## Regression:

**Modell:** Vi antar ett linjärt samband $y = \beta_0 + \beta_1 x$ som modelleras av $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ där

i) $x_1, x_2, \ldots, x_n$ är givna kända tal. De kallas för förklarande variabler,

ii) $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ är ober s.v. (störningar) s.a. $\mathbb{E}[\varepsilon_i] = 0$ och $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Vi antar att $\varepsilon_i \sim N(0, \sigma^2)$ iid

iii) Vi vill skatta $\beta_0$ = intercept och $\beta_1$ = lutningskoefft./slope.

iv) $Y_1, \ldots, Y_n$ kallas för <u>beroende variabler</u>

v) Linjen $y = \beta_0 + \beta_1 x$ kallas för den <u>teoretiska regressionslinjen</u>.

---

Vi skattar $\beta_0, \beta_1$ genom att välja $\hat{\beta}_0$ och $\hat{\beta}_1$ s.a.

# TMA 321

$$Q(\beta_0, \beta_1) = \sum_{k=1}^{n} (y_k - (\beta_0 + \beta_1 x_k))^2 \quad \text{minimeras.}$$

F21 ②

Här är $(x_k, y_k)$ våra datapunkter.

Ex:

| F-halt | 0.0 | 0.3 |
|---|---|---|
| Elast | 12.86 | 13.10 |

"bild 1"

Vi minimerar $Q(\beta_0, \beta_1)$ genom att lösa

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = 0.$$

Om vi låter $S_{xy} := \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$,

$S_{xx} := \sum_{k=1}^{n} (x_k - \bar{x})^2$ och $S_{yy} = \sum_{k=1}^{n} (y_k - \bar{y})^2$ får vi

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{och} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Vår anpassade regressionslinje blir

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

"bild 2"

---

Som ett mått på anpassning beräknas

förklaringsgraden $R^2 = 1 - \dfrac{\sum_{k=1}^{n} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2} = \dfrac{S_{xy}^2}{S_{xx} S_{yy}}$

Vi har att $R^2 \in (0,1)$ och att $R^2 \approx 1$ bra medans $R^2 \approx 0$ dåligt.

---

Under antagandet att $\varepsilon_i \sim N(0, \sigma^2)$ är iid. kan man visa att

$$\hat{\beta}_1 = \hat{\beta}_1(\varepsilon_1, \dots, \varepsilon_n) \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{och}$$

$$\hat{\beta}_0 = \hat{\beta}_0(\varepsilon_1, \dots, \varepsilon_n) \sim N\left(\beta_0, \sigma^2 \frac{\sum_{k=1}^{n} x_k^2}{n\, S_{xx}}\right).$$

OBS! $\mathbb{E}[\hat{\beta}_i] = \beta_i$ så båda är VVR.

Med kännedom om fördelningarna kan vi skapa $\alpha$.I.

$$1-2\alpha = \mathbb{P}(-z_\alpha < Z < z_\alpha) = \mathbb{P}\left(-z_\alpha < \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} < z_\alpha\right)$$

$$\Rightarrow \quad I_{\beta_1} = \hat{\beta}_1 \pm z_\alpha \frac{\sigma}{\sqrt{S_{xx}}}$$

---

Ex (forts): Med insättning av data och om $\sigma = 1$ får vi $R^2 = \dfrac{S_{xy}^2}{S_{xx} S_{yy}} \approx 0.9436$ och

$$\hat{\beta}_1 \approx N(\beta_1, 0.1257) \quad \text{medans}$$

$$\hat{\beta}_0 \approx N(\beta_0, 0.3627),$$

Oftast är $\sigma^2$ okänd. I sådfall skattar

Vi $\sigma^2$ med $\quad s_r^2 = \frac{1}{n-2} \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$

$$= \frac{1}{n-2} \sum_{k=1}^{n} (y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2$$

---

Anm: 1) $s_r^2$ är VVR

2) $s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2$ är ej bra här

"tar ej hänsyn till lutningen"

---

Som innan när $\sigma$ skattas får vi

$$\frac{\hat{\beta}_1 - \beta_1}{s_r / \sqrt{S_{xx}}} \sim t_{n-2} \qquad \text{och}$$

$$\frac{\hat{\beta}_0 - \beta_0}{s_r \sqrt{\frac{\sum_{k=1}^{n} x_k^2}{n S_{xx}}}} \sim t_{n-2}$$

dvs t-fördelning.

---

Ex forts: Om $\sigma$ är okänd skattar vi

och får $\quad s_r^2 \approx 0.239$

Ett 95 % K.I. för $\beta_1$ blir då

med $\quad t_6(0.025) \approx 2.45$

$$I_{\beta_1} = \hat{\beta}_1 \pm 2.45 \frac{\sqrt{0.239}}{\sqrt{S_{xx}}} = [1.31, 2.16]$$

---

Vi kan använda K.I. för att göra hypotestest:

---

Ex forts: Vi vill testa

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

på signifikansnivån 0.05,

Enligt förra gången kan vi välja RR s.a. $\hat{\beta}_1 \in RR \Leftrightarrow 0 \notin I_{\beta_1}$ där $I_{\beta_1}$ är ett K.I. för $\beta_1$ med samma $\kappa$,

Vi har att $0 \notin I_{\beta_1} = [1.31, 2.16]$

så vi förkastar $H_0$ på sign. 0.05

---

Rimlighetsanalys av modellen

Betrakta residualerna

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

Om vi ser mönster i residualerna bör vi vara skeptiska till modell-antagandet.

"Datamängd 1 med plotter"

Obs: $R^2 \approx 0.439$ också avslöjande.

Kanske är $Y_i = \beta_0 + \beta_1 (x_i)^2 + \varepsilon_i$ en bättre modell? Jämför med

"Datamängd 2 med plotter"

Här blir $R^2 \approx 0.835$.