Xinyuan Cao

Professor Nicole Dalzell

STA 381

9 December 2022

**Linear Mixed Effect Model**

**Section 1: Goal**

In the whole project, our second goal is to explain the variations in criminal trial rates among counties in North Carolina with models. To achieve this goal, we first need to collect all the variables that we need into a new dataset. Since our clients give us a list of explanatory variables that they are interested in, our group decided to collect the data for these variables by different persons and then combine them to create the data. I am responsible for calculating the acquittal rates and dismissal rates, organizing the prosecutorial district, judicial district into different years based on county, and calculating the mean volume of cases in different counties of each year. After the whole dataset is created, our whole group decides to achieve the analysis goal by using three models: decision tree, linear model, and linear mixed effect model, and I am responsible for the linear mixed effect model. To be more specific, decision trees are like a general map which gives us an overall view of all variables; a linear model takes a further step into the general map and analyzes the variables in numeric numbers; a linear mixed effect model is based on linear model results and makes some extensions. For example, in the decision tree and linear model, we find out that the County which is a significant variable for making the heat map in goal one is not significant, and we want to take County into account and see how important this variable is, so we decide to use the linear effect model. By clustering the data into different counties, we could consider counties as a random effect and get the result.

**Section 2: Data**

The dataset I use to make the linear mixed effect model is the whole dataset that contains all explanatory variables and response variables. Since I am responsible for some variables (acquittal rates, dismissal rates, prosecutorial district, judicial district, volume of cases, and defense counsel type) I use an excel called **Explanatory1** to collect all of them and upload it in RStudio for later combination. After other explanatory variables are fully collected and organized, we combine them with Y1 into a csv file in r called **Combine2** and combine with Y2 into another csv file in r called **Combine3**.

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

For the Explanatory1 Excel, I first calculate each year’s acquittal rates and dismissal rates separately in each cell in excels from fy10\_11 to fy 20\_21 and then I calculate the mean of these rates in different counties and combine these means together from year 2010 to year 2021 into a new columns called Acquittal Rates and Dismissal Rates. I also collect each county’s prosecutorial district and judicial district from 2010 to 2021 and combine them into two columns called Prosecutorial District and Judicial District in Explanatory1. Finally, I calculate the mean of the volume of cases in different counties every year and make it into a column called Volume of Cases in Explanatory1. After all the other explanatory variables are collected in r, I combine Y1 with these variables into a csv called Combine2, and I combine Y2 with these variables into a csv called Combine3.

**Section 4: Modeling/EDA**

For the response variable y1(average percentage of all convictions obtained by trials), I treat County as a random effect and other explanatory variables as fixed effect, so we could see what factors would affect our y1 within each County. I use RStudio to do this analysis, and the result shows that Acquittal Rates, Judicial District 21, Judicial District 29B, and Judicial District 30A are significant parameters in the fixed effects. Besides, the random effects tell us that:

Ø The average percentage of all convictions obtained by trial of each county may vary from the global average percentage by ± 0.0125.

Ø Within any given county, we expect the percentage of all convictions obtained by trial each year may vary from the county’s average percentage by ±0.021.

Ø Intraclass Correlation: 0.2691 (0.0122^2/0.0125^2+0.0206^2)

26.91% of the variation in Y1(percentage of all convictions obtained by trial) could be explained by differences in average percentage from county to county

Similarly, for the response variable y2 (average percentage of all dispositions obtained by trials), I treat County as a random effect and other explanatory variables as fixed effect, so we could see what factors would affect our y2 within each County. I use RStudio to do this analysis, and the result shows that Acquittal Rates, Judicial District 29B, Judicial District 30A are significant parameters in the fixed effects. Besides, the random effects tell us that:

Ø We expect that the average percentage of all dispositions obtained by trial of each county may vary from the global average percentage by ± 0.01183.

Ø Within any given county, we expect the percentage of all dispositions obtained by trial of each year may vary from the county’s average percentage by ±0.02515.

Ø Intraclass Correlation: 0.1812 (0.011832^2/0.02515^2+0.01183^2)

18.12% of the variation in Y2(percentage of all dispositions obtained by trial) could be explained by differences in average percentage from county to county

After that, I check the assumptions of these two models and find out that constant variance and normality of residuals are not satisfied, so I decide to transform the response variable. Since all the response variables are greater than or equal to 0, I decide to take the square root of y1 and y2 and make the linear mixed effect model again.

Now, for the response variable sqrt(y1), I treat County as a random effect and other explanatory variables as fixed effect, so we could see what factors would affect our sqrt(y1) within each County. I use RStudio to do this analysis, and the result shows that Acquittal Rates, Black, Volume of cases in County, and Judicial District 29B are significant parameters in the fixed effects. Besides, the random effects tell us that:

Ø The average sqrt percentage of all convictions obtained by trial of each county may vary from the global average percentage by ± 0.038.

