This part of the mini-project has three main goals: (1) test your predictive models against your previously held out test set; (2) try your hand at inference on a linear regression model built from your data; and (3) summarizing your work over the course of the project.

*Important note: Your report for this part of the project should not exceed **5** pages.* In particular, your report should consist only of written text, regression output as appropriate, and any visualizations you think are helpful or interesting. **You are welcome to attach appendices containing code, extra plots, etc., but you will be graded on the first 5 pages you submit.**

As in the previous part, you will be judged on the overall quality of your analysis and presentation, in addition to the whether you satisfy the minimum required steps we outline below. In particular, out of 10 total points, 6 points will be based on whether you satisfy the minimum required steps. The remaining 4 points will be based on subjective overall quality, calibrated across teaching assistants to ensure consistency.

The report is due on Gradescope (**due 11:59 PM, December 6, 2019**).

1 Prediction on the test set

Take the best model you built for regression and for classification in Part 1 of the project, and apply it to the test set you held out. Note the test error in each case; how does it compare to the estimate for test error that you derived previously (in Part 1)?

2 Inference

In this part of the project, you will try out some of the methods we learned in the class for *inference*. You should pick either a linear regression model (for the regression task you studied) OR a logistic regression model (for the classification task you studied).

Here are they key steps we want you to carry out:

a) For your chosen model, look at which coefficients are significant in the regression output, according to R. In words, describe what statistical significance means for these coefficients. Do you believe the results? Why or why not?

b) Now fit your chosen model on the test data, and look at whether the same coefficients are significant. Did anything change in doing so? If so, explain any differences that you found, and reflect on why they might be there.

c) Use the bootstrap to estimate confidence intervals for each of your regression coefficients. Do the results differ from what R gave you in its standard regression output? Again, explain any differences that you found, and reflect on why they might be there.

d) If your chosen model does not include all covariates you had available, compare the significant coefficients in your chosen model to the significant coefficients in a model that includes all covariates Did the significant coefficients change? If so, explain the differences.

e) Comment on potential problems with your analysis, including collinearity, multiple hypothesis testing, or post-selection inference; be specific. For example: what evidence suggests to you that collinearity is impacting your results? For multiple hypothesis testing, suggest how applying the Bonferroni correction would change your interpretation of significant coefficients. For post-selection inference, suggest aspects of your model-building and inference process that might bias your determination of which coefficients are significant.

f) For the relationships that you found that are significant, would you be willing to interpret them as causal relationships? If not, why not? What other covariates do you think might be confounding your ability to infer causal relationships?

3 Discussion

Finally, we would like you to summarize your work on the project. This is open-ended, and so you should choose what you think are the most interesting or important things to report. The following high-level questions are meant to guide your thinking.

a) How would you expect your models to be used practically? Do you think they would primarily be used for prediction, for inference, or for both? What decisions do you think your models would guide, and what pitfalls do you see in using your models to make these decisions?

b) How well would your models hold up over time (i.e., how often do you think they should be refitted)? Why?

c) Are there choices you made in your data analysis, that you would want to make sure any one (e.g., a manager, a client, etc.) that uses your models is aware of? Examples here might include approaches to data cleaning; data transformations that you chose; vulnerability to overfitting, multiple hypothesis testing, or post-selection inference; etc.

d) If you could, how would you change the data collection process? In particular, are there reasonable covariates you would like to collect, that were not present in the data?

e) If you were to attack the same dataset again, what would you do differently?