Predictors of COVID-19 Infection Rates in Virginia

Garrett N. Bushnell

April 2021

1 Introduction and Data

The first case of COVID-19 was documented in Virginia on March 7, 2020[1]. Just over a year later, the total cases in Virginia had ballooned to 651,000[2]. As a Virginia resident, I am interested to see what factors explain the differences between some counties experiencing widespread outbreaks and others having few cases over the entire county. Using a Bayesian framework, we can create several regression models before selecting the one that seems to explain our variances in COVID-19 cases the most.

1.1 Data

The data consists of socio-economic, demographic, and health data for each of the 132 counties in Virginia. Health and COVID-19 case data is from the Virginia Department of Health[3]. Socio-economic data comes from the US Census' American Community Survey[4]. These predictors were selected for their accessibility across all counties, as well as their ability to capture many of the largest differences amongst counties.

- cases_per_1k - Total COVID-19 cases per 1,000 county inhabitants
- total_cases - Total COVID-19 cases
- total_pop - Total county population
- pct_male - Percentage of county population identifying as male
- pct_over_65 - Percentage of county population over 65 years old
- pop_per_sqmile - County population per square mile
- total_hospitalizations - Total COVID-19 related hospitalizations
- total_deaths - Total COVID-19 related deaths
- pct_white - Percentage of county population identifying as white

- median_income - Median annual income of county's population
- pct_no_insurance - Percentage of county population with no insurance
- pct_in_poverty - Percentage of county population living in poverty
- pct_hs_degree - Percentage of county population with a high-school degree
- pct_bc_degree - Percentage of county population with a bachelor's degree
- pct_dem - Percentage of county population identifying as a Democrat
- pct_rep - Percentage of county population identifying as a Republican
- communicable_disease - Total cases per 1,000 county inhabitants of the top 10 most communicable diseases in Virginia.

Areas of Interest

- Political Affiliation The COVID-19 epidemic has been highly politicised. Two predictors (pct_rep and pct_dem) seek to determine if affiliations with a political party are more likely to be correlated with an increase in COVID-19 cases.
- Economic Disparity

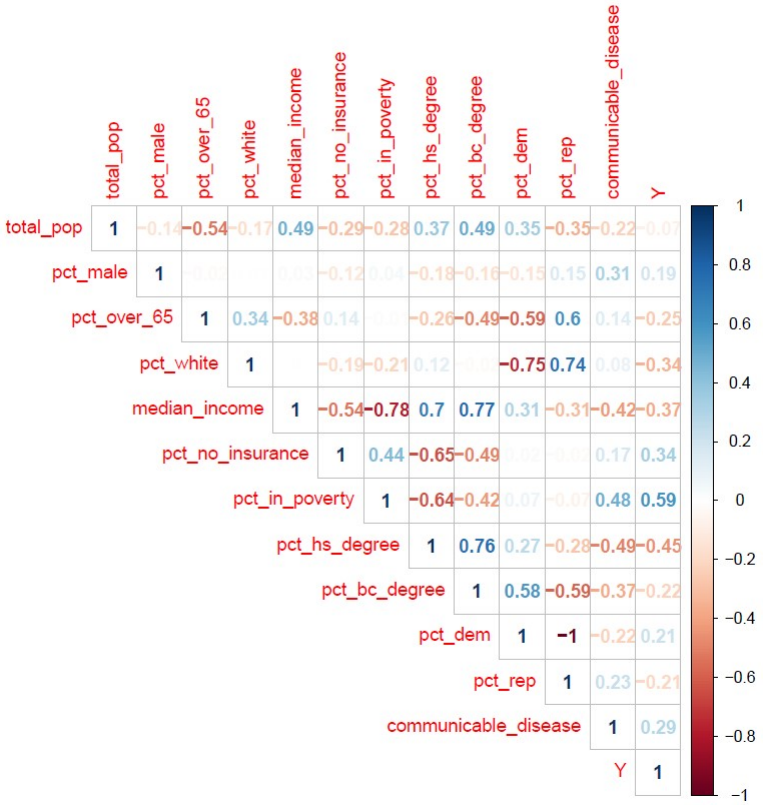
  A measure of the percentage of a county's citizens that fall under the poverty line is used to determine if counties with a poorer population are more correlated with higher COVID-19 cases.
- Education Levels

  Measures of the education levels of a county's citizens seek to determine if a correlation exists between the education level and COVID-19 cases.
- Demographics Measures of gender, age, and race seek to determine if any of these factors is associated with higher COVID-19 cases.
- Population Density Measures of population density are used to determine if more densely populated counties witness higher infection rates than those of rural counties.

