440 CS3 Final Report

Jake Epstein, Man-Lin Hsiao, Sahil Patel, Daniel Spottiswood, & Michael Tan

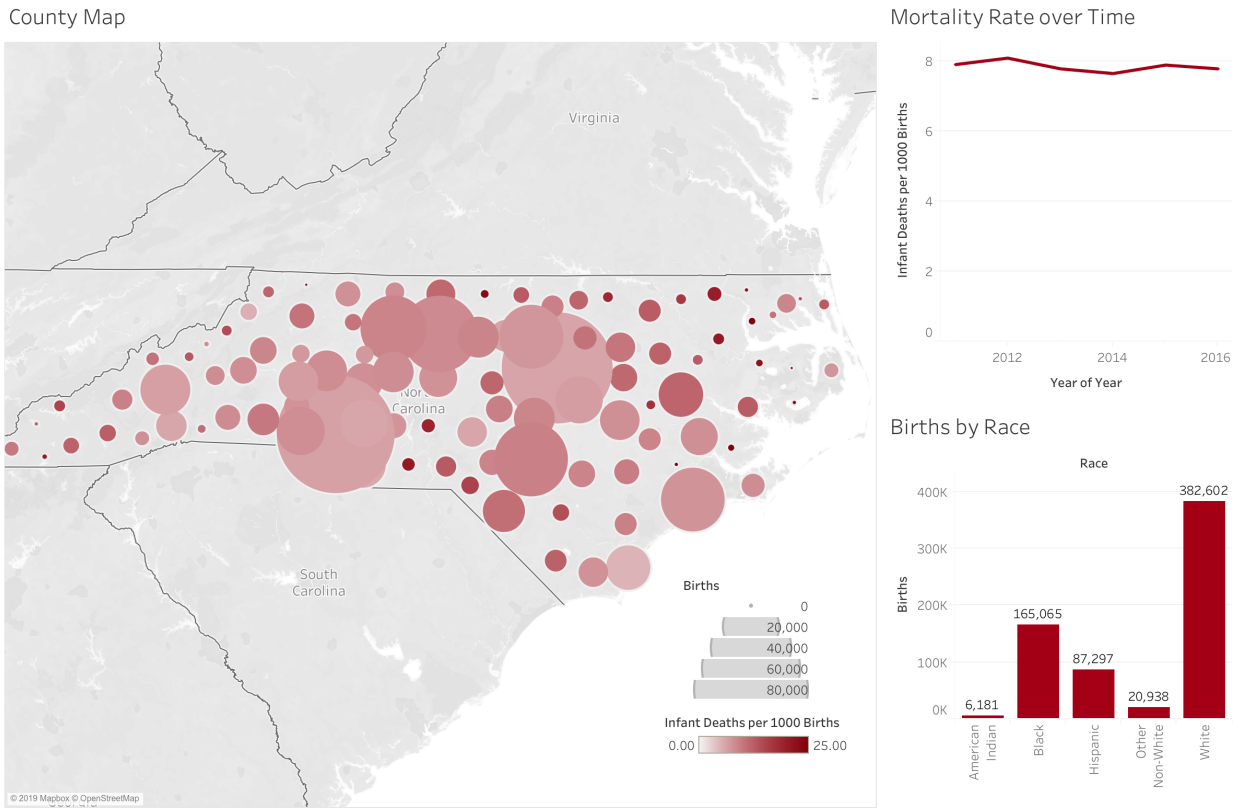
10/22/2019

Introduction

This case study aims to obtain robust estimates for mortality rates by race and year across North Carolina counties based on 2011-2016 mortality data from the North Carolina Center for Health Statistics. We supplemented this data with health and demographic data from a study performed by the Robert Wood Johnson Foundation. We explored correlations between mortality rate and percent insured, median household income, percent low birth weight, percent obese, and percent smokers. Through EDA and model fitting, we narrowed these features down and created a model that predict mortality rates by race, year and county, utilizing percent low birth weight, percent smokers, and median household income. We then comment on trends by group over time and compare those trends to the national averages.

