5th September, 2022

STK-IN4300 / STK-IN9300 Statistical Learning Methods in Data Science

Mandatory assignment 1 of 2

Submission deadline

Thursday 22nd September 2022, 14:30 at Canvas.

Instructions

The assignment must be submitted as a single PDF file. The submission must contain your name, course and assignment number.

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 have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

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!

Presentation of the data

It is really important that a statistical / data science analysis is not only performed correctly, but it is also clearly reported. The first step consists in describing the available data in a correct and exhaustive way. In this assignment you are asked to work on this aspect.

To perform this assignment, consider a dataset of your choice, with the only requirement that it contains ≥ 5 explanatory variables (covariates), with at least one categorical and one continuous. Suitable datasets can be found in the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets.php`), but it is not necessary to use a dataset from there.

The report must be written using R Markdown (`https://rmarkdown.rstudio.com/index.html`): for those who prefer coding in Python, the suggestion is to use `reticulate` (`https://cran.r-project.org/web/packages/reticulate/vignettes/r_markdown.html`) within R Markdown, but any alternative that produces reports in a comparable way will be accepted. Remember to include all the code in the report.

A quick introduction to R Markdown can be found at `https://rmarkdown.rstudio.com/articles_intro.html`.

Problem 1. Table

Summarize all information about the data in a table. A good example is the following table by Chan et al (2018):

Continuous (N = 49)	Mean (SD)	Median (IQR)	Min to Max
Age (years)	41.4 (11.6)	40 (32–54)	22–64
BMI (kg/m^2)	23.3 (2.9)	23.0 (21.6–24.0)	18.1–33.8
Categorical		**N/49[a] (%)**	
Female sex		36 (74)	
Regular exercise		30/48 (63)	
Smoking status			
Never-smoker		48 (98)	
Specialty			
Cardiology		12 (25)	
Endocrinology		12 (25)	
Neurology		25 (51)	
Occupational group			
Allied health		6 (12)	
Junior doctor		5 (10)	
Senior doctor		11 (22)	
Nurse		20 (41)	
Other		7 (14)	

BMI: body mass index; IQR: interquartile range; Max: maximum; Min: minimum; SD: standard deviation.
[a]Unless stated.

Problem 2. Bad Figures
While tables summarize very well the information, it is often more effective to present the result (including presenting the data) through figures. Data visualization, to this extent, is an important part of data science. Bad plots/graphs may be not only useless, but in the worst cases may also provide misleading information.
For at least a categorical and a continuous variable, provide a bad plot and describe in details its weakness(es), explaining why it should not be like you made it. It is up to you if you want to plot information about the variables alone or in relationship with the outcome.

Problem 3. Good Figures
Provide a good version of the plots you made in the previous point. One of the many resources about data visualization can be found at `https://clauswilke.com/dataviz`

References
Chan, L., McNaughton, H., & Weatherall, M. (2018). Are physical activity levels of health care professionals consistent with activity guidelines? A prospective cohort study in New Zealand. *JRSM Cardiovascular Disease, 7, 204800401774901*5.