North Carolina Infant Mortality

October 19, 2018

Introduction & Goals

In this case study, our goal is to create a hierarchical model to obtain estimates of infant mortality rates based on county and ethnic origin in North Carolina from 2011 to 2016. We will use this developed model to estimate infant mortality within each county-race combination over the 6 years. After obtaining these estimates, we will compare our results to the actual infant mortality rates and explore the discrepancies that occur. Overall, our goal is to create estimations that are representative of the population at a larger scale as opposed to just the points we were given.

Data & Cleaning

Before developing a model to predict infant mortality rates, we first examined the data sets that we were provided with. The first data set was on birth information from 2011-2016 in different counties within North Carolina. Features included baby birth characteristics (ie. birth weight, sex) and mother characteristics (ie. ethnic origin, race, parity, gestation, etc). The second data set was on death information from 2011-2016 in different counties within North Carolina. Features included death year, race, and Hispanic origin.

In order to create a data set that will be used for our model, we first cleaned the birth and death data sets so that there was uniformity within the following features: counties, year, race, and Hispanic origin. In other words, we wanted to ensure that all features were labeled the same way, specifically racial/ethnic origin, so that these data sets can be merged appropriately. We combined race and Hispanic origin as follows in the birth and death data set: 1 = white/Caucasian, 2 = black/African america, 3 = native american/Alaskan, 4 = other, 5 = Hispanic. In categories 1-4, people of these racial origins have non-Hispanic ethnic background and we placed everyone with a Hispanic background in category 5. Next, we aggregated the data based on all possible year/county/race combinations in order to obtain their respective birth and death counts. The data sets were then merged and we conducted our analysis as described in the following sections.

Model Selection & Description - Hierarchical Model

$$logit(Pr(y_i = 1)) = \beta_o + \beta_1 I(race_i = White) + \beta_2 I(race_i = African\ American)$$

$$+\beta_3 I(race_i = American\ Indian) + \beta_4 I(race_i = Other) + \beta_5 I(race_i = Hispanic)$$

$$+\beta_6 Year_i + \beta_7 County_i$$

$$for\ i = 1, ..., 100$$

In order to model the infant mortality rate for each year/county/race combination, we believed that major factors that would contribute to infant mortality rate would be year, county, and race. Year could be an important predictor because it can carry information about major events or circumstances that may have contributed to infant mortality in a particular race and/or county. Race could be an important predictor because it can carry information about socio-economic and financial status, which is known to have a correlation with infant mortality rate. Similar to race, counties will also carry information on socio-economic and financial status, which will also help with contribute to predicting infant mortality rate.

In order to incorporate these variables, we decided to use a multilevel (binomial) model as depicted in the above formulation. Specifically, we used the Bayesian form of the multilevel model over the Frequentist model so that we could get a probability distribution of what the infant mortality rates could be instead of just obtaining point estimates with a corresponding standard error. We felt that the multilevel model would be best to use as a way to handle correlated and clustered data, which is exactly what we are working with in this scenario. Additionally, since some county/race combinations have very little/no information provided on the birth and/or death counts, the multilevel model will be able to borrow information from other counties/races/years in order to make the appropriate estimates.

In our model specifically, we decided to make year and race as fixed effects and make counties as mixed effects for our model. We decided to make year a fixed effect because we wanted to study infant mortality trends over time within each race-county combination. We made race a fixed effect because each of the race categories don't need to borrow information from each other since each group is already relatively large in size. We made county a mixed effect because all of the North Carolina data is broken down into a total of 100 counties, which means that there will be counties with relatively small amounts of information and it is very likely that different counties will have different infant mortality rates in general. Something that we considered incorporating into our model is an interaction term between race and county, but in the end decided against doing so. That is because if we created an interaction term, there would be a total of 500 variables created (since there are 5 races and 100 counties) and the merged data set described in the previous section only has roughly 2700 rows consisting of birth and death counts for each year/county/race combination. Since this was the case, we found it most appropriate to exclude the interaction term.

