Regression

Modell: Vi antar ett linjärt samband $y = \beta_0 + \beta_1 x$ som modelleras av

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{där}$$

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

ii) $E_1, E_2, \ldots, E_n$ är ober. s.v. (störningar) sådana att $\mathbb{E}[E_i] = 0$ och $\mathrm{Var}(E_i) = \sigma^2$. Vi antar här att $E_i \sim N(0, \sigma^2)$.

iii) Vi vill skatta β_0 = intercept och β_1 = lutningskoeff. / 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) Mätningar kommer resultera i data $(x_1, y_1), \ldots, (x_n, y_n)$ där

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

och ε_i är en realisation av E_i.

Vi skattar β_0, β_1 genom att välja $\hat{\beta}_0$ och $\hat{\beta}_1$ så att

$$Q(b_0, b_1) = \sum_{k=1}^{n} (y_k - (b_0 + b_1 x_k))^2 = \sum_{k=1}^{n} e_k^2 \text{ minimeras.}$$

Ex:

F-halt	0.0	0.3
Elasticitet	12.86	13.10

"visa bild 1"

Om $b_0 = 14$ och $b_1 = 0.5$ så blir

$$Q(14, 0.5) \approx 68.1$$

Vi minimerar $Q(b_0, b_1)$ genom att lösa

$$\frac{\partial Q}{\partial b_0} = \frac{\partial Q}{\partial b_1} = 0.$$

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

att minimum uppnås av

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

Vår anpassade regressionslinje blir

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

där $\hat{\beta}_1 \approx 1.74$ och $\hat{\beta}_0 \approx 12.76$

Som ett mått på anpassning beräknas

förklaringsgraden (coefficient of determination)

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

där $y_k - \hat{y}_k = y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k)$

Vi har att $R^2 \in [0,1]$ och att $R^2 \approx 1$ är bra, men $R^2 \approx 0$ är dåligt.

Under antagandet att $E_i \sim N(0, \sigma^2)$ är i.i.d. kan man visa att (

$$\hat{\beta}_1 = \hat{\beta}_1(E_1, \dots, E_n) \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \text{ och}$$

$$\hat{\beta}_0 = \hat{\beta}_0(E_1, \dots, E_n) \sim N\left(\beta_0, \sigma^2 \frac{\sum_{k=1}^{n} x_k^2}{n S_{xx}}\right)$$

OBS! $E[\hat{\beta}_0] = \beta_0$ och $E[\hat{\beta}_1] = \beta_1$ så båda är VVR.

K.I. för β_1

Oftast är σ^2 okänd. I så fall skattar vi σ^2 med $S_r^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2$

$$= \frac{1}{n-2} SSE$$

Man kan visa att $\mathbb{E}[S_r^2] = \sigma^2$ så att S_r^2 är en VVR skattare av σ^2

OBS: $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \bar{Y})^2$ är ej bra här

"tar ej hänsyn till lutningen"

Liknande tidigare i kursen får vi att

$$\frac{\hat{\beta}_1 - \beta_1}{S_r/\sqrt{S_{xx}}} \sim t(n-2)$$

Vi kan då hitta ett K.I. för β_1.

$$1-\alpha = \mathbb{P}\left(-t_{\alpha/2}(n-2) \le \frac{\hat{\beta}_1 - \beta_1}{S_r/\sqrt{S_{xx}}} \le t_{\alpha/2}(n-2)\right)$$

$$= \cdots = \mathbb{P}\left(\hat{\beta}_1 - t_{\alpha/2}(n-2)\frac{S_r}{\sqrt{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2}(n-2)\frac{S_r}{\sqrt{S_{xx}}}\right)$$

så att

$$I_{\beta_1} = \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{S_r}{\sqrt{S_{xx}}}$$

är ett K.I. med konf.grad $(1-\alpha)100\,\%$ för β_1.

Tal: För exemplet ovan, bestäm ett 95 % K.I. för β_1.

L: Data ger oss att

$$S_r^2(x) = \frac{1}{n-2}\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2 \approx 0.239$$

Tabell ger oss att $t_{0.025}(6) \approx 2.45$ så ett 95 % numerisk k.i. för β_1 blir

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

Utifrån vårt k.i. ovan verkar det troligt att $\beta_1 \neq 0$. Tolkningen av detta är att halten F påverkar elasticiteten!

Vi vill göra hypotestest på detta.

Vi betraktar följande situationer:

I $H_0: \beta_1 = 0$, $H_1: \beta_1 > 0$

II $H_0: \beta_1 = 0$, $H_1: \beta_1 < 0$

III $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

Tal: Testa om halten F påverkar elasticiteten på signifikansnivån $\alpha = 0.05$.

Lösning: Vi skall alltså testa

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

Vi gör detta genom att beräkna p-värdet. Testet är tvåsidigt så att

$$\text{p-värde} = 2\,P(T(E) \geq |T(\varepsilon)|)$$

Här väljer vi test-statistikan

$$T(E_1,\dots,E_n) = \frac{\hat{\beta}_1(E_1,\dots,E_n)}{S_r(E_1,\dots,E_n)/\sqrt{S_{xx}}}$$

så att

$$T(\varepsilon_1,\dots,\varepsilon_n) = \frac{\hat{\beta}_1(\varepsilon_1,\dots,\varepsilon_n)}{S_r(\varepsilon_1,\dots,\varepsilon_n)/\sqrt{S_{xx}}} \approx \frac{1.74}{\sqrt{0.239}/\sqrt{S_{xx}}} \approx 10.$$

Vi ser från tabell (sid 699) att

$$2\,\mathbb{P}(T(E)\geq|T(\varepsilon)|) \approx 2\,\mathbb{P}(T(E)\geq 10) \leq 2\,\mathbb{P}(T(E)\geq 5.96)$$

$$\leq 2\cdot 0.0005 = 0.001 \quad \text{där } T(E)\sim t(6)$$

Då p-värdet $< \alpha$ så förkastar vi H_0 på signifikansnivån $\alpha = 0.05$. //

Kort rimlighetsanalys

Om förklaringsgraden är låg ($R^2 \lesssim 0.8$) så bör vi vara skeptiska till vårt modellantagande. Vi kan då betrakta residualerna

$$e_k = Y_k - \hat{Y}_k$$

och leta efter mönster.

"Visa figur"

I detta exempel blev $R^2 \approx 0.62$.

Kanske är

$$Y_i = \beta_0 + \beta_1 (x_i)^2 + E_i$$

en bättre modell?