Project Exam for STK4040/9040
Fall semester 2011

The exam in STK4040/9040 consists of this project exam and a written exam.

The written solution to the project exam (in Norwegian or English) must be handed in no later than Friday December 9th at 2 pm either by regular mail or by e-mail to
Ørnulf Borgan, Matematisk institutt, Universitetet i Oslo, P.B. 1053 Blindern, 0316 Oslo
e-mail: borgan@math.uio.no

You are not allowed to collaborate with others on the project exam

The written exam takes place Tuesday December 13th. Details are posted on the course web-page.

The project exam consists of four problems. In each problem you shall analyse a data set and interpret the analyses you have done. The written solution to the problems should be divided into two parts. In the main part you answer the questions and present the numerical results and plots that are necessary for your arguments. In an appendix you should document the computer code you have used to obtain the results in the main part. (You should only include the final code, not all trial and errors.)

If you have questions concerning the problems or technicalities in R, please send an email to borgan@math.uio.no

Remember to write your name, date of birth (cf. problems 3 and 4), and e-mail address on the solution.

Problem 1

It has been suggested that a change in skull size over time may be evidence of interbreeding of a resident population with immigrant populations. In this problem we will look at data on four measurements of Egyptian skulls for two different time periods, and investigate whether the (average) skull size differs between the periods. Period 1 is 4000 BC and period 2 is 1850 BC. For each of the two periods we have measurements for 30 skulls.

It is described on the course web-page how you may read the data into R.

The variables in the data set are the following:

- `maxb` maximum breadth of skull (mm)
- `bash` basibregmatic height of skull (mm)
- `basl` basialveolar length of skull (mm)
- `nash` nasal height of skull (mm)
- `per` period (1=4000 BC, 2=1850 BC)

(The exact definitions of these measurements are not important for our problem.)

a) Check if it is reasonable to assume that the four measurements of the skulls are multivariate normal (for each of the two time periods) or if a transformation of (one or more of) the measurements ought to be carried out. Also check if there are outliers in the data that ought to be excluded from the analysis. Make sure to comment on the interpretation of the plots and test statistics you use.

b) Investigate whether there is a difference in (average) skull size for the two time periods. If you find a difference, describe this difference by appropriate confidence intervals and/or confidence regions.

Problem 2

We will study weather data for Trondheim for the month of September 1997. A number of variables were measured each hour throughout the month, but we will only use the daily means of these variables.

On the course web-page it is described how you may read the data into R.

In the data set there is one line for each of the 30 days of September, and one column for each of the variables (daily means):

- `wind.dir` wind direction (degrees; 0 degrees corresponds to north)
- `wind.veloc` wind velocity (m/s)
- `air.press` air pressure (hPa)
- `precip` precipitation (mm)
- `air.temp` air temperature (^{o}C)
- `rel.hum` relative humidity (percent)

a) Plot the variables as a function of the day of the month. Also make a scatter plot matrix of the variables. Comment on what you may learn from the plots.

b) Perform a principal component analysis of the weather data using (i) the sample covariance matrix and (ii) the correlation matrix. Which of the two gives the most meaningful analysis?

c) Interpret the results of your preferred analysis focusing on the first 2-3 principal components.

Problem 3

Hemophilia is a hereditary bleeding disorder caused by a lack of a blood clotting factor. Without this factor, the blood cannot clot properly to stop bleeding. A woman can be a carrier for the recessive gene for hemophilia without suffering from the disease.

To construct a procedure for detecting potential hemophilia carriers, blood samples were taken for a number of women who did not have the hemophilia gene (non-carriers) and for a number of women who had the gene (carriers). From the blood samples measurement were taken on AHF activity and AHF antigen (where AHF denotes antihemophilic factor).

The original data contain measurements on 30 non-carriers and 45 carriers. However, for this problem you should use data for all the non-carriers and a sample of 30 of the carriers. Each student should use their own sample of 30 of the 45 carriers, and it is described on the course web-page how you should proceed to obtain your sample.

Make sure that you follow these instructions carefully.

The variables in the data set are the following:

- `group` group of women (1=non-carriers; 2=carriers)
- `activity` AHF activity (on log10-scale)
- `antigen` AHF antigen (on log10-scale)

a) Make a scatter plot of AHF activity and AHF antigen using different plotting symbols for non-carriers and carriers. Comment on what you may see from the plot.

b) Use the methodology of Section 11.3 in J&W to determine a rule that can be used to classify a woman as carrier or non-carrier of the hemophilia gene.

c) Determine the apparent error rate and (an estimate of) the expected actual error rate, and discuss what these error rates tell you about the classification rule.

d) Use logistic regression to make another classification rule, and compare this rule with the one you studied in questions b and c.

Problem 4

In a paper mill one makes paper from vegetable fibers such as wood pulp. We will consider data on pulp fiber characteristics and properties of the paper made from it.

Each student should use their own sample of 30 of the 62 observations, and it is described on the course web-page how you should proceed to obtain your sample.

Make sure that you follow these instructions carefully.

The variables in the data set are the following:

- `BL` breaking length of paper
- `EM` elastic modulus of paper
- `SF` stress at failure of paper
- `BS` burst strength of paper
- `AFL` arithmetic fiber length in pulp
- `LFF` long fiber fraction in pulp
- `FFF` fine fiber fraction in pulp
- `ZST` zero span tensile in pulp

(The exact definitions of these measurements are not important for our problem.)

Fit a two factor model to the data using maximum likelihood, and give an interpretation of the fitted factor model (without worrying about whether the model gives a good fit to the data or not). Make sure that you comment on the interpretation of all the (estimated) parameters of the model.