Appendix

Rena Cohen

12/8/2020

Data Cleaning and Compilation Process

We compiled data from a total of 8 different sources in order to obtain the combination of demographic, political, and public health data that we desired. We began with the data from the NYT, which was described in our main paper. Most of the cleaning was simply a matter of renaming variables, making raw counts into rates, creating indicators for countywide and statewide mandates, and merging based on county FIPS code. Once we had compiled our predictor variables and done exploratory data analysis, we transformed them within the dataset in order to help us keep track of them, renaming accoordingly (so a variable like density, which required a log transformation, became log_density in the dataset). For more information on each of our predictors and what datset they came from, see the table below:

Name	Source
countyfp	New York Times
always	New York Times
frequently	New York Times
sometimes	New York Times
rarely	New York Times
never	New York Times
cases_02	New York Times
cases_14	New York Times
cases_27	New York Times
pop_2019	United States Census Bureau
ru_continuum	United States Census Bureau
density	county_level_election.csv from class
pct_less_than_hs	2014-18 American Community Survey
pct_hs	2014-18 American Community Survey
pct_some_college	2014-18 American Community Survey
pct_college	2014-18 American Community Survey
pct_poverty	U.S. Census Bureau, Small Area Income and Poverty Estimates (SAIPE) Program
pct_female	U.S. Census Bureau
pct_black	U.S. Census Bureau
pct_native	U.S. Census Bureau
pct_hispanic	U.S. Census Bureau
pct_seniors	U.S. Census Bureau
pct_trump_2016	county_level_election.csv from class
pct_trump_2020	Scraped by GitHub user tonmcg from Fox News, Politico, and New York Times
dem_governor	National Governor's Association
state_mandate	Axios
county_mandate	Harris Institute of Public Policy

Handling Missing Data

As mentioned in our paper, we had 127 rows with at least one missing predictor value. 28 of these rows were from Alaska, all of which were missing county-level data for both the 2016 and 2020 elections due to the fact that Alaska reports election results using boroughs instead of counties. Because partisanship was such a key variable in many of our models and because this data was not easily accessible for Alaska (even the cleaned data sets we used in class did not have it), we decided to exclude Alaska from our models entirely, understanding that any conclusions we came to would not be able to be reasonably generalized to this state.

Beyond the total omission of political data from Alaska, the next most worrying part of our predictor data was the fact that case rates were missing for five extremely populous counties in New York City with over a million residents each. Upon further investigation, we realized that this was because (for reasons unknown to us) New York City reports its COVID-19 data in most sources as a full unit rather than as the 5 different counties it can be broken up into. In order to remedy this, we manually inputed values for current December case rates by searching for them online, allowing us to include them in our mixed model for predicting current case rates from mask-wearing in July. Unfortunately, we could not easily find COVID-19 data for these counties available in July at the county level, but we ultimately did not end up using this predictor in any of our final models, limiting the adverse impact of this missing data.

The rest of the missing data primarily came in the form of missing case/death rates from July in small, rural counties; more than likely, these counties simply were not publishing information about their case rates at that time, or they did not yet have any cases at all. After conducting a two sample t-test, we determined that these counties were more rural, male, Republican, older, and slightly more college educated, more white, and less likely to wear masks than the counties in the full data set. There was not a statistically significant difference in the poverty rate. Although this finding implicates that the data we ended up using slightly underestimates small, rural, Republican counties, the entire premise of analyzing data at the county level inherently *overepresents* these counties, because they have far smaller populations than urban, Democratic counties. To illustrate, even though counties with missing data (not including NY or AK) represented 2.96% of the *counties* in our dataset, they contained just 0.11% of the current US *population.* While this could have been addressed by weighting our data by county population, doing so negatively affected many of our diagnostic plots (i.e. made them nonlinear) and prevented our mixed model from converging. All that is to say, we did not end up weighting by population, so our missing county-level data slightly underestimates the influence of rural counties, but puts us a little closer to the reality of the American population at large.

Models and Diagnostic Plots

Section 2: Mask Mandate Models and Diagnostics

Here is the summary and output for our logistic model (called `logit_1`)

