Regression Project

(a) explanation of the problem

The problem we were assigned for this final project was to build a model that works to predict the outcome Y. We were given a simple data set with one response variable Y and 22 X predictor variables. The goal of the project is to create a parsimonious linear regression model that works to effectively predict Y with only the predictor variables that seem to be fit for the final model. In addition, our given objective is to make thorough use of the content and techniques learned in class in order to create our final models.

(b) description of the data

The data that we were given for this project is a rather large set of data with over 2100 observations and 22 explanatory variables that were to be regressed on the response variable Y. Applying the summary() function to the data set gives me a simple five number summary of all the different variables. The variable Y for example has a minimum value of 85.8, the first quartile is 241.4, its median exists at 279.3, it has a mean of 278.8, the third quartile is at the value 315.1, and the max value for Y is 481.1. Just like this, a summary of all the variables in the data set can be pulled up in order to check different values.

One of the first functions that I applied to the data set in order to gain insight is pairs(). Although the large number of observations in the data set made it much more challenging to properly interpret the data, it was clear to see that some linear relationships exist outside of the response vs explanatory variable which would suggest multicollinearity in the data set. This practically goes against the Gauss-Markov theorems, and in order to check for this I applied the vif() function to my data set. In order to run a vif() test I created a simple OLS model that regresses Y on all of the X variables in the data set. Then applying vif() to my model gives me a value of 10.606 which would suggest multicollinearity between the variables. In order to double check my process, I decided to calculate the condition number for the model which gave me a value <30 but >15 which would suggest that the model has potential problems with collinearity but nothing that would require external attention. I then checked to see if the residuals/errors of the model were constant and I did that by plotting the residuals which indicated that the residuals were indeed independent of each other or the errors were constant which means the model has homoskedasticity.

(c) description of modeling process

When it came down to building an effective model, there were a lot of steps and approaches that could be taken, some of which were taught in this class others of which I applied using external knowledge sources. My intuition when it came to the modeling process was to first run through the StepReg library in RStudio which gave me tools like stepwise, forward, and backward selection via which I was able to narrow down my X variables by providing a sle cutoff value and an sls cutoff value. I was not surprised to see that forwards and backwards selection both produced the same model but stepwise created a model with a couple more variables. This variability in the model selection process pushed me to try out other techniques to create my model. My second approach to the problem included running checks for MallowsCP and ADJR2. According to this technique, I was able to see a pattern in the variables that were being picked by these different techniques. The common variables that have been relevant in the process so far are X4,X11,X12,X14,X22. A little more confident in my model I decided to double check if my models are similar if I use a standardized data set. I standardize my data using the scale() function and then re-run my modeling process thus far but with standardized data. To my surprise, the models produced using the non-standardized data and standardized data were similar, which was enough assurance to keep these variables as part of the final model. Although I used lasso regression later in the process as a means to verify my variables, I did not use the output it gave me as a part of my final model.

(d) explanation of decisions concerning transformations and which variables to keep

When it came time to consider which of the variables I wanted to transform it became apparent that this was going to be the hardest part of the project. The first method I decided to implement into the transformation process was to simply create the plots of the X variables against Y and check for linearity. If linearity was present, the variables would not require transformations but if the plot is not linear, transformation is due for the said variable. But looking at the plots alone proved to be a challenge when it came to interpreting what kind of transformations were required because of the large amount of observations. In order to combat this I used the function boxTidwell() in order to find the lambda values for my variables and then transformed the variables that required transformations. For example, boxTidwell() gives me a lambda score of -0.173 for X22 which would then mean for me to take the log of X22 and enter that into my final model. It was simple to decide which variables to keep as there were common variables that seem to be reproduced by multiple different techniques and the same variables seemed to be removed by these processes.

(e) your final model

Yhat = -355.2154+0.6308*X4+21.1933*X11+1.4004*X12+1.5732*X14+7.3344*X22