**Nursing Homes and the COVID Death Rate**

**Connor Krenzer[[1]](#footnote-1)**

**The Story:**

Old people in nursing homes are dying en masse to the coronavirus, skewing the death rate upward and providing public officials greater justification for prolonged lockdowns. This is a common talking point I have heard over the past 6 months and I think econometrics provides a unique opportunity to test it out.

**Why This Story?**

There are two reasons I am researching the coronavirus:

1. The patients are running the asylum these days, so I have no idea what—or who—to believe. One statistic I heard from a particular political pundit is that the average age of death with COVID is older than life expectancy in the US, which suggests the virus isn’t a substantial problem for healthy twenty-somethings who do not have pre-existing conditions. But is that true? What is false? Did Governor Cuomo kill Granny? I have to find that out myself.
2. This topic provides me with an excellent introductory project on web scraping. The coronavirus is a highly publicized issue, meaning there have been countless articles and reports published on the matter. The data is all there in the articles, but it is not organized in the .csv format I have grown all too accustomed to. Most data collected for this project was pulled directly off webpages or read in from .PDF files that were created from images downloaded from webpages, all with the use of R. I have the opportunity to learn some new R packages and get hands-on experience tidying data from its raw, messy form.

**My Thoughts:**

Politicians rank highly among the handful of people who have benefitted from the coronavirus. I think they love the extra attention a little too much. Did you know who Dan McCoy was at the beginning of March this year [2020]? Me neither. I buy into the story that the nursing home fiasco we’ve heard so much about caused the number of deaths from the coronavirus to explode. Therefore, I expect variation in the COVID death rate (deaths *from* COVID) to be explained reasonably well by the rate of nursing home patients in the adult population.

**The Model:**

Using regression analysis subject to the BLUE method of Least Squares, I will test cross-sectional data from the 50 states and DC using the following equation:

***Deaths = β1 + β2Nursing + β3Tests + β4Elderly+ β5Mask + β6Density + β7Lockdown***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Definition** | | | **Source** | |
| **Y**  ***Deaths*** | The sum of the increase in deaths from the previous day from October 19th, 2020 to October 23rd, 2020 as a rate per 100000 persons, where deaths are defined as the "Total fatalities with confirmed OR probable case diagnoses (per the expanded CSTE case definition of April 5th, 2020 approved by the CDC) In states where the information is available, it only tracks fatalities with confirmed OR probable COVID-19 case diagnosis where on the death certificate, COVID-19 is listed as an underlying cause of death according to WHO guidelines." | | | The COVID Tracking Project for the death data  Kaiser Family Foundation analysis of Certification and Survey Provider Enhanced Reports (CASPER) data for the population data. | |
| **β2 of Interest**  ***Nursing*** | The number of people in nursing homes per 100000 adults | | | Kaiser Family Foundation analysis of Certification and Survey Provider Enhanced Reports (CASPER) data; The CASPER system includes data for all certified nursing facilities in the U.S. | |
| **β3**  ***Tests*** | The total number of tests completed on October 23rd per 100,000 residents | | | Pulled from predictcovid.com, who pulled their data from The COVID Tracking Project (who, of course, compiles their information from many local, regional, state, and federal sources) | |
| **β4**  ***Elderly*** | The percent of the population 65 years and older | | | Kaiser Family Foundation analysis of Certification and Survey Provider Enhanced Reports (CASPER) data | |
| **β5**  ***Mask*** | The percentage of state residents who say they wear a mask in public all or most of the time | | | The Washington Post (who got it from Delphi CovidCast, Carnegie Mellon University) | |
| **β6**  ***Density*** | The state's 2020 population divided by the land area (miles squared) | | | Pulled from World Population Review's website, who pulled the data from the US Census State Population Estimates Program | |
| **β7**  ***Lockdown*** | Dummy Variable determining the severity of government lockdowns as determined by the Washington Post as of September 11th, 2020 on a scale of "No Restrictions", "Some Minor Restrictions", "Moderate Restrictions", and "Major Restrictions" with zeros being No/Minor Restrictions and ones being Moderate/Major Restrictions. Note: Any state with a "\_CONDITION\_ restrictions vary by region" caveat are pooled in with the severity of the lockdown representative of the entire state (Ex. Illinois says "Moderate restrictions vary by region" but is still considered "Moderate" and therefore is a one in this model) | | | Washington Post | |
| **Variable** | | **Expected Sign** | **Expected Statistical Significance** | | **Expected Practical Significance** |
| **β2 of *Interest* *Nursing*** | | + | Low Significance | | Moderate Significance |
| **β3**  ***Total Tests Per 100,000 Residents*** | | **-** | Moderate Significance | | Moderate Significance |
| **β4 *Elderly*** | | + | Moderate Significance | | Moderate Significance |
| **β5**  ***% Wearing Masks*** | | **-** | High Significance | | High Significance |
| **Β6 *Population Density (people/mi sq.)*** | | + | Low Significance | | Low Significance |
| **β7**  ***Has Strict Lockdown*** | | **-** | Moderate Significance | | Moderate Significance |
| **Additional Variable Introduced Further into the Investigation** | | **Definition** | | | **Source** |
| **Urban** | | 2010 urban land area (the “physical city”), measured in square miles, divided by urban population. | | | Data was pulled from newsgeography.com; they collected the data from the U.S. Census Bureau |
| Expected Sign | | Expected Statistical Significance | | | Expected Practical Significance |
| + | | Moderate Significance | | | Moderately Significant |

