Ekstraoppgave 4

Vi ser på den multiple linære regresjonsmodellen, der de stokastiske variablene $Y_1, \ldots, Y_n$ er gitt ved

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \tag{1}$$

og ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte. I denne oppgaven vil vi anta at forklaringsvariablene er sentrerte, dvs.

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{for} \quad j = 1, \ldots, k \tag{2}$$

og ortogonale (ukorrelerte), dvs.

$$\sum_{i=1}^{n} x_{ij} x_{i\ell} = 0 \quad \text{for alle} \quad j \neq \ell. \tag{3}$$

a) Vis at minste kvadraters estimatorer er gitt ved

$$\hat{\beta}_0 = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4}$$

og

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij} Y_i}{\sum_{i=1}^{n} x_{ij}^2} \quad \text{for} \quad j = 1, \ldots, k. \tag{5}$$

Sammenlign (5) med den estimatoren du ville fått for den enkle linære regresjonsmodellen med x_{ij} som eneste forklaringsvariabel.

b) Forklar at $\hat{\beta}_j$-ene er normalfordelte, og vis at

$$E(\hat{\beta}_j) = \beta_j \quad \text{og} \quad V(\hat{\beta}_j) = \frac{\sigma^2}{\left(\sum_{i=1}^{n} x_{ij}^2\right)^2} \quad \text{for} \quad j = 1, \ldots, k$$

og at

$$\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_\ell) = 0 \quad \text{for alle} \quad j \neq \ell$$

De tilpassede verdiene er gitt ved

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}.$$

Fra den generelle teorien for multippel lineær regresjon har vi at

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

der

$$\mathrm{SST} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2, \quad \mathrm{SSR} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \quad \text{og} \quad \mathrm{SSE} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

c) Vis at vi kan skrive SSR på formen:

$$\text{SSR} = \hat{\beta}_1^2 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2^2 \sum_{i=1}^{n} x_{i2}^2 + \cdots + \hat{\beta}_k^2 \sum_{i=1}^{n} x_{ik}^2.$$

d) Vis at vi kan skrive R^2 som

$$R^2 = r_1^2 + r_2^2 + \cdots + r_k^2,$$

der

$$r_j = \frac{\sum_{i=1}^{n} x_{ij} Y_i}{\sqrt{\left(\sum_{i=1}^{n} x_{ij}^2\right)\left(\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right)}}$$

er den empiriske korrelasjonen mellom responsen og den j-te forklaringsvariabelen.