Untitled

Brandon Kill

5/6/2021

Executive Summary

Introduction

Our primary research interest was in predicting the number of COVID-19 deaths for counties in Indiana (as of 4/28/2021). We had a variety of predictor variables at our disposal. The information we felt could be most helpful in predicting deaths were data related to mask wearing; vaccine distribution; COVID-19 tests and cases; and county demographic information such as total population, political affiliation (based on 2016 Presidential Election results), and both the number and proportion of Senior Citizens (65+) in the county. Although demographic information related to race were available, the truth of the matter is that increased racial diversity would highly correlate to county size, and would be unlikely to be a significant predictor. We leave it as an exercise to the reader to see if this is truly the case.

Data Gathering and Manipulation

We had to merge several .csv files and do quite a bit of data manipulation in order to get everything we needed. You will find the URLs to the data on the final page of this document. Most of the data sets were merged by county FIPS code, known as `LOCATION_ID` in our primary data frame.

The mask data from the *New York Times* consists of the proportion of respondents in each county that wear a mask never, rarely, sometimes, frequently, and always. The file `covid_report_county.csv` contained the number of deaths, positive cases, and tests administered in each county up to the date it was accessed. The county vaccination demographics file contains information about total number of vaccine doses administered (1st dose and 2nd dose for two-dose vaccines, or just Single Dose for one-dose vaccines), as well as the number of people by race who are fully vaccinated. Many counties had "Suppressed" their vaccine dose counts for non-white citizens, but for small, rural counties, it is possible that zero non-white citizens have been vaccinated in those counties, and thus they had nothing to report. Because of this, we replaced "Suppressed" with a 0, and then aggregated vaccination figures across race in each county. `idwd_data_31.csv` contained the number of senior citizens residing in each county. The elderly were at much higher risk of death due to COVID-19 than any other age bracket (according to the CDC). The `pres_votes.csv` file contains the number of votes for Donald Trump, Hillary Clinton, and Other candidates in the 2016 election. We felt either the number or the proportion of the vote for each candidate could be significant in predicting death counts, as Trump supporters have been quite vocal about not taking the COVID-19 pandemic seriously. And finally, we wanted to try and categorize each county by it's size, and we stumbled upon the 2013 Urban-Rural Classification Scheme (URCS) for U.S. counties. It uses an ordinal 1-6 scale to categorize a county as a Large Central Metropolitan Area (1) down to a Non-Core (fully rural) area (6). Most of the counties in Indiana are rated 4 through 6; however, a few counties containing Indiana's largest cities do rate as a 2 or 3. Marion County (Indianapolis) is the only level 1 county in the state.

Model Selection

In every model, the count of the number of deaths due to COVID-19 is the response. We made two different fixed effects models and two different mixed models, one of each being a Binomial count and a Poisson count model. When viewing a scatter plot of the original predictor variables, we noticed that all of the variables that were counts (number of cases, number of votes for Clinton, etc) were right-skewed and would probably benefit from a log transformation. The log-counts were much closer to normally distributed and were thusly added to our data frame as possible predictors. **All raw counts are listed in the data frame as var_name.x, and all log-counts are listed in the data frame as var_name.y** The summaries and plots of our four models are in the last few pages, right before the Data Sources page.

Fixed-Effects Models

When modeling counts, the two primary general linear models to use are Binomial Counts and Poisson Counts. Both models map to logit function, but follow different probability distributions. At first, we thought it would be most appropriate to use a Binomial model, since one either does or does not die of COVID, and the number of "failures" is easy to calculate when total population is known. However, and we will see this later, the Poisson model tends to fit better. This makes sense to us, since the Poisson distribution is used to model counts of sparse or rare events, and dying of a new disease would probably fall under the "rare event" umberella.

When we made our Fixed-Effects models, we were able to get a Binomial model teeming with significant predictors. The model we are referencing is `mod1.1.3`. However, we felt we could do better, and decided to look for interactions in a few specific places (based on gut instinct). We had a feeling that the county URCS code could be interacting with support for Trump, the proportion of the county that is older, and reluctance to wear a mask. As we whittled down our model, all but two subsets of the four-way interaction between `RARELY:olderprop:TrmpProp:2013 code` were highly significant in our model. Unfortunately, every interaction containing URCS code 6 gave had NA values due to singularities. We were unable to diagnose the problem, which is unfortunate as exactly 1-in-4 counties in Indiana are assigned code 6. The Poisson model was generated in a similar way, but with slightly better results and a few extra predictors. This gave us an AIC of 798.23 which turned out to be the lowest of all four models.

Mixed-Effect Models

We decided to see if the variation between death rates in counties could be better explained as a random effect. We tried both the county itself (`LOCATION_ID`) and the URCS code (`2013 code`) as possible random effects. We even used a bootstrap likelihood ratio test to see if a model where `LOCATION_ID` was nested inside of `2013 code` was better than a model with just `2013 code`. The p-value was 0, so as we were model selecting, our models included both `2013 code` and `LOCATION_ID` nested inside `2013 code`. Once we whittled down the predictors, the `2013 code` random effect had a standard deviation of zero, so we removed it, and our final Poisson Mixed-Effects model (`mod26.off`) appears better fitting without it. This was the case for both the Binomial and Poisson regression procedures. When we offset the Poisson model by the log of the population, it made the model regress to a Poisson proportion instead of a count - good when the county population sizes are very different. This generated better-fitting Poisson models across the board.

Conclusion

The standard deviation of our random effect of `LOCATION_ID` in our Poisson model was most definitely non-zero (~0.2). However, the mixed-effects Poisson model had an AIC that was about 30 points higher than the fixed-effects Poisson model, tempting us to default to the fixed effect model. On the other hand, the mixed effect model did not need any of the interaction terms that the fixed effect model required. Because of this, the mixed effect Poisson model was much simpler than the fixed effect model, temping us to want to default to the mixed effects model instead.

In summary, due to the NA's in the output of the best-fitting Fixed Effects Poisson Model, in addition to the sheer complexity of the interactions in the model (suggesting possible over-fitting), we will use the Poisson Mixed Effects Model with `LOCATION_ID` as a random effect with $\sigma_{LID} = 0.1972$. Here is our regression equation, where $\hat{y}_i$ is the log-odds of a person dying in Indiana county i.

$\hat{y}_i$ = -8.8337 + 13.9832×`prop_cases` - 1.1864×`COVID_COUNT.y` + 1.5559×`Older (65 plus).y` + 0.4408×`TrmpVote.y` - 1.0827×`TotalVote.y` + 0.3913×`fully_vaccinated.y`

So, the log-odds of a person dying from COVID-19 in each county increases substantially with the proportion of people in the county with a positive test (prop_cases), increases quite a bit with the log-count of senior citizens, and increases some as the log-count of votes for Trump and log-counts of the number of fully-vaccinated citizens increases. It seems weird that vaccination counts would be positively correlated with deaths, but then again, vaccination counts are probably more indicative of the size of the overall population and even the size of the at-risk population in each county (since at-risk folks were vaccinated first). The log-odds of dying from COVID-19 decrease as the log of COVID count and log of Total Vote increase. Similar to the reverse signage of log-fully-vaccinated, the log COVID count and log Total Vote are probably more indicative of the size of the county, and larger counties tend to vote Democratic and take COVID precautions more seriously.