MATH 372: Linear Regression Analysis
Final Project - Creating Regression Software

Overview

Suppose that you are given a predictor matrix (or data frame) $\mathbf{X}$ and a continuous-valued response y. The purpose of this homework is to create an end-to-end linear regression function that will fit a model using the techniques you've learned in class. Create a function that takes as input X and y and performs the following tasks:

- Preprocesses all data and gets rid of missing observations
- Develop a predictive model for y
- Fit a parsimonious explanatory model

That is, the user of your code can choose up front if they are most interested in explanation or in prediction. All models should be chosen using cross-validation (if needed), and grid-searches should be run where appropriate. Be sure to incorporate the following aspects of linear regression steps in your code:

1. When appropriate, compare Lasso, Ridge and OLS Regression techniques
2. Check (both visually and with appropriate hypothesis tests) for and account for outliers, influential points, and points of high leverage
3. Perform model selection using various metrics (MSE, AIC, BIC, Mallow's Cp, Adjusted R^2)
4. Formal F-tests to check nested models when appropriate
5. Diagnostics - normality, homoscedasticity, and linearity
6. Determine which transformations, if any, on y are appropriate

Explanatory plots should be reported that provide a reason for any parameters chosen and explanatory plots should be provided that will help the user analyze the data. Include plots that compare the methods, and inform the user which model to use and why.

Apply the function to a data set of your choice that is **not** available in R or Python. Show the output of your function and how you interpret the results. Note that this function should be a "one-stop shop" for applying regression techniques. Keep in mind that this function will be incredibly useful for you in the future. So this is to give you an incredibly handy function to run as a burgeoning data scientist :)

Rubric

- Including each of the 6 objectives listed above – each is worth 10 points. (60 points)
- Testing on a data set - 10 points
- Having clear concise output that does not list extraneous information (e.g., listing a summary when not needed)- 10 points
- Wrapping all components into one major function - 10 points
- Discussion of what your linear regression fit says about the data you analyze - 10 points

Note: You can use pre-built functions from R or Python to wrap in your function. I'm not asking you to recreate the wheel here, just to put everything together to have a useable, production worthy function for linear regression.

What to Turn In

- A function file with your executable regression function
  - Please enumerate in comments each of the above 6 components in your code
- A notebook or markdown file and output showing the use of the function on a data set of your choice
  - Include a summary of what your regression function says about your data (best predictive / explanatory model, what is significant, etc.)
  - Show output of the function and interpretation