Bayesian Multilevel Model Validation

In order to validate our multilevel model, we analyzed the trace plots for each of the parameters (fixed and mixed effects). From the trace plots, nothing appeared out of the ordinary and all of the chains appeared to converge to a mean value (as expected). Therefore, we continued to use our model in order to assess our predictions for each race/county/year combination and compare the estimated values of infant mortality rate to the actual infant mortality rate.

Model Assessment

After validating the model by checking the trace plots for each of the parameters, we then assessed the model by checking its predictive capabilities. Since we trained the model on 2011-2015 data, we wanted predict the 2016 infant mortality rates and compare our predictions to the actual rates for the available data for 2016 infant mortality rates. Based on the mean squared error between the predicted infant mortality rate and the actual mortality rate, we got a value of 1570.1214688, which indicates that our predictions are not as accurate as they could be. Further investigation shows that the major reason for such a high mean squared error value is because when the actual infant mortality rate is 0, the predicted values tend to be much greater than 0, which explains the penalization.

After validating the model by checking the trace plots for each of the parameters, we then assessed the model by checking its predictive capabilities. Since we trained the model on 2011-2015 data, we wanted predict the 2016 infant mortality rates and compare our predictions to the actual rates for the available data for 2016 infant mortality rates.

Findings

County Level Infant Mortality Rate

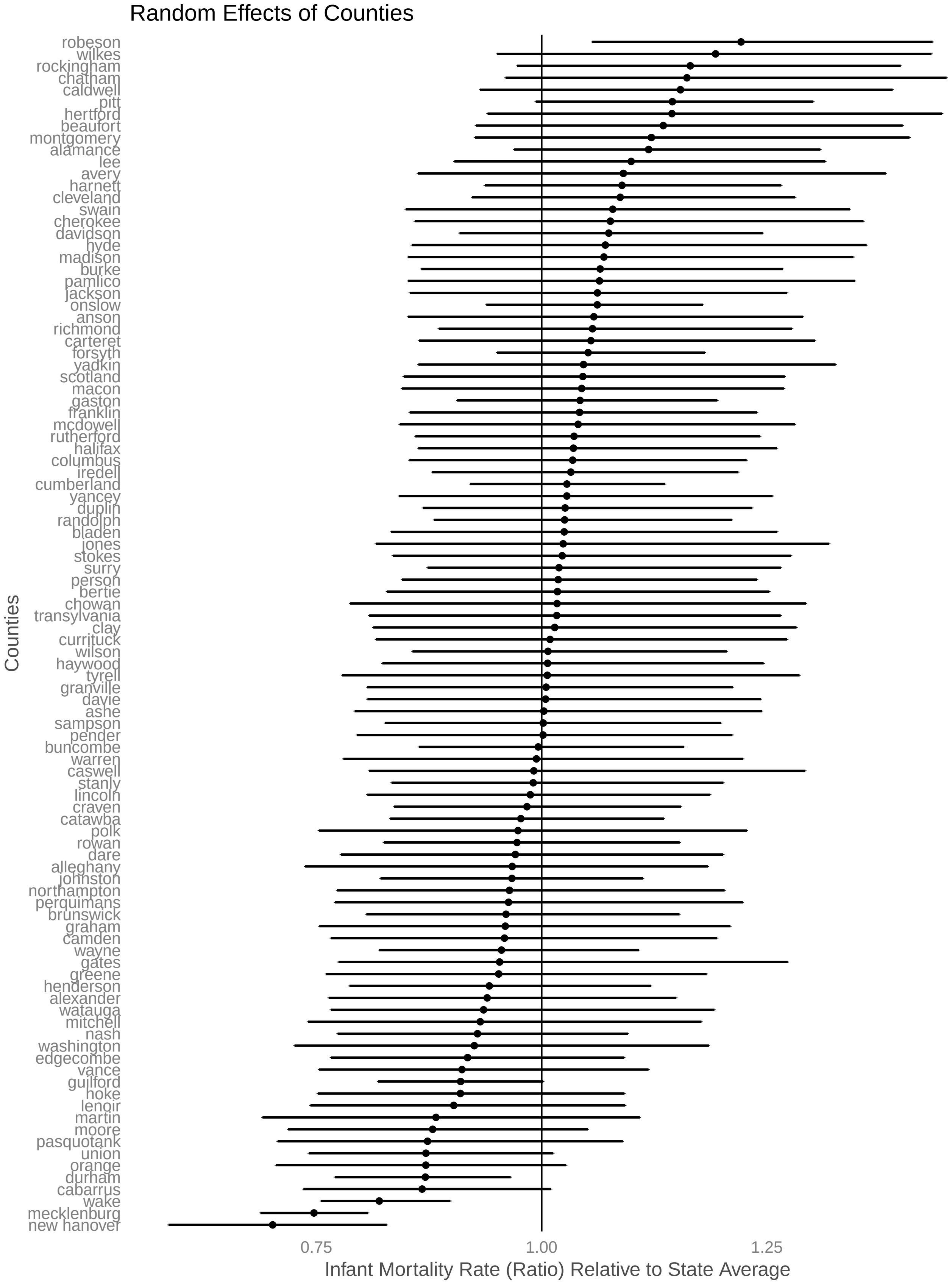
After training the model on the all the data from 2011 to 2015, the first thing that we investigated was the odds ratio between the pooled infant mortality rate with the infant mortality rate of each county. Since we

