440 CS3 IR2

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

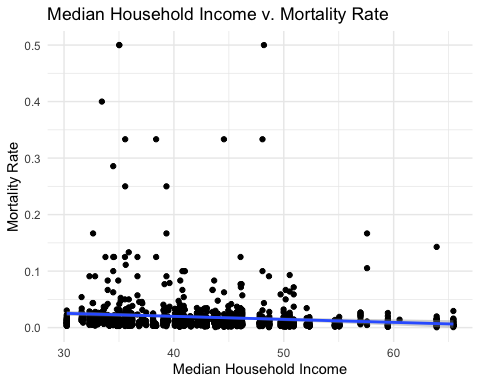
10/16/2019

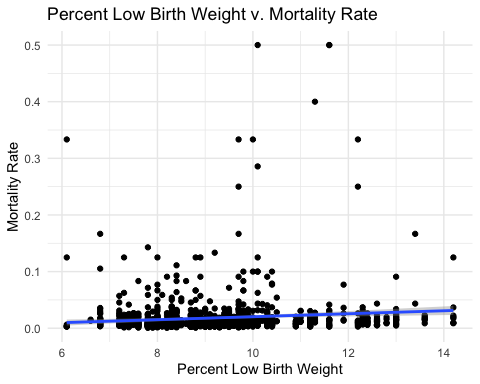
### EDA

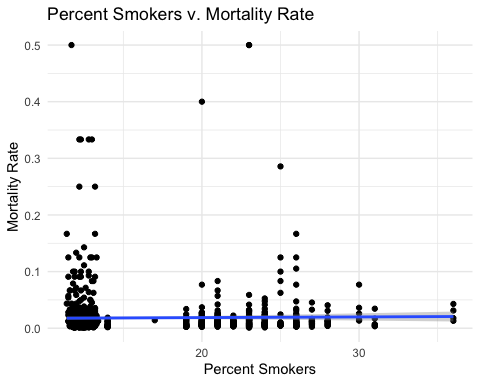
XXXtableau screenshot

The bar graph above clearly depicts the variance in the sample size for each race in North Carolina. As shown, there are substantially more white births (382,602) than any other race. Nonetheless, the number of births in each year from 2011 to 2016 stays fairly constant.

In addition, the variance in the sizes of each circle on the map above is indicative of the high variation in the number of births in each county. This information, 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. Furthermore, the colour of each circle on the map shows that the mortality rate, in fact, varies greatly by county.

 As illustrated by the scatter plot above, there is distinct 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 -0.1060368. This implies that counties with higher median income tend to have lower mortatility rates.

 As illustrated by the scatter plot above, there is distinct negative relationship between the percent 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 0.0978085. This implies that counties with a percent of births that had a low birth weight tend to have higher mortatility rates.

 As illustrated by the scatter plot above, there is distinct negative relationship between the proportion of a county that smokes and the mortality rate in that county. Specifically, the two variables have a negative correlation of 0.0159182. This implies that counties with more smokers tend to have lower mortatility rates.

### Model Selection

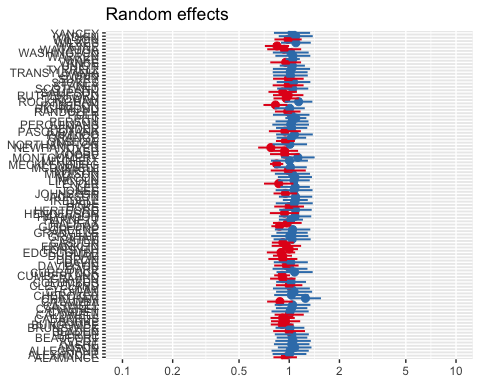
In order to accurately predict the mortality rates, we elected to use a multilevel logsitic model. We allowed for random effects by county due to the large number of counties (100), as well as by race (as there are large samples across races). We implemented fixed effects for median household income, the percent of low birth weights in a county, and the proportion of a county that smokes. Therefore, our model allows us to account for clustering of mortality rates at the county level and by race, while allowing for variation due to the aforementioned fixed effects.