It is important to note that many of these factors suffer from multicollinearity. This has not been addressed within this paper.

2 Exploratory Data Analysis

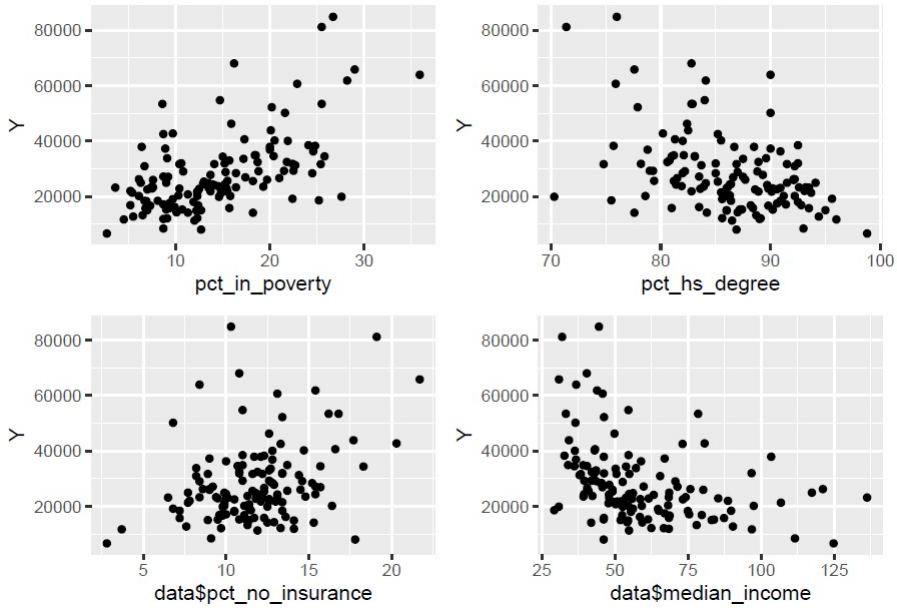
Several of our predictors are moderately correlated with the number of COVID-19 cases. As we can see in the following figure, the percentage of the county's population living in poverty is the highest correlated variable with COVID-19 cases. From this metric alone, we can hypothesize that poverty is correlated with COVID-19 cases. However, this does not imply causation.



We can also note that counties with higher percentage of high-school educated citizens are associated with lower COVID-19 cases. Interestingly, we can also see that there is a moderate negative correlation with increases in the median income and a positive correlation with the percentage of a county's population with no health insurance (which is likely highly correlated with our poverty measure).

Finally, we can see that there exists a slight negative correlation among counties with higher Republican citizens and a slightly positive correlation amongst those with higher Democratic populations. Based on the signs of these correlations alone, this is somewhat counter-intuitive, as media often reported Republicans as those more likely to not follow CDC precautions. However, it must be noted that this again is likely correlated with our other economic predictors.

Selected scatterplots are included below:



3 Modeling

Our modeling approach consists of two main methodologies - multiple linear regression (MLR) and poisson regression. For each, several models were fit. We then select the best model based on the deviances.

In MLR, we assume that our response is best modeled as a linear combination of our predictors. We assume our response is normally distributed and assign non-informative priors to each predictor's coefficient.

$$y_i \sim X^T\beta \tag{1}$$

where our β's are given non-informative priors of N(0,τ/p^2). Where τ is given a non-informative prior of Gamma(.001, .001) and p is the number of predictors.

For each iteration, we also fit a poisson regression model that assumes the log of the expectation of the count of cases is a linear combination of our predictors.

$$y_i \sim Pois(\exp(X^T\beta)) \tag{2}$$

We then assign a non-informative prior to our β's of N(0,.001). It's important to note that Poisson regression requires our response be a count (as opposed to a normalized cases per thousand citizens) that our MLR model uses.

