2021 Covid-19 Vaccine Completion Prediction Using 2020 Presidential Election Vote Share

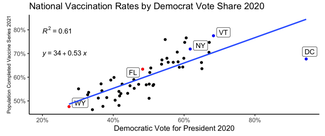
John Markowicz

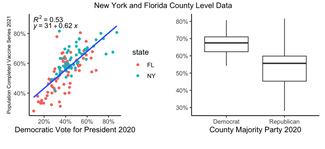
Research Question

Was the 2020 presidential election vote share a reliable predictor for 2021 Covid-19 population vaccination completion percentage? The arrival of Covid-19 in the United States quickly became a major talking point during a turbulent 2020 presidential election cycle, and it seemed Republican and Democratic voters had conflicting strategies to deal with its presence. Florida and New York garnered constant media coverage for their contrasting approaches to lockdown procedures and vaccine mandates. I analyzed these states at the county level in order to create a reliable predictive model that can assess the variability in vote share and vaccine completion rate.

To establish that the 2021 Covid-19 vaccine completion percentage was dependent on the 2020 presidential election at the national level, I performed a country-wide OLS regression. Key statistics measured were R-squared being .61, meaning 61% of the total variability in the dependent variable was explained by the model, and the Pearson's correlation coefficient of .783 meaning a strong positive correlation between variables. For measuring prediction accuracy at the county level, I calculated the RMSE that evaluated predicted values within 8.27 percent of the observed values on average. The included box plot depicts the upper quartile range for majority Republican counties falling below the lower quartile range for Democratic majority counties again emphasizing that most Republican counties had lower vaccination completion rates than most Democratic counties.

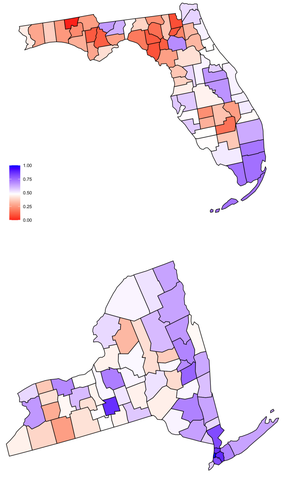
I relied on CDC 2021 Covid-19 data and 2020 presidential election survey data at both county and state levels, and 2020 demographic data at the county level for New York and Florida





Variation Between States

To visually distinguish between New York and Florida, I formulated a metric to weigh both independent and dependent variables for scaling each county's color scheme. I added both variables' percentages and performed a min-max normalization procedure in order to scale the metric between 0 and 1. This metric is used to effectively display the variation between New York and Florida in terms of their correlation between variables, measured at .77. Counties with a scaled color closer to 0 had a higher Republican vote share and a lower percentage of the population with completed vaccine series while counties with a color closer to 1 had a higher Democratic vote share and a higher percentage of the population with completed vaccine series.



Controlling for Demographic Features

Utilizing demographic data for Florida and New York, I built a multivariable regression model that improves upon the original simple OLS model. To carry out variable selection, I performed a stepwise regression using minimum AIC as the stopping criterion and later removed a few variables to reduce the model's complexity. To check for multicollinearity, I measured the variance inflation factor that showed most independent variables were close to 1 and none greater than 3 indicating that multicollinearity was not a factor. In order to account for overfitting, I split the data 70/30 between training and test sets and measured a regression model for each. After running models for each set multiple times, I discovered no discrepancies in the adjusted R-squared values as the test set closely approximated the training set's results suggesting overfitting was a nonissue. The regression coefficients show the change in the vaccine completion rate for each unit increase in a given independent variable while holding all other independent variables constant. The RMSE equated to 5.91, a significant decrease from the simple OLS model's RMSE of 8.27.

	Dependent variable:
	Pct. Pop. Completed Vaccination Series
Democratic Vote Share	0.613*** (0.049)
Median Housing Price	-0.00001** (0.00001)
Median Age	0.733*** (0.103)
Median Household Income	0.0004*** (0.0001)
Constant	-19.316*** (5.681)
Observations	129
R^2	0.762
Adjusted R^2	0.754
Residual Std. Error	6.032 (df = 124)
F Statistic	98.988*** (df = 4; 124)
Note:	*p<0.1; **p<0.05;* p<0.01

Conclusion

My analysis depicts 2021 Covid-19 vaccine completion rates as a politicized topic that can be accurately predicted using the 2020 presidential vote share. While the variation in vaccination rates may have decreased in subsequent years, my analysis focuses on the initial divisive reaction. I demonstrate variance in the chosen features between New York and Florida and given the country-wide correlation, my analysis can be replicated to forecast Covid-19 vaccination completion rates between other states. I finish by improving upon the simple linear regression for Florida and New York county data by including demographic data, verifying the multivariable model's adequacy, and achieving both a lower RMSE and higher R-squared deeming the model's prediction as both reliable and accurate.