440 CS3 IR2

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

10/16/2019

### Introduction

This case study aims to estimate mortality rates across North Carolina counties. We compare our predictions to that of the SCHS in 2015. In order to better inform our predictions, we obtained North Carolina health data from a study performed by the Robert Wood Johnson Foundation. Initially 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 to percent low birth weight, percent smokers, and median household income.

### EDA

#### Tableau

— INSERT TABLEAU SCREENSHOTS —

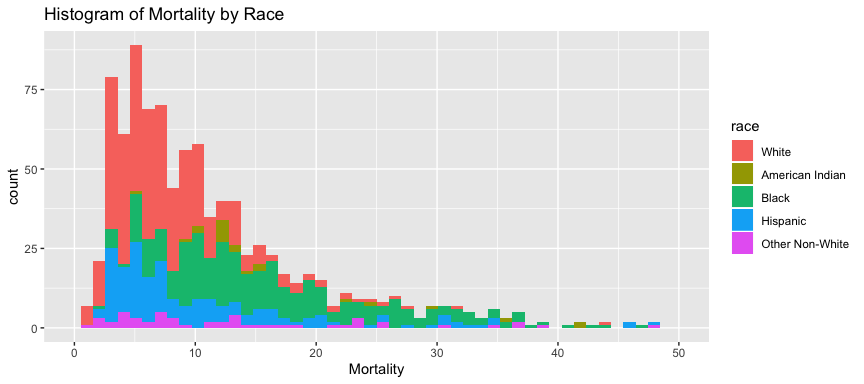
The visualizations above are captured from our tableau dashboard. The map depicts the **large variance in number of births and mortality rates across the different counties.** The largest counties, based on births, appear in lighter colors - they have relatively lower mortality rates, while we see darker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

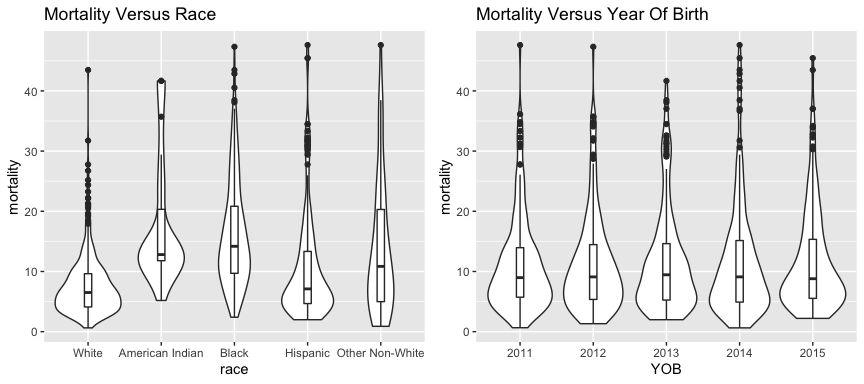
We also compare North Carolina versus National mortality rate trends throughout time by race. We note that **the trend of White mortality rate in NC is largely similar to that on a National level**. On the other hand, there are **large divergences in American Indian mortality rates**. While it has stayed relatively constant throughout the national level, there was a persistent rise in American Indian mortality rates throughout 2011 to 2013, then followed by a particularly sharp decline 2013 onwards.

#### Exploring Categorical Variables

ggplot(eda, aes(x = mortality, fill = race)) +  
 geom\_histogram(bins = 50) +  
 scale\_x\_continuous(name="Mortality", limits=c(0, 50)) +  
 ggtitle("Histogram of Mortality by Race")



