Statistisk Modellering ST523, ST813
Exercise Session 3

The first section contains exercises that will be addressed during the exercise session. All those exercises should be prepared BEFORE the exercise session.

The second section contains self study exercises which will not be covered. But make sure you know how to solve these.

3 Exercises

Exercise 3.1

Simple linear regression:

Let $\mathbf{Y} \in \mathbb{R}^n$ be a random (response) vector and consider a linear model for $\mathbf{Y}$ with an intercept and a single predictor x. The entries in $\mathbf{Y}$ can be expressed as

$$Y_i = \beta_1 + \beta_2 x + \varepsilon_i, \text{ for } i = 1, \dots, n$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent error terms with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ for $i = 1, \dots, n$. Equivalently we have

$$\mathbf{Y} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \boldsymbol{\varepsilon},$$

with $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$.

The aim of this exercise is to derive an explicit formula for the least-square estimator $(\hat{\beta}_1, \hat{\beta}_2)^T$ and to discuss its precision.

a) State and solve the normal equations, and show that $(\hat{\beta}_1, \hat{\beta}_2)^T$ is given by

$$\hat{\beta}_1 = \bar{\mathbf{Y}} - \bar{x}\hat{\beta}_2 \quad \text{and} \quad \hat{\beta}_2 = \frac{S_{xy}}{S_x^2}.$$

b) What is the variance matrix of $(\hat{\beta}_1, \hat{\beta}_2)^T$ and what does the precision of the estimated slope $\hat{\beta}_2$ depend on?

Exercise 3.2

Interpretation of regression coefficients, fitted values and residuals:

Install and load the library ”faraway” in **R** by typing

```
install .packages(”faraway”)
library (faraway)
data(”uswages”)
```

Complete the following tasks

a) Fit a model with weekly ages as the response and years of education and experience as predictors.

b) Report and give a simple interpretation of the regression coefficient for years of education.

c) Calculate the fitted values $\hat{Y}$ and the vector of residuals $\hat{\boldsymbol{\varepsilon}}$ and check their orthogonality in **R** .

d) Plot the residuals $\hat{\boldsymbol{\varepsilon}}$ versus the fitted values $\hat{\boldsymbol{Y}}$. Does the plot support the assumptions of centered errors and homogeneous error variances (i.e. $Var(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$)?

e) Discuss which properties of the least-square estimator could be impacted if the above assumptions are invalid.

Exercise 3.3

Part II: Simulations and bias of LS estimator

Next, we use simulations in **R** to investigate bias of the LS estimator under different residual scenarios. In order to do this, we use the same design matrix $\mathbf{X}$ as given by the data *uswages* (cp. last exercise) and assume that the data follows the below linear model

$$wages_i = -250 + 50educ_i + 10exper_i + \varepsilon_i \tag{1}$$

For the error terms ε_i we assume independence and consider the following three different error distributions:

- Setting 1: ε_i iid $\sim N(0, \sigma^2)$ with $\sigma^2 = 400^2$
- Setting 2: $\varepsilon_i \sim N(0, c(\mu_i)^2 \cdot \sigma^2)$, i.e. residual variance depending on the mean value $\mu_i = -250 + 50educ_i + 10exper_i$ and where $c(\mu_i) = \frac{\mu_i - \min_j \mu_j}{\max_j \mu_j - \min_j \mu_j}$
- Setting 3: $\varepsilon_i \sim N(1000\sin(\mu_i), \sigma^2)$, i.e. non-centered residuals depending on μ_i

In which setting would you expect a bias for the LS-estimate?

Then for Setting 1, use **R** and repeat the following steps $N_{sim} = 10000$ times: (If necessary, you might work on a subset of the data to increase computational performance or adapt N_{sim}.)

1. Simulate new wages according to (1).
2. Calculate the LS-estimate $\hat{\beta}$ for the simulated data.
3. Save these estimates in a suitable structure in **R** .

You should now have a sample of N_{sim} different LS-estimators. Calculate the average of these estimators over the different simulations. What does this tell you about the bias?

Repeat the above considerations for the other two settings.

Exercise 3.4

One-way layout

Consider a study with the aim of comparing response distributions for different groups, e.g. comparing mean crop yields for four fertilizers. For k groups of independent observations, let Y_{ij} denote the response observation j in group i for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. We regard the k groups as categories of a qualitative factor. This data structure is called the *one-way layout.* Especially, we assume that

$$Y_{ij} = \beta_i + \varepsilon_{ij}$$

where ε_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$, are i.i.d. with zero mean and variance σ^2.

1. Write the model as a linear model in matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, and identify the design matrix.
2. Calculate the least-squares estimator $\hat{\boldsymbol{\beta}}$.
3. In which way does the variance $Var(\hat{\boldsymbol{\beta}})$ depend on $n_1, \ldots, n_k$?
4. Now assume $k = 2$ and that the main interest of the study is to compare the two groups, i.e. to estimate the difference in means $\beta_1 - \beta_2$. Which estimate would you use and how would you choose n_1 and n_2 in order to achieve best possible precision of your estimate?

Exercise 3.5

Proof of Lemma 5.2

Prove that for a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a random vector $\mathbf{Z} \in \mathbb{R}^n$ the following equality holds

$$\mathrm{E}(\mathbf{Z}^T\mathbf{A}\mathbf{Z}) = \mathrm{tr}(\mathbf{A}\mathrm{Var}(\mathbf{Z})) + \mathrm{E}(\mathbf{Z})^T\mathbf{A}\mathrm{E}(\mathbf{Z}).$$

(We assume that all considered moments exist.)

Hints:

- $\mathrm{Cov}(Z_i, Z_j) = \mathrm{E}(Z_i Z_j) - \mathrm{E}(Z_i)\mathrm{E}(Z_j)$
- Start to transform the expression $\mathrm{E}\left((\mathbf{Z} - E(\mathbf{Z}))^T\mathbf{A}(\mathbf{Z} - E(\mathbf{Z}))\right)$

4 Self-study at home

Finish exercise 2.4 (Expectation and variance of random vectors) from the last exercise sheet in selv-study.