Exploratory Data Analysis

Initial Visualization

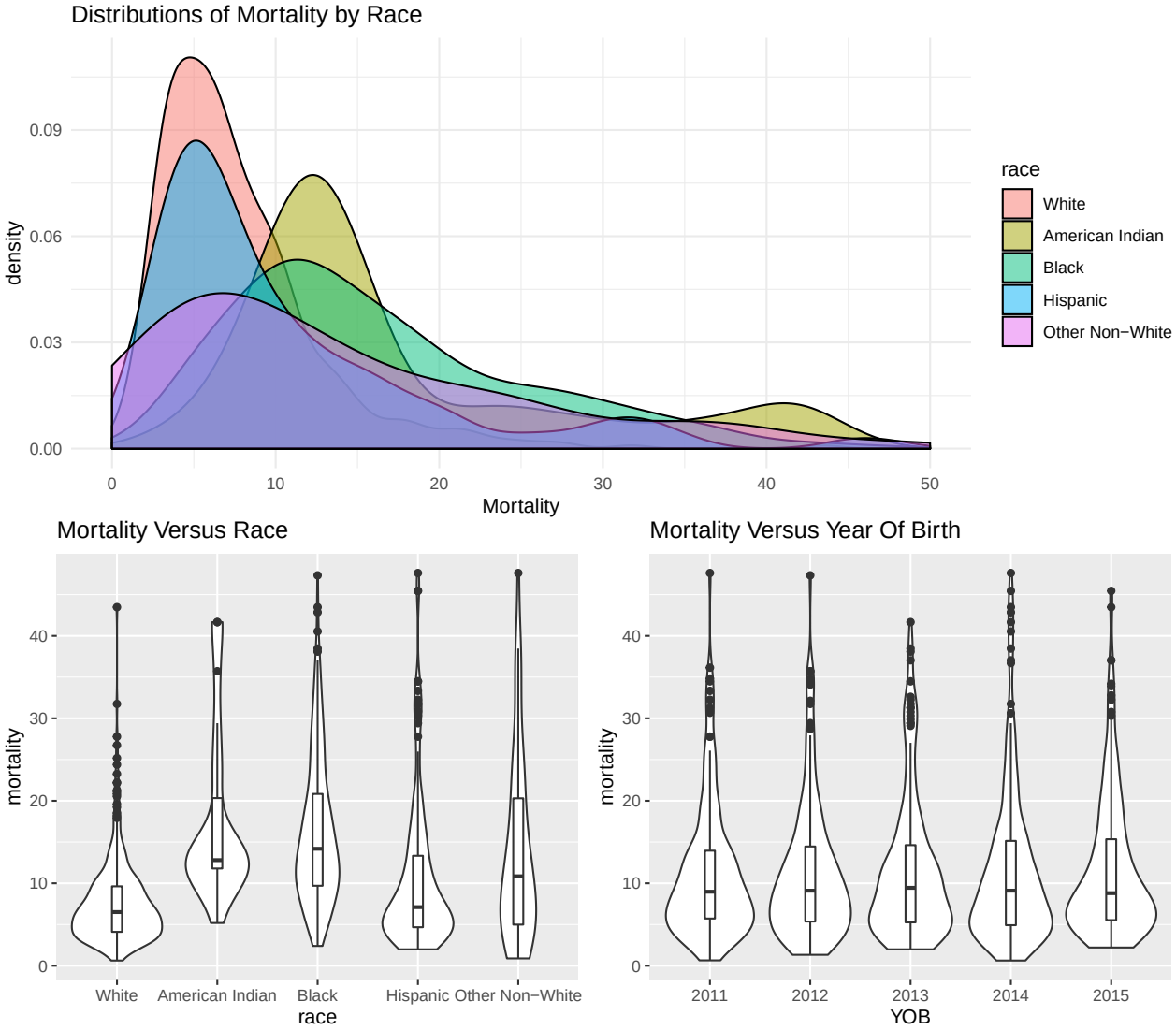


The visualizations above are captured from our tableau dashboard. The map depicts the **large variance in number of births and mortality rates across the different counties.** The largest counties, based on births, appear in lighter colors - they have relatively lower mortality rates, while we see darker colors, which are spikes in mortality rate, in many of those with smaller populations. This variation in mortality rates is likely the result of smaller sample sizes in these counties, but higher mortality rates in small counties could

possibly be attributed to differences in medical resources in these areas. From the bar charts, we also see that there are **substantially more white births** (382,602) than any other race, specifically, we have very few data points for American Indians and people that identify as other non-white. On the other hand, **the overall mortality rate for each year from 2011 to 2016 stays fairly constant**.

The high variance in births per county paired with the aforementioned variance in births by race, underscores **the need to share some, but not all, information at the county level and by race**.

Exploring Categorical Variables



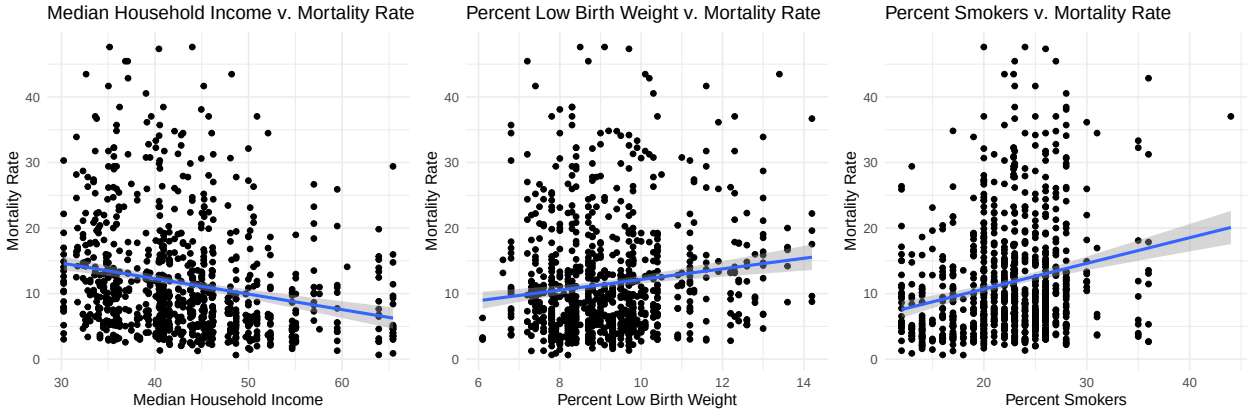
The above plots confirm several observations we were able to make in our tableau dashboard, while also providing additional visualizations on distributional information.

The distribution plot of mortality by race again shows a significantly more white births than other races. In addition, we can particularly see that white births appear to be skewed towards aggregating on the left side of the x-axis on the plot - where mortality rate is low. We can make a similar observation in the first violin plot, where the widths are particularly wide between a mortality rate of 0 and 10. Both these plots indicate that **white births generally contribute to lower mortality rates, whereas high mortality rates are mostly composed of births of races that are not white**.

The second violin plots show similar distributions of mortality rates across time. Here we can understand that

not only the number of births, but also mortality rates on an aggreggated basis within North Carolina throughout the years do stay fairly consistent and have no notable swings.

Exploring Numerical Continuous Variables



As shown in the scatterplots above, there is a positive relationship between mortality rate with percentage of smokers (correlation of 0.2182623) as well as percentage of births that had a low birth weight (correlation of 0.1377997). On the other hand, there is a negative relationship between the median household income and the mortality rate in that county (correlation of -0.2169295). These relationships suggest a possibility of incorporating these factors as appropriate predictors in our model.