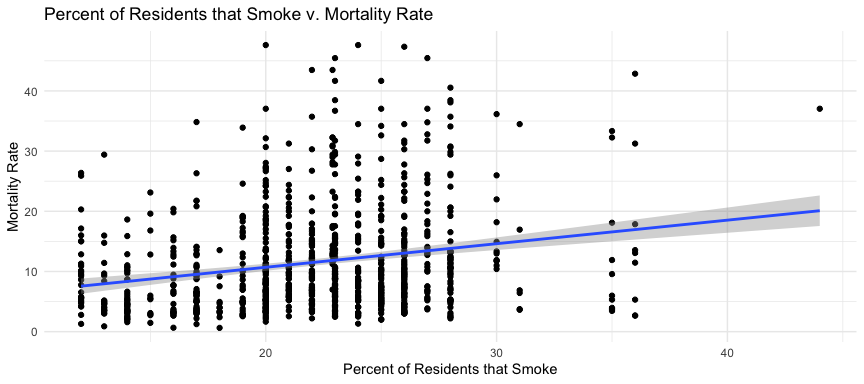
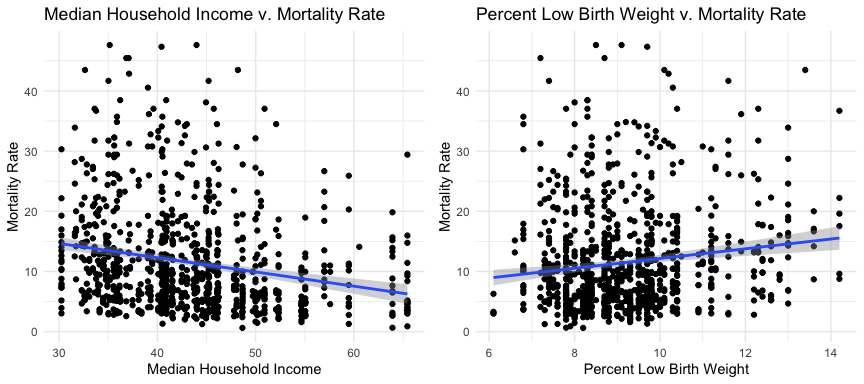
p2 <- ggplot(eda, aes(x=race, y=mortality)) +   
 geom\_violin(trim=TRUE)   
vio1 = p2 + geom\_boxplot(width=0.1) + ggtitle("Mortality Versus Race")  
eda$YOB <- as.factor(eda$YOB)  
p <- ggplot(eda, aes(x=YOB, y=mortality)) +   
 geom\_violin(trim=TRUE)   
vio2= p + geom\_boxplot(width=0.1) + ggtitle("Mortality Versus Year Of Birth")  
grid.arrange(vio1, vio2, nrow = 1)

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

The histogram again shows a significantly more ample amount of data points on White births compared to 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 histogram - 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

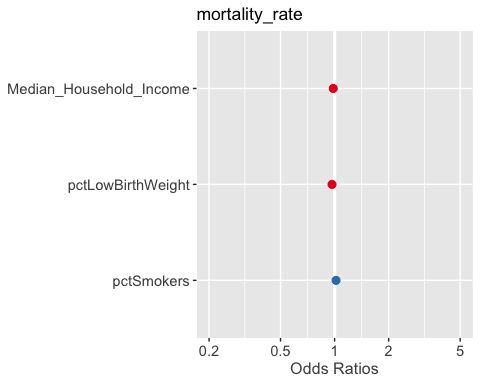
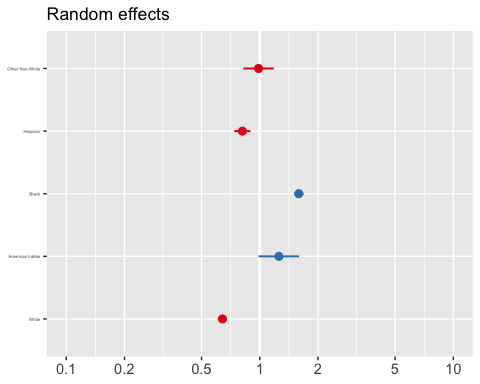
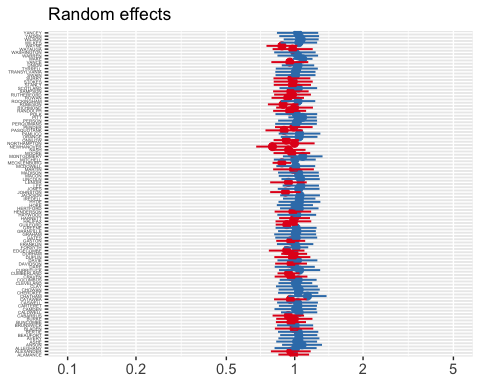
 Median Household Income: There is a negative relationship between the median income of a county and the mortality rate in that county. Specifically, the two variables have a negative correlation of -0.2169295. This suggests that counties with higher median income tend to have lower mortatility rates.

Percentage Low Bith Weight: There is a positive relationship between the percentage of births that had a low birth weight in a county and the mortality rate in that county. Specifically, the two variables have a positive correlation of 0.1377997. This suggests that counties with a higher percent of births that had a low birth weight tend to have higher mortatility rates.

Percentage Low Bith Weight: There is a positive relationship between the percentage of people that smoke in a county and the mortality rate in that county. Specifically, the two variables have a positive correlation of 0.2182623. This suggests that counties with a higher percent of smokers tend to have higher mortatility rates.

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =  
## control$checkConv, : Model failed to converge with max|grad| = 0.119203  
## (tol = 0.001, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unidentifiable: very large eigenvalue  
## - Rescale variables?

