25th March, 2020

STK4900/9900 — Statistical methods and applications

Mandatory assignment 2 of 2

Submission deadline

This assignment must be handed in electronically using the **Canvas** system **no later than 2:30 pm Thursday April 16th 2020**.

Instructions

You are allowed to collaborate and discuss the problems with other students, but each student has to formulate her or his own answers. You should give the names of the students you collaborate with, so that it is possible to compare the written solutions.

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LATEX or Word). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

The solution may be divided into two parts, in particular if it is handwritten. In the main part you answer the questions and present the numerical results and plots that are necessary for your arguments. It is not sufficient to present the numerical results and the plots, you should also discuss what you can learn from them. In an appendix you may include computer printouts and other technical material that do not fit nicely into the main part (you should only include the final code, not all trial and errors).

You may use a 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 data files are provided on the course web page. If you have problems reading the data, please send an email to osamuels@math.uio.no.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you understand the content you have handed in, we may request that you give an oral account.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics, e-mail: studieinfo@math.uio.no) well before the deadline.

All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1

In this problem you will work on data from a study of horseshoe crabs on an island in the Gulf of Mexico. During spawning season, a female migrates to the shore to breed. With a male attached to her posterior spine she burrows into the sand and lays clusters of eggs. The eggs are fertilized externally, in the sand beneath the pair. During spawning, other male crabs may cluster around the pair and may also fertilize the eggs. These male crabs are called *satellites*.
The response outcome for each of the $n = 173$ female crabs is a binary indicator y of whether one or more satellites were present. Explanatory variables are the female crab's color, spine condition, weight and carapace width. The data set is available in the file **crabs.txt** on the course web page. The variables in the file are

y	Indicator for one or more satellites (0=no, 1=yes)
width	width of carapace of female crab (in cm)
weight	weight of female crab (in kg)
color	color of female crab (1=medium light, 2=medium, 3=medium dark, 4=dark)
spine	conditions of spine (1=both good, 2=one worn or broke, 3=both broken)

a) Choose a suitable regression model for studying how the probability of presence of satellites depends on the explanatory variable width. Give the reasons for your choice of regression model.
b) In particular find the odds ratio of presences of satellites between crabs that differ one cm in width, and explain what this odds ratio means. Can the odds ratio be considered as an approximation to a relative risk in this situation? Also find a confidence interval for the odds ratio and determine whether width influences presence of satellites significantly.
c) Then consider the other explanatory variables weight, color and spine as covariates, one at a time. Discuss whether these covariates should be included as categorical or numerical. Determine which variables has a significant influence on the presence of satellites.
d) Next use all variables in the regression (as main effects), and describe your findings. Try to simplify the model only using the significant covariates. In particular discuss the covariates weight and width.
e) Finally investigate whether there are interactions between covariates.

Problem 2

Facts or not? It is often claimed that participants from larger and wealthier nations are more likely to win medals in competitions like the Olympic games. In the file **olympics.txt** on the course web page you find the total number of medals each nation won under the Olympic games in Sydney in the year 2000. Only the 66 nations that won at least one medal in both 2000 and 1996 (in Atlanta) are considered here. In addition to the number of medals for each nation in 2000, the file contains information on

Total1996	Number of medals won by the nation in the previous game
Log.population	Logarithm of the nation's population size per 1000
Log.athletes	Logarithm of the number of athletes representing the nation
GDP.per.cap	The per capita Gross Domestic Product of the nation

a) Since the outcome total medals in 2000 is a count variable it might be reasonable to analyze the data by Poisson regression. Present and explain such a model. Often Poisson regression models include offset terms. Explain why Log.athletes is a sensible choice for an offset. In the following analyses you should include Log.athletes as offset.
b) Fit a model for the rate of medals won per athlete, using (possibly only some of) the predictors above. Explain how you arrived at your final model. Give an interpretation of the results and write a summary of your findings. Do you confirm the statement above?

Problem 3

488 patients with liver cirrhosis at various hospitals in Copenhagen were included in a randomized clinical trial with several years of follow up. The purpose of the study was to investigate whether patients treated with the hormone prednisone had better survival than patients who got an inactive placebo treatment. 251 of the patients received prednisone while 237 received placebo.
At the course web page you find the data set **cirrhosis.txt** containing the data for the study. The data are organised with one line for each of the 488 patients who took part in the study, and with the following variables in the seven columns:

status	Indicator for death/censoring (1=dead; 0=censored)
time	Time in days from start of treatment to death/censoring
treat	Treatment (0=prednisone; 1=placebo)
sex	Gender (0=female; 1=male)
asc	Ascites at start of treatment (0=none; 1=slight; 2=marked)
age	Age in years at start of treatment
agegr	Age group (1=<50; 2=50-65; 3=>65)

You are supposed to study the effect of treatment with prednisone, sex, age, and ascites (excess fluid in the abdomen).
a) Make Kaplan-Meier plots for the survival function for each level of the covariates treatment, sex, ascites, and grouped age (so 4 plots in total). 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) Then do multiple Cox regression where the effects of all the covariates are studied simultaneously. Use age in years (not grouped). Summarize (and interpret) your findings. For this 'full' model with all covariates as main effects, find a 95% confidence interval for the hazard ratio for men versus women when all other covariates are constant. Write a conclusion about the effect of prednisone in this trial.