Model Selection

Model Selection			
Model	**AIC**	**BIC**	**AUC**
Model 1	3894.3	3943.1	0.6388
Model 2	3934.7	3968.9	0.6385
Model 3	3877.2	3940.7	0.6387
Model 4	3917.8	3966.7	0.6385

Model 1: Mortality Rate ~ (1 | County) + (1 | Race) + Year + Median Household Income + Percent Low Birthweight + Percent Smokers

Model 2: Mortality Rate ~ (1 | County) + (1 | Race) + Year

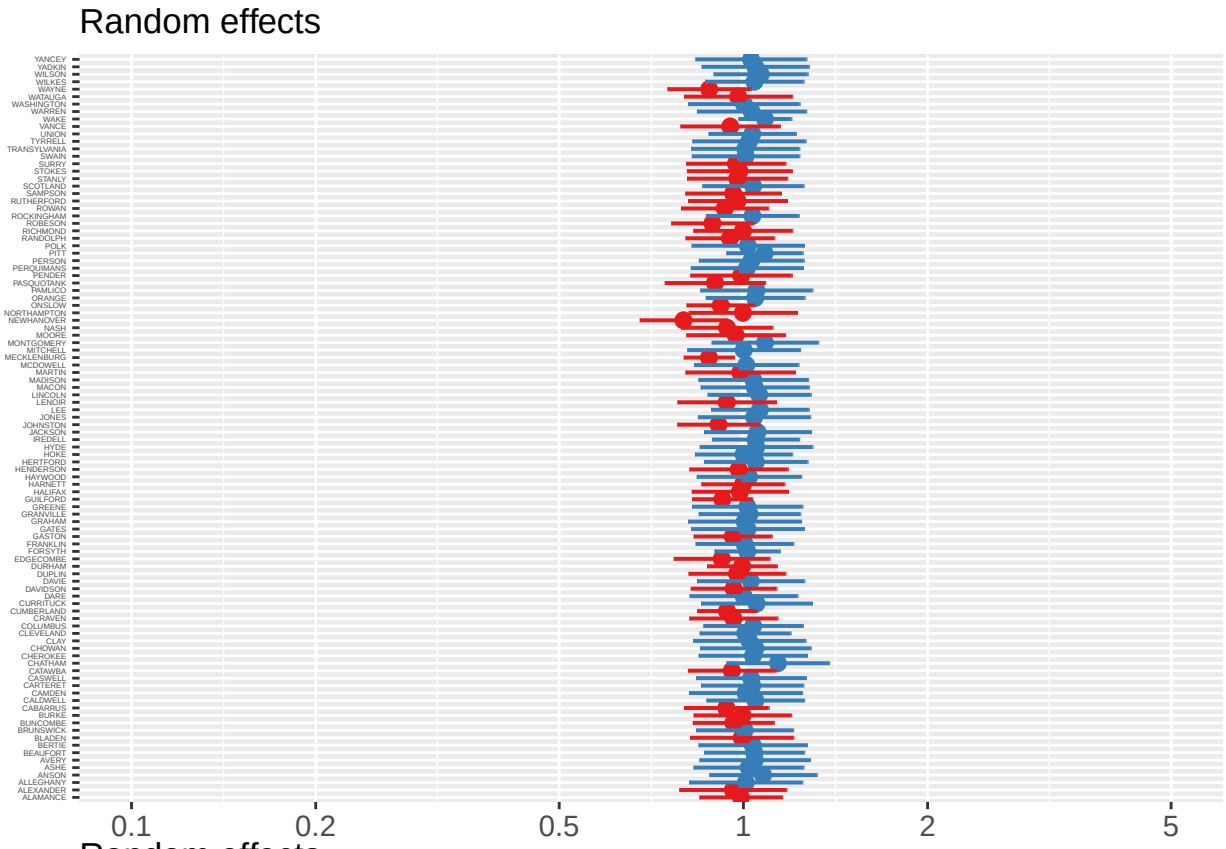
Model 3: Mortality Rate ~ (1 | County) + Race + Year + Median Household Income + Percent Low Birthweight + Percent Smokers

