

# Sovellettu todennäköisyytlaskenta B

Antti Rasila

22. marraskuuta 2007

Antti Rasila ()

TodB

22. marraskuuta 2007

1 / 15

## 1 Yhden selittäjän lineaarinen regressiomalli

- Virhetermin standardioletukset
- Pienimmän neliösumman menetelmä
- Pienimmän neliösumman menetelmä
- Sovitteet ja residuaalit

## 2 Sovituksen tunnuslukuja

- Neliösummat SST, SSE ja SSM
- Selitysaste
- Jäännösvarianssi

## 3 Testejä

- Testi regressiosuoran kulmakertoimelle
- Testi regressiosuoran vakiokertoimelle
- Testi korrelaatiokertoimelle
- Testi korreloimattomuudelle

Antti Rasila ()

TodB

22. marraskuuta 2007

2 / 15

## Yhden selittäjän lineaarinen regressiomalli

- Yhden selittäjän lineaarinen regressiomalli on muotoa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n,$$

jossa

- $y_j$  = Selitettävän muuttujan havaittu arvo havaintoyksikössä  $j$ .
- $x_j$  = Selittäjän muuttujan havaittu arvo havaintoyksikössä  $j$ .
- $\beta_0$  = Vakioselittäjän regressiokerroin, joka on tuntematon vakio.
- $\beta_1$  = Selittäjän  $x$  regressiokerroin, joka on tuntematon vakio.
- $\varepsilon_j$  = Satunnainen virhetermi havaintoyksikössä  $j$ .

Antti Rasila ()

TodB

22. marraskuuta 2007

3 / 15

## Virhetermin standardioletukset 1/2

- Regressiomallin virhetermit  $\varepsilon_j$  ovat satunnaismuuttujia, joiden ns. *standardioletukset* ovat:

- $E(\varepsilon_j) = 0, j = 1, 2, \dots, n.$
- $\text{Var}(\varepsilon_j) = \sigma^2, j = 1, 2, \dots, n.$
- $\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, j \neq l.$

- Tavallisesti tehdään myös normaalisuusoletus

- $\varepsilon_j \sim N(0, \sigma^2), j = 1, 2, \dots, n.$

Antti Rasila ()

TodB

22. marraskuuta 2007

4 / 15

## Virhetermin standardioletukset 2/2

- Jos regressiomallin virhetermejä  $\varepsilon_j$  koskevat standardioletukset (i)-(iii) pätevät, on selitettävän muuttujan havaituilla arvoilla seuraavat stokastiset ominaisuudet:

- $E(y_j) = \beta_0 + \beta_1 x_j, j = 1, 2, \dots, n.$
- $\text{Var}(y_j) = \sigma^2, j = 1, 2, \dots, n.$
- $\text{Cor}(y_j, y_l) = 0, j \neq l.$

- Jos myös normaalisuusoletus (iv) pätee, niin

- $y_j \sim N(\beta_0 + \beta_1 x_j, \sigma^2), j = 1, 2, \dots, n.$

Antti Rasila ()

TodB

22. marraskuuta 2007

5 / 15

- Kuten aikaisemmin, pienimmän neliösumman menetelmässä yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n,$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  estimaattorit määrätään minimoimalla virhetermien  $\varepsilon_j$  neliösumma  $F(\beta)$

$$F(\beta) = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  suhteen.

Antti Rasila ()

TodB

22. marraskuuta 2007

6 / 15

## Regressiokertoimien PNS-estimaattorit

- Määritellään havaintojen  $x_j$  ja  $y_j, j = 1, 2, \dots, n$  aritmeettiset keskiarvot ( $\bar{x}$  ja  $\bar{y}$ ), otosvarianssit ( $s_x^2$  ja  $s_y^2$ ), otoskovarianssi ( $s_{xy}$ ) ja otoskorrelaatiokerroin ( $r_{xy}$ ) tavanomaisilla kaavoilla.

- Yhden selittäjän lineaarisen regressiomallin regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit ovat

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$$

Antti Rasila ()

TodB

22. marraskuuta 2007

7 / 15

## Sovitteet ja residuaalit

- Olkoot  $b_0$  ja  $b_1$  yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Estimoidun mallin *sovitte*

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

on estimoidun regressiosuoran arvo havaintopisteessä  $x_j$ .

