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:

Show in New WindowClear OutputExpand/Collapse Output

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

Dealing with Missing Data: Writeup

```
clean_data_complete_ny <- read_csv("raw_data_masks/clean_data_complete.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   county_name = col_character(),
##   state = col_character()
## )

## See spec(...) for full column specifications.
```

```
data_complete <- na.omit(clean_data_complete_ny)
```

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

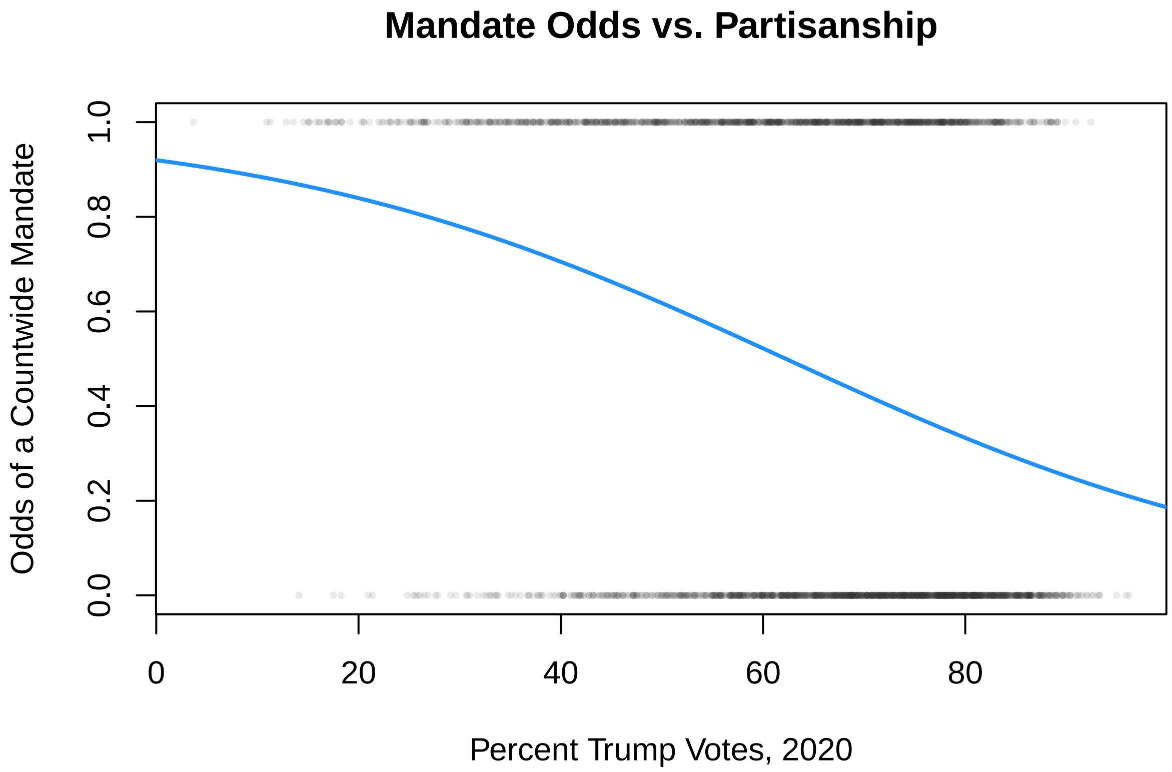
Mask Wearing Section

```
# Running logistic regression

data_train = readRDS("~/Desktop/STAT 139/stat139_project/data_train.RDS")
logit_1 = glm(county_mandate~pct_trump_2020, data = data_train, family = "binomial")
dummy.pct = seq(0,100,1)
yhat = predict(logit_1, new = data.frame(pct_trump_2020 = dummy.pct))
phat = exp(yhat)/(1+exp(yhat))
```

```r
# Making a plot

plot(county_mandate~pct_trump_2020, data = data_train, cex = 0.5, pch = 16, col = rgb(0.2,0.2,0.2,0.1),
     xlab = "Percent Trump Votes, 2020",
     main = "Mandate Odds vs. Partisanship")
lines(phat~dummy.pct, col = "dodgerblue", lwd = 2)
```

