ST523/ST813 - Statistical Modelling

E2024

Home Assignment 1

This assignment should be **submitted electronically** as a single **pdf file** via **itslearning.** See itslearning for the submission deadline. It is expected that you **work independently** on the assignment. Interactions will be considered as exam fraud.

Unless stated differently, **R** can be used to solve the exercises. The code that you use has to be included in the solution and has to be documented in a detailed way. For example:

The following code creates a vector x that contains even integers from 2 to 10 and calculates the mean of x.

```
x <- 2 * 1:5
m <- mean(x)
```

The mean of x is equal to 6.

Code with no explanation will not be accepted as a valid answer.

Your submission should not exceed 8 pages.

Exercise 1

This exercise focuses on the analysis of data from house sales in Odense municipality in the period from August 9th, 2024 until September 30th, 2024. The data relies on free housing trades published on `www.boligsiden.dk`.

To access the data for this exercise, save the file `Data_ST523_813_E2024.rdata` in your **R** working directory and type the following command `load("Data_ST523_813_E2024.rdata")` in **R** .

The main interest in this exercise lies in the statistical modelling of sales prices and their relation to selected characteristics of the sold property. We focus on sales prices relative to the surface, i.e. price per square meter. Due to the recent change of rules concerning maintenance and replacement of roofs made from fibre cement containing asbestos our main interest will further lie on the role of the roof material for the sales price. The data includes the following variables:

`Adress`	adress of the property
`District`	part of town
`Postcode`	postal code (5000, 5200+5210, 5220+5250+5260, 5230, 5240+5320, 5270)
`Price`	achieved sales price in DKK
`Pricesqm`	corresponding price per m^2 (in DKK/m^2)
`Dato`	date of the sale
`Type`	type of property (villa, row house, apartment)
`Energyclass`	energy label of the property (A,B,C,D,E,F)
`Surface`	living area (in m^2)
`Rooms`	number of rooms
`Toilets`	number of toilets
`Roof`	type of roof (roof tile, fibre cement including asbestos, other material)
`Yearc`	year of construction

1. (Data Exploration) Explore the dataset.

   - What are the dimensions of the data? (nr. of variables and nr. of observations)
   - Specify the type of each variable. (categorical vs. continuous)
   - Create a table containing relevant summary measures for each of the variables individually.
   - In case a variable contains missing values, report the percentage of missing values.

2. (Two-way ANOVA) Consider now a linear model with `Pricesqm` as response and `Roof` as well as `Postcode` as only predictors. The aim of this exercise is to perform a two-way analysis of variances.

   a. Create a scatter plot of the response versus the two factors. Reflect both factors simultaneously in your plot(s).

   b. Conduct a two-way analysis of variances including an interaction to analyze the data: Investigate first the presence of an interaction between the two predictors wrt the response.

      Comment on the results and specify the following:

      – underlying statistical model
      – investigated null hypothesis and alternative hypothesis
      – observed value of the test statistic
      – null distribution and corresponding degrees of freedom
      – p-value

      Conclude whether there is statistical evidence for an interaction or not.

   c. Continue the statistical analysis and test the presence of additive effects of `Roof` and `Postcode` onto the sales price and interpret your results.

   d. According to your resulting model, which mean sales prices per m^2 are estimated for the different combinations of the two factors. Present your results in a table.

e. How much variation in **Pricesqm** is explained by the two factors?

   And what is the corresponding (absolute) reduction in the residual sum of squares, i.e. indicate the residual sum of squares from your resulting two-way model and the total sum of squares?

3. (Model extensions) Consider now an extended statistical model for **Pricesqm** including additionally **Type**, **Yearc** as well as **Surface** and perform a suitable test to compare the extended model with the two-way model.

   As before, report the corresponding hypothesis, test statistic, p-value, null distribution and degrees of freedom. What is your conclusion?

4. (Residual diagnostics) Perform a graphical check of the most inclusive model (i.e. including all available predictors as well as an interaction between **Roof** and **Postcode**).

   - Calculate standardized residuals and fitted values.
   - Create two residual plots (residuals versus fitted values and a QQ-plot)

   Comment on these plots. Do they indicate that the underlying assumptions of your prior analyses are corrupted? Mention one possible consequence of invalid assumptions.

Exercise 2

Assume a normal linear model for y with a continuous predictor x and with intercept:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

The errors $\varepsilon_1, \dots, \varepsilon_n$ are assumed to be iid $N(0, \sigma^2)$ distributed.

You performed a corresponding linear regression on a dataset with $n = 20$ observations. The estimated parameters were $\hat{\beta}_1 = -10.3$ and $\hat{\beta}_2 = 2.5$, the residual variance $\hat{\sigma}^2 = 5.0$ and

$$(X^T X)^{-1} = \begin{bmatrix} 2.00 & -0.10 \\ -0.10 & 0.05 \end{bmatrix},$$

where X is the design matrix. The first column of X corresponds to the intercept and the second column to x.

1. Calculate a confidence interval for the mean value of y at $x = 2.0$ with confidence level 0.90.

2. Does the data provide evidence for a mean value above 0 at significance level $\alpha = 0.05$, i.e. investigate $H_0 : \beta_1 + 2\beta_2 \leq 0$ vs. $H_1 : \beta_1 + 2\beta_2 > 0$?

Exercise 3

Consider a linear regression model with response y, an intercept and two explanatory variables x_1 and x_2, i.e.

$$y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \varepsilon_i, \quad i = 1 \dots, n, \tag{1}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent, centered and have a common variance σ^2, and let $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ be the least square estimators for β_1, β_2 and β_3.
Assume that the sample correlation of the two explanatory variables x_1 and x_2 is positive, i.e.

$$r_{x_1,x_2} > 0. \tag{2}$$

1. Show that the correlation between $\hat{\beta}_2$ and $\hat{\beta}_3$ is negative.

2. Write a small simulation program in **R** that confirms the above point 1. Especially this should involve the following steps: For a model matrix of your choice corresponding to (1) and (2),

   - simulate the response
   - calculate the least square estimates
   - repeat these steps a sufficient number of times
   - confirm 1. using suitable plots and summary measures of the obtained sample of LS-estimates

   Provide your **R** code well-documented as appendix. Moreover, use your own commands for the repeated calculation of LS-estimates during the simulation, **i.e. do not use built-in functions for model fitting such as `lm`**.

   It should be possible to **reproduce** your simulation.

   NB: For the simulation, you might reuse in a suitable way the data from exercise 1 and e.g. consider `Surface, Rooms` as predictors for `Pricesqm`.

Exercise 4
(only ST813)

Assume we fit the following model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{3}$$

although the true model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Assume that the model errors ε satisfy $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$.

- Derive the expected value and variance of the least-squares estimate $\hat{\boldsymbol{\beta}}_1$ calculated based on (3).

- In which situation is $\hat{\boldsymbol{\beta}}_1$ unbiased?