Table 1: Counties with significant impacts on infant mortality

Worse Counties	Better Counties
Robeson	Durham
	Mecklenburg
	New Hanover
	Wake

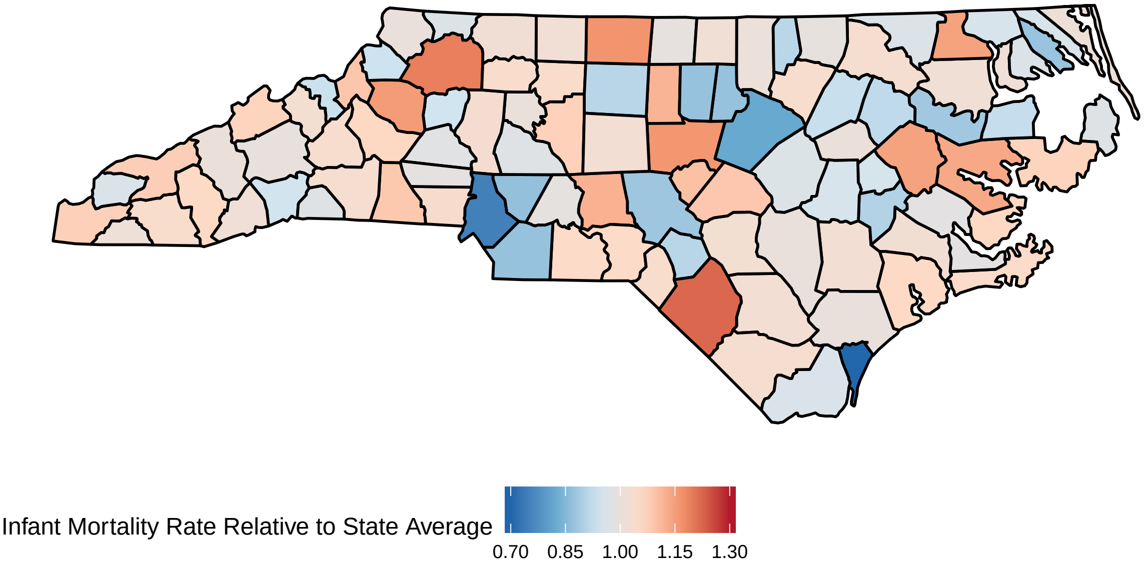
used a Bayesian multilevel model, we were provided with a series of credible intervals of odds ratios that reflected the comparison of county level infant mortality rate to the pool infant mortality rate. Not to our surprise, most of the counties had credible intervals that contained 1, meaning that there might not be a significant difference between the pooled mortality rate and the county level infant mortality rate. However, there were some counties that had credible intervals that did not contain 1 – indicating a significant difference. Table 1 shows which of these counties stood out from the rest of the data.