- Full Models

  We begin by modeling our responses as a linear combination of all predictors. This will give us a base-line and show us (given all other predictors in the model), what predictors may be statistically significant. This will also serve as a baseline for comparison amongst future models.

- MLR

  Our full MLR model produces a deviance of 2820.53807 and an R^2 measure of .49. The model suggests the coefficient of pct_over_65 as statistically negative at the 95% level. This result is somewhat counter intuitive. On one hand, the elderly population were significantly affected by COVID-19 - particularly in nursing homes. However, the negative correlation could be due to the fact that elderly, in general, are more health-conscious and are likely followed guidelines more strictly - including social distancing.

  It also suggest a higher percentage of poverty is associated with higher COVID-19 cases.

- Poisson

  The full Poisson model produces a significantly higher deviance of 13503.47537. It also concludes that a negative correlation exists between COVID-19 cases and counties with higher population over 65 years old. This model also suggests a negative relationship exists between counties with a high percentage of citizens having high-school degrees and a positive relationship with counties with high percentages of the population having no insurance.

- Variable Selection

  We now use stochastic search to perform variable selection. In short, we fit a linear model with an additional variable (δ) denoting if the predictor was chosen or not for a particular model. Given the included/not included options, a Bernouilli prior is assigned.

$$y_i \sim \delta * \beta * X \tag{3}$$

  We then select all variables with δ's greater than .5 - meaning they were in at least half of the models tested. The selected variables are: pct_over_65, pct_white, pct_hs_degree, pct_rep, and communicable_disease.

  - MLR

    This MLR model produces a deviance of 3108.4852, which is higher than our base MLR model. This model suggests both high-minority counties may be associated with higher cases and populations with higher proportions of high-school diplomas and higher prevalence of communicable diseases is associated with higher cases.

  - Poisson

    Again, the poisson model produces a higher deviance of 275175.73383. This model suggests that the coefficients for all but high-school degrees are statistically negative. Although this model shows all coefficients as statistically significant (as their CI's don't include zero), the deviance shows the model is a poor choice.

- Variable Selection with Interactions

  As a final step, interaction terms are added between all predictors, resulting in 144 predictors. We again use a stochastic search for variable selection. However, due to the numbers of predictors, we limit our results to the top 5 predictors. Keeping with regression norms, we also include the bases of interaction terms (Eg. If our interaction Term A.B is selected, we also model A and B separately in our model).

  The variable selection with interactions selected the following predictors: pct_over_65.pct_no_insurance, pct_in_poverty.pct_hs_degree, median_income.pct_dem, communicable_disease, pct_white.pct_rep.

  - MLR

    This model produced the lowest deviance, thus far, of 2819.83085. However, the 95% intervals for all coefficients includes zero - which is rather disheartening. Although the deviance is better and overall our model does suggest predictive power as a whole, we cannot statistically conclude each coefficient is different from zero.

  - Poisson

    One again, the poisson model produces a large deviance measure of 222029.3911. However, it does suggest all coefficients are statistically different from zero at a 95% level. The model suggest no insurance and poverty play the strongest roles in increasing the number of cases and that a high Democratic population is correlated with lower levels of cases.

4 Conclusion

After testing all models and methods, the original MLR regression is selected as it provides the second lowest deviance and a number of predictors are statistically different from zero. Our final model is. This model's R^2 is .494, suggesting that this combination of predictors accounts for 49.4% of the variance in covid cases per 1,000 residents.

$$Y = -0.0604 - 483.11 total\_pop + 904.65 pct\_male - 740.12 pct\_over\_65$$

- 218.75pct_white -57.43median_income + 507.88pct_no_insurance

+ 825.79pct_in_poverty - 476.24pct_hs_degree + 237.91pct_bc_degree

+ 275.25pct_dem + 444.64pct_rep - 7.05communicable_disease
(4)

Two variables are sadistically significant: Pct_over_65 is statistically negative, while pct_in_poverty is statistically positive. Although we cannot conclude

