440 CS3 IR2

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

10/16/2019

### Introduction

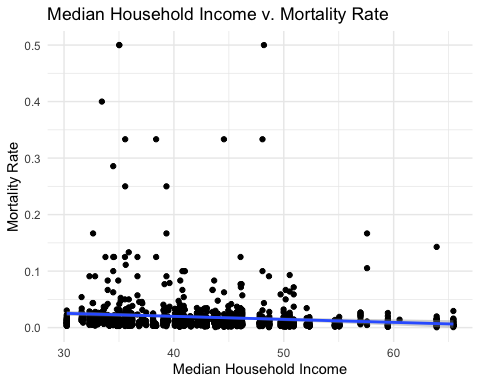
This case study aims to estimate mortality rates across North Carolina counties. We compare our predictions to that of the SCHS in 2015. In order to better inform our predictions, we obtained North Carolina health data from a study performed by the Robert Wood Johnson Foundation. Initially we explored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

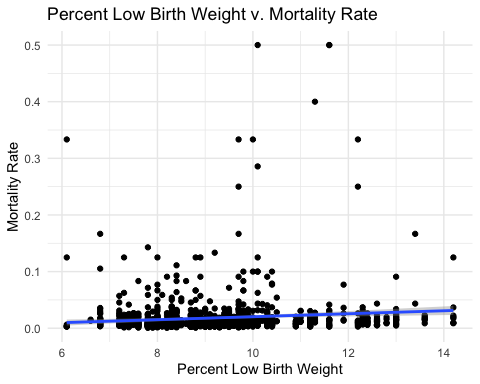
### EDA

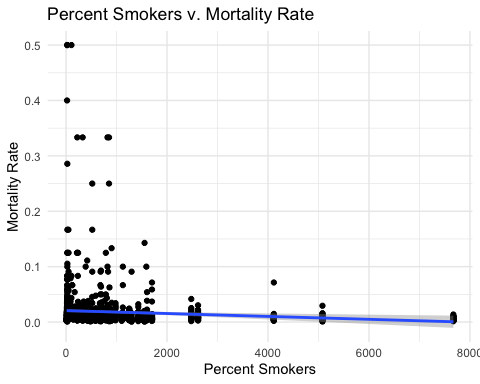
XXXtableau screenshot

The map above clearly depicts the large variance in number of births and mortality rates across the different counties. The largest counties, based on births, appear to have relatively low mortality rates, while we see spikes in mortality rate in many of those with smaller populations. As shown, there are substantially more white births (382,602) than any other race, specifically we have very few data points for American Indian’s and people that identify as Other Non-White. It is important to note that 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 -0.0934415. 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.

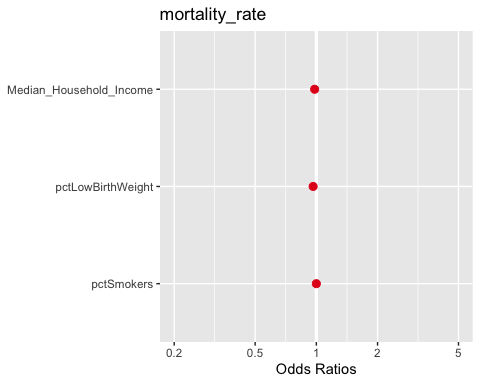
We chose to use a logit model as it aptly bounds the outcome variable (the mortality rate) between 0 and 1. Furthermore, the multilevel logistic model incorporates some pooling and thus allows us to benefit from borrowing information across counties and races, while still accounting for variation across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

## Warning: Some predictor variables are on very different scales: consider  
## rescaling

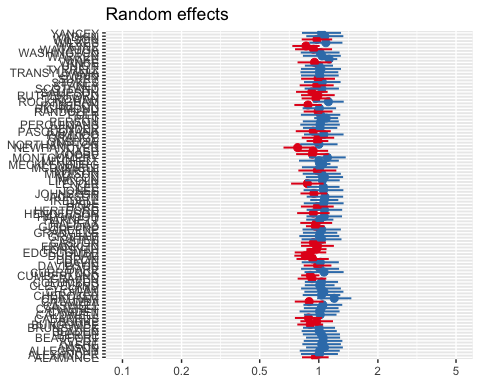
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =  
## control$checkConv, : Model failed to converge with max|grad| = 0.0119545  
## (tol = 0.001, component 1)