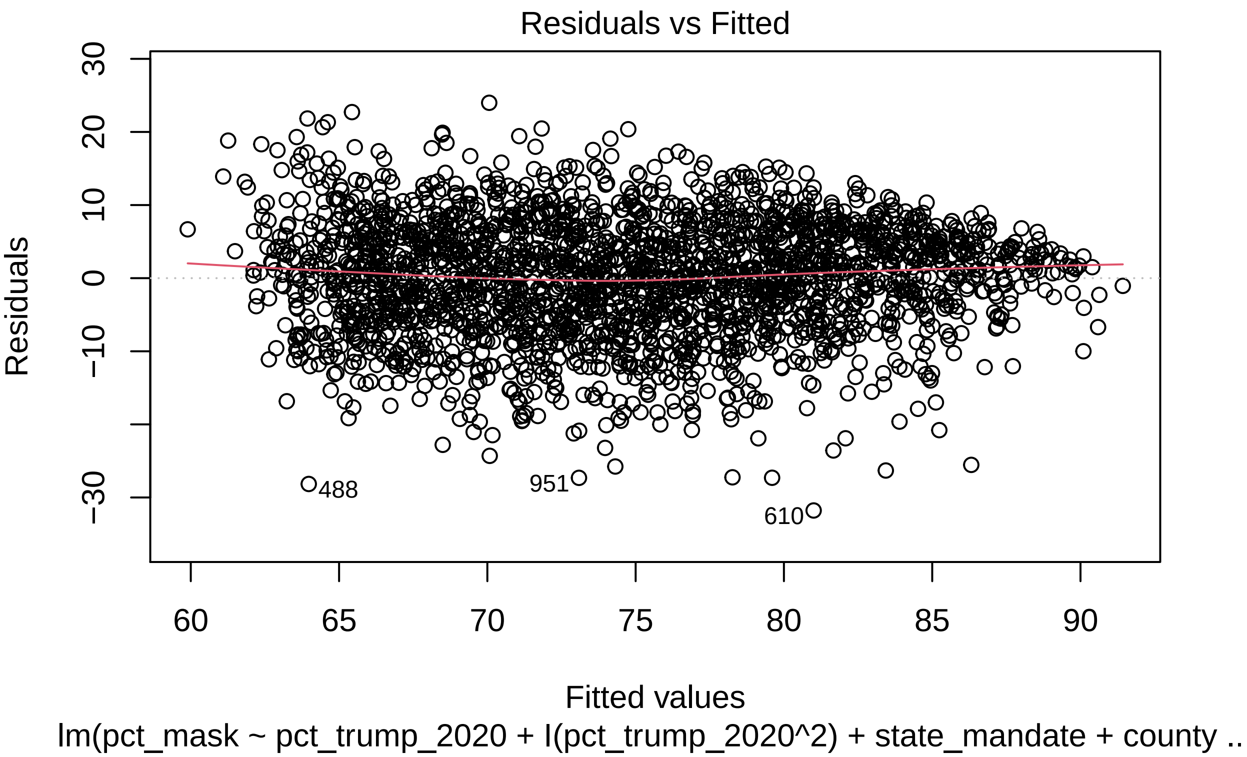


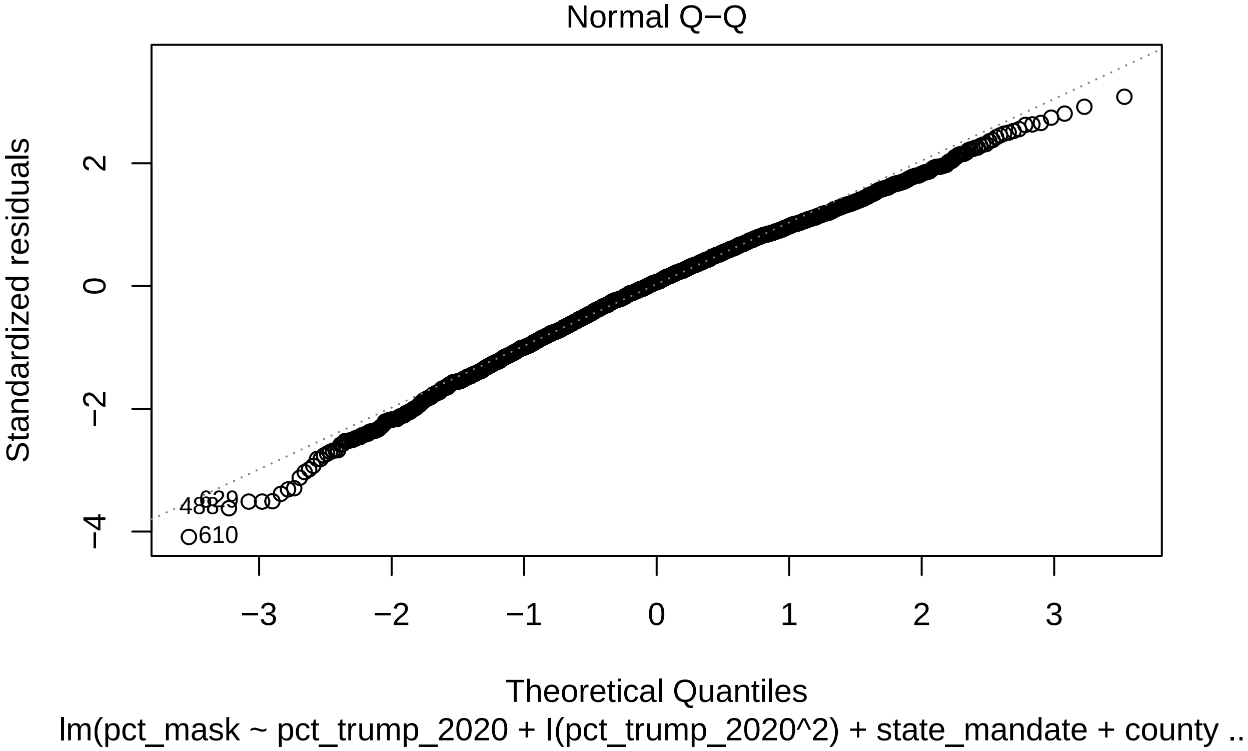
Formula and output for our final model to predict mask wearing from mandates;

