

## Lab 3: Multiple Linear Regression in practice

Create a short reproducible document (using RMarkdown) that describes the basic structure of a dataset and summarizes some key features of the data using a few key tables and figures. Choose a dataset that you have not used before from [these datasets](#), [these datasets](#), the ones in the class Google Drive, or some other dataset that interests you. *Be sure to pick a dataset that has a continuous variable that you can use as an outcome variable in a linear regression model.* Your write-up should address each of the following.

The report should be less than 6 pages, including all figures, and should be submitted as both PDF and Rmd formats. You do not need to show your code in the PDF report.

### Data description

- What is the background/context for this data?
- Data management: How many observations are there? What is the unit of observation? What are the key response variable(s) and explanatory variables? Is there any missing data? If so, are there any obvious patterns to the missingness?
- Choose 4 to 10 key variables from your dataset (including the outcome variable).
  - Include a table that lists for each chosen variable the name, definition, type of variable (i.e. categorical, continuous, binary), and the number of missing observations.
  - Choose at least two of these variables and provide figures that show their univariate distributions. Describe the plotted distributions in words.
  - Provide a pairs plot that provides a visual overview of the chosen variables.

### Simple Linear Regression

Run two simple linear regressions each with different predictor variables. Interpret the results. Plot a scatterplot of each regression and include the fitted line on the graph. Rescale your explanatory variables if necessary to obtain a meaningful interpretation of  $\beta_0$ .

### Multiple Linear Regression

Build 1 or 2 multiple linear regression models. If appropriate, use dummy variables to model categorical predictors. Interpret some of the MLR model coefficients in the context of your particular dataset.