Lineaariset mallit, kevät 2014

Harjoitus 2, viikko 13

1. Tarkastellaan lineaarista mallia
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0,\ \mathrm{r}(\boldsymbol{X}) = p.$$
Osoita, että $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} = \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ ja että $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_p\bar{x}_p$, kun mallissa on vakio eli $x_{i1} = 1$ kaikilla $i = 1, \ldots, n$. Tässä $\bar{x}_j = (x_{1j} + \cdots + x_{nj})/n$, jossa x_{ij} on matriisin $\boldsymbol{X}$ yleinen alkio ($i = 1, \ldots, n, j = 1, \ldots, p$), $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \cdots \hat{\beta}_p]'$.

2. Olkoon $Y_{11}, \ldots, Y_{1n_1}$, $Y_{21}, \ldots, Y_{2n_2}$, …, $Y_{p1}, \ldots, Y_{pn_p}$ riippumattomia ja $Y_{ji} \sim N(\mu_j, \sigma^2)$ ($\mu_j \in \mathbb{R}$). Esitä tilanne lineaarisen mallin erikoistapauksena käyttäen lineaarisen mallin matriisiesitystä. Mikä on matriisin $\boldsymbol{X}$ aste eli mikä on $\boldsymbol{X}$:n lineaarisesti riippumattomien sarakkeiden lukumäärä?

3. Tarkastellaan aineistosta $y_1, \ldots, y_n$ laskettua otoskeskiarvoa $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ ja otosvarianssia $s^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$. Osoita, että
$$(n-1)s^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J})\boldsymbol{y},$$
jossa $\boldsymbol{y} = [y_1 \cdots y_n]'$ ja $\boldsymbol{J} = \boldsymbol{1}_n(\boldsymbol{1}_n'\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n'$ ($\boldsymbol{1}_n = [1 \cdots 1]'$, $n \times 1$). Osoita lisäksi, että $\boldsymbol{J}$ (ja siten $\boldsymbol{I}_n - \boldsymbol{J}$) on symmetrinen ja idempotentti.

4. Esitä yhden selittäjän lineaarisen regressiomallin $Y_1, \ldots, Y_n$ $\perp\!\!\!\perp$, $Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$ normaaliyhtälöt komponenttimuodossa (ilman matriiseja) ja osoita, että niiden ratkaisuna saatavat PNS-estimaatit voidaan lausua muodossa
$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{ja} \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x},$$
jossa esimerkiksi $\bar{y} = (y_1 + \cdots + y_n)/n$. Esitä $\hat{\beta}_2$ käyttäen havainnoista (y_i, x_i), $i = 1, \ldots, n$, laskettuja keskihajontoja ja korrelaatiokerrointa. (Huom.: Havaintojen korrelaatiokeroimen määritelmä löytyy monisteen s. 11 alaviitteestä ja esim. y-havaintojen keskihajonta on $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}$)

5. Tehtävässä 3 on todettu, että $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J})\boldsymbol{y}$, jossa $\boldsymbol{y} = [y_1 \cdots y_n]'$ ja $\boldsymbol{J} = \boldsymbol{1}_n(\boldsymbol{1}_n'\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n'$ eli monisteen s. 10 merkinnöin $SST = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J})\boldsymbol{y}$. Osoita, että $SSR = \boldsymbol{y}'(\boldsymbol{P} - \boldsymbol{J})\boldsymbol{y}$ ja $SSE = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}$, jossa $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ ja kuten monisteen vastaavassa kohdassa on myös tässä matriisin $\boldsymbol{X}$ ensimmäinen sarake ykkösvektori $\boldsymbol{1}_n$. Tästä saat vaihtoehtoisen perustelun monisteessa s. 10 esitetylle tulokselle $SST = SSR + SSE$.