Project Exam for STK4900 and STK9900
Spring semester 2014

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

The answer to the project exam (in Norwegian or English) must be handed in no later than Thursday May 15th at 2:30 pm in Niels Henrik Abels hus, 7th floor (inside the reception room) or by regular mail to Department of Mathematics, University of Oslo, PO.Box. 1053 Blindern, 0316 Oslo. You must hand in **two copies** of your solution. The copies should be marked with course and your **candidate number**, not by name.

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

The written exam takes place on Monday June 4th. Details are posted on the course web-page.

The project exam consists of three problems. STK9900 students should answer all questions. STK4900 students are exempt from question d) in Problem 2 and question c) in Problem 3.
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.)

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

PROBLEM 1

Data have been collected to check whether the presence of urea formaldehyde foam insulation (UFFI) has an effect on the formaldehyde (CH_2O) concentration inside a house. Twelve houses with and 12 houses without UFFI were studied, and the average weekly CH_2O concentration was measured. It was thought that the CH_2O concentration could also be influenced by the amount of air that can move through the house via windows, cracks, chimneys, etc. For this purpose, a measure of "air tightness" was determined for each house.

At the course web-page you find the data set `uffi.txt` containing the measurements for the 24 houses. The variables in the data set are coded as follows:

- `CH2O`	Average weekly CH_2O concentration in parts per billion
- `AIR`	Measure of air tightness on a scale from 0 to 10 (high values correspond to a tight house)
- `UFFI`	Indicator for formaldehyde foam insulation (0=absent, 1=present)

a) Study the relation between CH_2O concentration and air tightness and between CH_2O concentration and formaldehyde foam insulation using suitable plots and univariate regression. Discuss the results.

b) Study the relation between CH_2O concentration, air tightness and formaldehyde foam insulation using multiple regression. In this connection you should also investigate if a model with second order term for air tightness and/or interaction between air tightness and formaldehyde foam insulation gives a better fit.

c) Give an interpretation of the model you arrive at in question b), and check the fit of the model using suitable plots of the residuals. Make sure that you comment on what each of the plots tells you.

PROBLEM 2

The data for this problem stems from an investigation of whether a health reform in Germany in 1997 led to reduced number of doctor visits. Some individuals were interviewed in 1996 (before the reform) and others in 1998 (after the reform). The data for this problem are restricted to women working full time and are given in the file `drvisits.txt` found at the course webpage. It has the following variables

- `numvisits` number of doctoral visits during the three months prior to the interview
- `age` age in years
- `educ` education in years
- `married` indicator variable for being married
- `badh` indicator for self-reported current health being classified as “very poor” or “poor” versus “very good”, “good” and “fair”
- `loginc` logarithm of household income (in 1995 Gemarn Marks, based on OECD weights for household members)
- `reform` indicator for the interview being undertaken in the year following the reform compared to the year preceding the reform

a) Explain why Poisson regression may be reasonable for analyzing how the number of doctoral visits depends the covariates. In particular explain the concept of a rate ratio. Carry out a Poisson regression using only the health reform variable and find 95% confidence intervals for the rate ratio. Interpret the result.

b) Investigate how all the covariates together affect the tendency to visit the doctor. Include only main effects. Discuss why the effect of the health reform variable is only modestly changed after including the other covariates.

c) Study whether there is a need to include quadratic effects of covariates or interactions between the covariates in how they affect the number of doctoral visits.

d) [Only to be answered by STK9900 students] Discuss the concept of over-dispersion relative to the Poisson assumption. How can the analyses be corrected for such over-dispersion. Carry out such an analysis using only main effects of the covariates. Compare with the previous analysis. Comment on the differences.

PROBLEM 3

In this problem we will consider the risk of wheezing (a whistling sound produced in the respiratory airways during breathing) among children according to whether their mother smokes and their age. The file “wheezing.txt” found at the course website contains 2148 such responses. It has the following variables

- `smoking` smoking status of the mother (1=yes, 0=no)
- `age` age of the child in years
- `wheezing` wheezing status of the child (1=yes, 0=no)

a) Calculate the proportions of children with wheezing according to whether the mother smokes. Test whether the probabilities of wheezing are significantly different in these two groups. Also calculate and compare the relative risk and odds-ratio of wheezing among children with smoking mothers compared to children of non-smoking mothers.

b) Analyze the wheezing data using logistic regression. In particular, demonstrate how the odds-ratio from question a) can be obtained from logistic regression. Furthermore investigate whether age has an influence on the outcome wheezing and whether the association with smoking is changed when taking age into account.

c) [Only to be answered by STK9900 students] In the two first questions in this problem we have omitted the fact that the wheezing information is obtained on the same children at ages 7, 8, 9 and 10 years. In an extended file “wheezing1.txt” a variable “`id`” is included which identifies each child.

Discuss in general terms an issue that should handled with such longitudinal data. Carry out an analysis an analysis that is appropriate for the data. Comment on the differences with the analysis in question b). (You may note that each mother is recorded as a smoker or a non-smoker at every age).