We chose to use a logit model as it aptly bounds the outcome variable (the mortality rate) between 0 and 1. Furthermore, the multilevel logistic model incorporates some pooling and thus allows us to benefit from borrowing information across counties and races, while still accounting for variation across counties and races.

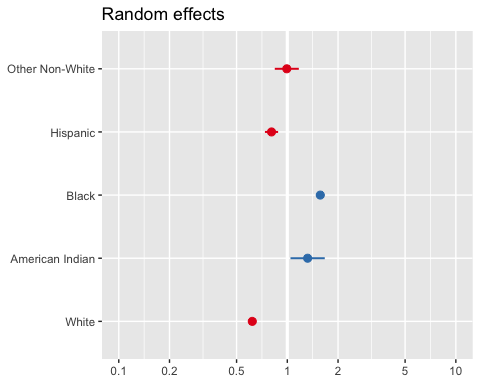
## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: binomial ( logit )  
## Formula:   
## mortality\_rate ~ (1 | COUNTY) + (1 | race) + Median\_Household\_Income +   
## pctLowBirthWeight + pctSmokers  
## Data: mortality  
## Weights: births  
##   
## AIC BIC logLik deviance df.resid   
## 4680.8 4711.3 -2334.4 4668.8 1174   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.3396 -0.5813 0.0613 0.7934 9.5817   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## COUNTY (Intercept) 0.01723 0.1313   
## race (Intercept) 0.11584 0.3403   
## Number of obs: 1180, groups: COUNTY, 100; race, 5  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.281285 0.342604 -9.577 < 2e-16 \*\*\*  
## Median\_Household\_Income -0.023500 0.003474 -6.764 1.34e-11 \*\*\*  
## pctLowBirthWeight -0.043913 0.019542 -2.247 0.0246 \*   
## pctSmokers 0.007530 0.004383 1.718 0.0858 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) Md\_H\_I pctLBW  
## Mdn\_Hshld\_I -0.741   
## pctLwBrthWg -0.770 0.523   
## pctSmokers -0.287 0.136 0.055



## [[1]]



##   
## [[2]]

  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).

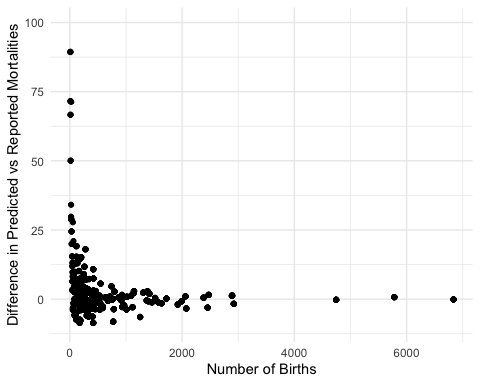
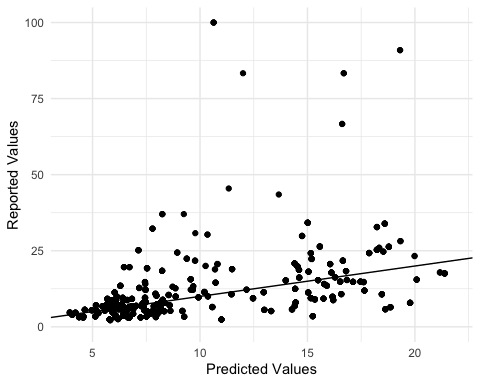
The explanation of this seems intuitive. It could be the case that becoming wealthier leads to being able to afford better obstetrics (birthing and pregnancy) services, which would lead to a decreased probability of a baby dying.

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.

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

### Comparison

 We see that our predictions do a strong job of estimating the reported values, and are centered around a slope of 1, so there is no clear bias The large residual points are mostly indicative of reported values that draw on a small quantity of data points. We must 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.

### 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.

test test