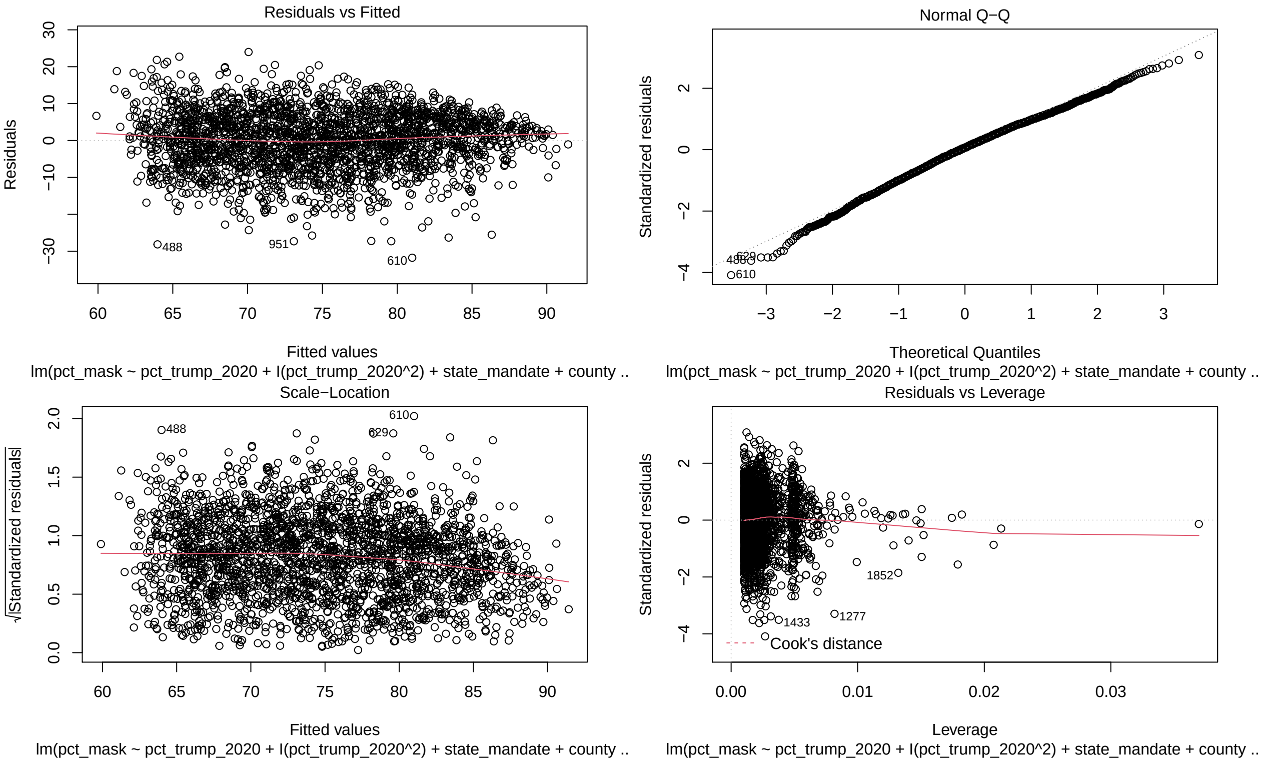
```
##
## Call:
## glm(formula = county_mandate ~ pct_trump_2020, family = "binomial",
##     data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0133  -1.0387  -0.8004   1.1614   1.7020
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.437969   0.189008   12.90   <2e-16 ***
## pct_trump_2020 -0.039170   0.002823  -13.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 3432.9  on 2480  degrees of freedom
## Residual deviance: 3214.4  on 2479  degrees of freedom
##   (33 observations deleted due to missingness)
## AIC: 3218.4
##
## Number of Fisher Scoring iterations: 4
```

Here is the formula and output for our final model to predict mask wearing from state and county mask mandates with an interaction term (called `lm_trump_state_interact`):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.8495	1.7211	48.14	0.0000
pct_trump_2020	-0.0864	0.0596	-1.45	0.1475
I(pct_trump_2020^2)	-0.0016	0.0005	-3.12	0.0018
state_mandate	7.4699	0.5883	12.70	0.0000
county_mandate	6.3295	0.4351	14.55	0.0000
state_mandate:county_mandate	-4.8829	0.7464	-6.54	0.0000

Next, we needed to check the assumptions for this model: diagnostic plots are shown below



Overall, diagnostics look fairly good, particularly linearity and constant variance. There is some slight deviation in the qqplot in the tails, but nothing too worrisome. We do have an outlier in the x direction, as shown on the leverage vs. standradized residual plot, but it does not have a large residual and is well within the Cook's distance lines, suggesting it is not too influential.

