -\frac{15}{50} \\ -\frac{15}{50} & \frac{5}{50} \end{pmatrix} = \begin{pmatrix} 1,1 & -0,3 \\ -0,3 & 0,1 \end{pmatrix}$$

$$t_{kp} = t_{0,975,3} = 3,182 \Rightarrow \text{отклонение} \Rightarrow b \neq 0$$

$$\hat{\sigma}_{\hat{\beta}_j}^2 = \hat{\sigma}^2 \cdot c_{jj} \rightarrow \text{элемент } jj \text{ матрицы } (X^T X)^{-1}$$



4) Точечная и интервальная оценки

Точечная оценка

$$X = 6$$

$$\hat{y}(6) = \hat{a} + \hat{b} \cdot 6 = 6,5 - 0,7 \cdot 6 = 2,3$$

Интервальный прогноз

$$\hat{y}(x_0) \pm t_{\alpha/2, n-2} \cdot SE_{pred}$$

$$SE_{pred} = \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

$$n_{пр} \hat{\sigma}^2 = 0,1, x_0 = 6, \bar{x} = 3, \sum (x_i - \bar{x})^2 = 10, n = 5$$

$$SE_{pred} = \sqrt{0,1 \left( 1 + \frac{1}{5} + \frac{(6-3)^2}{10} \right)} = \sqrt{0,1(1+0,2+0,9)} = \sqrt{0,1 \cdot 2,1} \approx 0,458$$

$$2,3 \pm 3,182 \cdot 0,458 \approx 2,3 \pm 1,46 \quad [0,84; 3,76]$$

3. Имеются данные (из «Основы химии» Д.И. Менделеева) о количестве (Y) азотнатриевой соли  $\text{NaNO}_3$ , которое можно растворить в 100 граммах воды в зависимости от температуры t

t	0	4	10	15	21	29	36	51	68
Y	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Оцените по МНК параметры регрессионной модели  $Y_i = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 + \varepsilon_i$ , где вектор  $\varepsilon \sim N(0, \sigma^2 I)$ . Проверьте гипотезу о том, что параметр  $\theta_2 = 0$ . Следует ли для описания данного явления перейти к более простой регрессионной модели? Оцените коэффициент детерминации для «короткой» и «длинной» модели.

$$\sigma_{\varepsilon}^2 = \frac{RSS}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$

$H_0: \theta_2 = 0$  (не знаем)

$H_a: \theta_2 \neq 0$  (знаем)

$$T = \frac{\hat{\theta}_2 - \theta_2}{\sigma_{\hat{\theta}_2}} \sim t(6)$$

$$\sigma_{\hat{\theta}_2}$$

$$t_k = \frac{-0,085 - 0}{0,015} = -\frac{85}{15} = -5,6$$

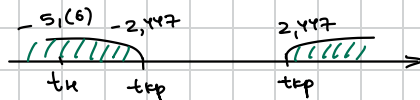
$$t_k = t_{0,975,6} = 2,447$$

$$(X^T X)^{-1} = \begin{pmatrix} 0,4879 & -0,0299 & 0,00036 \\ -0,0299 & 0,00288 & -0,00004 \\ 0,00036 & -0,00004 & 0,0000006 \end{pmatrix} = C \theta_{22}$$

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} 66,7062 \\ 0,87934 \\ -0,0085 \end{pmatrix} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}$$

$$\hat{\sigma}_{\varepsilon}^2 = \frac{2253,7}{9-2-1} = 19,3808^2$$

$$\sigma_{\hat{\theta}_2}^2 = 19,3808^2 \cdot 6 \cdot 10^{-7} \approx 0,015$$



Ответ: знаем