Economics 430: Project 1
Fall 2023, UCLA
Instructor: Dr. Rojas

Due Date: Oct 25, 2023

The document that y sts of a written report which includes the discussion of results (e.g., interpret data description, motivation for the project (as well as the problem you are on, references and respective Python source code. You only need to submit c group but please make sure that every group member's name is included. Pl oload your submission in PDF or HTML format only. Lastly, please review the "Assignment Policies" discussed in the course syallbus.

Identify a dataset of your choosing. Make sure it has at least 5 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) Discuss your findings from doing an exploratory analysis using Pandas Profiling. Did you discover anything new?

   (c) Estimate density distributions (e.g., Cullen & Frey) for all your variables, and show the plots with the respective fits.

   (d) 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?

   (e) Comment on any outliers and/or unusual features of your variables.

   (f) 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 (including R^2), and comment on the findings. In particular how do these estimates compare against your LS estimates?
- Use cross-v your model's performance