Report 1

Emily Bellis

This report is due on 9/19/22. Bring a hard copy of the .Rmd code to class as well as the rendered report for the peer code review.

Introduction

Choose a dataset from an external source (i.e. NOT from the the textbook or included with base R) that allows you to ask a compelling research question in the context of the statistical learning techniques presented in class. I recommend a dataset with a minimum of 100 observations and at least 5 continuous predictors (more is even better). If you don't know where to start, Kaggle is a good source of publicly available datasets, and I also like a lot of the datasets from Tidy Tuesday.

This project is expected to be the independent work of each individual student and you should not use the same dataset as anyone else. You can share the dataset you are using with other members in the class by posting to the Blackboard discussion board. Spend some time thinking carefully about this! At the end of the course, you will ideally end up with a very nice project report you can show off in a portfolio or that you can build on for a graduate research project.

For Report 1, write an introduction section (no more than a short paragraph) that briefly describes your research question. Include a citation to the dataset itself.

Methods

For Report 1, provide a short Methods section that describes your dataset. Be sure to indicate the number of samples, the number of predictors, and to describe briefly the response variable and predictors. Do you plan to analyze this dataset as a regression or classification problem?

Include all code you used to answer the above questions in the .Rmd file. By default, **knitr** displays all possible output from a code chunk, including the source code, text output, messages, warnings, and plots, but you can hide them individually. For example use `echo = F` to hide source code or `results = F` to hide text output.

Don't forget to cite R R Core Team (2020)! Check out the R Markdown Cookbook for more info on in-line citations(https://bookdown.org/yihui/rmarkdown-cookbook/write-bib.html)

We also used Wickham (2016).

Results

For Report 1, provide ONE visual representation of your dataset. This can be a figure demonstrating the distribution of your response variable, scatterplot of pairwise relationships among variables, etc. The data should be clearly presented, including labels for x- and y-axes and any necessary legends.

Give at least one sentence to interpret your figure and what it tells us about your dataset.

References

By default these are provided at the end in the rendered document for anything you cited in-line.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.