Some of the counties that performed the best in infant survival were the largest and highest-income counties in the state. These included Mecklenberg County, which contains Charlotte; Wake County, which contains Raleigh; Durham County, which contains Durham; and New Hanover County, which contains Wilmington. There are several plausible reasons for this, none of which can be definitievely proven by our model. For one, large urban centers tend to be home to the best healthcare facilities, acces to which can play a role in infant and maternal health. These urban counties also have higher median incomes, which has been shown in other studies to be correlated with better health outcomes. It's also possible that this is simply an effect of the credible intervals of these counties being smaller due to having larger populations and thus more data available. However there are some urban counties that don't make appear to have a such positive outcomes, like Guilford (Greensboro), Forsyth (Winston-Salem), and Cumberland (Fayetteville). Conversely, some of the worst performing counties like Robeson are smaller and more rural, with low median incomes compared to the state average. It's also notable that Robeson county is one of the few majority-minority counties in the state, which may have some effect despite race being accounted for in other parts of the model.

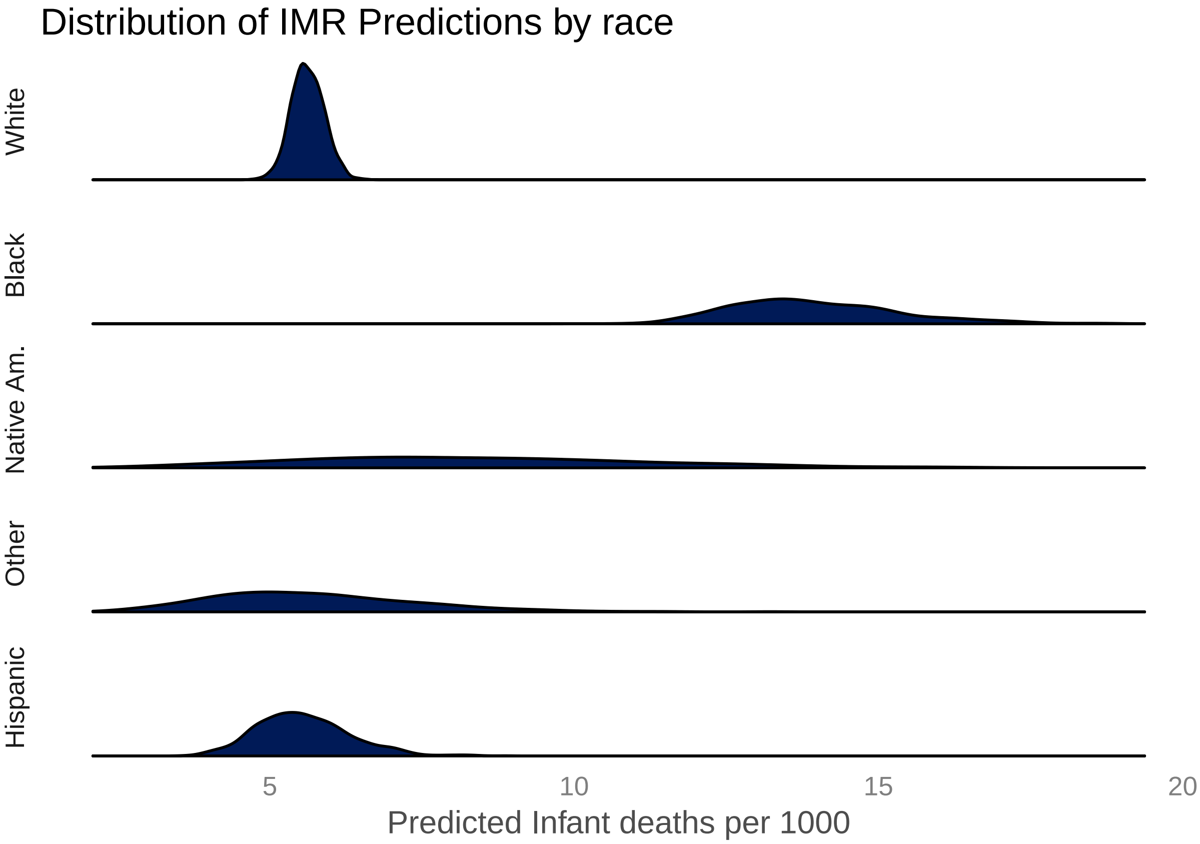


Random Effects of Counties
Counties
robeson
wilkes
rockingham
chatham
caldwell
pitt
hertford
beaufort
montgomery
alamance
lee
avery
harnett
cleveland
swain
cherokee
davidson
hyde
madison
burke
pamlico
jackson
onslow
anson
richmond
carteret
forsyth
yadkin
scotland
macon
gaston
franklin
mcdowell
rutherford
halifax
columbus
iredell
cumberland
yancey
duplin
randolph
bladen
jones
stokes
surry
person
bertie
chowan
transylvania
clay
currituck
wilson
haywood
tyrell
granville
davie
ashe
sampson
pender
buncombe
warren
caswell
stanly
lincoln
craven
catawba
polk
rowan
dare
alleghany
johnston
northampton
perquimans
brunswick
graham
camden
wayne
gates
greene
henderson
alexander
watauga
mitchell
nash
washington
edgecombe
vance
guilford
hoke
lenoir
martin
moore
pasquotank
union
orange
durham
cabarrus
wake
mecklenburg
new hanover
0.75
1.00
1.25
Infant Mortality Rate (Ratio) Relative to State Average

In order to get a better idea of the odds ratio of infant mortality rates for each county, below is a visualization of the map of North Carolina and its counties with the corresponding odds ratio.

