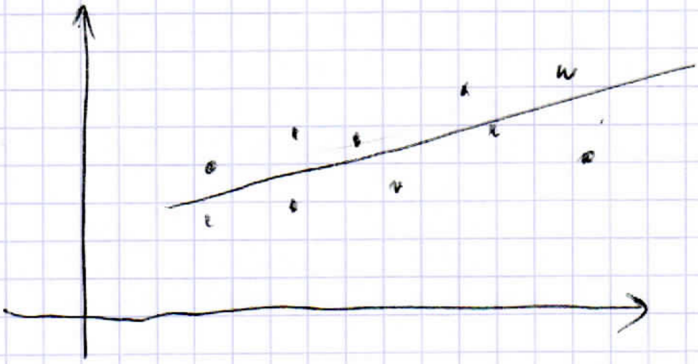
# Regresja liniowa prosta



(2)



$$y_i = \beta_0 + \beta_1 X_i + \xi_i \qquad , \quad i = \overline{1,n}$$

wyraz wolny — intersept

współczynnik kierunkowy — slope

błąd losowy

$E\xi_i = 0$

$Var\,\xi_i = \sigma^2$

niezależne

$(\xi_i \sim N(0, \sigma^2))$

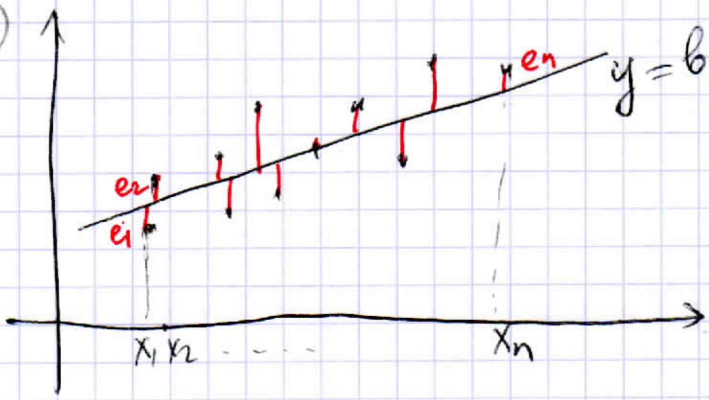
$cov(\xi_i, \xi_j) = 0$

$(X_i, Y_i)$ ← wiadomo

zmienna objaśniająca — explanatory; zmienna objaśniana — response

$\boxed{\beta_0, \beta_1, \sigma^2 ???}$ ← nie wiadomo

(2a)




$$\sum e_i \to \min.$$

$$\Rightarrow \quad \hat\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^{n} (x_i - \bar x)^2} = b_1 \qquad \bar x = \frac{1}{n}\sum x_i$$

$$\bar y = \frac{1}{n}\sum y_i$$

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x = b_0$$

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 = s^2$$

(2b) $T = \frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t(n-2)$

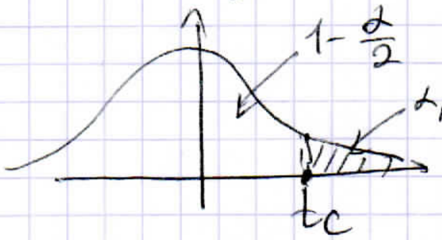
rozkład Studenta z $(n-2)$ stopniami swobody

$S^2(\hat{\beta}_1) = \frac{S^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$ ; $/ S^2 = \hat{\sigma}^2$

$\Rightarrow$ Przedział ufności dla $\beta_1$:

$\hat{\beta}_1 \pm t_c \, S(\hat{\beta}_1)$

kwantyl rzędu $1-\frac{\alpha}{2}$ z $t(n-2)$ $\qquad t_c = t^*(1-\frac{\alpha}{2}, n-2)$




qt( $\frac{\alpha}{2}$, df = n-2, lower.tail=T)

(2c) Test istotności dla $\beta_1$:

$H_0: \beta_1 = 0$; $\qquad H_1: \beta_1 \neq 0$

$T = \frac{\hat{\beta}_1 - 0}{S(\hat{\beta}_1)}$

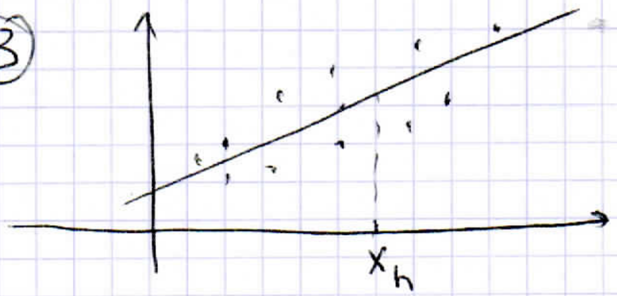
$|T| > t_c \Rightarrow$ odrzucamy $H_0$

<u>P-wartość</u>: $\quad p = P(|z| > |T|)$, gdzie

$z \sim t(n-2)$

$T = \frac{\hat{\beta}_1 - 0}{S(\hat{\beta}_1)}$ - liczymy na podstawie danych

3




$$y_h = \beta_0 + \beta_1 X_h + \xi_h$$

$$EY_h = \mu_h = \beta_0 + \beta_1 X_h$$

$\Rightarrow \quad \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ – estymator $EY_h = \mu_h$

<u>Przedział ufności dla $EY_h = \mu_h$</u>

$$T = \frac{\hat{\mu}_h - \mu_h}{s(\hat{\mu}_h)} \sim t(n-2)$$

$$s^2(\hat{\mu}_h) = s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

<u>Przedział ufn.</u> $\Rightarrow$ dla $\mu_h$: $\quad \hat{\mu}_h \pm t_c\, s(\hat{\mu}_h)$

