Nicolson C. Panos

Professor Nicole Dalzell

STA 381

9 December 2022

Write Up on Linear Models

**Section 1: Goal**

The second goal of our project was to explore the possible explanatory variables that can explain the variations in criminal trial rates in North Carolina. We wanted to tackle this goal by looking at different models that suited the explanatory variables we were interested in. We also wanted to find a model that fits well for interpretation, thus making it easier for our clients to understand the findings. We started by working on a decision tree model to best look at the variation aspect of our goal. Team member Jiayi Zhou mainly worked on this aspect of the project and her work proved important for the next steps to come. What we know about decision trees is that they are undoubtedly helpful for finding how much variability is covered by flowing through each choice possible. For example, we could comprehend how one option, such as a county that did vote Republican, versus the choice of the county not voting Republican, thus creating some comparison of interest. However, the issue we know with the decision tree is the lack of helpful information in terms of interpretation, as well as a lack of coefficients for each of the explanatory variables, thus making it difficult to give our clients something substantial in terms of finding the correct explanatory variables that prove important. To deal with this issue, we agreed upon looking at linear regression models for our Y1 and Y2 that can give us coefficient estimates and also find out which of the explanatory variables can prove significant as predictors of the response variables. After agreeing to pursue the project in this direction, I took over responsibility for the creation and interpretation of the linear models.

**Section 2: Data**

The datasets I used for finding the linear models are the holistic excel files Combine2 and Combine3. Combine2 is a combination of all the explanatory variables of interest, which include: county name, total trial rate, percent of people that are Hispanic/Latino in a county, percent of people that are Black in a county, the violent crime rate, the property crime rate, the acquittal rates, the dismissal rates, the prosecutorial district distinctions, the judicial district distinctions, the volume of cases in a county, the defense council type, and finally the percent a county votes Republican. Besides this long list of explanatory variables, Combine2 also has the response variable Y1, the percentage of trials as a percentage of all convictions obtained, thus the first linear model will be dealing with the Y1 variable. For Combine3, the data set has all the same data on the explanatory variables listed above, but instead of having the response variable Y1, we replaced that with the response variable Y2, the percentage of trials as a percentage of total case disposal. The Combine3 data set will be important for the second linear model.

**Section 3: Data Cleaning, Processing, and Creation**

I was already working with cleaned datasets in the form of Combine2.xlsx and Combine3.xlsx, thus I did not need to complete any data-cleaning steps. All the data had been processed, but I have included the processing steps in the R code associated with this aspect of the project to make sure it is easier for the clients to get the information all in the R code, instead of having to find the processing steps from the other R code files created by my teammates in the Google Drive folders. Finally, there was no need for any data creation as the data sets we needed had already been finalized with the explanatory variables we were interested in.

**Section 4: Modeling/EDA**

Methods Applied:

