Seminar 3

Start Assignment

- Due 10 Nov by 23:59
- Points 4
- Submitting a file upload
- File types doc, docx, pdf, and odf

General instructions

Written solutions to all the tasks must be submitted before the deadline [DATE-TIME SEMINAR 3: Nov 10, 23:59] (as pdf files in Canvas, preferably including the most important parts of your code).

You are also expected to prepare an oral presentation of your solution for each of the tasks. The presentation should aim at taking 15 minutes, in order to leave room for questions and discussion. At the seminar, a member of your group (randomly chosen by the teacher) will be asked to present the solution.

We expect each of you to spend around 15h on these tasks; please plan your time and meetings keeping this in mind.

Please note that the course book (and the lectures) are not your only sources of information. There are lots of information available on the Internet about statistics as well as R and its possibilities, and you are very much encouraged to try and find methods not mentioned in the book or the lectures.

Background: colorectal cancer

Colorectal cancer (CRC), one of the most deadly cancer forms, has been projected to increase in incidence in the coming decades. Further, an increase has been observed in the incidence of CRC in younger adults.

Treatment success and cure of CRC is highly dependent upon early detection. However, early detection remains challenging due to inability to distinguish symptoms from other disease (e.g., gastrointestinal disease). CRC has a high rate of reoccurence and emergence of treatment resistance.

You are part of a team of clinical data modellers in the theme of CRC. It is your task to analyse and interpret generated clinical study data.

Task 1: detecting colorectal cancer (CRC)

A routine screening tool for CRC could aid early detection of disease and potentially lead to improved treatment success. A study has been carried out in 710 individuals to develop a screening method for colorectal cancer based on routine bio-analysis data from primary care.

The dataset (available in data_task1.csv (https://canvas.kth.se/courses/49059/files/8405575?wrap=1) ⤓ (https://canvas.kth.se/courses/49059/files/8405575/download?download_frd=1)) contains a classifier of CRC (termed dependent variable, 'DV', in the dataset) and a number of biomarkers explained in the table below.

Your task is to:

- develop a model for diagnosing CRC based on the collected information.
- diagnose model performance.
- explore what aspects could inform a threshold value for the model.
- what do you need to consider for the viability of implementing the model in clinical practice.

Table Task 1. Variable explanations.

Variable [unit]	**Explanation**
DV	Dependent variable, colorectal cancer classifier. 1: yes, 0: no.
ALT [U/L]	Alanine transaminase
AST [U/L]	Aspartate transaminase
GGT [U/L]	γ-glutamyltransferase
TC [mmol/L]	Total cholesterol
TG [mmol/L]	Triglycerides
HDL [mmol/L]	High-density lipoprotein
LDL [mmol/L]	Low-density lipoprotein
CRP [mg/L]	hs-CRP: high-sensitivity C-reactive protein
APOA1 [g/L]	ApoA1: apolipoprotein A1
LPA [g/L]	Lp(a): lipoprotein A
CEA [ng/mL]	Carcinoembryonic antigen
WBC [10^9/L]	White blood cells
RBC [10^{12}/L]	Red blood cells
NEU [10^9/L]	Neutrophils
LYM[10^9/L]	Lymphocytes
MONO [10^9/L]	Monocytes
HGB [g/L]	Haemoglobin

PLT [10^9/L]	Platelets

Task 2: colorectal cancer staging

CRC prevalence has increased in younger adults and is expected to continue to increase in the coming years.

To further investigate causes of late stage CRC detection, a study was carried out in 200 individuals to assess the impact of age (in years), ethnicity, living status (partnered, 1, or alone, 0), and site of detected cancer (right colon, left colon and rectum) and how these impact CRC stage (I-IV) at the time of diagnosis. The dataset is available in **data_task2.csv** **(https://canvas.kth.se/courses/49059/files/8405576?wrap=1)** ⤓ **(https://canvas.kth.se/courses/49059/files/8405576/download?download_frd=1)** .

- Analyse the dataset for any trends in CRC stage.
- What recommendations would you propose based on the data analysis?

Task 3: clinical study on drug exposure

CRC is commonly treated with surgery followed by chemotherapy. Previous studies have observed that women experience toxicity to higher degree than men during chemotherapy.

A clinical study was carried out to assess the concentration-time profiles of the drug fluorouracil in a patient group consisting of 45 participants receiving a fixed intravenous bolus dose of 100 mg the drug.

The data file, **data_task3.csv (https://canvas.kth.se/courses/49059/files/8405577?wrap=1)** ⤓ **(https://canvas.kth.se/courses/49059/files/8405577/download?download_frd=1)** , contains the following information: individual concentrations (DV; mg/L), time of observation (Time; hours), Dose (Dose; mg), body weight (BW; kg), sex (SEX; 0 - female and 1 - male) and age (AGE; years).

- Develop a model of the log-transformed concentration over time.
- Investigate if there are any effectors on (variables that influence) drug exposure.
- What conclusions can you draw?

Task 4: clinical study of an experimental treatment vs. control

A clinical study was carried out to investigate the impact of a novel experimental treatment on overall survival as compared to the gold-standard control treatment.

In the data file, **data_task4.csv (https://canvas.kth.se/courses/49059/files/8405578?wrap=1)** ⤓ **(https://canvas.kth.se/courses/49059/files/8405578/download?download_frd=1)** , you will find survival data (indicated by the variable status where 1 indicates death at a given Time, Months) for the control and treatment arms (Group variable).

Carry out analysis of the efficacy of the experimental treatment vs. control.