Economics 104: Project 1
Winter 2023, UCLA
Due Date: June 26, 2023 by 5PM (PST)

For this project, you will work any dataset you like, however, it must contain at least 10 different predictors and one response variable which you will aim to predict. Your task will be to find a reasonable model by following the 11 steps outlined below.

As an illustration of a good dataset (you cannot use this dataset), the file `diamonnds.csv` contains the prices and other attributes of almost 54,000 diamonds. The data description and file can be accessed directly from kaggle and the goal is to predict *diamond prices* . There are many datasets that are publicaly available in kaggle but you can also get data from FRED, BLS, and so on.

1. Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.
2. Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates.
3. Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.
4. Use Mallows Cp for identifying which terms you will keep in the model (based on part 3) and also use the Boruta algorithm for variable selection. Based on the two results, determine which subset of predictors you will keep.
5. Test for multicollinearity using VIF on the model from (4) . Based on the test, remove any appropriate variables, and estimate a new regression model based on these findings.
6. For your model in part (5) plot the respective residuals vs. $\widehat{y}$ and comment on your results.
7. For your model in part (5) perform a RESET test and comment on your results.
8. For your model in part (5) test for heteroskedasticity and comment on your results. If you identify heteroskedasticy, make sure to account for it before moving on to (9).
9. Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison.
10. Evaluate your model performance (from 9) using cross-validation, and also by dividing your data into the traditional 2/3 training and 1/3 testing samples, to evaluate your out-of-sample performance. Comment on your results.
11. Provide a short (1 paragraph) summary of your overall conclusions/findings.