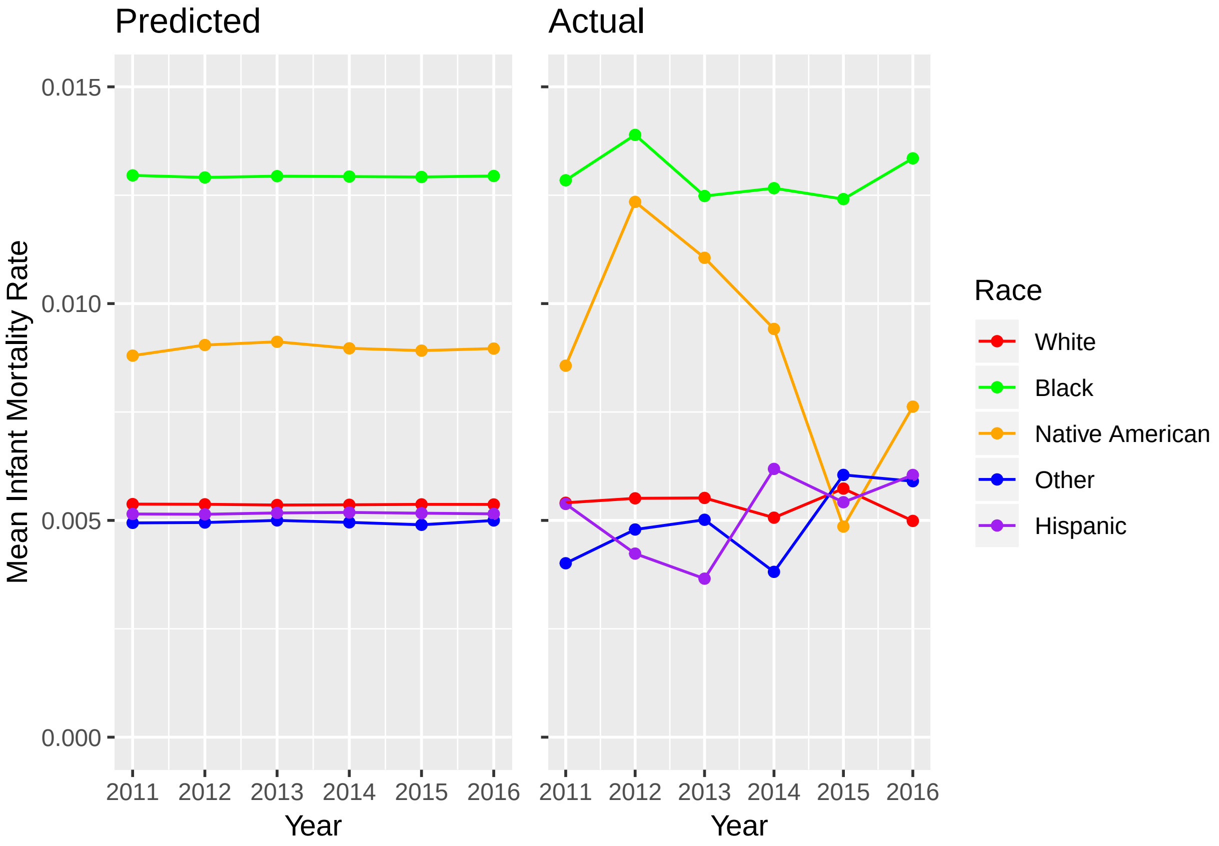


Race



In the above visualization, we can see the probability distribution of infant mortality rates by race. From these distributions, what we can see here is that the White, Other, and Hispanic populations have a relatively low infant mortality rate in comparison to the AFrican American and Native American population (when comparing the distribution means). Something that is also interesting to note is that the African American, the Native American, and the Other populations seem to have a much wider probabilitiy distribution in comparison to the White and Hispanic populations, giving the impression that there is a lot more variability in the infant mortality rates within ethnic groups.

Infant Mortality Rate by Year



Based on our multilevel model, we were able to aggregate the infant mortality rates by county per year basis as a way to assess the changes of infant mortality rate over time by race. Based on the graph below, we see that the infant mortality rate within the African American/Black population is the highest (ranging from 13.92 to 13.98), which is followed by American Indian/Alaskan population with the second highest infant mortality rate (ranging from 8.38 to 8.42). The remaining three racial groups appear to have very similar infant morality rates, which all range from 5.5 to 5.8 deaths for every 1000 births. Though not immediately obvious based on the appearance of this graph, there is a very small upward trend in infant mortality rate over the span of 2011 to 2016, hinting that in future years, there may be continuous marginal increases in the mortality rates.

