Preliminary Planning

Joshua Ingram

5/3/2020

Project Overview

Although the LendingClub loan data is technically time series (released on a quarterly basis), there is not a large enough sample to view more than an up-trend in the market cycle. For this reason, I will be selecting 3-5 years between 2007 and 2019 and treating each selection as cross-sectional data. My main variable of interest is the quantity of loans (loan amount in dollars). I will create a multiple linear regression model for each of the 3-5 selected years that predicts the quantity of loans based on demand-side variables such as income and the price of the loan (interest rate). To work around the endogeneity bias that comes from the price of loans being a predictor, I will create a second model that will predict the price of loans based on several supply-side variables such as credit score, collateral (whether someone owns a house), etc. The price predicted in this model will then be placed as the predictor variable in the quantity of loans model, as opposed to the directly observed value. Then I will compare the effects of each explanatory variable on the price of loans across each year to see how they change from post-Great Recession to Pre-Corona Virus market. A theoretical model will be developed based on economic theory.

Multiple Linear Regression Model

After the final variables are selected to predict the quantity of loans demanded, I will fit the model. I'll observe the individual t-statstics for variable significance and the F-statistic for overall model significance. Next, I'll look for collinearity using the variance inflation factor and correlation plots. If everything looks well initially (no need to drop//add variables), I'll move on to check for model assumptions (heteroscedasticity, normality, non-linearity) using the plots from the lm function. Depending on the outcome, I'll implement the proper fixes such as transformations or bootstrap sampling (using my bootstrap function).

I'll also check for influential outliers using different methods like hat-values, standardized residuals, and Cook's distance. If all looks well, I'll move forward to model comparison. If there are issues, I'll compare the models with and without the otliers and see how they change. Depending on the outliers' effects, I may drop them entirely.

Once model assumptions are checked and outliers are worked around, I'll begin to compare several models. Some may have transformations, some may have more/less variables, but I will compare them by looking at the R^2, RSE, and at how they compare to the theoretical model. The final form of the MLR model will be used for interpretation and variable effect comparison over the 3-5 seperate years.

Multiple Logistic Regression Model

For the multiple logistic regression model, I will be using the loan grade as the response variable. This is a multinomial response variable, but I have taken categorical data analysis and have worked with the model for this before. (If I need to use a binary response variable, I can create one that indicates whether a loan is a high-risk or low/moderate-risk loan) Since this is similar to the price of the loan, I will use the same variables that I used to predict in the MLR model for price.

After fitting, I'll check for collinearity if anything looks wrong. I'll of course look at the individual variable significance and overall model significance. I may add some transformations to the predictors if necessary and I'll compare models. After the final model is selected, I'll interpret and visualize the results.