

Economics 103: Project 2
Fall 2024, UCLA
Due Date: De 3, 2024 by 5PM (PST)

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 R Code. You only need to submit one project report per group but please make sure that every group member's name is included. All the analysis and discussion needs to be done in R Markdown. Remember to knit the .RMD file to PDF/HTML as opposed to "save as" PDF/HTML or cutting and pasting into a Word document.

Identify a dataset of your choosing. Make sure it has at least 10 predictor variables, and that it includes a mix of quantitative and factor variables (they could be binary). Remember that Boruta doesn't work on indicator variables, therefore, when you use it, subset your data accordingly, and evaluate the significance of the indicator variables separately.

In your introduction, please clearly state the objective of your analysis, and discuss what types of questions you wish to answer with your model. Also discuss the source(s) of your data.

1. Variable Selection:

- (a) Using the Boruta Algorithm identify the top ~5 quantitative predictors (could be fewer 5 depending on your finding).
- (b) Using standard techniques, identify at least 2-3 factor variables to include as predictors. The combined predictors from parts (a) and (b) are the ones you will work with going forward in the analysis for parts (2) and (3).

2. Descriptive Analysis: Perform a univariate analysis of your variables following the steps below.

- (a) Begin by providing a descriptive analysis of your variables. This should include things like histograms, quantile plots, correlation plots, etc.
- (b) Estimate density plots for all your variables, and show the respective fitted distributions.
- (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, and then justify their removal, exclusion or imputation.
- (e) If you have any NAs, remove them or impute them using any of the methods discussed in class, but make sure to justify your choice.

3. Model Building: Explore several competing multiple-regression models and decide on one model only. 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. Keep in mind that the order of the required checks below may vary as you may need to perform some of the tests more than once.

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

- Evaluate transformations of variables.
- Test for multicollinearity.
- Test for model misspecification.
- Make sure to test for the need of interaction terms.
- Look at Cook's distance Plot, Residuals Plot, QQ-Plot, etc.
- Use AIC and BIC for model selection.
- Evaluate the robustness of your estimates by bootstrapping your model. Provide a histogram of the bootstrapped estimates, and comment on the findings.
- Use cross-validation to evaluate your model's performance.
- Make sure to also discuss any relevant marginal effects estimated.
- If you identify any model issues (e.g., multicollinearity, etc.) make sure to resolve them before finalizing your proposed model.
- Once you have finalized your model, please provide an interpretation of your model parameters, and any economic insights you learned from it (this includes answering the questions you proposed in the introduction).