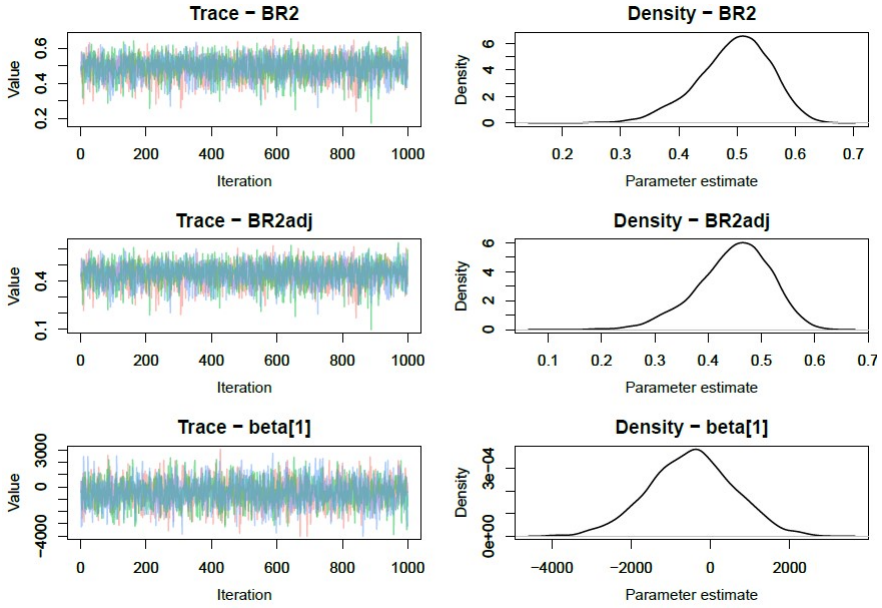
causation, we can hypothesize on why these two factors may be correlated with COVID-19 cases.

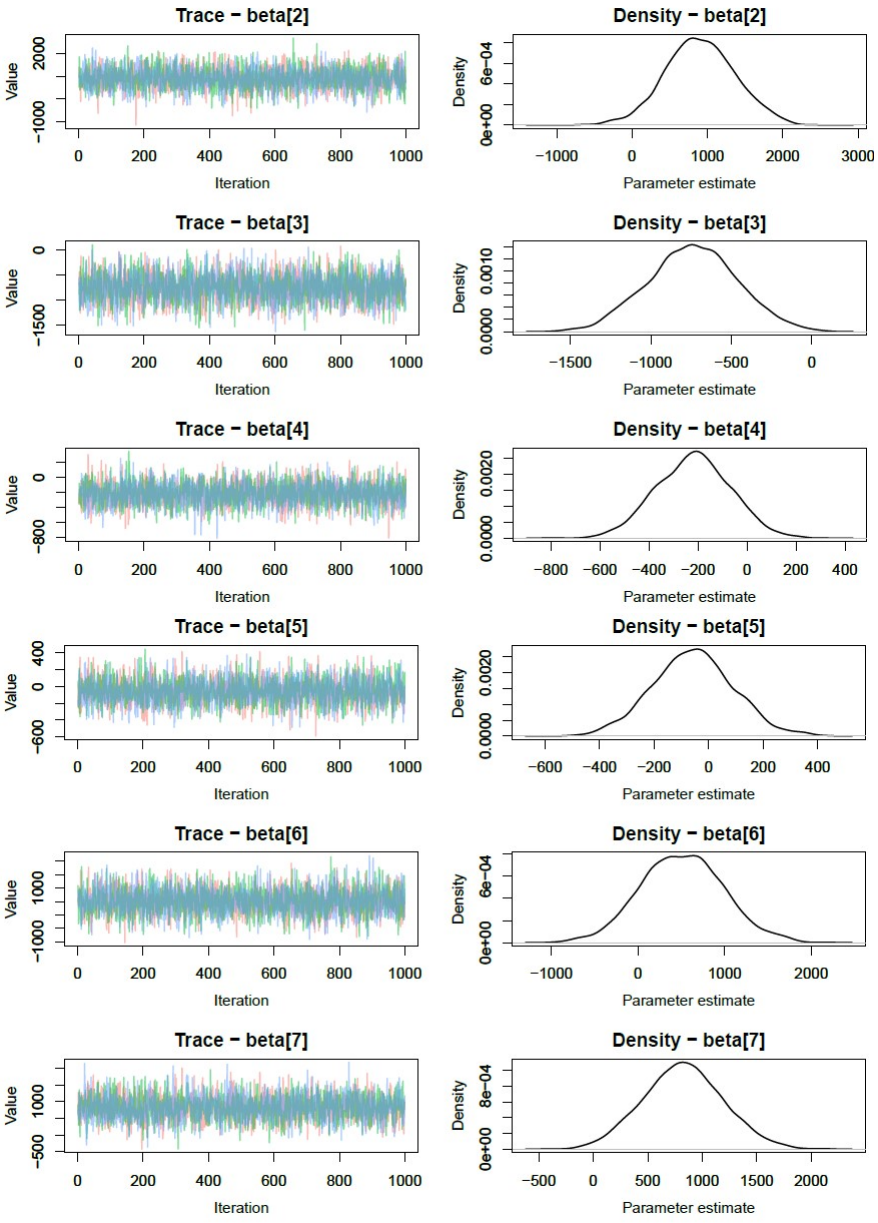
As mentioned previously, the elderly population is most likely retired and spends a greater portion of their day at home and not exposed. They also are probably more likely to follow CDC guidance and practice social distancing, given that many already likely suffer from some type of health conditions associated with the elderly.

