Lineaariset mallit, kevät 2014

Harjoitus 3, viikko 14

1. Monisteen sivun 7 alareunassa esitetään yhtäsuuruus

$$\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

   Perustele huolellisesti suureiden dimensiot ja yhtälön paikkansapitävyys komponenteittain.

2. Tutkitaan yksinkertaista yhden selittäjän lineaarista regressiomallia $Y_i \sim N(\beta x_i, \sigma^2)$, $i = 1, \ldots, n$. (Ainoa selittävä muuttuja ei ole tässä välttämättä vakio.) Osoita, että PNS-estimaattori parametrille β on

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

3. Tarkastellaan yhden selittävän muuttujan lineaarista regressiomallia, jolloin malliyhtälö on havaintoyksiköittäin ilmaistuna $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, \ldots, n$. Merkitään $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ \hat{\beta}_2]'$ ja $\bar{x} = (x_1 + \cdots + x_n)/n$.

   (i) Muodosta Lauseen 2.1(i) tulosta käyttäen $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ ja edelleen $\mathrm{Var}(\hat{\beta}_2)$ ja $\mathrm{Var}(\hat{\beta}_1 + \hat{\beta}_2\bar{x})$.

   (ii) Oletetaan, että selittävien muuttujien arvot $x_1, \ldots, x_n$ voidaan valita vapaasti väliltä $[c, d]$. Miten ne on valittava, jos $\mathrm{Var}(\hat{\beta}_2)$ halutaan minimoida? Onko tämä valinta muuten järkevä?

   (Huom.: Kohdassa (ii) ei vaadita yksityiskohtaisia matemaattisia todistuksia. Niissä voit myös olettaa, että n on parillinen.)

4. Tarkastellaan yhden selittäjän lineaarista regressiomallia kahden ryhmän tapauksessa:

$$Y_1, \cdots, Y_n \perp\!\!\!\perp, \ Y_i \sim \begin{cases} N(\beta_1 + \beta_2 x_i, \sigma^2), & \text{kun } i = 1, \ldots, n_1 \\ N(\beta_3 + \beta_4 x_i, \sigma^2), & \text{kun } i = n_1 + 1, \ldots, n_1 + n_2 = n, \end{cases}$$

   jossa $n_1, n_2 > 2$.

   a) Osoita, että kysymyksessä on lineaarinen malli käyttäen lineaarisen mallin matriisiesitystä (eli $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$).

b) Osoita, että parametrien β_1 ja β_2 (vastaavasti β_3 ja β_4) PNS-estimaatit saadaan kuten harjoituksen 2 tehtävässä 4 käyttäen vain havaintoyksiköiden $i = 1, \cdots, n_1$ (vastaavasti $i = n_1 + 1, \cdots, n$) havaintoja.

c) Miten tulkitset ehtoa $\beta_2 = \beta_4$, kun y on lääkärien kuukausipalkka, x on työvuosien määrä ja sukupuoli määrää ryhmäjaon?

5. Halutaan tutkia koehenkilöiden sukupuolen ja iän yhteyttä hapenottokykyyn. On saatu seuraavanlainen aineisto:

Ikä (x_2)	Sukupuoli (x_3)	Hapenkulutus (y)
0	0	44
1	0	45
0	0	54
1	1	59
0	1	50
1	1	45

jossa

$$x_2 = \begin{cases} 0 & \text{kun ikä} \geq 45, \\ 1 & \text{kun ikä} < 45, \end{cases}$$

ja

$$x_3 = \begin{cases} 0 & \text{kun sukupuoli=nainen}, \\ 1 & \text{kun sukupuoli=mies}. \end{cases}$$

Sovitetaan aineistoon lineaarinen regressiomalli, jossa hapenkulutus on selitettävä muuttuja (y) ja selittävinä muuttujia ovat ikä (x_2) ja paino (x_3).

a) Muodosta mallimatriisi $\boldsymbol{X}$ ja laske $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Huom! mallissa on nyt regressiovakio.

b) Estimoi mallin parametrit β_1, β_2, β_3 ja σ^2 suurimman uskottavuuden menetelmän avulla.

c) Miten tulkitaan regressiokertoimet β_2 ja β_3?