

## Yleistetyt lineaariset mallit

7. harjoitukset, 10. vko 2002

- 7.1. Yleistetyt lineaarisen mallin parametrivektorin  $\beta$  suurimman uskottavuuden estimaattorin  $\hat{\beta}$  asymptoottinen kovarianssimatriisi on  $E(\mathbf{U}\mathbf{U}^T)$ , missä

$$\mathbf{U} = \sum_{i=1}^n \frac{Y_i - \mu_i}{g'(\mu_i)V_i} \mathbf{x}_i,$$

$g$  on linkkifunktio ja  $V_i = \text{var}(Y_i)$ . Olkoon  $g(\mu_i) = \mathbf{x}_i^T \beta$ ,  $i = 1, 2, 3$   $\beta^T = (\beta_1, \beta_2)$  ja selittäjien arvovektorit ovat  $\mathbf{x}_1^T = (1, 1)$ ,  $\mathbf{x}_2^T = (1, 2)$  ja  $\mathbf{x}_3^T = (1, 3)$ . Osoita, että

$$E(\mathbf{U}\mathbf{U}^T) = \sum_{i=1}^3 w_i \mathbf{x}_i \mathbf{x}_i^T,$$

missä  $w_i = \frac{1}{V_i [g'(\mu_i)]^2}$ . Kirjoita tulos myös matriisimuodossa  $\mathbf{X}^T \mathbf{W} \mathbf{X}$ , missä  $\mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$  ja  $\mathbf{W}$  on diagonaalimatriisi, jonka diagonaalialkiot ovat  $w_1, w_2$  ja  $w_3$ .

- 7.2. Oletetaan edellisessä tehtävässä Poissonin jakauma ja logaritmilinkki. Kirjoita tässä tapauksessa "auki" estimaattorin  $\hat{\beta}$  asymptoottisen kovarianssimatriisin lauseke. Totea laskemalla, että kovarianssimatriisi on  $-E\left(\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right)$ .

- 7.3. Laske trinomijakauman

$$f(n_1, n_2, n_3) = \frac{n!}{n_1! n_2! n_3!} \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3}$$

parametrien  $\pi_1$ ,  $\pi_2$  ja  $\pi_3$  suurimman uskottavuuden estimaatit derivointikeinolla ( $n_1 + n_2 + n_3 = n$ ,  $\pi_1 + \pi_2 + \pi_3 = 1$ ). Mitä ovat luonnollisten parametrien  $\theta_1 = \log\left(\frac{\pi_1}{\pi_3}\right)$  ja  $\theta_2 = \log\left(\frac{\pi_2}{\pi_3}\right)$  suurimman uskottavuuden estimaatit?

- 7.4. Tarkastellaan automieliämyksiä koskevaa aineistoa <http://mt1.uta.fi/tilasto/glim/datat/car.preferences.txt>, josta on luennolla jaettu taustaselvitys (kaksi sivua, Dobson 2001). Koehenkilöiltä (300 kpl) kysyttiin sitä, pitävätkö he ilmastointia ja ohjaustehostinta tärkeänä (Vastausvaihtoehdot: ei tärkeä/tärkeä/hyvin tärkeä). Vastauskäyttäytymistä selitettiin sukupuolella ja vastaajan iällä (ks. aineisto). Tarkastellaan nominaalista logistista regressiomallia

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \quad j = 1, 2,$$

missä  $x_1 = 1$ , kun mies, muutoin  $x_1 = 0$ . Muuttujilla  $x_2$  ja  $x_3$  on koodattu ikä:  $x_2 = 1$ , kun ikä on 24 – 40, muutoin  $x_2 = 0$  ja  $x_3 = 1$ , kun ikä on yli 40, muutoin  $x_3 = 0$  (Muuttujien koodaus aineistossa `... /glim/datat/car.pref.d.txt`. Estimoi malli tästä aineistosta). Laske vastaustodennäköisyyksien  $\pi_1, \pi_2, \pi_3$  estimaatit selittäjien eri arvokombinaatioilla (`multinom, library(nnet)`).

- 7.5. Aineistossa `car.preferences.txt` selittävät muuttujat `sex` ja `age` on koodattu nominaaliasteikollisina. Estimoi nyt tästä aineistosta funktiolla `multinom` malli `response ~ sex + age`. Selvitä estimoidun mallin ja edellisen tehtävän mallin välinen yhteys. Molemmista malleistaan saadaan samat vastaustodennäköisyyksien estimaatit.
- 7.6. Testaa, onko edellä tarkastellussa mallissa
- (a) sukupuoli
  - (b) ikä merkitsevä selittäjä?
- 7.7. Tarkastellaan edelleen aineistoa `... /glim/datat/car.pref.d.txt` ja selitetään vastauksia sukupuolella ja iällä. Olkoon ikä nyt kvantitatiivinen selittäjä, joka saa arvot 0, 1, 2 (Aineistossa valmiina). Estimoi malli. tulkitse tulos. Vertaa mallin antamaa devianssia tehtävän 4 mallin antamaan devianssiin. Mistä ero johtuu? Mieti, mitä ovat näihin kahteen malliin liittyvät täydet mallit.