CS Modeling

Chloe Shawah

12/7/2020

Is Mask Wearing Inherently Political?

To decide if mask-wearing behavior is inherently political, we will divide our predictors, removing state and county mandates, into two categories: political and apolitical. Even though these predictors can be correlated across categories as we saw in our correlation table, this does not undermine our goal of determining whether only demographic factors can explaining mask wearing more or less than only political factors.

political:	pct_trump_2020, dem_governor
apolitical:	pop_2019, log_pct_seniors, ru_continuum, log_pct_minority, pct_anycollege, log_density, pct_female, log_pct_poverty

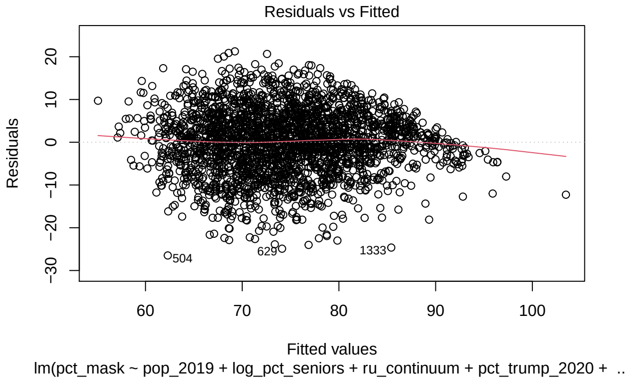
First, we will run 3 models with these two categories of predictors *combined* to predict mask-wearing score so we have an idea of our base ability explain mask-wearing behavior.

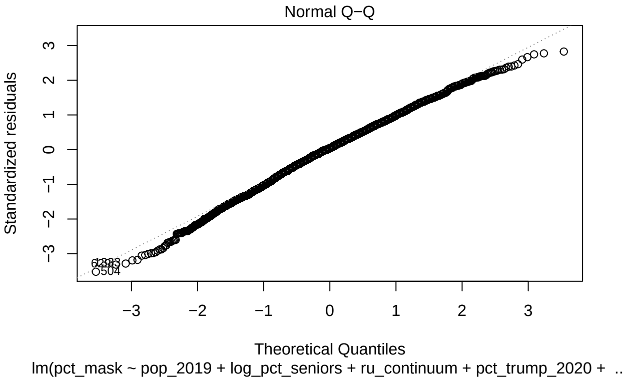
1) `linear full model`: an OLS model with single effects only

2) `linear interaction model`: an OLS model with single effects and all interaction terms

3) `linear selected model`: an OLS model using step-wise selection in both directions with an intercept-only model as an lower bound and starting point and the interaction model as an upper bound.

Included here are the assumptions checks for the first model `linearfullmodel`. The normality and linearity assumption appear to be okay based on the residual and Q-Q plot, however we do see that the variance of residuals appears lower for very high $\hat{y}$, but does not seem severe enough to warrant a more robust method than OLS. The plots for the second two models are very similar to these plots for `linearfullmodel`.





Now that we feel as though our assumptions are met, we will report our R^2 values for these 3 models. At the end of this section, we will include a full table with all models and their train and test RMSE.

Model Name	R^2
linear full model	0.4751
linear interaction model	0.5329
linear selected model	0.5309

This tells us that a few things: adding the interaction terms does improve the power of our model, and cutting our full interaction model with 55 predictors to 31 well-chosen predictors only decreases the R^2 of our model slightly. We compared the coefficients for the single effects in the `linear full model` and `linear selected model` and noticed that when we added the interaction terms, some of the coefficients flipped signs for the single effect: `log_pct_seniors`, `pct_trump_2020`, and `pct_female`. For `pct_trump_2020`, this coefficient was negative in `linear full model`, but flipped to positive in the `linear selected model` while almost all of the interaction terms involving `pct_trump_2020` were negative. This is still a little surprising: when we control for many interaction terms such as `pct_trump_2020:pct_anycollege`, an increase in Trump votes while holding all of these other interactions constant is associated with an increase in mask-wearing score. Very interesting! Perhaps this indicates that the relationship between politics and mask-wearing is a more complex than we had thought.

Next, we will fit two more models:

4) `linear political model`: an OLS model with just our political predictors and their interactions

5) `linear apolitical model`: an OLS model with just our apolitical predictors and their interactions

Model Name	R^2
linear political model	0.3222
linear apolitical model	0.4287

Based on these two R^2 scores, it seems like our apolitical model is better able to explain variance in mask-wearing behavior by county than using our two political predictors. Even though we only have two political predictors, this still helps us debunk the theory that mask wearing behavior is completely tied to support for President Trump, because we are able to explain just as much variability, if not more, using only our demographic variables.

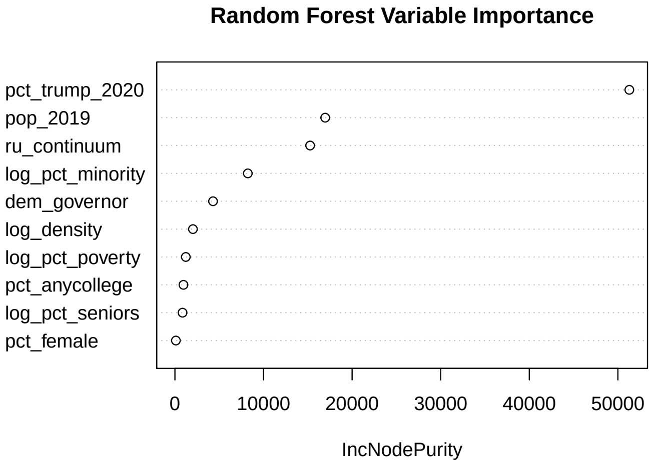
For our final models for this section, we will fit 2 random forest models and 1 decision tree that will have an added advantage of being able to catch nonlinear trends in our data. They will be:

6) `rf full`: a random forest with all predictors

7) `rf apolitical`: a random forest with only apolitical predictors

8) `dt political`: a decision tree with the two political predictors

Both random forests were fit with mtry = 6 and maxmodes = 10, while the decision tree was fit with maxdepth = 10. After fitting the `rf full` model, we checked the variable importance plot pictured below.



We found that `pct_trump_2020` was most often used to split the nodes. This does not contradict our early discussion of apolitical factors being able to explain more than political factors, but certainly adds some nuance: on its own, `pct_trump_2020` may be the most the most important predictor of mask-wearing score, but at least in the linear models, when we include lots of other, less individually important demographic predictors, we can get a model that is just as good if not better. We also noticed that pop_2019 is the second most important predictor according to this plot, even though it was deemed unimportant in some of the linear models above; this indicates to us that there may be a nonlinear relationship between pop_2019 and pct_mask.

Final RMSE Table:

Model Name	Train RMSE	Test RMSE
linear full model	7.437	7.953
linear interaction model	7.011	7.609
linear selected model	7.032	7.648
linear political model	8.451	8.634
linear apolitical model	7.745	8.120
rf full model	7.601	8.127
dt political model	8.497	8.683
rf apolitical model	7.959	8.465

From this final RMSE table, we can make a few observations. The model that performed best on the test data was the linear model with all interaction terms. Some of our models appear to be a little bit overfit, but nothing drastic. Finally, and most importantly, in both the linear models and decision tree/random forest models, the apolitical models outperformed the political models on train and test data.

This tells us, tentatively, that we can explain more variability in mask-wearing behavior between counties with demographic predictors as we can with politcal predictors. Mask-wearing may not be inherently political.