Economics 430: Project 1
Fall 2022, UCLA
Instructor: Dr. Rojas

Due Date: Oct 19, 2022

The document that you will submit, consists of a written report which includes the discussion of results (e.g., interpretation of coefficients), data description, motivation for the project (as well as the problem you are addressing), conclusion, references and respective Python source code. You only need to submit one project report per group but please make sure that every group member's name is included.

Identify a dataset of your choosing. Make sure it has at least 10 predictor variables.

1. Descriptive Analysis: Perform a univariate analysis following the steps below.
   (a) Begin by providing a descriptive analysis of your variables (include all predictors and response variable). This should include things like histograms, quantile plots, correlation plots, etc.
   (b) Estimate density plots for all your variables.
   (c) Identify if there are any non-linearities within your variables. What transformations should you perform to make them linear? What would happen if you included non-linear variables in your regression models without transforming them first?
   (d) Comment on any outliers and/or unusual features of your variables.
   (e) If you have any NAs, impute them using any of the methods discussed in class but make sure to justify your choice.
2. Variable Selection:
   (a) Using the Boruta Algorithm identify the top 2 predictors
   (b) Using Mallows C_p identify the top 2 predictors
3. Model Building: Explore several competing OLS models (based on part 2) and decide on one model only (with just one predictor). You will need to explain in detail how you arrived at your preferred model. Discuss the economic significance of your parameters, and overall findings. Make sure you discuss your main conclusions and recommendations.

   At a minimum. you need to include the following checks:

   - Evaluate transformations of variables
   - Look at Cook's distance Plot, Residuals Plot
   - Evaluate the robustness of your coefficient estimates by bootstrapping your model. Provide a histogram of the bootstrapped estimates, and comment on the findings.

- Use cross-validation to evaluate your model performance
- Evaluate your model's out of sample performance by splitting the data into testing and training sets, and predicting on the testing set