Applied Data Analysis

Exercise Sheet 4

Exercise 14

Consider a normal linear model

$$\boldsymbol{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

as given in Definition I.4.3 with $n \geq \operatorname{rank}(B) = p \geq 2$, (unknown) parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ and error term $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, where $\sigma > 0$ is also unknown.

(a) Explain, how to test the following hypotheses

(i) $H_0\colon \beta_1 = \beta_2 \quad \longleftrightarrow \quad H_1\colon \beta_1 \neq \beta_2$,

(ii) $H_0\colon \beta_1 = \dots = \beta_p \quad \longleftrightarrow \quad H_1\colon \beta_j \neq \beta_k$ for at least one pair (j,k) with $j \neq k$,

by using general linear hypothesis testing and nested model comparison.

(b) For $\alpha \in (0,1)$, construct two confidence regions for $\beta_1 - \beta_2$ of confidence level $1-\alpha$, one based on the F-distibution and one based on the t-distribution.

Hint: Confidence regions can be constructed by inverting a statistical hypothesis test procedure in the sense of constructing the non-rejecting region corresponding to that statistical test.

(c) If for the testing problem (a),(ii) H_0 is rejected, one might ask, which pairs in

$$M := \big\{(j,k) \in \{1,\dots,p\} \,\big|\, j \neq k\big\}$$

have significantly different effects.

To answer this question, the multiple pairwise comparison problem

$$H_0^{(j,k)} : \beta_j = \beta_k \quad \longleftrightarrow \quad H_1^{(j,k)} : \beta_j \neq \beta_k$$

for all $m := \binom{p}{2}$ pairs $(j,k) \in M$ could be considered.

Show for $\alpha \in (0,1)$ that if each single hypothesis $H_0^{(j,k)}$, $(j,k) \in M$, is tested on significance level $\frac{\alpha}{m}$, the multiple pairwise comparison tests all together hold a family-wise error rate of α (so-called *Bonferroni correction*).

To formalize this task, for $\boldsymbol{\beta} \in \mathbb{R}^p$, let $I_0(\boldsymbol{\beta}) \subseteq M$ denote the subset of all pairs (j,k) for which $H_0^{(j,k)}$ is true (i.e. $\beta_j = \beta_k$) and for $(j,k) \in M$, let $A^{(j,k)}$ denote the event that the corresponding testing procedure decides in favour for $H_1^{(j,k)}$.

Then, you have to show for for $\boldsymbol{\beta} \in \mathbb{R}^p$:

$$P_{\boldsymbol{\beta}}\Bigg(\bigcup_{(j,k)\,\in\, I_0(\boldsymbol{\beta})} A^{(j,k)} \Bigg) \leq \alpha$$

(with $P_{\boldsymbol{\beta}}$ denoting the underlying probability distribution corresponding to $\boldsymbol{\beta}$).

Exercise 15

Consider the model of simple linear regression as given in Example I.4.6 with $n \geq 2 = \operatorname{rank}(B)$ and corresponding sums of squares as given in Theorem I.5.11:

$$\text{SST} = \sum_{i=1}^{n} \big(Y_i - \overline{Y}\big)^2 \ , \ \text{SSR} = \sum_{i=1}^{n} \big(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{Y}\big)^2 \ , \ \text{SSE} = \sum_{i=1}^{n} \big(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\big)^2 \ .$$

According to Definition I.5.43, the corresponding coefficient of determination is defined by

$$R^2 := \frac{\text{SSR}}{\text{SST}} \overset{\text{Th. I.5.11,(1)}}{=} \frac{\text{SST} - \text{SSE}}{\text{SST}} \ .$$

Show:

(a) $\big(\widehat{\beta}_0 \,, \widehat{\beta}_1\big)' = \left(\overline{Y} - \frac{s_{\boldsymbol{x}\boldsymbol{Y}}}{s_{\boldsymbol{x}\boldsymbol{x}}}\, \overline{x} \,, \frac{s_{\boldsymbol{x}\boldsymbol{Y}}}{s_{\boldsymbol{x}\boldsymbol{x}}}\right)'$,

(b) $\text{SSE} = (n-1)\, s_{\boldsymbol{Y}\boldsymbol{Y}} \big(1 - r^2_{\boldsymbol{x}\boldsymbol{Y}}\big)$,

(c) $R^2 = r^2_{\boldsymbol{x}\boldsymbol{Y}}$,

using the usual notations of the corresponding sample variances, sample covariance and Pearson correlation coefficient

$$s_{\boldsymbol{x}\boldsymbol{x}} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \ , \ s_{\boldsymbol{Y}\boldsymbol{Y}} := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \ ,$$

$$s_{\boldsymbol{x}\boldsymbol{Y}} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(Y_i - \overline{Y})^2 \ , \ r_{\boldsymbol{x}\boldsymbol{Y}} := \frac{s_{\boldsymbol{x}\boldsymbol{Y}}}{\sqrt{s_{\boldsymbol{x}\boldsymbol{x}}\, s_{\boldsymbol{Y}\boldsymbol{Y}}}} \ .$$

Exercise 16

Consider the one factorial analysis of variance model in effect representation as given in Definition I.6.15 with the side condition

$$\sum_{i=1}^{p} n_i\, \alpha_i = 0 \ .$$

For the corresponding sum of squares $\text{SSR} = \sum_{i=1}^{p} n_i\, (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$ (defined on slide 165), show:

$$\text{E(SSR)} = (p-1)\,\sigma^2 + \sum_{i=1}^{p} n_i\, \alpha_i^2 \ .$$

Exercise 17

Consider the following two classes of univariate probability distributions:

(i) The class of all univariate (regular) normal distributions $\big\{\mathcal{N}(\mu, \sigma^2) \,\big|\, \mu \in \mathbb{R} \,, \sigma > 0\big\}$,

(ii) the class of all Poisson distributions $\big\{\mathcal{P}(\mu) \,\big|\, \mu \in (0, \infty)\big\}$.

(a) Show that each of these two classes is a member of the (univariate) Exponential Dispersion Family (EDF) by finding suitable representations of the corresponding density functions or probability mass functions, respectively, as in Definition II.2.3.

In particular, for each class, confirm the representations for the natural parameter θ and the functions b, a and c given in Example II.2.5.

(b) Using the results of (a), confirm the expressions for the mean $\text{E}(Y)$ and for the variance $\text{Var}(Y)$ given in Example II.2.5 for each of the two classes.