 These charts show the differing intercepts for the differing counties and races.

### Model Interpretations

The model revealed the following statistically significant effects:

For a given county, race, and percentage of births in a county that have low birth weight, a one thousand dollar increase in median household income leads to 0.975x the probability of mortality (changed by a multiplicative factor of e^-0.025122  = 0.975).

The explanation of this 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, and median household income in a county, a one percent increase in the percentage of births that have low birth weight leads to 0.954x the probability of mortality (changed by a multiplicative factor of e^-0.046546  = 0.954).

The explanation of this 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.

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.

The county level variability is 0.01728 and the race level variability is 0.11702.

### Comparison of Models

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

## Setting levels: control = 0, case = 1

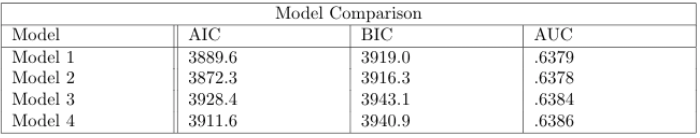
## Setting direction: controls > cases

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases



Models:

Model 1: mortality\_rate ~ (1 | COUNTY) + (1 | race) + Median\_Household\_Income + pctLowBirthWeight + pctSmokers

Model 2: mortality\_rate ~ (1 | COUNTY) + (1 | race)

Model 3: mortality\_rate ~ (1 | COUNTY) + race + Median\_Household\_Income + model\_racefixed: pctLowBirthWeight + pctSmokers

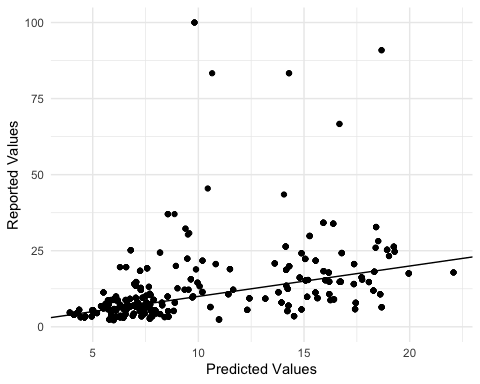
Model 4: mortality\_rate ~ (1 | COUNTY) + race

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. 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 tace as a fixed effect and keeping the additional features.

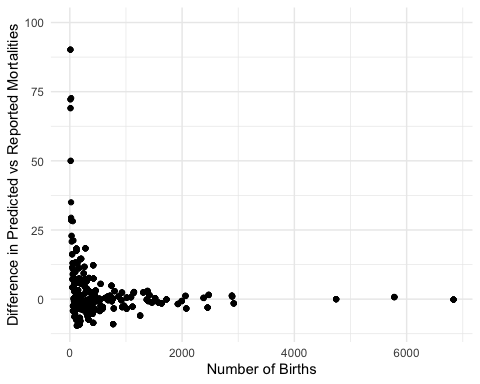
our\_table = mortality\_fitted %>%  
 filter(YOB == 2015) %>%  
 select(fitted.values, race, COUNTY) %>%  
 mutate(predicted\_mortality = 1000\*fitted.values) %>%  
 select(-fitted.values) %>%  
 spread(race, predicted\_mortality)  
  
  
our\_table\_as\_list = mortality\_fitted %>%  
 select(COUNTY, race, fitted.values) %>%  
 mutate(predicted\_mortality = 1000\*fitted.values) %>%  
 select(-fitted.values)  
  
schs\_table\_as\_list = SCHS\_Table\_2015 %>%  
 gather(race, mortality, 2:length(SCHS\_Table\_2015)) %>%  
 na.omit()  
  
table\_comparisons = merge(our\_table, SCHS\_Table\_2015, by = "COUNTY")  
  
comparisons = merge(our\_table\_as\_list, schs\_table\_as\_list, by = c("race", "COUNTY"))  
  
comparisons = merge(comparisons, mortality %>% filter(YOB == 2015))  
  
comparisons = comparisons %>%  
 mutate(difference = mortality - predicted\_mortality)  
  
write.csv(comparisons, "comparisons.csv")  
  
ggplot(comparisons, aes(x = predicted\_mortality, y = mortality)) +  
 geom\_point() +  
 ylim(c(0, 100)) +  
 geom\_abline(slope = 1, intercept = 0) +  
 xlab ("Predicted Values") +  
 ylab ("Reported Values") +  
 theme\_minimal()