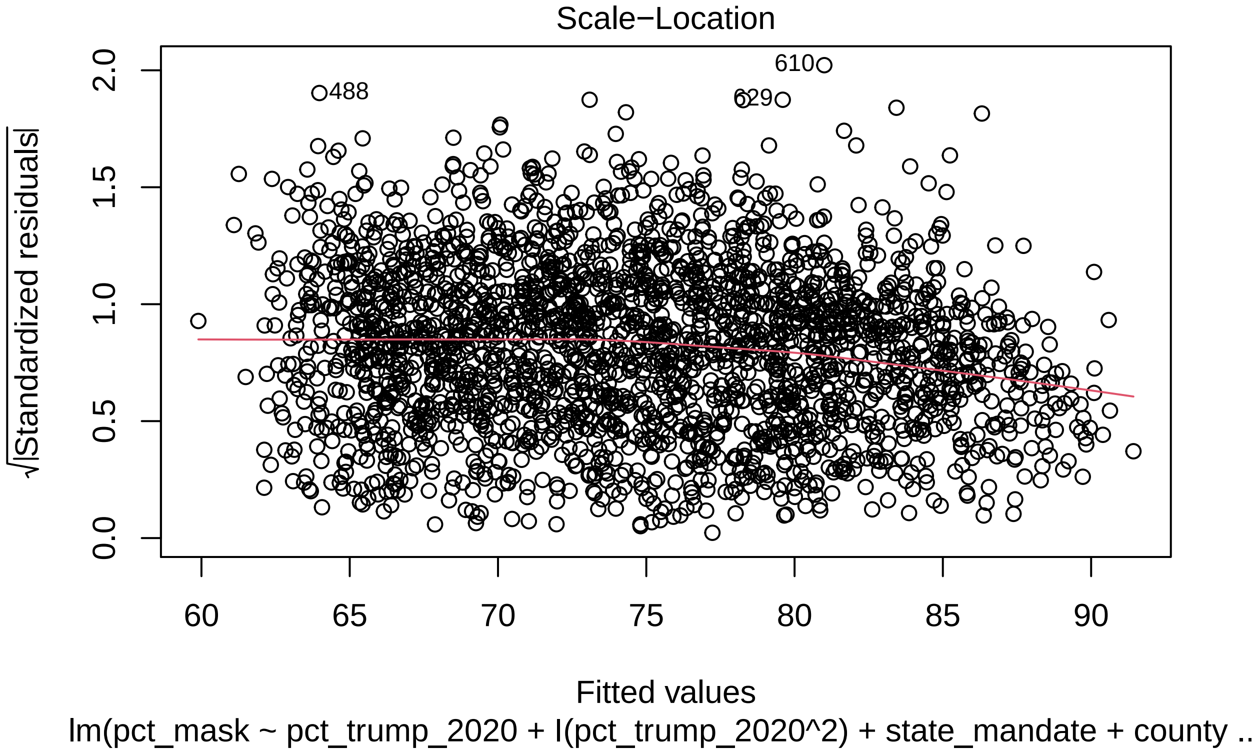
% latex table generated in R 4.0.2 by xtable 1.8-4 package % Tue Dec 8 21:48:34 2020