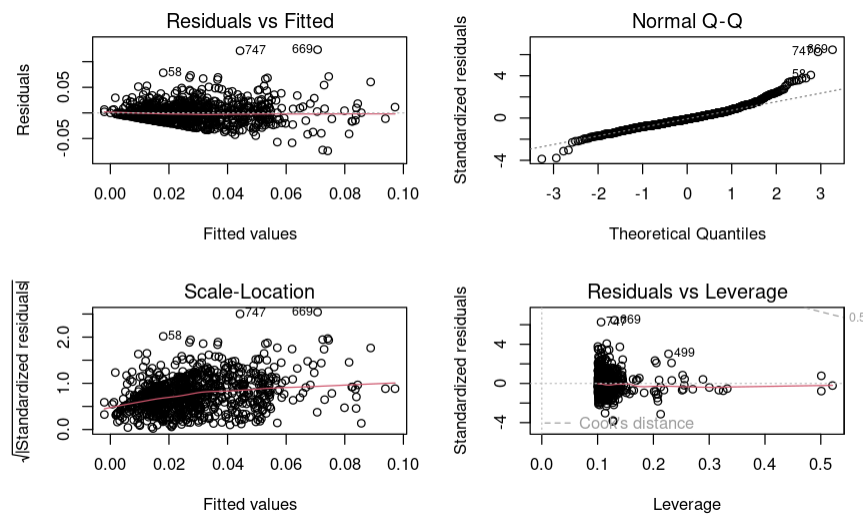
Overall, creating the two linear models for predicting the Y1 response variable, the Trial Conviction Rate, and Y2, the Trial Disposal Rate, was a simple procedure of inputting everything correctly. The thinking for applying the linear model was about just looking at everything and seeing if we should pursue the linear model further or does there need to be another direction. Overall, one item of information I knew would be important for this part of the project is helping to explain the difference between the categorical variables versus the easily understood quantitative variables when explaining to the clients. Essentially, with quantitative variables like the percent of a county’s population that is Black, that is an easily findable number that creates a coefficient estimate that would make sense to a person with a little bit of knowledge on statistics. However, for a categorical variable like North Carolina county, it may be more difficult to understand, so it was important to make sure the clients understood that county acts as an indicator variable where there are 100 different coefficient estimates, each representing a North Carolina county. Essentially, when given a prediction that says it is from one of those counties, the final prediction will be added with the coefficient estimate, while the other 99 estimates will not be factored into the final estimated value. In R, one of the categorical variables, Defense Council Type, was being treated as numeric due to just being a list of numbers, though these numbers meant a certain type instead of a specific value, thus in both linear models, we had to make sure to tell R to treat it as categorical and not a quantitative variable. Finally, after finding the linear regression models and running a simple summary of the coefficient values as well as other calculations possible such as the R-Squared value, I also had to check the validity of assuming the linear regression model works for this problem. This was done through the specific plots R does when asking it to plot the model, which to R, means giving the specific graphs that pertain to the assumptions for linear models, and from there I will interpret the conclusions that can be drawn from there.

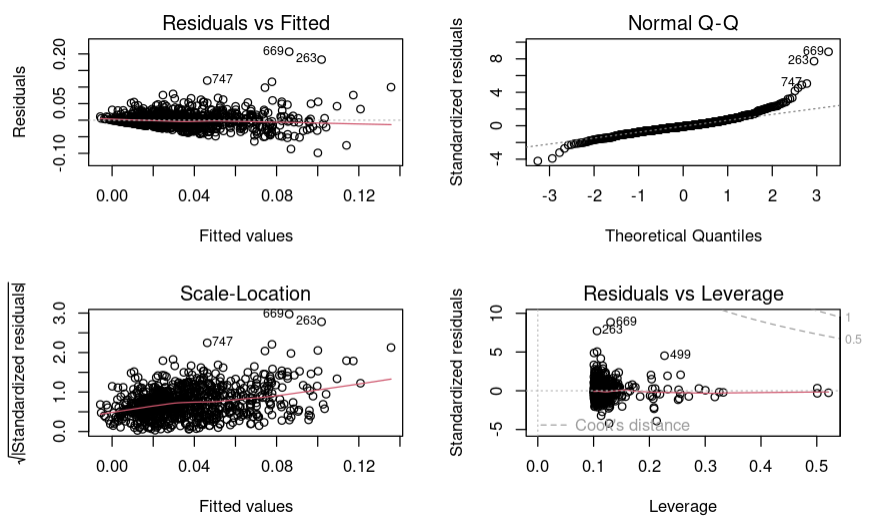
Results:

For the first linear model, after inputting all the information correctly and running a summary of the model, it was found that only 3 of the 100 indicator variables representing each North Carolina county, (Forsyth, Johnston, and Pamlico), had coefficients that prove to have some measure of passing the bar of a 0.05 level of significance. Also, for our other explanatory variables, only the percentage of people in a county’s population that are Hispanic/Latino as well as the Acquittal Rates proved to have significance for predicting Trial Conviction Rate, thus many of the explanatory variables proved to not be very valid in our model. Overall, this first model was able to amass a decently high R-squared value of 0.4073, given the complexity of estimating for this situation, which means that 40.73% of the variability of the data can be explained by the model. For the second linear regression model, a similar story was told. For predicting the Trial Disposal Rate, it was again found that only 3 of the 100 indicator variables for county proved to have significance; this time, the counties of significance were Duplin, Johnston, and Montgomery, completely different from the counties that were significant in the first model. Again, the percentage of people in a county’s population that are Hispanic/Latino as well as the Acquittal Rates were significant for our model, though Hispanic/Latino was very borderline since its p-value was slightly above 0.05, we felt comfortable with saying its coefficient was significant based on the previous work done, and also the harsh approach we needed, given the complexity of predicting the response variable. For this model, we were able to get an even higher R-squared value of 0.4732, which was certainly encouraging. However, we still needed to check if the linear model assumptions were met.