Ø Within any given county, we expect the sqrt percentage of all dispositions obtained by trial of each year may vary from the county’s average percentage by ±0.0664.

Ø Intraclass Correlation: 0.2526 (0.03859^2/0.03859^2+0.066378^2)

25.26% of the variation in sqrt Y1(percentage of all dispositions obtained by trial) could be explained by differences in average sqrt percentage from county to county

Similarly, for the new response variable sqrt(y2), I treat County as a random effect and other explanatory variables as fixed effect, so we could see what factors would affect our sqrt(y2) within each County. I use RStudio to do this analysis, and the result shows that Acquittal Rates, Black, Volume of cases in County, Judicial District 29B, and Judicial District 30A are significant parameters in the fixed effects. Besides, the random effects tell us that:

Ø We expect that the average sqrt percentage of all dispositions obtained by trial of each county may vary from the global average percentage by ± 0.0367.

Ø Within any given county, we expect the sqrt percentage of all dispositions obtained by trial of each year may vary from the county’s average percentage by ±0.066.

Ø Intraclass Correlation: 0.2329 (0.0367^2/0.0367^2+0.0666^2)

23.29% of the variation in sqrt Y2(percentage of all dispositions obtained by trial) could be explained by differences in average sqrt percentage from county to county

**Section 5: Conclusions**

Based on the non-transform models, we conclude that acquittal rates, judicial district 21, judicial district 29B, judicial district 30A are important indicators for **y1**, and they could be inferred as follows:

Ø Holding other variables constant, for every 1 increase in the acquittal rates, we expect that the average conviction rates obtained by trials would increase by 0.275.

Ø Holding other variables constant, judicial district 21 will increase the convictions rate by 0.0484 relative to judicial district 1.

Ø Holding other variables constant, judicial district 29B will increase the convictions rate by 0.0335 relative to judicial district 1.

Ø Holding other variables constant, judicial district 30A will increase the convictions rate by 0.0217 relative to judicial district 1.

For **y2**, we conclude that acquittal rates, judicial district 29B, judicial district 30A are significant indicators. We can infer them as followings:

Ø Holding other variables constant, for every 1 increase in the acquittal rates, we expect that the average dispositions rates obtained by trials would increase by 1.032.

Ø Holding other variables constant, judicial district 29B will increase the dispositions rate by 0.039 relative to judicial district 1.

Ø Holding other variables constant, judicial district 30A will increase the dispositions rate by 0.0328 relative to judicial district 1.

Based on the transform models, we conclude that acquittal rates, judicial district 29B, black, and volume of cases are important indicators for **sqrt(y1)**, and they could be inferred as followings:

Ø Holding other variables constant, for every 1 increase in the acquittal rates, we expect that the square root of average conviction rates obtained by trials would increase by 0.824.

Ø Holding other variables constant, judicial district 29B will increase the square root of convictions rate by 0.0926 relative to judicial district 1.

Ø Holding other variables constant, black people will decrease the square root of convictions rate by 0.1404 relative to people who are not black or Hispanic Latino.

Ø Holding other variables constant, for every 1 increase in the volume of cases, we expect that the square root of average conviction rates obtained by trials would decrease by 0.081.

For **sqrt(y2)**, we conclude that acquittal rates, judicial district 29B, judicial district 30A, black, and volume of cases are significant indicators. We can infer them as followings:

Ø Holding other variables constant, for every 1 increase in the acquittal rates, we expect that the square root of average dispositions rates obtained by trials would increase by 2.688.

Ø Holding other variables constant, judicial district 29B will increase the square root of dispositions rate by 0.0948 relative to judicial district 1.

Ø Holding other variables constant, black people will decrease the square root of dispositions rate by 0.1348 relative to people who are not black or Hispanic Latino.

Ø Holding other variables constant, for every 1 increase in the volume of cases, we expect that the square root of average dispositions rates obtained by trials would decrease by 0.0933.

Although we transform y1 and y2 into sqrt(y1) and sqrt (y2), based on the QQ plot we still have problems in normality of error, which may make our linear mixed effect regression model not so valid.

**Section 6: Next steps**

First, due to the limitations that we have now, there could be some following steps to deal with the normality problems. Additionally, now our model only considers county as a random effect, we can only treat year as a random effect or treat year and county together as random effects. Then we run the model and see which variables are significant, and we can also compare the results in different regressions. Finally, since the judicial district has many categorical levels, we can run the LME regression with and without this variable and compare the results.