# Project exam for deferred/new exam in STK4900/9900

The deferred/new exam in STK4900/9900 consists of this project exam and a written exam. The written exam took place Friday August 15th.

*The written solution to the project exam (in Norwegian or English) must be handed in no later than Friday August 29th 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*

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

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 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

**PROBLEM 1**

In this problem we are interested in studying how the fuel consumption of a car depends on its size and engine characteristics. At the course web-page you find the data set `fuel` that gives fuel consumption, size and engine characteristics for a sample of 38 different makes of cars. The data are organised with one line for each car, and with the following variables in the five columns:

- `cons` fuel consumption (in litres per 10 km)
- `wt` weight (in kg)
- `disp` engine displacement (in litres)
- `cyl` number of cylinders
- `hp` horsepower of the engine

**a)** Study the relation between fuel consumption and weight and between fuel consumption and engine displacement using simple linear regression. Which of the two variables weight and engine displacement is by itself the best predictor for fuel consumption?

**b)** Study the relation between fuel consumption, weight and engine displacement using multiple linear regression. Give an interpretation of the fitted model.

**c)** Assess the fit of the model in question b using suitable plots of the residuals. Make sure that you comment on what each of the plots tells you.

**d)** Investigate if you get a better model for predicting fuel consumption by also including the number of cylinders and the horsepower of the engine in your regression model.

**PROBLEM 2**

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 a part of these data.

At the course web-page you find the data set `birthweight` 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.

**PROBLEM 3**

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` 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 log rank 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.