STA321 Week #3 Assignment

Due: 11:30 PM, Sunday

Contents

1 Introduction

This week's assignment has two components.

- Finding a data set for **project #1**. The detailed requirements of the data will be described in the next section.
- Fitting a simple linear regression (SLR) by selecting a numerical explanatory variable from the data set using the least square approach and then constructing 95% bootstrap confidence intervals for the regression coefficients.

2 Data Requirements and Sources

The selected data set will be used for multiple linear regression analysis and bootstrap analysis as well. In order to implement the commonly used regression techniques, the desired data must meet some requirements. You also have the flexibility to choose a data set that you are interested in so you can easily formulate the analytic questions and tell a better story from the analytic results.

2.1 Data set requirements

The desired data set must have

- the response variable **must** be continuous random variables.
- at least two categorical explanatory variables.
- at least one of the categorical variables has more than two categories.
- at least two numerical explanatory variables.
- at least 15 observations are required for estimating each regression coefficient. For example, if your final linear model has 11 variables (including dummy variables), you need $12 \times 15 = 180$ observations.

2.2 Data Sources

The following websites contain many links to sites that have different types of data sets (some of the links may not be active).

- 10 open data sets for linear regression https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/
- UFL Larry Winner's Teaching Data Sets http://users.stat.ufl.edu/~winner/datasets.html
- The suggested data repository for this class http://stat321.s3.amazonaws.com/w00-datasets.html
- Datasets for Teaching (Univ. Sheffield, UK) https://www.sheffield.ac.uk/mash/statistics/datasets
- Data.World https://data.world/datasets/regression

2.3 Post Your Selected Data on D2L

Before you start searching your data set, check the D2L discussion board and make sure you will not select the data set your classmates have already chosen for their project. After you identify your data set, please post your data set name and the link to that data set.

3 This Week's Data Analysis

Please prepare an RMarkdown document to include the following two parts. Please start your work earlier to

3.1 Description of the Data Set

Write an essay to describe the data set. The following information is expected to be included in this description.

- How the data was collected?
- List of all variables: names and their variable types.
- What are your practical and analytic questions
- Does the data set have enough information to answer the questions

3.2 Simple Linear Regression

Make a pair-wise scatter plot of all variables in your selected data set and choose an explanatory variable that is linearly correlated to the response variable.

- Make a pairwise scatter plot and comment on the relationship between the response and explanatory variables.
  - If there is a non-linear pattern, can you perform a transformation of one of the variables so that the transformed variable and the other original variable have a linear pattern?
  - If you have a choice to transform either the response variable or the explanatory variable, what is your choice and why?
- Fit an ordinary least square regression (SLR) to capture the linear relationship between the two variables. If you transformed one of the variables to achieve the linear relationship, then use the transformed variable in the model. and then perform the model diagnostics. Comment on the residual plots and point out the violations to the model assumptions.
- Using the bootstrap algorithm on the **previous final linear regression model** to estimate the bootstrap confidence intervals of regression coefficients (using 95% confidence level).
- compare the p-values and bootstrap confidence intervals of corresponding regression coefficients of the final linear regression model, make a recommendation on which inferential result to be reported, and justify.