Regression Analysis Final Project Guidelines

The goal of the final project is to apply what you have learned in this course to conduct a statistical analysis. It should be an in-depth regression or logistic regression analysis of a question that interests you. This question may come from your personal interests, your career interests, etc.

The final project should be individual or in a group of 2 students. You can form groups yourselves. I will randomly form pairs if you would like to work with another student and are unable to find one. Edit the following Google Sheet to indicate your preference:

<https://docs.google.com/spreadsheets/d/1HiEa84ZbtzFVjZB62ZpBKB_A_ZEDq_kVLJDEqzAk06s/edit?usp=sharing>

# 

# Summary of Deliverables

* Project Proposal: Due 12/01
* Final Project: Due 12/16

# Data

Your dataset can come from

* [data.gov - The home of the U.S. Government’s open data](http://www.data.gov).
* [Google dataset search](https://datasetsearch.research.google.com/)
* BLS data: <https://www.bls.gov/data/>
* Other reliable data sources such as large firms, universities, or non-profit organizations

In order for you to have the greatest chance of success with this project, it is important that you choose a manageable dataset. This means that the data should be readily accessible and large enough that multiple main effects and interactions may be explored for your model. As such, your dataset should have at least 100 observations and at least 5 variables.

# Project Proposal

## Components

This is a draft of the introduction section of your project as well as a regression analysis plan and your dataset. The proposal should be no more than 4 pages, 1.5 line spacing (including figures and tables, which can be single-spaced).

The proposal should include the following:

* You will introduce the research question you wish to explore. This includes the motivation for the question (citing any relevant literature), and your hypotheses regarding your question of interest. The research question can take the following format:
  + What is the effect of X on Y?
  + What are the most important factors for Y?
  + Does X have a significant association with Y?
  + What is the best linear model to predict Y?
* You will describe your data set, including the data source. This section will include
  + Description of the response (Y) variable and its variable type
  + Description of the predictor variables and their meaning
  + Preliminary exploratory data analysis, including some summary statistics and visualizations.
  + Description of the regression methods you plan to use for the analysis (this may be updated later as you conduct the analysis)

# Final Project Report

The final report should be no more than 15 pages, 1.5 line spacing (including figures and tables, which can be single-spaced).

The goal of the write up is to demonstrate your ability to ask meaningful questions and answer them with the results from regression analysis, that you are proficient in using Python, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions. You do NOT have to apply every statistical procedure we learned. Pay attention to your writing and the presentation of figures. Neatness, coherency, and clarity will count. At a minimum, your write up should have the following sections:

* Section 1: Introduction
  + This is basically a revised version of what is in the project proposal. This should include your research question, hypotheses, and a description of the data. It can also include the exploratory data analysis.
* Section 2: Regression Analysis
  + This section includes the results of your final regression model. In addition to displaying the model output, you should include a brief description of why you chose that type of model and any interpretations/ interesting findings from the coefficients. You should also include a discussion of the model assumptions, variable selection, and model fit analysis.
* Section 3: Discussion & Limitations
  + This section should include any relevant predictions and/or conclusions drawn from the model. Also, critique your own methods and provide suggestions for improving your analysis. For example, a paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project can be included.
* Section 4: Conclusion
  + In this section, you should summarize your project and highlight any final points you wish the reader to get from the project.

# Tips

* Read [sample regression reports](https://www.dropbox.com/s/4zcwvxf8pjc21rf/Sample%20Regression%20Reports.zip?dl=0).
* In your report, round your numbers to **at most 4 decimal places**. Most of the time, 2-3 decimal places are sufficient.
* **Avoid using screenshots if possible**. Use proper tables and figures.
* Figures and tables need to have numbers and captions (e.g., Figure 1. Scatter plot of xxx and yyy).
* Number your pages.
* Use basic fonts and be consistent, such as Arial, Calibri, or Times New Roman.

*\* This guide is adapted from* [*https://www2.stat.duke.edu/courses/Spring19/sta210.001/project/project.html*](https://www2.stat.duke.edu/courses/Spring19/sta210.001/project/project.html)*.*