# Линейна регресия

Нека имаме наблюдения $\vec{x}, \vec{y}$.
Предполагаме, че $y_k = b_0 + b_1 x_k + \epsilon_k$
Нека $f(b_0, b_1) = \sum\limits_{k=1}^{n} \epsilon_k^2 = \sum\limits_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2$
По метода на най-малките квадрати (МНМК) намираме коефициентите, за да минимизираме грешката:

$$\begin{cases} \dfrac{\partial}{\partial b_0} f(b_0, b_1) = 0 \\ \dfrac{\partial}{\partial b_1} f(b_0, b_1) = 0 \end{cases} \iff \begin{cases} b_1 = \sum\limits_{k=1}^{n} \dfrac{y_k (x_k - \overline{x})}{\sum\limits_{j=1}^{n} (x_j - \overline{x})^2} \\ b_0 = \sum\limits_{k=1}^{n} (\frac{1}{n} - \frac{(x_k - \overline{x})\overline{x}}{\sum\limits_{j=1}^{n} (x_j - \overline{x})^2}) y_k \end{cases}$$

Това са и точковите оценки на $b_0$ и $b_1$ - $\hat{b_0}$ и $\hat{b_1}$.

**Формално:**
$Y_k = \beta_0 + \beta_1 X_k + \epsilon_k$ за $1 \leq k \leq n$
$Y_k$ - отклик и $X_k$ - предикатор

- $(\epsilon_0, \ldots, \epsilon_n)$ са независими в съвкупност
- $\mathbb{E}[\epsilon_k] = 0$ за $1 \leq k \leq n$
- $\mathbb{D}[\epsilon_k] = \sigma^2$ за $1 \leq k \leq n$
- $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ за $1 \leq k \leq n$

Нека
$\overline{X} = \frac{1}{n} \sum\limits_{k} X_k$
$A = \sum\limits_{k} (X_k - \overline{X})^2$
$\hat{\beta}_1 = \sum\limits_{k} \frac{(X_k - \overline{X})}{A} Y_k$
$\hat{\beta}_0 = \sum\limits_{k} (\frac{1}{n} - \frac{(X_k - \overline{X})\overline{X}}{A}) Y_k$
$Y_k \sim \mathcal{N}(\beta_0 + \beta_1 X_k, \sigma^2)$

Откъдето
$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{A})$
$\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma^2(\frac{1}{n} + A))$

Понеже:

1)

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= \sum_k \frac{(X_k - \overline{X})}{A}\mathbb{E}[Y_k] \\
&= \sum_k \frac{(X_k - \overline{X})}{A}(\beta_0 + \beta_1 X_k) \\
&= \frac{1}{A}\beta_0 \sum_k (X_k - \overline{X}) + \frac{1}{A}\beta_1 \sum_k (X_k - \overline{X})X_k \\
&= \frac{\beta_1}{A}(X_k - \overline{X})^2 \\
&= \beta_1 \rightarrow \text{неизместена оценка}
\end{aligned}
$$

2)

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\overline{Y}] - \mathbb{E}[\hat{\beta}_1 \overline{X}] \\
&= \frac{1}{n}\sum_k (\beta_0 + \beta_1 X_k) - \beta_1 \overline{X} \\
&= \beta_0 + \beta_1 \overline{X} - \beta_1 \overline{X} \\
&= \beta_0 \rightarrow \text{неизместена оценка}
\end{aligned}
$$

3)

$$
\begin{aligned}
\mathbb{D}[\hat{\beta}_1] &= \mathbb{D}\left[\sum_k \frac{(X_k - \overline{X})}{A} Y_k\right] \\
&= \sum_k \frac{(X_k - \overline{X})}{A}\mathbb{D}[Y_k] \\
&= \sum_k \frac{(X_k - \overline{X})}{A^2}\sigma^2 \\
&= \frac{A}{A^2}\sigma^2 \\
&= \frac{\sigma^2}{A}
\end{aligned}
$$

4)

$$\begin{aligned}
\mathbb{D}[\hat{\beta_0}] &= \sum_k \left(\frac{1}{n} - \frac{(X_k - \overline{X})\overline{X}}{A}\right)\sigma^2 \\
&= \sigma^2 \sum_k \left(\frac{1}{n^2} - \frac{2\overline{X}}{An}(X_k - \overline{X}) + \frac{(X_k - \overline{X})^2\overline{X}^2}{A^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \overline{X}^2 \frac{\sum\limits_k (X_k - \overline{X})^2}{A^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} + A\right)
\end{aligned}$$

Също така имаме и следните зависимости:

- Ако е известно $\sigma^2$ (малко вероятно)

$$Y_k \sim \mathcal{N}(\beta_0 + \beta_1 X_k, \sigma^2)$$

$$\frac{Y_k - \beta_0 - \beta_1 X_k}{\sigma} \sim Z = \mathcal{N}(0,1)$$

$$\sum_{k=0}^{n} \frac{(Y_k - \beta_0 - \beta_1 X_k)^2}{\sigma^2} \sim \chi^2(n)$$

$$\sigma^2 \sim \frac{1}{n}\sum_{k=0}^{n}(Y_k - \beta_0 - \beta_1 X_k)^2$$

- Ако $\sigma^2$ не е известно

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=0}^{n}(Y_k - \beta_0 - \beta_1 X_k)^2 \sim \sigma^2$$

$$\sum_{k=0}^{n} \frac{(Y_k - \hat{b_0} - \hat{b_1} X_k)^2}{\sigma^2} \sim \chi^2(n-2)$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{k=0}^{n}(Y_k - b_0 - b_1 X_k)^2 \approx \sigma^2$$

$$\lim_{n \to \infty} \frac{\hat{\sigma}^2}{\sigma^2} \overset{\text{п.с.}}{=} 1$$

***Проверка на хипотези:***

$$
\begin{aligned}
H_0: &\quad \beta_1 = \beta = \theta \\
H_1: &\quad \beta_1 = \beta^* \neq \beta
\end{aligned}
$$

$T$ е централна статистика

- при известно $\sigma^2$

$$T = \frac{\hat{b_1} - \beta}{\frac{\sigma}{\sqrt{A}}} \sim \mathcal{N}(0, 1)$$

- при неизвестно $\sigma^2$

$$T = \frac{\hat{b_1} - \beta}{\frac{\hat{\sigma}}{\sqrt{A}}} \sim t(n-2)$$

Понеже $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$
И тогава $T = \dfrac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2(n-2)}{n-2}}}$