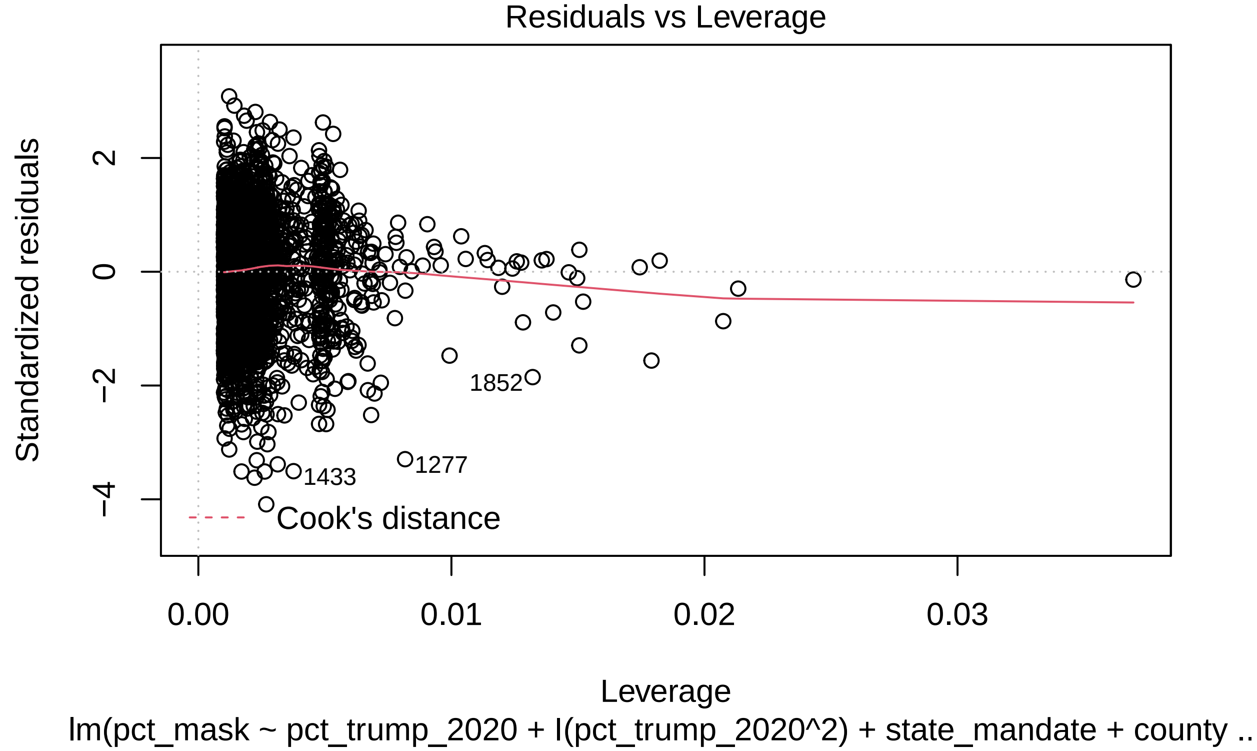
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.8495	1.7211	48.14	0.0000
pct_trump_2020	-0.0864	0.0596	-1.45	0.1475
I(pct_trump_2020ˆ2)	-0.0016	0.0005	-3.12	0.0018
state_mandate	7.4699	0.5883	12.70	0.0000
county_mandate	6.3295	0.4351	14.55	0.0000
state_mandate:county_mandate	-4.8829	0.7464	-6.54	0.0000

```r
# Checking assumptions
plot(lm_trump_state_interact)
```


Residuals vs Fitted
Residuals
30
20
10
0
−10
−30
488
951
610
60
65
70
75
80
85
90
Fitted values
lm(pct_mask ~ pct_trump_2020 + I(pct_trump_2020^2) + state_mandate + county ..


Normal Q−Q
Standardized residuals
2
0
−2
−4
488
629
610
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(pct_mask ~ pct_trump_2020 + I(pct_trump_2020^2) + state_mandate + county ..





Overall, diagnostics look fairly good, particularly linearity and constant variance. 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, suggesting it is not too influential.

Code and Assumption Checking Mixed Effects

```
# Assumption Checking for final mixed effects
```