Keaton Spiller

STAT 364  
Modern Regression Analysis

Instructor: Ge Zhao

Final - Writeup

Starting with 474 rows and 498 columns I attempted to build a model that would be reasonable to make predictions and easy to explain. However, with real data that is much harder than it seems.

What to keep and what to remove often comes down to how close a value is to 0, or how much variation is in the model, but having a higher goodness of fit doesn’t always mean a model is better. With large data there could be clustering and clumping in the influential points that incorrectly dealt with could distort the real path of the model.

The first thing I did was remove 477 variables that showed high collinearity using vif (). I checked multiple variations of subsets to find the largest F value (984.8142) and smallest p value (2.154454e-255), then compared the new model with anova to see if there were improvements.

Side note, throughout most of this project the p values were extremely large, I don't know if that meant something was wrong, but from the homework I was used to seeing some data I could fail to reject ( > 0.05 ).

Because some of the data was in the thousands and others were in decimals I decided to scale the data, hopefully this would correct relevant axis I'm comparing each other too, and I could truly see how far significant each variable is from each other.

To further simplify I used ( ols\_step\_forward\_p ) because I found that the different methods were relatively the same from homework 4, and as long as it can run on the large data it should be valid in simplifying the data further.

After that I decided to simplify the model using PC components, and used only the components with 95% confidence at a 0.05 threshold. This was tricky figuring out how grab Λ from prcomp and translate that transformation with the specific components into a linear model Y= ( X\*Λ )B + ϵ. I don't know if I did this correctly because after they are in this form I have a difficult time explaining the variables that are inside the components.

I decided to go with using pcr over pls because I don't understand how pls deals with X, and Y simultaneously. I would rather deal with X separately because I understand how the components are representing each proportion of data and it's easier to explain.

After doing diagnostics on the data I could easily see that there were patterns in the data that I wasn't accounting for, the variance wasn't constant, there were slight tails in the normality, and there were patterns of correlation still shown in the data. However, I had already filtered the data pretty heavily, and more manipulation could result in higher bias that would further invalidate the model I'm trying to build.

I tried removing influential points and transforming the model but there wasn't anything significant that justified it's change to the model.

One thing I did that was interesting was run a prediction for each linear model every step of the way. What was significant about this experiment was all of the significance range predictions hovered around the ~ (1 to 3). This makes me wonder how important it was for all the heavy filtering I did on the data when the predictions without any filtering were roughly the same. However if this was a prediction on a project worth millions of dollars I could see how a small difference in response could make a substantial difference.

{Answering Questions from the midterm for the Final}

(a) If you fit a multiple linear model, analyze your result using what we have discussed.

From the results I can't make any valid predictions.

After filtering and diagnosing the data, there are no transformations that clearly show a constant variance and normality.

And without knowing what the X values are there could be deeper relationships invalidating my model.

(b) Doing real experiment is very expensive for all of the properties. Assuming

they are of equally likely expense. If I can only afford one experiment, which property do you suggest me do? Give me your reason.

To choose the best experiment I would choose between the most significant p value of the lm\_reduced\_further model.

The two to choose from are either X203 or X464 which both have p values < 2e-16.

In order to choose between the two I would remove both of them from the linear model individually and see which one affected the model more significantly

Checking the models difference in summaries with

X464 being removed made r^2, F statistic and RSE worse than X203 being removed

Therefore, X464 is the experiment I would use if I was given only 1 experiment.

(c) What other ideas do you have from this data set?

I believe there are some relationships in how the data was collected that was created when the data was collected that I can't account for and will always throw off any model I create unless I can guess the type of error and account for it accordingly, but I couldn't find any way to reasonable guess.