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

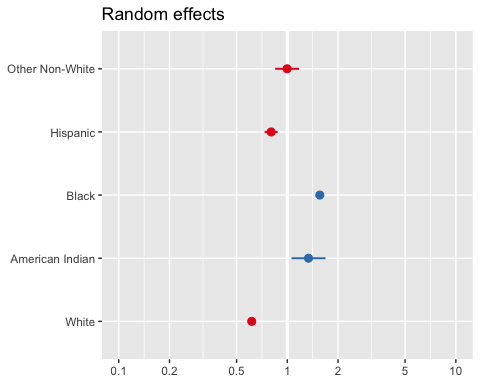
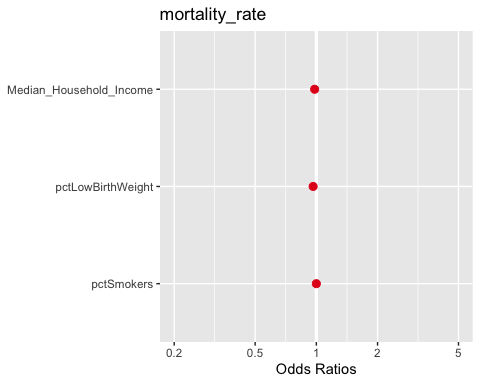
## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: binomial ( logit )  
## Formula:   
## mortality\_rate ~ (1 | COUNTY) + (1 | race) + Median\_Household\_Income +   
## pctLowBirthWeight + pctSmokers  
## Data: mortality  
## Weights: births  
##   
## AIC BIC logLik deviance df.resid   
## 4674.3 4704.7 -2331.1 4662.3 1174   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.3349 -0.5922 0.0507 0.7978 9.2802   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## COUNTY (Intercept) 0.01439 0.1200   
## race (Intercept) 0.11788 0.3433   
## Number of obs: 1180, groups: COUNTY, 100; race, 5  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.363e+00 3.292e-01 -10.218 < 2e-16 \*\*\*  
## Median\_Household\_Income -1.928e-02 3.645e-03 -5.289 1.23e-07 \*\*\*  
## pctLowBirthWeight -3.629e-02 1.913e-02 -1.897 0.05782 .   
## pctSmokers -4.759e-05 1.589e-05 -2.994 0.00275 \*\*   
## ---  
## 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 -0.788 0.533   
## pctSmokers 0.199 -0.407 -0.136  
## fit warnings:  
## Some predictor variables are on very different scales: consider rescaling  
## convergence code: 0  
## Model failed to converge with max|grad| = 0.0119545 (tol = 0.001, component 1)  
## Model is nearly unidentifiable: very large eigenvalue  
## - Rescale variables?  
## Model is nearly unidentifiable: large eigenvalue ratio  
## - Rescale variables?



## [[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.

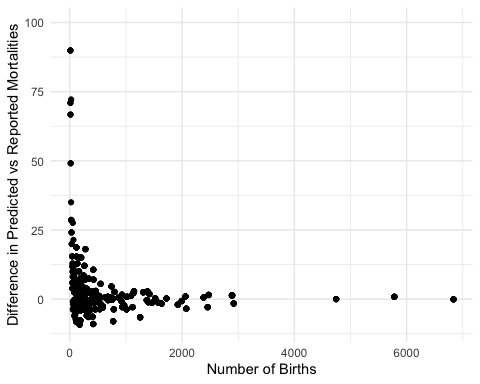
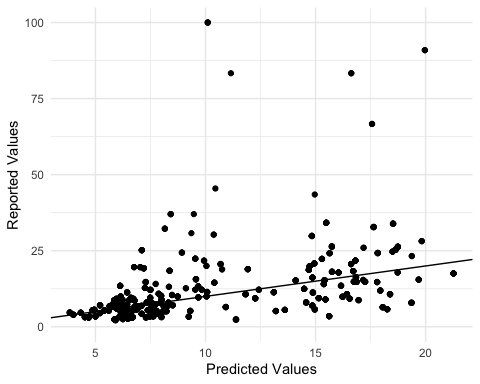
For a given county, race, and median household income in a county, a one percent increase in the percentage of births that have low birth weight leads to 0.954x the probability of mortality (changed by a multiplicative factor of e^-0.046546  = 0.954).

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.

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

 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.