In addition to analyzing our predictions of infant mortality by race from 2011 to 2016, we compared our predictions to the actual infant mortality rates for each by year. In the plot below, we can see the squared error rate for between what our model predicts and what the actual infant mortality rates are. From the plot, we can see that all of our predictions are very close to the actual value infant mortality rate, with very little variation in error per year. The only thing that sticks out is the fact that our squared error rate when predicting infant mortality for the Black/African-American population is not as accurate as the other races, which is depicted through the two spikes in the graph. This means that overall, our model does well in predicting infant mortality rate.

Conclusion

From our analysis of infant mortality over the 2011 to 2016 time period, we have discovered 3 keys findings. The first finding is that year does not seem to play an essential role in predicting infant mortality, which is something that we had originally anticipated. Over the years, infant mortality seemed to stay relatively constant for most race/county combinations. The second finding is that there are empirical differences in infant mortality rates between the different races, which makes sense because of the underlying socio-economic and financial statuses of these racial groups. The third finding is that most counties seem to have very similar infant mortality rates, which are close to the overall North Carolina mean infant mortality rate. However, there are a few counties (as specified before) that have an “exceedingly” low or high rate.

For future analysis, we could potentially try to incorporate other data sets such as income levels, disease incidence rates, population/density, urban area, and more to get a better idea of factors that are indicative of infant mortality. At the same time, incorporating more data into the multilevel model would be able to help us better predict future infant mortality rates in the upcoming years.