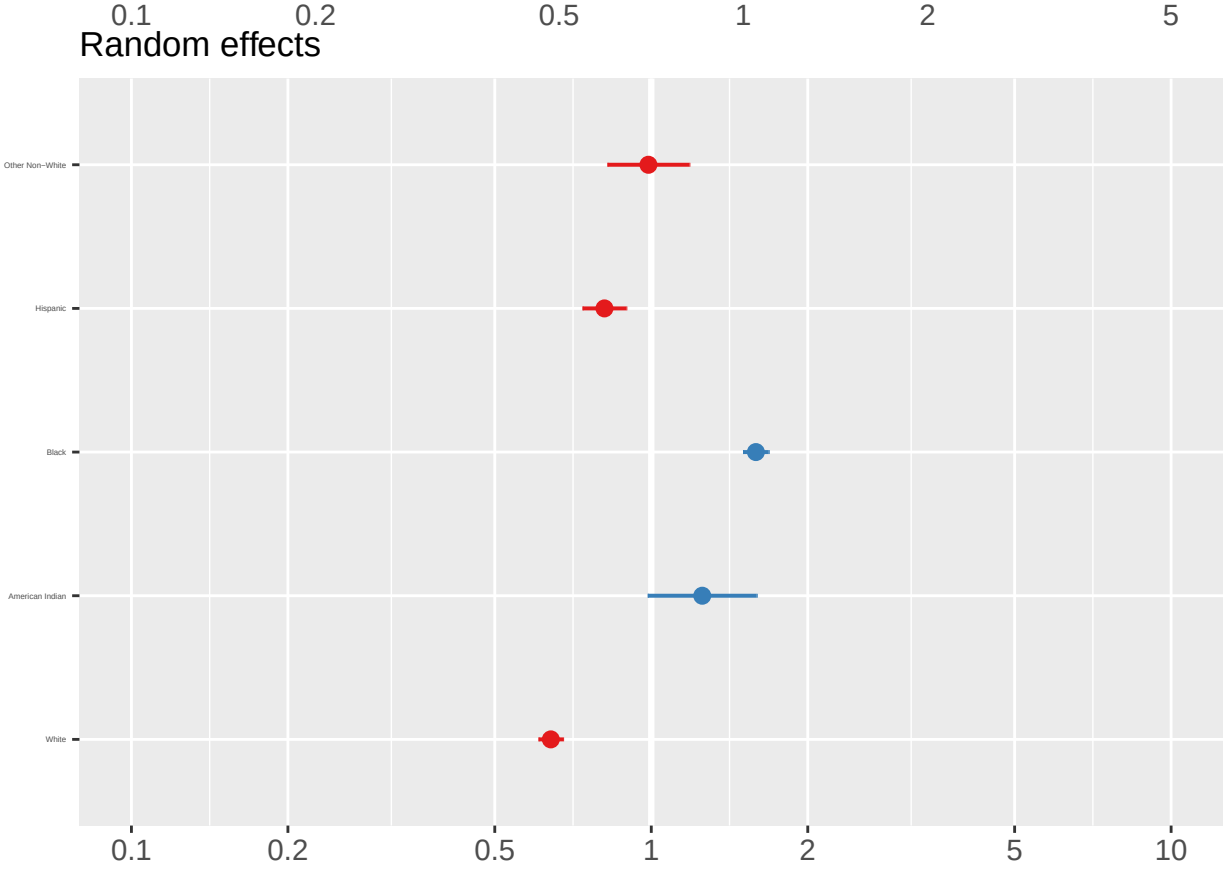
Model 4: Mortality Race ~ (1 | County) + Race + Year

We trained four separate models on the 2011-2015 data and then tested on the 2016 data. We used three statistics to assess model fit/accuracy: AIC, BIC, and AUC, which are shown above. All four models use random effects at the county level, but differ in how they model race and whether or not they make use of the additional features: median income, smoking, and low birth weight. Models 1 and 2 use a random effect to model race, while 3 and 4 use a fixed effect. Models 1 and 3 make use of the additional features, while 2 and 4 do not.

