Statistisk Modellering (ST523,ST813)
Exercise Session 5

All exercises should be prepared BEFORE the exercise session.

5 Exercises

Exercise 5.1

Confidence interval for simple linear regression

For simple linear regression, consider the confidence interval for the mean outcome at specific values x^* of the predictor, cp. lecture 7.

- Derive/check the formula for the confidence interval.
- How does the width of the interval depend on x_*?

Consider now the corrosion dataset from the faraway library in **R** . The dataset can be loaded as follows

```
library(faraway) # loads external library
data(corrosion) # loads dataset
help(corrosion) # to show some information
```

Copper-nickel alloys are widely used in marine applications due to its good resistance to seawater corrosion. The latter effect is enhanced by further addition of iron to the alloy. Our example dataset contains observations of weight loss (in units of milligrams per square decimeter per day) for a number of independent alloy samples with varying iron content which have been submerged in sea water during 60 days.

- Plot the data.
- Fit a simple linear regression model and interpret the estimates.
- Which weight loss do you have to expect for 1.5% iron content?
- Calculate a confidence interval for the weight loss at 1.5% iron content.

You want to test the null hypothesis that the same expected weight loss is equal to 95mg/dm^2/day.

- Explain and perform the corresponding test and provide a p-value.
- Can you reject H_0?
- Perform the test in **R** , e.g. using the linearHypothesis command from the car package.
- Would your conclusion change if you had considered the null hypothesis H_0 : "weight loss of at least 95mg/dm^2/day"?

Exercises with focus on linear model theory

Exercise 5.2

(Sums-of-squares decomposition) Assume a normal linear model. The decomposition of the sums of squares is given by

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SS_{tot}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SS_{reg}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{RSS}$$

The global F-test uses the test statistics

$$F = \frac{SS_{reg}/(p-1)}{RSS/(n-p)}$$

a) Prove that SS_{reg} and RSS are independent.
Hint: Use the same approach as in the proof of point 3 from Proposition 7.9.

b) Use **R** and repeated simulations from a linear model of your choice, to check (graphically) that $SS_{reg}/(p-1)$ and $RSS/(n-p)$ are χ^2-distributed and independent, and check that their ratio is F-distributed under the null-model.
The corresponding sums of squares can be calculated directly or obtained from e.g. the anova-function in **R** using e.g. `anova(fit0, fit1)$RSS`.

c) Is also the residual sum-of-squares corresponding to the null model, i.e. SS_{tot}, independent of SS_{reg}?

d) * Choose one specific parameter vector from the alternative (i.e. not satisfying the global null hypothesis) and try to determine the power of the global F-test adapting your simulations.

Exercises focussing on data analysis and application of F-tests

The exercises use the economic dataset on 50 countries, `savings`, from the `faraway` package. The data represents averages from 1960 to 1970 (to remove business cycle or other short-term fluctuations):

- `sr` is aggregate personal saving divided by disposable income;
- `pop15` is the percentage of population under 15;
- `pop75` is the percentage of population over 75;
- `dpi` is per capita disposable income in U.S. dollars;
- `ddpi` is the percentage rate of change in per capita disposable income.

Exercise 5.3

- Start out inspecting the dataset e.g. what is the sample size, how many variables and of which type. Produce a scatterplot matrix using the command `pairs`.
- Model `sr` as reponse variable depending on the remaining variables using a normal linear model. Fit a regression model and interpret the output.

Exercise 5.4

(Test of all the predictors)

- Perform an overall F-test. Does any of the predictors have significance in the model? In other words, can you reject $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$?
- Verify the results for the F-test from **R** 's regression summary by repeating the underlying calculations on your own.