Dr. Tim Brereton
Lisa Handl

Winter Term 2014/15

Stochastics III

Problem Sheet 4

Deadline: December 10, 2014 at noon, before the practical

Exercise 1 (4 + 2 + 4 points)

The stopping distance y of a car (in m) depends on its velocity v (in km/h). This relationship has been investigated experimentally for one specific type of car by measuring the stopping time for $n = 20$ different velocities. The following three models shall be compared regarding how well they describe the relationship between v and y:

(i) $y = \beta_1 + \beta_2 v$,

(ii) $y = \beta_1 + \beta_2 v + \beta_3 v^2$ und

(iii) $y = \beta_1 v + \beta_2 v^2$.

The measured values can be found in the file *velocity.txt* on the lecture website.

a) Calculate the least squares estimator for the regression coefficients β_i and draw a scatter plot of the data and the regression function for each model using **R**. For reasons of comparability generate all three plots next to each other in one figure.

   Hint: This can be done using `par`, e.g., `par(mfrow = c(2,2))` prepares a 2-by-2 matrix of plots.

b) The coefficient of determination R^2 is a number that indicates how well a model fits given data, and is defined as

$$R^2 = 1 - \frac{\widehat{\varepsilon}^\top \widehat{\varepsilon}}{y^\top y - n(y^*)^2},$$

   where $y^* = \frac{1}{n}\sum_{i=1}^n y_i$ if the model has an intercept and $y^* = 0$ otherwise. It can be interpreted as the fraction of variance in the data which is explained by the model and takes values in $[0, 1]$. Large values indicate a good model fit, low values indicate a bad one. Calculate R^2 for each model (using **R** as a calculator) and use it to decide which model seems to be the best.

c) Assume that the measurement errors in the experiment described above are independent and normally distributed with known variance $\sigma^2 > 0$. Determine the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ from a) for model (i) and provide confidence intervals with level $\alpha = 0.05$ for the true β_1 and β_2. You may use **R** to compute quantiles of the standard normal distribution.

Exercise 2 (4 points)

Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$ be independent and (centrally) chi-squared distributed with m and n degrees of freedom. Show that

$$W = \frac{U/m}{V/n} \sim F_{m,n},$$

i.e., show that W has density

$$f_W(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}} \mathbb{I}_{[0,\infty)}(x).$$

Exercise 3 (3 + 2 points)

The number of visitors last season was recorded at 10 different ski resorts. It is suspected that the number of visitors depends linearly on the total length of pistes and on the capacity of ski lifts. The data is available in the file *ski.txt* on the lecture website. You may assume that the error terms are normally distributed. Work through the following exercises **using R only as a calculator**:

a) Write out a suitable linear model and determine $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^\top$. Conduct a test for the hypotheses

- $H_0^{(1)}$: The number of visitors does not depend on any of the predictor variables.
- $H_0^{(2)}$: The number of visitors does not depend on the length of pistes.
- $H_0^{(3)}$: The number of visitors does not depend on the capacity of ski lifts.

with significance level $\alpha = 0.05$ and interpret the results.

b) Now assume that the number of visitors depends only on the capacity of ski lifts, i.e., we now consider a simple linear regression model. Determine $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2)^\top$ for this model and test the hypothesis $H_0^{(4)}$, that the number of visitors does not depend on the lift capacity. Interpret the result.

Exercise 4 (2 + 3 points)

Consider the data in the file *water.txt* on the lecture website. It contains the experimentally determined boiling points (in °F) of water at different air pressures (in inches of mercury). The model we consider is that the boiling point depends linearly on the air pressure. You may assume that the error terms are normally distributed.

a) Fit the regression model using **R**. Output the regression coefficients and plot the data together with the regression line.

b) Test if the boiling point grows twice as fast as the air pressure and the intercept is 150 (in one common test) **using R only as a calculator**.