ST523/ST813 - Statistical Modelling

E2025

Home Assignment 1

This assignment should be **submitted electronically** as a single **pdf file** via **itslearning.**

R (or another statistics software) can be used to solve the exercises. The **output from the software should be explained in detail**, and all **replies to a question should be stated separately in the text**. Copied output alone without further explanation will not be accepted as valid answer.

Your final submission **must not exceed 13 pages** (additional pages will be disregarded). The page count is number of pdf pages (A4) using normal font size 11 or 12.
You can choose to include pieces of code, but it is recommended to NOT include unimportant code (e.g. for generating figures, data handling or formatting). Avoid long text passages, formulate shortly and precisely.

It is expected that you **work independently** on the assignment. Interactions will be considered as exam fraud.

Generative AI:

- It is **NOT ALLOWED** to use generative AI **for carrying out any part of the statistical analysis** of the datasets that are studied during this assignment.

Exercise 1

This exercise analyses data from a gas turbine. The data stems from `https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set`, where more information is available. In this exercise we work with a smaller subset of the published data.

To access the data for this exercise, save the file `Data_ST523_813_E2025_Exam.rdata` in your **R** working directory and type the following command in **R** :
`load("Data_ST523_813_E2025_Exam.rdata")`

We are interested in emissions of carbon-monoxide, which are considered harmful pollutants, and our aim is to investigate how different ambient variables as well as process parameters of the gas turbine relate to the emission. The data includes the following variables

`AT`	ambient temperature ($^\circ C$)
`AP`	ambient pressure (mbar)
`AH`	ambient humidity (%)
`GTEP`	gas turbine exhaust pressure (mbar)
`TIT`	turbine inlet temperature ($^\circ C$)
`TAT`	turbine after temperature ($^\circ C$)
`CDP`	compressor discharge pressure (mbar)
`TEY`	turbine energy yield (MWh)
`CO`	emission of carbon monoxide (mg/m^3)

The first three variables describe the conditions in the surrounding of the turbine, the subsequent variables represent technical process parameters related to operating the turbine and obtained from sensors placed inside/at the turbine. The last variable is our response variable describing the emissions of the considered pollutant.

1. Explore the dataset.

    - What are the dimensions of the data? (nr. of variables and nr. of observations)
    - Create a table containing relevant summary measures for each of the variables individually.
    - Make pairwise scatterplots.

2. Fit a linear model using `CO` as response and the ambient variables `AT, AP, AH` as well as the process parameters `GTEP, TIT, TAT, CDP, TEY` as predictors.

    - Report the estimated parameters.
    - According to this model, which change in CO-level do you expect when the ambient temperature increases by $1^\circ C$ while all other parameters stay unchanged?

3. Perform an overall F-test and specify the following:

    - investigated null hypothesis and alternative hypothesis
    - observed value of the test statistics
    - p-value
    - null distribution and corresponding degrees of freedom

    What is your conclusion at significance level $\alpha = 0.05$?

4. Apply suitable tests to investigate whether the full model can be reduced to one of the following two submodels:
   - Model A containing only ambient variables `AT, AP, AH`
   - Model B containing only the process variables `GTEP, TIT, TAT, TEY`

   Report the corresponding hypotheses, test statistics and p-values. What is your conclusion?

5. How much variation in `CO` is explained by your final model from the last step?

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

Exercise 2

A Danish labour economist is analysing how different types of employment contracts affect the annual income of individuals with a mathematics-related education (e.g., statisticians, actuaries, data scientists), while controlling for years of relevant full-time work experience. The employment contracts considered are:

- *Fastansættelse* (Permanent) – used as the reference category
- *Tidsbegrænset – forskningsansat* (Temporary – Research/Academic)
- *Tidsbegrænset – privat konsulent* (Temporary – Private Consultant)
- *Freelancearbejde* (Freelance)

A linear model was fitted with annual income in DKK as response and type of contract as categorical and years of experience as numerical covariate:

$$Y_i = \mu + \alpha_{j(i)} + \beta \cdot X_i + \varepsilon_i, \qquad i = 1, \ldots, 45,$$

where:

- Y_i: Annual income (in DKK),
- α_j: Effect of contract type j relative to permanent contracts,
- X_i: Years of full-time experience,
- ε_i: Independent, normally distributed errors with mean zero and common variance.

Model output from statistical software:

Coefficient	Estimate (DKK)	Std. Error
(Intercept)	520,000	20,000
ContractTempResearch	-40,000	24,000
ContractTempPrivate	-10,000	23,000
ContractFreelance	15,000	21,000
Experience (years)	18,000	2,000

Estimated covariances between selected coefficient estimates:

$$\widehat{\mathrm{Cov}}(\hat{\alpha}_{\mathrm{TempResearch}}, \hat{\alpha}_{\mathrm{TempPrivate}}) = 22,000,000$$

$$\widehat{\mathrm{Cov}}(\hat{\alpha}_{\mathrm{TempResearch}}, \hat{\alpha}_{\mathrm{Freelance}}) = 20,000,000$$

$$\widehat{\mathrm{Cov}}(\hat{\alpha}_{\mathrm{TempPrivate}}, \hat{\alpha}_{\mathrm{Freelance}}) = 21,000,000$$

Number of observations: 45.

1. Compute a 90% confidence interval for the difference in adjusted mean income between:

    - Temporary - Research/Academic, and
    - Temporary - Private Consultant

    That is, compute a confidence interval for $\alpha_{\text{TempResearch}} - \alpha_{\text{TempPrivate}}$

2. Does the data provide statistical evidence for temporary researchers earning less than temporary private consultants at significance level $\alpha = 0.05$?

Exercise 3

Assume you are modelling a straight line relationship using simple linear regression and your aim is to estimate the slope as precisely as possible. Assume further the region of interest for the x variable is $-2 \leq x \leq 2$. Where should the observations $x_1, x_2, \ldots, x_n$ be taken? Prove your answer.

Exercise 4
(only ST813)

Verify that the $n \times n$ identity matrix I is a projection matrix and describe the linear model which it corresponds to.