- Estimoidun mallin *residuaali*

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

on selitettävän muuttujan  $y$  havaitun arvon  $y_j$  ja sovitteen  $\hat{y}_j$  arvon erotus.

Antti Rasila ()

TodB

22. marraskuuta 2007

8 / 15

## Neliösummat SST, SSE ja SSM

- kokonaisneliösumma:

$$SST = \sum_{j=1}^n (y_j - \bar{y})^2$$

- jäännöseliösumma:

$$SSE = \sum_{j=1}^n e_j^2$$

- mallineliosumma:

$$SSM = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2$$

- Näille neliösummille pätee

$$SST = SSM + SSE$$

## Selitysaste

- Tunnuslukua

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

käytetään regressiomallin hyvyyden mittarina.

- Tunnuslukua  $R^2$  kutsutaan *selityasteeksi* ja se mittaa regressiomallin selittämää osuutta selitettävän muuttujan  $y$  havaittujen arvojen kokonaisvaihtelusta.

- Yhden selittäjän lineaarisessa regressiomallissa pätee:

$$R^2 = r_{xy}^2$$

- Selitysasteelle pätee aina

$$0 \leq R^2 \leq 1$$

## Jäännösvarianssi

- Jos yhden selittäjän lineaarisen regressiomallin virhetermejä  $\varepsilon_j$  koskevat standardioletukset (i)-(iii) pätevät, *jäännösvarianssin*  $Var(\varepsilon_j) = \sigma^2$  harhaton estimaattori on

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2,$$

missä

- $e_j$  on estimoidun mallin residuaali ja
- $n$  on havaintojen lukumäärä

## Testi regressiosuoran kulmakertoimelle

- Nollahypoteesi  $H_0 : \beta_1 = 0$

- Vaihtoehtoiset hypoteesit:  $H_1 : \beta_1 < 0$  tai  $H_1 : \beta_1 > 0$  tai  $H_1 : \beta_1 \neq 0$

- Testisuure:

$$T_1 = \frac{b_1}{s/(\sqrt{n}\hat{\sigma}_x)}$$

- Nollahypoteesin pätiessä testisuure noudattaa t-jakaumaa vapausasteilla  $n-2$ :

$$T_1 \sim t(n-2)$$

## Testi regressiosuoran vakiokertoimelle

- Nollahypoteesi  $H_0 : \beta_0 = 0$

- Vaihtoehtoiset hypoteesit:  $H_1 : \beta_0 < 0$  tai  $H_1 : \beta_0 > 0$  tai  $H_1 : \beta_0 \neq 0$

- Testisuure:

$$T_0 = \frac{b_0}{s\sqrt{\sum x_j^2/(n\hat{\sigma}_x)}}$$

- Nollahypoteesin pätiessä testisuure noudattaa t-jakaumaa vapausasteilla  $n-2$ :

$$T_0 \sim t(n-2)$$

## Testi korrelaatiokertoimelle

- Nollahypoteesi  $H_0 : \rho_{xy} = \rho_0$

- Vaihtoehtoiset hypoteesit:  $H_1 : \rho_{xy} < \rho_0$  tai  $H_1 : \rho_{xy} > \rho_0$  tai  $H_1 : \rho_{xy} \neq \rho_0$

- Testisuure:

$$Z = \frac{\frac{1}{2} \log\left(\frac{1+r_{xy}}{1-r_{xy}}\right) - \frac{1}{2} \log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}$$

- Nollahypoteesin pätiessä testisuure noudattaa approksimatiivisesti standardinormaalijakaumaa

$$Z \sim_a N(0, 1)$$

## Testi korreloimattomuudelle

- Nollahypoteesi  $H_0 : \rho_{xy} = 0$

- Vaihtoehtoiset hypoteesit:  $H_1 : \rho_{xy} < 0$  tai  $H_1 : \rho_{xy} > 0$  tai  $H_1 : \rho_{xy} \neq 0$

- Testisuure:

$$T = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}$$

- Nollahypoteesin pätiessä testisuure noudattaa t-jakaumaa vapausasteilla  $n-2$ :

$$T \sim_a t(n-2)$$