We found that AUC was fairly similar across the four models, but was slightly higher for models 3 and 4 which treat race as a fixed effect, suggesting this may help prevent overfitting. We also saw that AIC and BIC were slightly lower for the models that make use of the additional features, suggesting better fit. Because of the low AIC and BIC combined with a high AUC, we chose to move forward with model 3, treating race as a fixed effect and keeping the additional features.

Model Interpretation





	Estimate	Std. Error	Z	P-Value
Intercept	(4.393)	0.383	(11.474)	0.000
American Indian	0.710	0.136	5.217	0.000
Black	0.914	0.036	25.578	0.000
Hispanic	0.236	0.053	4.440	0.000
Other Non-White	0.434	0.098	4.429	0.000
2012	0.022	0.048	0.458	0.647
2013	(0.023)	0.049	(0.465)	0.642
2014	(0.038)	0.048	(0.775)	0.438
2015	(0.001)	0.048	(0.014)	0.989
Median Household Income	(0.017)	0.004	(4.410)	0.000
Percent Low Birth Weight	(0.034)	0.020	(1.681)	0.093
Percent Smokers	0.018	0.006	3.294	0.001

Random Effect	Number of Groups	Variance
County	100	0.0113

The model with race fixed revealed the following statistically significant effects:

For a given county, race, year of birth, percentage of births in a county that have low birth weight, and percentage of people in a county who are smokers, a one thousand dollar increase in median household income leads to 0.983x the probability of mortality (changed by a multiplicative factor of eˆ-0.0174654 = 0.983). This is a significant effect at a 0.001 significance-level.

The explanation of this effect (median income) seems intuitive. It could be the case that becoming wealthier leads to being able to afford better obstetrics (birthing and pregnancy) services, which would lead to a decreased probability of a baby dying.

For a given county, race, year of birth, median household income in a county, and percentage of people in a county who are smokers, a one percent increase in the percentage of births that have low birth weight leads to 0.967x the probability of mortality (changed by a multiplicative factor of eˆ-0.0337263 = 0.967). This effect has a p-value of 0.093, and does not meet the signficance threshold of 0.05.

The explanation of this effect (% low birth weight) seems counterintuitive. It could be the case that families that know they will have a baby with low birth weight (mothers that know they will give birth early) will prepare for such a situation by picking a higher-quality obstetrics service, which would lead to a decreased probability of a baby dying. The effect is also not significant.

