Project Exam for STK4900 and STK9900
Spring semester 2012

The exam in STK4900 and STK9900 consists of this project exam and a written exam.

The written solution to the project exam must be handed in no later than 2.30 pm Thursday May 31st at Niels Henrik Abels hus, 7th floor (by the reception).

Note that you are not allowed to collaborate with others on the project exam.

The written exam takes place June 12th. Details are posted on the course web-page.

The project exam for STK4900 consists of the first three problems given below, while the project exam for STK9900 consists of all four problems.

In each problem you shall analyse a data set and interpret the analyses you have done. You may use the software package of your choice, but whether you use R or not, you must be able to answer all questions. We recommend that you use R.

The written solution to the problems should be divided into two parts. In the main part you should 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.)

Instructions on how you should read the data into R are given on the course web-page. If you have problems reading the data or other problems concerning technicalities in R, please send an email to `borgan@math.uio.no` or `osamuels@math.uio.no`

PROBLEM 1 (both STK4900 and STK9900)
The forced vital capacity (FVC) is the total amount of air an individual can expel after a maximum inhalation and so is a measure of lung capacity. In this problem we will look at data on FVC for 602 Norwegian children with the purpose of studying how FVC depends on the covariates height, weight, age, and sex of the child.

The data are found in the file `lungfunction.txt` at the course web-page. The file contains the following variables for all children:

• `fvc`	Forced vital capacity in litres
• `weight`	Weight in kg
• `height`	Height in cm
• `sex`	Sex (0 for girls, 1 for boys)
• `age`	Age in years (decimal numbers)

a) Carry out a simple linear regression with FVC as the response variable and height as the covariate. Interpret the estimated regression coefficient and the R^2 measure.

b) Compare the model in a) with a model with all covariates included. Interpret the results of this "full model".

c) Do a residual analysis of the full model from question b). Is the model fit adequate?

d) Investigate whether the full model may be improved by including second order or interaction terms. Also investigate whether log-transforms of the response variable and covariates seem to improve the model.

e) Make a summary of your findings.

PROBLEM 2 (both STK4900 and STK9900)

In this problem you shall analyse data on accidents (in 2004 and 2005) in a portfolio of private cars in an Australian insurance company. At the course web-page you find the data set `insurance.txt` which contains the number of accidents according to the age of the car, the sex of the driver, and the age of the driver. The data set also contains information on the number of person-years in each group (defined by age of car, sex of driver, age of driver).

The variables in the data set are as follows:

• `car.age`	Age of car: 1 = 0-1 years 2 = 2-5 years 3 = 6-10 years 4 = more than 10 years
• `sex`	Sex of driver: 1 = Male 2 = Female
• `driver.age`	Age of driver: 1 = less than 25 years 2 = 25-34 years 3 = 35-44 years 4 = 45-54 years 5 = 55-64 years 6 = more than 65 years
• `accidents`	Number of accidents in the group
• `exposure`	Number of person-years in the group

a) Explain why it is reasonable to assume that the number of accidents in a given group is Poisson distributed, and describe a type of regression model that is suitable for analysing the data.

b) Perform an analysis of the data using the regression formulation described in question a). As part of the analysis, you should investigate whether all the factors (age of car, sex of driver, age of driver) are needed to describe the accident rate, and whether there are interactions between the factors.

c) Give an interpretation of the model you arrive at in question b).

PROBLEM 3 (both STK4900 and STK9900)

In a study one considered 575 drug addicts who were under treatment for their drug problems, and who were not using drugs at the start of the treatment. For each drug addict, one recorded time to relapse (when drug abuse started again) or censoring.
A number of covariates were recorded for each drug addict at the start of the study.
We will restrict attention to the covariates age, number of earlier treatments, and use of heroine/cocaine the last three months before start of treatment.

At the course web-page you find the data set `drug.txt` containing the data we will consider. The data are organized with one line for each of the 575 drug addicts, and with the following variables in the five columns:

- `days` number of days from start of study to relapse to drug abuse or to censoring
- `relapse` relapse to drug abuse or censoring (0: censoring; 1: relapse)
- `age` age at start of study (in years)
- `no` number of earlier treatments for drug abuse
- `use` use of heroine/cocaine last three months before treatment (1: both heroine and cocaine, 2: only heroine, 3: only cocaine; 4: neither heroine nor cocaine)

When analysing the data, we start out with univariate analyses where we group according to the values of the numeric covariates age and number of earlier treatments.
It is described on the course web-page how you may derive categorical covariates by grouping the numeric covariates. These categorical covariates should be used in questions a) and b). In question c) you may decide yourself whether you will use the original numeric covariates (or some transformation of these) or the categorical covariates.

a) Make Kaplan-Meier plots for the survival function for each level of the covariates grouped age, grouped number of previous treatments, and drug use the last three months before treatment. Discuss what the plots tell you.

b) For each of the covariates, use the logrank test to investigate if the covariate has a significant effect on survival.

c) Use an appropriate regression model to study how the risk of relapse is influenced by age, number of previous treatments, and drug use the last three months before treatment.

d) Give an interpretation of the model you arrive at in question c).

PROBLEM 4 (only STK9900)
A child is said to have low birth weight if its weight at birth is below 2.5 kg. In an American study one collected information on variables that could influence the probability that a woman will give birth to a child with low birth weight. In this problem, we will consider part of these data.

At the course web-page you find the data set `birthweight.txt` containing the data we will investigate. The data are organised with one line for each of the 189 women who took part in the study, and with the following variables in the five columns:

• `LOW`	low birth weight (0: no; 1: yes)
• `AGE`	age of mother (in years)
• `LWT`	weight of mother at last menstrual period before pregnancy (in pounds)
• `RACE`	race of mother (1: white; 2: black; 3: other)
• `SMOKE`	smoking status of mother during pregnancy (0: no; 1: yes)

a) Use an appropriate regression model to study how the covariates age, weight, race, and smoking status influence the probability that a woman will give birth to a child with low birth weight.

b) Give an interpretation of the model you arrive at in question a).