HW02 - Simple Regression Analysis

Stat 159, Fall 2016, Prof. Sanchez

Simple Regression Analysis

So far we have been focused on introducing and learning the basic tools typically used in computational reproducible workflows (e.g. bash, git, github, Make, markdown, pandoc, and some text editor). However, we haven't done any statistical data analysis... yet.

The purpose of this assignment is to give you the opportunity to start applying the computational toolkit (plus R) to reproduce a simple regression analysis. More specifically, the idea is to reproduce the analysis from Section 3.1 (pages 59 to 71), of *Chapter 3. Linear Regression*, from the book "An Introduction to Statistical Learning" (by James et al):

http://www-bcf.usc.edu/~gareth/ISL/data.html

The data set is in the `Advertising.csv` file available here:

http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv

The analysis involves carrying out a simple linear regression of `TV` advertising on `Sales` of a particular product. The overall idea is to write a report in which you are able to replicate the following results:

- Figure 3.1 (page 62) Scatterplot with fitted regression line (the vertical distances of each point to the line are optional).
- Table 3.1 (page 68) Summary of regression coefficients.
- Table 3.2 (page 69) Quality indices RSE, R^2, and F-statistic.

Mindset: To generate the report, you should use an R markdown file (`.Rmd`). Because you are going to be using a `Makefile` to automate all the workflow, you may need to learn how to run R from the command line: how to run R scripts, how to render Rmd files into PDF or HTML format, etc. I've created a tutorial on how to run R in non-interactive mode:

https://github.com/gastonstat/tutorial-R-noninteractive

The way you are going to work with the `.Rmd` file is a bit different from the usual way you've been using them so far. Instead of simply writing all the narrative and code in the `.Rmd` file, you will use this file mainly to write the narrative. The R code for most of the analysis will be written in separate `.R` script files, with the main outputs being generated outside the `.Rmd` file.

File Structure

The file-structure for this assignment is the following:

```
stat159-fall2016-hw02/
    .gitignore
    README.md
    Makefile
    data/
      README.md
      Advertising.csv
      eda-output.txt
      regression.RData
    code/
      README.md
      eda-script.R
      regression-script.R
    images/
        histogram-sales.png
        histogram-sales.pdf
        histogram-tv.png
        histogram-tv.pdf
        scatterplot-tv-sales.png
        scatterplot-tv-sales.pdf
    report/
        report.Rmd
        report.pdf
```

Makefile targets

Your `Makefile` should have the following three Phony targets:

- `all` will be associated to the production of `report.pdf`, `eda-output.txt` and `regression.RData`
- `data` will download the file `Advertising.csv` to the `data/` folder
- `clean` will delete the generated report (pdf and/or html)

In addition to the phony targets, `Makefile` should have targets that allow you to generate the following files:

- `report.pdf` which depends on `report.Rmd`, `regression.RData`, and `scatterplot-tv-sales.png`
- `regression.RData` which depends on `regression-script.R` and `Advertising.csv`
- `eda-output.txt` which depends on `eda-script.R` and `Advertising.csv`

Files

- Code scripts:
  - `eda-script.R` reads in the `Advertising.csv` data set, and computes summary statistics and histograms of `TV` and `Sales`. The summary statistics should be clearly labeled, and will be saved in a file `eda-output.txt`. The charts are saved in both PNG and PDF formats.
  - `regression-script.R` reads in the `Advertising.csv` data set, and computes a `"regression"` object—via `lm()`—as well as the summary of such regression object—via `summary()`. This script also produces the scatterplot with the regression line. The R objects from the regression analysis are saved in the file `regession.RData`. In turn, the scatterplot is saved in both PNG and PDF formats.
- Data Files:
  - `Advertising.csv` is the main data set.
  - `eda-output.txt` is a text file containing the summary statistics of `TV` and `Sales`. This file should be produced via `sink()` from the `eda-script.R` file.
  - `regression.RData` is an R's binary format file containing the regression objects obtained when running `regression-script.R`. This file should be produced via `save()` from the `regression-script.R` file.
- Image files:
  - `histogram-tv.png` and `histogram-tv.pdf` contain the histogram for `TV`. These files are an output of `eda-script.R`.
  - `histogram-sales.png` and `histogram-sales.pdf` contain the histogram for `Sales`. These files are an output of `eda-script.R`.
  - `scatterplot-tv-sales.png` and `scatterplot-tv-sales.pdf` contain the chart of the scatterplot between `TV` and `Sales`, with the fitted regression line. These files are an output of `regression-script.R`.
- Report files:
  - `report.Rmd` is the source Rmd document used to generate the pdf report. It reads in the objects of `regression.RData`, produces the tables, and includes the image in `scatterplot-tv-sales.png`.
  - `report.pdf` is the generated pdf file from the Rmd document. Alternatively you could have a `report.html` file. Or even better, you can try to have both types of output: pdf and html. It is not mandatory to have both types of files.
- README files:
  - Your project should include one readme file at the top level. In addition, folders `data/` and `code/` should also contain their own readme files with a brief description of the files in such directories.

Report

You should write a PDF report (HTML is fine if you don't have LaTeX) in the form of a paper with the following sections:

- Abstract
- Introduction
- Data
- Methodology
- Results
- Conclusions

Make sure all the images and tables have captions.

A minimalist sample of a report is in HW02's folder (in the github repo): see file `report.pdf`. You report must be longer.

Grading

In addition to checking whether you meet all the listed project requirements, for this and subsequente assignments, we will evaluate the following core competencies of your reports:

- Computation: Perform computations correctly.
- Analysis: Carry out analysis appropriate for data and context.
- Synthesis: Identify key features of the analysis, and interpret results.
- Visual presentation: communicate findings graphically clearly.
- Verbal: communicate findings in writing clearly, precisely and concisely.

What you need to "turn in" is basically the **public** github repository of this project—don't use a private repo (keep in mind *Open Science*).