To preempt, one of the assumptions of a linear model is that the data have independence from each other, which we assume based on our collection process of the data, as well as the information of the data collection given by the trusted sources we used. For the first linear model, we looked at the assumption of constant variance, which means that at every level of x values, the residuals are fairly consistent from the smaller to bigger values. By looking at the residual plot vs fitted values, the plot of the residuals, the actual Y1 values minus the estimated Y1 values from the model, versus the estimated values for Y, we found a problem where though most of the data was agreeing with this assumption, there was an issue whereas the fitted values got bigger, we saw a wider range of residual values, thus meaning that the assumption was broken. Next, for the assumption of linearity, we looked mainly at the same plot as explained previously as well as the square root of the standardized residuals, the square root of the residual divided by the standard deviation, vs fitted values plot. Based on the second plot, we found concerns with the red line shown on the graph which demonstrates the mean line straying away from its expected value of 1, thus meaning concerns that the assumption of linearity could be broken. Finally, for the assumption of normality, we looked at the QQ plot, as it is commonly known as, or to explain in more literal terms, the standardized residuals, the residuals divided by the standard deviation, versus theoretical quantiles, what R assumes as the theoretical levels of the data for the linear assumption, plot. We want the values of this plot to fit perfectly along the line with little to fluctuation above or below it, but based on the QQ plot for the first model, we can not accept the assumption of normality since the data points tail way up as the theoretical quantities get much bigger, especially after 2 theoretical quantile value. For the second linear model, we found the same exact problems in the assumptions that we had with the first model, thus we knew that some transformation process was going to be needed. Certainly, the linear model results were widely infrequent in their success.

Graphs that Correspond to the Linear Assumptions:

Linear Model for Y1: Linear Model for Y2:



**Section 5: Conclusions**

Overall, while we did think creating linear models for our 2 response variables would be a good place to start with working on finding the true explanatory variables that would be pertinent to our clients’ interests, we found many problems in assuming it was valid to work with such linear models as well as having issues with how to properly factor in the county variable, which was pertinent to how we would discuss our findings for the two response variables. Based on the results of the modeling, while we did find some fairly decent R-squared values, these calculations did prove deceptive as we were left with nothing too conclusive as it was tough to conclude how the indicator variables for the North Carolina counties played a role in the model. Thus, we felt comfortable with moving on to working on the linear mixed effects model, as it related to county, which was mainly worked on by our teammate Evangeline (Xinyuan) Cao. Some limitations of finding the linear models were the amount of time we had to find all the possible explanatory variables that we were interested in, which meant some listed variables got left on the “cutting room floor.” Also, overall, we were able to only look at three different model situations, but there were loads of other model ideas we could have looked at, thus I would say we were limited by too few people and time to find other models of interest.

**Section 6: Next Steps**

For a person approaching my work on the linear models and approaching my analysis, I would obviously recommend continuing further with transforming the linear model, which may actually meet the assumptions. We had too little time to go further on the linear model as we believed the mixed effects models would prove more important and interesting for the work, but we still struggled with the assumption of normality with that model, thus there may be some future work on transforming the response variables or explanatory variables. Also, there could be more work in trying to find the best model instead of just looking at all the explanatory variables, thus there can be more work with looking at the AIC or BIC values and seeing what may come with finding those values given the situation. Another area of work that could be good to look into was the explanatory variables that could prove valuable to the model. Some ideas we had that we/I did not have time to work on included the median/mean income of the county residents, the percentage of people with access to transportation (which I thought would be tough to find data on), and finally, the number of criminal law attornies in a county based on where their practice locates, just to see if this provided any good information. Certainly, there are many steps to go in the future of this project and though we concluded that linear models were not the best for this project, there could be more exploration done to see if this conclusion was, in fact, the correct one to make.