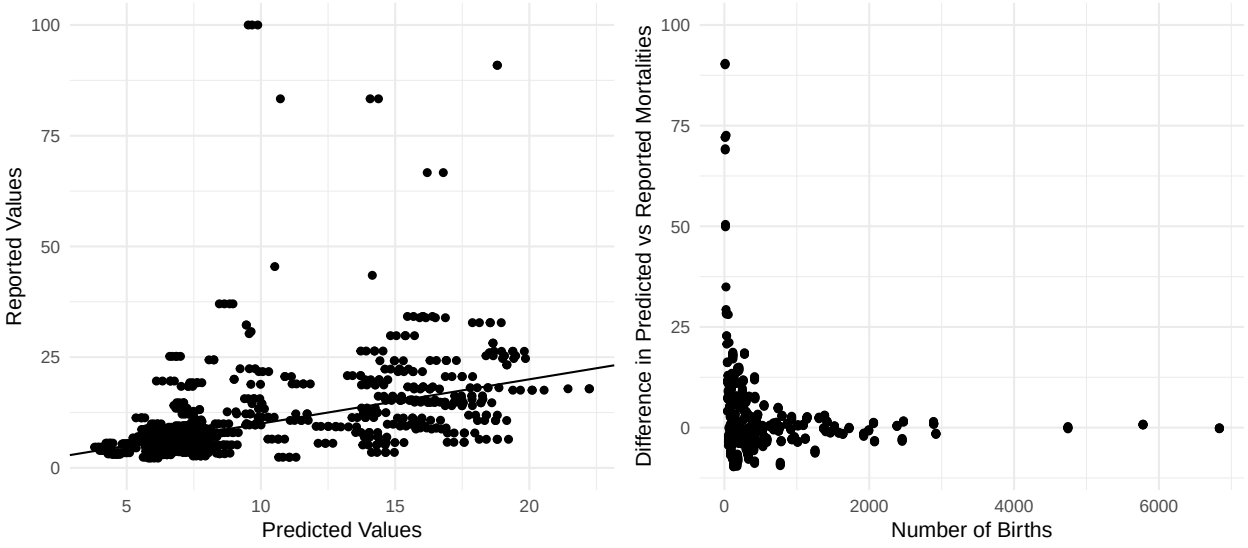
For a given county, race, year of birth, median household income in a county, and percentage of births in a county that have low birth weight, a one percent increase in the percentage of people who are smokers leads to 1.019x the probability of mortality (changed by a multiplicative factor of eˆ0.0183569 = 1.019). This is a significant effect at a 0.001 significance-level.

The explanation of this effect (% smokers) seems intuitive. Smoking is detrimental to health. It has scientifically been shown that smoking by a pregnant mother is especially bad for the unborn fetus.

In addition, we see that holding all else constant, we predict American Indian's and African American's to

experience the largest mortality rate, while we predict whites to experience the least. For example, holding all else equal, an African American would experience a mortality rate that is 2.5 times that of a white person (e^0.9138114 = 2.5).

The county level variability is 0.01127.