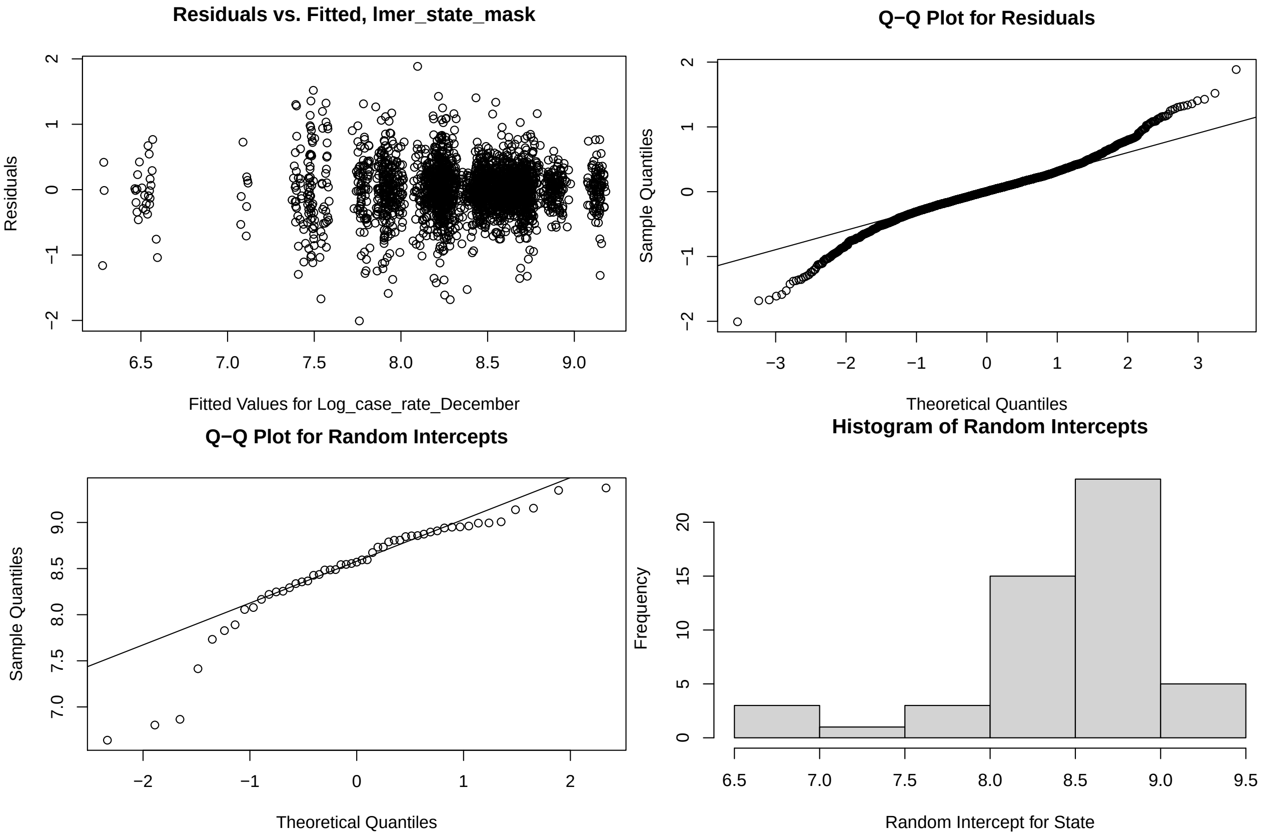
December Case Prediction Assumption Checking and Models

Here is the summary and output for our final mixed effects model, `lmer_state_mask`

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log_case_rate_december ~ pct_mask + (1 | state)
##    Data: data_train
```

```
##
##      AIC      BIC   logLik deviance df.resid
##   2498.8   2522.2  -1245.4   2490.8     2507
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2847 -0.5209  0.0077  0.5433  4.9627
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.3526   0.5938
##  Residual             0.1443   0.3798
## Number of obs: 2511, groups:  state, 51
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  8.487461   0.121356  69.939
## pct_mask    -0.003855   0.001134  -3.399
##
## Correlation of Fixed Effects:
##          (Intr)
## pct_mask -0.720
```

And here are the diagnostic plots for the mixed model. Note that in addition to the standard assumptions of linearity, normality of residuals, and constant variance, we must also check the assumption that the random intercepts for states are normally distributed.



Linearity seems to be a reasonable assumption, as points are clustered randomly around 0 (note: the visible vertical clusters appear because the magnitude of the random effect of state is much bigger than that of the fixed effect of pct_mask, thus ensuring that the points for a state fall more or less in a line). There does not appear to be evidence of heteroskedasticity. Normality of residuals looks fairly good; they deviate a bit more than expected at the tails, but do so in a symmetric way. The qqplot for the distribution of the random effects for state is slightly concerning in that it is left skewed (you can also see this in the histogram); more than anything else, this suggests that a log transformation on case rates may have been an over-correct. While this will not bias our estimates, it could lead to unrealistic standard errors for the random effect of state, which we were wary of when interpreting this model.

```
# Calculating RMSE for the two mixed effects models on the test set

RMSE = function(model, newdata, y){
  yhat = predict(model, newdata = newdata)
  RMSE = sqrt(sum((y-yhat)^2/nrow(newdata)))
  return(RMSE)
}
data_test <- readRDS("~/Desktop/STAT 139/stat139_project/data_test.RDS")
RMSE(lmer_state_mask, data_test, data_test$log_case_rate_december)
```

```
## [1] 0.3905824
```

```
RMSE(lmer_state, data_test, data_test$log_case_rate_december)
```

```
## [1] 0.3925893
```