STAT GU4205 Project
10% Final Grade + EC
Due TBA

The STAT GU4205 Project is an open ended case study intended to be a capstone on the Linear Regression Models course.

1 Types of Projects

Students should choose a project that suits their personal interests and directly utilizes topics presented in Linear Regression Models (GU4205). Project categories:

I. Inferential procedures based on linear regression models

II. Predictive modeling based on linear regression and related techniques

III. Other

The above categories allow flexibility and creativity while constraining students to the linear regression setting. Please communicate with Prof. Young for approval if a student would like to choose a different category (**other**). Do not hesitate to contact Prof. Young for recommendations on any project type.

Students are required to work individually, i.e., this is **not** a group project. Students are allowed to communicate with their classmates but all work must be completed on their own. If students have similar topics, please set-up a meeting with Prof. Young for approval.

1.1 Inferential Procedures

The first choice fits most naturally to this class. For example, suppose that your project is focused on income disparity of males based on their race. More specifically, you might answer the research questions:

1. Do African American males have statistically different wages compared to Caucasian males?
2. Do African American males have statistically different wages compared to all other males?
3. Or similar..

There are several steps required to answer the above research question(s), including: data collection, exploratory analysis, choosing a model, diagnostics, potential remedial measures, and running the hypothesis testing procedure(s).

Other considerations to stimulate thought: What covariates should be included in your analysis so that the model adequately controls for important sources of variation? Are there any

functional forms (other than linear) that need to be included? Is your model additive or does it include interaction terms? Does your model suffer from any negative impacts of collinearity? Are influential observations impacting your hypothesis testing procedure? Can you rule out all or most confounding variables? Can you replicate your experiment on a new unseen dataset and achieve similar conclusions?

1.2 Predictive Modeling

The second choice requires students to build a highly predictive model using techniques introduced in GU4205. For example, suppose you would like to build a statistical model used to predict wages ($\hat{y}$) as a function of several features or covariates: age (x_1), race (x_2, x_3), work experience (x_4), ... etc. For substance, maybe compare the predictive performance of several candidate models on an un-observed test set. The candidate models might include different subsets of selected covariates and/or the use of different loss functions for estimation. In this case, your research question is focused on the predictive performance of your model and not focused on variable relationships.

Other considerations to stimulate thought: Should you use K-fold cross validation as a measure of model performance or choice of hyper parameters? What metrics should you use to perform variable selection, i.e., AIC, BIC, adjusted R^2, ... etc? Does you model suffer from overfitting? What is your baseline model for comparison? Are there any functional forms (other than linear) that need to be included? Is your model additive or does it include interaction terms? Are influential observations impacting your predictive performance? What loss-functions are you choosing and why?

1.3 Other Projects (Less Structured)

Students are allowed to choose a topic that doesn't fit the mold of **Inferential Procedures** and **Predictive Modeling**. These topics must be approved by Prof. Young and must heavily relate to the Linear Regression Models (GU4205) course. The following recommendations are not *"set in stone"* and are provided as a reference.

- Run an extensive simulation study investigating important properties of regression estimators and/or related testing procedures. Provide some real-word data examples to help synthesize your simulation study with empirical results.
- Investigate the *"replication crisis"* by attempting to reproduce one or several several published studies.
- Investigate the impact *p-hacking* has on the replication crisis. Show this through simulation and attempting to reproduce one or several several published studies.

- Study several *"multiple comparison procedures"* through simulation and real-world data. For example, how does the Bonferonni procedure compare to Dunnett's, Tukey, .. etc? In what situations should you use each procedure? What metrics are you basing your comparisons on?
- Compare the performance and utility of traditional parametric linear regression testing procedures versus non-parametric testing procedures. Study the performance through simulation and real-world data examples.
- Follow the **Inferential Procedures** recommendations using a Bayesian linear regression model to answer a research question of interest. Compare the traditional linear regression model versus the Bayesian model for reference. Note that estimating the Bayesian regression model requires MCMC and/or the use of other software such as `Stan`. Are the diagnostic tools the same in this setting?
- Many more...

Students should try to avoid choosing pure machine learning models such as neural networks, decision-trees, random forests, or similar because these topics are reserved for more advanced classes. Again, run any ideas by Prof. Young before committing.

2 Write up

Students are required to type a final report. The structure of the report is subjective. For reference, below describes a template for **Inferential Procedures**. Feel free to adjust your final report as you see fit.

I. **Introduction:** Include a brief description of the goals of your project coupled with some exploratory data analysis. Be creative with your exploratory analysis and only include items that you feel are informative.

II. **Data Collection and Data Description:** Explain your data collection process and briefly describe your data-set of interest, i.e., variables, sample size,.. etc.

III. **Statistical Model:** In this section, clearly state your final model along with the `R` summary output. Be sure to describe all interactions, functional forms and transformations of your model.

IV. **Research Question:** Perform the relevant testing procedures to answer your research question(s). Also include a brief written summary of your results.

V. **Appendix**

   a. **Model Selection:**

i. Here you will explain in detail what interactions, functional forms and variables you decided to include in your model. Describe if and why a transformation is applied to the response variable. Without overwhelming the instructor, include relevant `R` output and plots that helped you arrive at your final model.

b. **Diagnostics and Model Validation:**

i. Include all relevant diagnostic plots on the final model and any important diagnostic plots leading up to your final model.

ii. Include the computed $MSPR$ (test error) and compare this number to the computed MSE of your final model. This could be included in Section (III) instead.

iii. Include a section on influential observations. Describe any negative impacts of the influential observations and if remedial measures were required.

iv. **Anything you Feel Necessary:**

3 R Code (or Similar)

- Students should prepare an organized `R script` (or `Rmd`) file that complements the written report. This should have a `.R` or `.rmd` extension.
- **Do not** copy and paste the `R` code into your appendix. Only include very important `R` code, or no code, in your final report. Please upload the `R script` (or `.Rmd`) file on Canvas by the due date.
- You are allowed to use other programming languages for this project, i.e., `SAS`, `Python`, `MATLAB`, ... etc. Note that other statistical programming languages often use different parameterizations of common models. You are **not** allowed to use spreadsheet based software such as `Excel`, `Minitab` and `SPSS`.

4 Grading

- This project will be graded on:

i. Completeness (don't forget to turn in your `R` file also)

ii. Correctness

iii. Organization/neatness

iv. Creativity

- I want to see a nice organized final report. It must be typed with graphs labeled. Please do not make the report too long! No more than around 10 pages.
- Students who submit high quality work earn extra credit.

5 Websites for Finding Data

- Kaggle
- UCI Machine Learning Repository
- NYC Open Data
- London DATASTORE
- Search on your own..