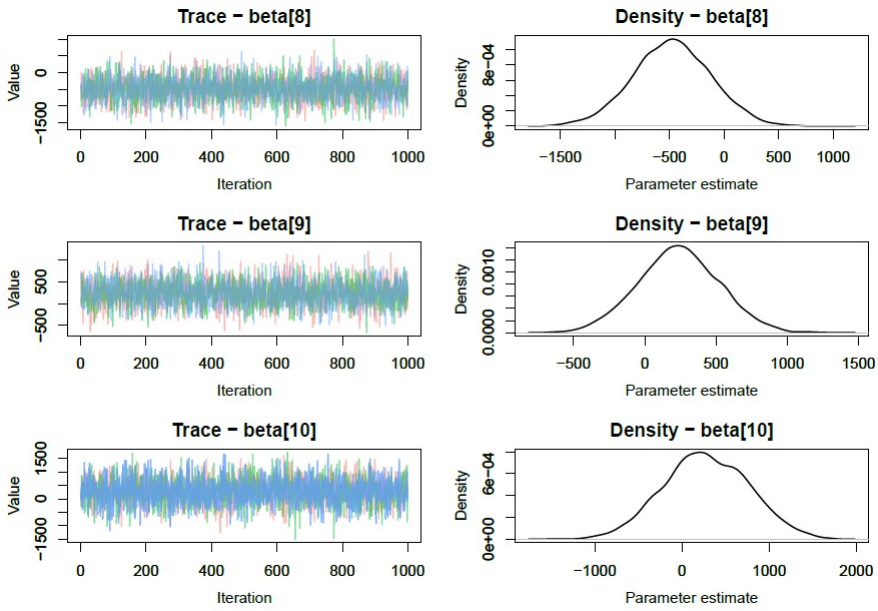
Counties with a higher portion of residents living below the poverty line could experience higher COVID-19 cases due to the fact that this portion of the population has less financial stability and cannot as readily afford to quarantine as wealthier populations. Their economic conditions may have also hampered their ability to secure masks and cleaners that others used to protect themselves.

(Note that the above are only hypothesized explanations for the statistical significance and are unfounded.)

As noted, it's likely our models suffered from multicollinearity, which often significantly affects standard errors and confidence intervals. It's likely this is why variables seemed to switch signs and significance between models. Variance Inflation Factors (VIF) tests and other methods could have been used to account for this. This study was a baseline for future improvement.




Trace − beta[2]
Density − beta[2]
Trace − beta[3]
Density − beta[3]
Trace − beta[4]
Density − beta[4]
Trace − beta[5]
Density − beta[5]
Trace − beta[6]
Density − beta[6]
Trace − beta[7]
Density − beta[7]
Value
Density
Iteration
Parameter estimate



5 Resources

- 1. Virginia Department of Health. www.vdh.virginia.gov/news/2021-news-releases/first-virginia-case-of-covid-19-confirmed-at-fort-belvoir/.
- 2. The New York Times. https://www.nytimes.com/interactive/2021/us/virginia-covid-cases.html
- 3. Virginia Department of Health - Data. https://www.vdh.virginia.gov/data/
- 4. US Census - American Community Survey. https://www.census.gov/programs-surveys/acs

6 Code