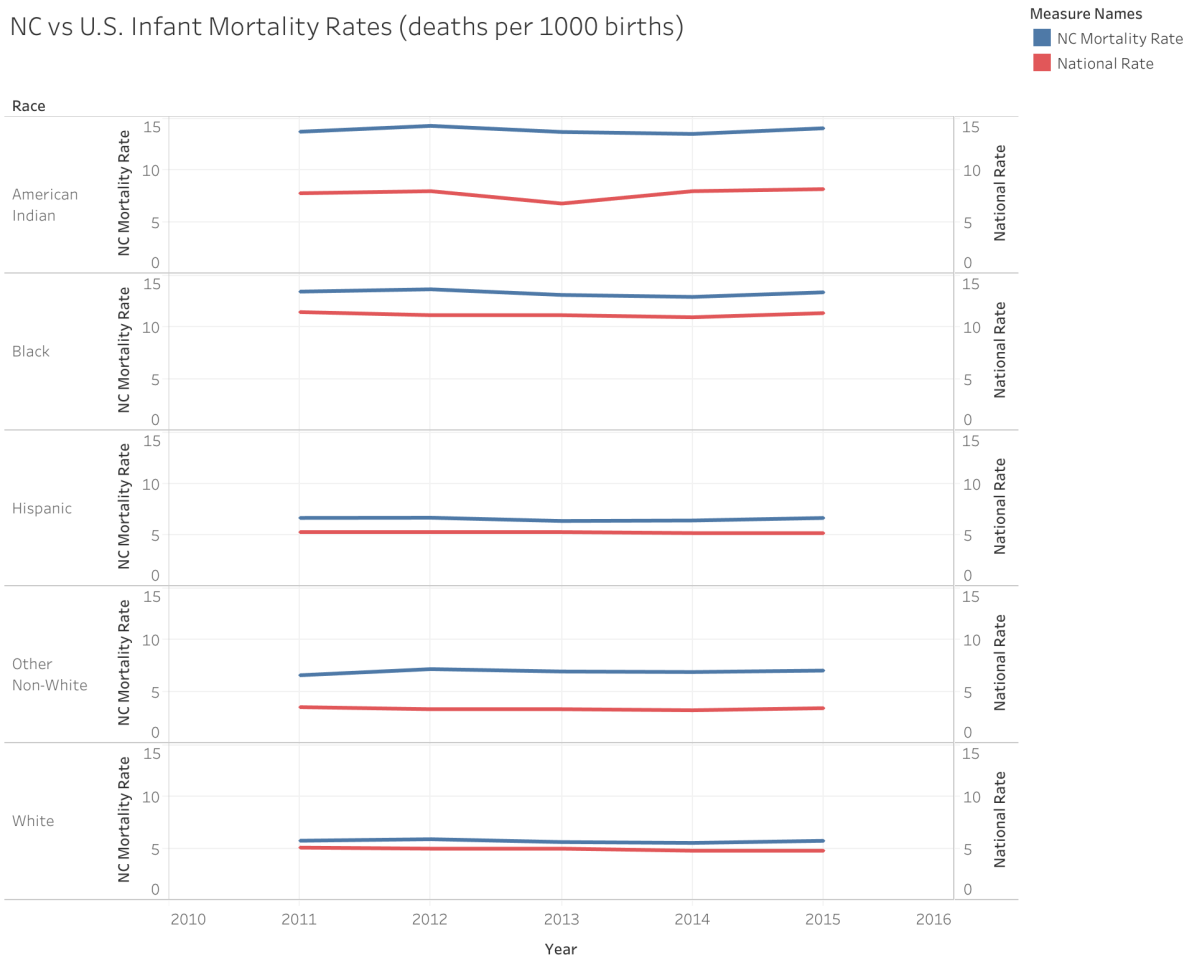
Comparison to NCHS Estimates

[todo: shift the focus of this section– it seems like we're saying that we're trying to predict the sample estimates NCHS uses, but in reality we want to talk more about how our predictions shrink the sample estimates in a robust way]

From the first chart, we see that our predictions do a strong job of estimating the reported values, and are centered around a slope of 1 (where predicted mortality rates equal reported mortality rates), so there is no clear bias.

Zooming into the chart (chart 2 is on a smaller scale), it seems that predicted mortality rates are most accurate when they are low, but become more inconsistent when they are larger. These large residual points are mostly indicative of reported values that draw on a small quantity of data points. We should also note that there are some points that lie outside of this graph, but again are mostly caused by a lack of data. This is demonstrated in the third chart that shows that as the number of births increase our predictions converge with the sample means. There is larger variance along the y-axis with points nearer to the left hand side of the x-axis (sample size is small). On the other hand, points nearer the right hand side (sample size is large) of the x-axis are consistently closely aligned to 0. Therefore, As the sample size increases, the difference in the predicted mortality rate versus reported mortality rate of that county converges to zero.

Conclusion and Trends



In the above visualization, we compare our predictions by race and year to U.S. rates overall. We obtained this national data from the Centers for Disease Control and Prevention. As shown by our exploratory analysis and our model, year dpoes not have a significant relationship with infant mortality, and unsurprisingly there are no large directional trends in infant mortality for any racial group over time in North Carolina. This is consistent with what has been seen in the U.S. in a whole, as the mortality rate for each race has remained essentially constant over this time period.

Infant mortality rates in North Carolina are higher than those of the U.S. overall for every category of race. Our North Carolina estimates are relatively similar to those for the U.S. overall for white, black and hispanic populations, but our predictions diverge greatly in the other non-white category and the American Indian category. One possible explanations for this large divergence is the relatively small sample size of American Indian and other non-white individuals in our data in comparison to white individuals. Further research is advised to explore why North Carolina has experienced higher rates of infant mortality. One possible routes of exploration would be looking into differences in poverty rates and access to medical care across various states.