23rd March 2026

STK4900/9900 - Statistical Methods and Applications

Mandatory assignment 2 of 2

Submission deadline

Monday 13th APRIL 2026, 14:30 in Canvas (canvas.uio.no).

Instructions

The assignment must be submitted as a single PDF file. If you are scanning handwritten pages, they must be clearly legible. The submission must contain your name, course and assignment number. Note that you have **one attempt** to pass the assignment. This means that there are no second attempts. All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

For the report, it is expected that you give a clear presentation with all necessary explanations. In particular, in addition to presenting all relevant numerical results and plots from an analysis, you should also discuss the conclusions you make based on them. Note also that you need to include the code used to run the analysis, either as part of the main text of the report or as an appendix (you should only include the final code, not all trial and errors). You may use any software package of your own choice, as long as you are able to answer the questions.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject.

The relevant data files are available on the course web page. If you have problems accessing the data, please send an email to johanpen@math.uio.no.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, see the part about extensions in the following page:

https://www.uio.no/english/studies/examinations/compulsory-activities/mn-math-mandatory.html

which contains general guidelines for mandatory assignments at the Department of Mathematics.

Note: To pass the assignment, the presented solutions do not have to be entirely correct, but you are expected to make a genuine effort at solving the problems (which should also be reflected by the submitted report).

Problem 1. Horseshoe Crabs

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 on the course page through the link:

https://www.uio.no/studier/emner/matnat/math/STK4900/data/crabs.txt

Each row (except the first) represents an observation of measurements over the following variables:

`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`: condition of spines (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) 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 a good approximation of the relative risk in this situation? Also find a confidence interval for the odds ratio and determine whether width influences presence of satellites significantly.

(c) 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) Use all variables in the regression (as main effects), and describe your findings. Try to simplify the model by using only the significant covariates. In particular, discuss the covariates weight and width.

(e) Investigate whether there are interactions between covariates.

Problem 2. Olympic games - 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. We are going to study a data set that contains the total number of medals won by each nation under the Olympic games in Sydney in the year 2000, and it is available on the course web page through the following link:

https://www.uio.no/studier/emner/matnat/math/STK4900/data/olympic.txt

Only the 66 nations that won at least one medal in both 2000 and 1996 (in Atlanta) are considered here. We are going to consider the following variables:

`Total2000`: number of medals won by the nation in 2000,

`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`: per capita gross domestic product of the nation.

Since the outcome Total2000 is a count variable, we are going to analyze the data by a Poisson regression model.

(a) Briefly present and explain the Poisson regression model. Explain why Log.athletes is a sensible choice for a so-called offset in this case.

(b) Including Log.athletes as an offset, 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. Is the claim stated in the beginning of the problem supported by the data?

Problem 3. Liver cirrhosis

In a randomized clinical trial, 488 patients with liver cirrhosis at various hospitals in Copenhagen were followed over several years. 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. Out if the 488 patients, 251 received prednisone while 237 received placebo. At the course web page you find the data set through the link:

https://www.uio.no/studier/emner/matnat/math/STK4900/data/cirrhosis.txt

Each row (except the first) represents a patient taking part in the study, and the columns represent the following variables:

`status`: indicator for death/censoring (0: censored; 1: dead),

`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 (that is, 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) Perform 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.

Good luck!