Statistics for Data Science, Homework #1

Veronica Vinciotti, Marco Chierici

Released: 19/03/2025. Due: 02/04/2025

You should submit a single PDF file of the homework via Moodle, with the PDF rendered directly from a Quarto or RMarkdown source file (using `output: pdf_document`; see Guidelines) and not converted from any other output format.

You should write your report like a mini scientific paper. In particular, you should: introduce the analysis, discuss/justify the choices that you make, provide comments on the results that you obtain and draw some conclusions.

Guidelines

- Show only what is relevant to the analysis and the results that you obtain:
  - Include the *main code* for reproducing the analysis, but do not expect the reader to actually read the code for understanding what methods you have used (and why);
  - Visualize/summarize the results with a *selection* of informative output, tables and figures.
- Note that by default Quarto/RMarkdown will repeat (echo) the R code of a code chunk in the final output. If you want to display only the output of a code chunk without echoing the corresponding code, use the option `echo=FALSE` in the code chunk. For example, in RMarkdown,

~~~
```{r, echo=FALSE}
my_function <- function(a, b) {
    # ...
    # ...
}

glm.fit <- glm(Output ~ ., data=dataf, family=binomial)
# ...
```
~~~

In Quarto, you can use the YAML-style chunk options identified by `#|` at the beginning of the line:

~~~
```{r}
#| echo: false

glm.fit <- glm(Output ~ ., data=dataf, family=binomial)
# ...
```
~~~

- To add a plot, just insert the plotting code in a code chunk: then, Quarto/RMarkdown will output the resulting plot below the corresponding chunk. You can use code chunk options `fig.width` and `fig.height` to customize the size of plots, `fig.align` to change the horizontal alignment of a plot, and `fig.cap` to add a caption:

~~~
```{r, fig.width=10, fig.height=4, fig.cap="Figure 1: your caption here.", fig.align="center"}
...
plot(...)
...
```
~~~

- Specify PDF as the output file for rendering. With Quarto:

```
---
title: <title>
date: <date>
author: <name and ID>
format:
    pdf:
        latex_engine: xelatex
---
```

With RMarkdown:

```
---
title: <title>
date: <date>
author: <name and ID>
output:
    pdf_document:
        latex_engine: xelatex
---
```

- Include the following code chunk at the beginning of your Quarto/RMarkdown file, and adjust `width.cutoff` to avoid that long lines of code go beyond the margins in the output file (typical values are 50, 60, 70, 80):

~~~
```{r setup, include=FALSE}
knitr::opts_chunk$set(warning=FALSE,
                      message=FALSE,
                      tidy.opts=list(width.cutoff = 80),
                      tidy = TRUE)
```
~~~

Markdown cheat sheet

Markdown is a combination of regular text (like this), combined with tags that change the way the text is formatted. Here are the most common formatting tags:

- To make headers, use one or more # (pound) symbols at the beginning of the line: the number of # dictates the level of the header;
- To make text *italic*, wrap the text between *;
- To make text **bold**, wrap the text between **;
- To make numbered lists, just use a number followed by a dot (`1.`) at the beginning of each line:
  1. First element
  2. Second element
  3. Third element
- To make unordered lists, use a dash (`-`) or an asterisk (*) followed by a space at the beginning of each line:
  - item
  - item
  - item

Additional resources

- Ten simple (empirical) rules for writing science, Cody J. Weinberger, James A. Evans, and Stefano Allesina, PLOS Computational Biology 11(4): e1004205, 2015.
- RMarkdown intro and reference guide

Assignment

The data for this homework are available at `chd.csv`. The data come from a cardiovascular study conducted in the US, investigating the possible risk factors associated to the development of coronary heart disease (CHD) within 10 years.

The possible risk factors, that are available in the data, are the patients' sex (`sex`), age (`age`), education level (`education`, coded as 1: no high school degree, 2: high school graduate, 3: college graduate, 4: post-college), current smoking status (`smoker`), number of cigarettes per day (`cpd`, continuous), previous occurrence of strokes (`stroke`) or hypertension (`HTN`), presence of diabetes (`diabetes`), cholesterol levels (`chol`, continuous), diastolic blood pressure (`DBP`, continuous), body mass index (`BMI`, continuous), and heart rate (`HR`, continuous). The variable `CHD` records whether the patient developed CHD in the next 10 years or not.

In your report, you should:

- Explore the data: what is the nature of the response variable (`CHD`)? Are there potential issues with any of the predictors? Can you find some useful visualization of the discriminative power of each predictor? Do you need pre-processing or can you proceed with the data set as it is?
- Split the data into (reproducible) training and test sets. Given the class imbalance, you could aim for sets that have the same imbalance with respect to the outcome variable. In order to do this, you could either perform the splitting manually on each class, or use dedicated functions (for example, `caret::createDataPartition(labels, p=train_size)`, with `train_size` a number between 0 and 1 representing the percentage of data you would like to use for training.
- Fit the following GLM model:

$$\text{logit}(\text{E}(\text{CHD})) = \beta_0 + \beta_1\text{sex} + \beta_2\text{age} + \beta_3\text{education} + \ldots + \beta_{12}\text{HR}$$

  Discuss the `summary` and the interpretation of the regression coefficients in the context of the study.
- Fit a k-nn classifier, by performing a careful selection of the tuning parameter k.
- Evaluate the performance of the two methods.
- Draw some overall conclusions about which method may be more suitable in answering the question of interest and discuss possible limitations of the study that you have conducted which may have had an impact on these conclusions.

Please note that the **maximum allowed number of pages is 8**.