4 Predykcja punktowa dla <u>nowej</u> wartości $y_h$.

$y_h = \beta_0 + \beta_1 X_h + \xi_h$ – wartość nowej zmiennej $y_h$

$\hat{y}_h = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ – estymator punktowy dla $y_h$

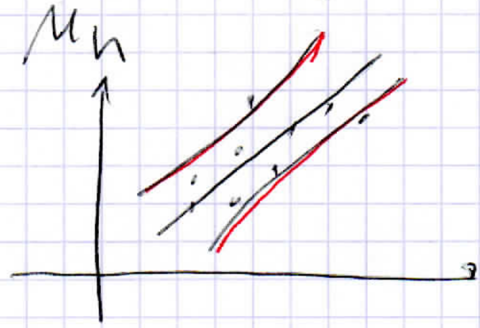
$$T = \frac{y_h - \hat{\mu}_h}{s(pred)} \sim t(n-2),$$

$$s^2(pred) = s^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

Przedział predykcyjny dla $y_h$:

$$\hat{\mu}_h \pm t_c \, s(pred)$$

⑤ Przedziały ufności dla wszystkich możliwych wartości $\mu_h$ (pasmo ufności)

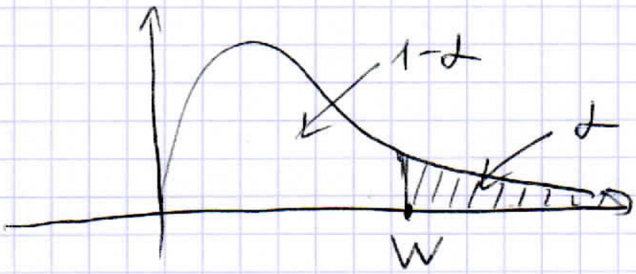


$$\hat{\mu}_h \pm W \cdot s(\hat{\mu}_h)$$

$$W^2 = 2\underbrace{F(1-\alpha, 2, n-2)}$$

kwantyl rzędu $1-\alpha$ z rozkładu Fishera-Snedekora z $2, n-2$ stopniami swobody

$qf(1-\alpha, df1=2, df2=n-2)$



## R komendy

`DataF <- data.frame(... )`

DataF:

| y | X |
|---|---|
| | |

`LinM <- lm(y ~ X, DataF)`

`LinM` / $\hat{\beta}_0, \hat{\beta}_1$

`summary(LinM)` / wszyskie rodzaje cech modeli liniowej

`LinM$coefficients`
`LinM$residuals` / wartości $\hat{\beta}_0, \hat{\beta}_1, \{e_i\}$

`confint(LinM)` — 95% przedziały ufności dla $\beta_0, \beta_1$.

(6) $H_0: \beta_1 = 0$ ; $H_1: \beta_1 \neq 0$

$$T = \frac{\hat{\beta_1}}{S(\hat{\beta_1})}$$

$$s^2(\hat{\beta_1}) = \frac{s^2}{\sum_i (x_i - \bar{x})^2} \quad ; \quad s^2 = \frac{1}{n-2} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Funkcja mocy testu:

$$\pi(a) = P_{\beta_1 = a}(|T| > t_c), \quad \forall a \neq 0$$

$$a = 0 \Rightarrow \quad T \sim t(n-2, \delta)$$

niecentralny rozkład Studenta z $(n-2)$ st. sw. i parametrem niecentralności

$$\delta = \frac{\beta_1}{\sigma(\hat{\beta_1})}$$

$$\begin{matrix} p \\ q \\ r \\ d \end{matrix} t(\ldots, df = , \quad ncp = \underset{\delta}{\shortparallel} )$$

---

$n$, $SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $\sigma^2$ – są wiadome

$$\Rightarrow \sigma^2(\hat{\beta_1}) = \frac{\sigma^2}{SSX} \quad ;$$

$$\delta = \frac{\beta_1}{\sigma(\hat{\beta_1})}$$

$$\Rightarrow T \sim t(n-2, \delta)$$

$$\pi(\beta_1 = a) = P_{\beta_1 = a}(|T| > t_c) =$$

$$= P_{\beta_1 = a}(T < -t_c) + P_{\beta_1 = a}(T > t_c)$$

7 $X = (X_1 \ldots X_{200})^T \sim N(0, \frac{1}{200} I)$

$Y = 5 + \beta_1 X + \xi$

---

$$\begin{pmatrix} X_1 \\ \vdots \\ X_{200} \end{pmatrix} \longmapsto \begin{pmatrix} Y_1 \\ \vdots \\ Y_{200} \end{pmatrix} \longmapsto T = \frac{\hat{\beta}_1 - 0}{S(\hat{\beta}_1)} \longmapsto$$

$$\longmapsto |T| > t_c ? \longmapsto \begin{cases} 1, & |T| > t_c \\ 0, & |T| \leq t_c \end{cases}$$

1000 razy

$\Rightarrow$ $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \leftarrow 1000 \times 1 \longmapsto \hat{p} = \frac{\#\{1\}}{1000}$

częstość odrzuceń

---

$Y = 5 + \beta_1 X + \xi$

$H_0: \beta_1 = 0$ $\quad$ $H_1: \beta_1 \neq 0$

(a) $-(\xi \sim N(0, I))-$ (c)

(b) $-(\xi \sim \cancel{N(0,I)})-$ (d)

$\downarrow$ $\quad$ $\downarrow$

$\hat{p}$ – estymator $P$(błąd I r.)

$\hat{p}$ – estymator mocy testu ($1 - P$(błąd II r.))

$\Rightarrow$ porównać wyniki z wartościami teoretycznymi dla $P$(błąd I r.) oraz mocy test (liczonymi przy założeniu $\xi \sim N(0,I)$)