## Warning: Removed 4 rows containing missing values (geom\_point).



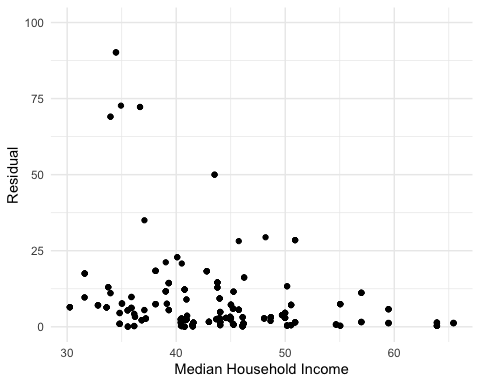
ggplot(comparisons, aes(x = births, y = difference)) +  
 geom\_point() +  
 ylim(c(-10, 100)) +  
 xlab("Number of Births") +  
 ylab("Difference in Predicted vs Reported Mortalities") +  
 theme\_minimal()

## Warning: Removed 18 rows containing missing values (geom\_point).



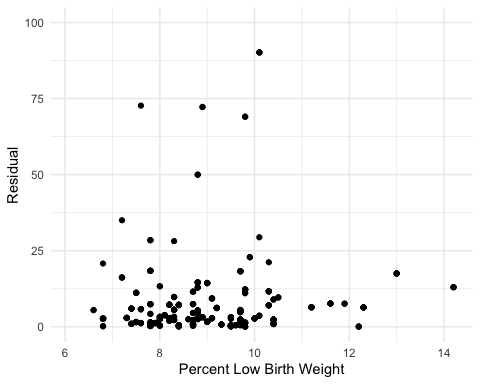
ggplot(comparisons, aes(x = Median\_Household\_Income, y = difference)) +  
 geom\_point() +  
 ylim(c(0, 100)) +  
 xlab ("Median Household Income") +  
 ylab ("Residual") +  
 theme\_minimal()

## Warning: Removed 429 rows containing missing values (geom\_point).



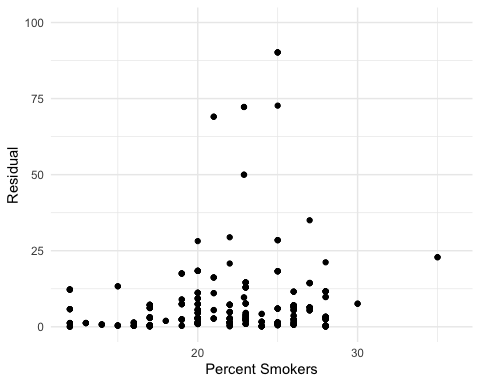
ggplot(comparisons, aes(x = pctLowBirthWeight, y = difference)) +  
 geom\_point() +  
 ylim(c(0, 100)) +  
 xlab ("Percent Low Birth Weight") +  
 ylab ("Residual") +  
 theme\_minimal()

## Warning: Removed 429 rows containing missing values (geom\_point).



ggplot(comparisons, aes(x = pctSmokers, y = difference)) +  
 geom\_point() +  
 ylim(c(0, 100)) +  
 xlab ("Percent Smokers") +  
 ylab ("Residual") +  
 theme\_minimal()

## Warning: Removed 429 rows containing missing values (geom\_point).

 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.

### Discussion of Differences

XXXtableau screenshot

Our table of predicted mortality was pretty close to the table given by the 2015 NC Infant Mortality Report. For all counties except for Gates (23 births observed), Alleghany (22 births observed), Pamlico (30 births observed), Hyde (22 births observed), and Avery (24 births observed), the absolute difference between predicted and actual is less than 16. These aforementioned counties with absolute differences greater than 16 all have a very low number of births observed. The graph “Difference in Predicted vs Reported Mortalities” vs “Number of Births” also shows this trend. The model is much less accurate when number of births is low. As number of births gets higher (past about 500 births observed), the difference becomes very small (<~10). This makes sense, as there is less data to train the model on if there are fewer births observed. Of note is that there are about 30 counties where the difference was less than 1.

These charts suggest that SCHS puts too much weight on the sample mean of counties where there are relatively few data points. We urge the SCHS to change their practices and use pooling in order to get a more realistic estimate amidst counties with large deviations. For example, simply looking at Gates county, we see a standard deviation of ~272 across the three years of mortality data(mortality rate given 1000 births). This variance can be handled using pooling in which one weights the credibility of the sample mean by the total number of births.

Furthermore, we strongly urge the SCHS to include the features, percent smokers, median household income, and percent low birth weight in their model. These features help further eliminate variance and improve upon our predictions of mortality rates. These features all have been shown to have strong correlations with mortality rates and help us further account for county level differences beyond simply the sample estimates. They provide a systematic way to categorize/understand the mortality rates at the county level.