**Hesitations**

A weakness key to my variables deals with time—the timeframes on several of my variables do not line up properly. The most notable of these variables is the lockdown variable, which was last updated on September 11th, 2020, while several other variables use data from October 23rd, 2020. The urban variable used in later iterations of the model contains data from 2010; though it is unlikely drastic urban expansions have happened from one state to another in the last ten years, more recent data would only increase the reliability of the results. Smaller problems dealing with time are with regards to the Nursing (B2) and Elderly (B4) variables. The data in those are from 2019 and—although the difference will probably not matter with elderly—if people are worried about patients in nursing facilities catching the coronavirus, it is possible that many who were in nursing facilities a year ago have decided to go elsewhere, reducing the accuracy of the nursing variable. Another problem is that DC will heavily skew the population density results. Regardless of the accuracy of this model, the coronavirus’ dangers cannot be dismissed if all the variables explain little of the variation in the COVID death rate. Other factors, such as lingering symptoms or permanent health problems from the virus also need consideration in discussions of policy.

**Analyzing the Dataset**

*Descriptive Statistics*

An initial look at the data indicates that the range of values for each variable is fit for regression analysis:

**Descriptive Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Maximum | Range | Median | Mean | Std. Deviation |
| Y, Death | 0.000 | 7.97 | 7.97 | 1.92 | 2.11 | 1.54 |
| B2, Nursing | 0.096 | 0.97 | 0.87 | 0.54 | 0.54 | 0.23 |
| B3, Tests | 18721.500 | 75190.00 | 56468.50 | 34102.90 | 36776.48 | 13348.62 |
| B4, Elderly | 11.500 | 21.30 | 9.80 | 17.00 | 16.98 | 2.02 |
| B5, Mask | 63.000 | 97.00 | 34.00 | 84.00 | 83.39 | 7.01 |
| B6, Density | 1.000 | 11815.00 | 11814.00 | 108.00 | 431.57 | 1647.28 |

*Simple Regression Models*

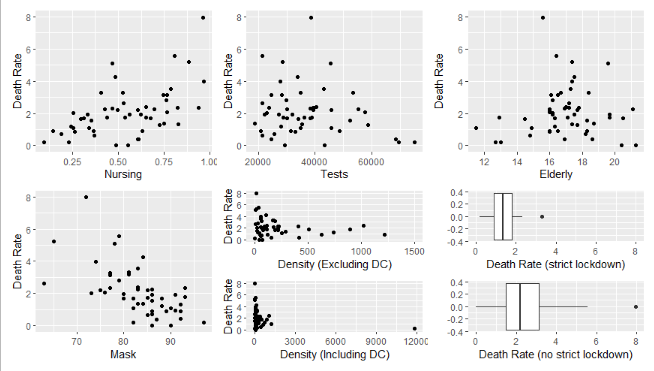
The beginning of the analysis applied each explanatory variable in a simple linear regression against the dependent variable, “Death” with a 95 percent confidence interval. The calculations are provided in the table below:

**Simple Regression Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | R-Square | t-stat | p-value | Beta |
| B2, Nursing | 0.2892 | 4.465 | 0.000047 | 3.6829 |
| B3, Tests | 0.03175 | -1.268 | 0.211 | -0.00002060 |
| B4, Elderly | 0.001852 | 0.302 | 0.764 | 0.03287 |
| B5, Mask | 0.3694 | -5.358 | 0.0000022457 | -0.13372 |
| B6, Density | 0.0446 | -1.512 | 0.137 | -0.0001979 |
| B7, Lockdown | 0.09842 | -2.313 | 0.025 | -1.0331 |

The table above shows that two of the explanatory variables have statistically significant p-values at the 0.001 level of significance, and three of the six explanatory variables have statistically significant p-values at the 0.05 level. The explanatory variable this study is most interested in, the number of people in certified nursing facilities per 100,000 adults (Nursing), has an R-square just below 0.29, meaning roughly 29 percent of the variation in the number of new deaths to the coronavirus—controlled for population size—is explained by changes in the number of nursing home residents. The coefficient for the nursing home population, 3.6829, suggests that an increase of one person per 100,000 state residents in nursing facilities will increase the number of people who die from the coronavirus by 3-4 per 100,000 state residents. Among the three variables that are statistically significant, only the lockdown and nursing home variables are practically significant with a beta value greater than one. A one percent increase in the percent of people wearing masks will only decrease the number of deaths due to the coronavirus by roughly one in one million.

Further, it is important to note that all variables were assumed to have a straight-line relationship with the dependent variable. The figure below plots all explanatory variables against the dependent variable. No obvious relationship emerges from any of the variables, so a straight-line relationship will be assumed for each variable. In fact, preliminary tests performed showed regression statistics are slightly improved by adopting a quadratic relationship between the variables, yet no clear, logical explanation why this may be the case is apparent from the data, and no obvious story can be told using a different linear relationship.

**Explanatory Variables vs. Dependent Variable (New Deaths from COVID)**

*Multivariate Regression Models*

The following table displays the preliminary results estimating the COVID death rate (the dependent variable) using a linear regression model containing a 95 percent confidence interval with all six explanatory variables included.

**Initial Multivariate Regression Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-square | Adjusted R-square | Residual Standard Error |
| 1 | .709 | 0.5028 | .4351 | 1.159983 |

\*Predictors: Percent Wearing Masks, State Population Density, Total COVID Tests Per 100k, Percent of Population over 65, Nursing Home Residents Per 100k Adults

**Initial Multivariate Regression Model Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std.Error | Estimate |
| (Intercept) | 9.98919995 | 2.67947762 |  | 3.73 | 0.00055 |
| Nursing | 2.67388455 | 0.79789873 | 0.390469966 | 3.35 | 0.00166 |
| Tests | -0.00000353 | 0.00001387 | -0.030513324 | -0.25 | 0.80042 |
| Elderly | -0.01334721 | 0.09052120 | -0.017474948 | -0.15 | 0.88345 |
| Mask | -0.10789524 | 0.02991967 | -0.490400902 | -3.61 | 0.00079 |
| Density | 0.00003328 | 0.00011703 | 0.035519053 | 0.28 | 0.77748 |
| Lockdown | 0.02512581 | 0.41076692 | 0.007629509 | 0.06 | 0.95150 |

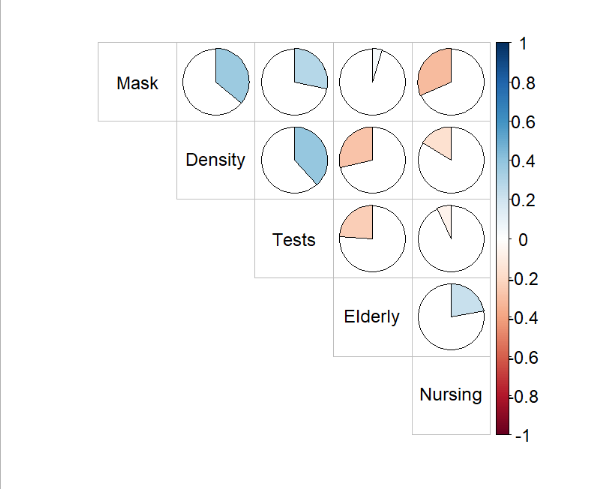
The R-square value of .5028 suggests that the variables included in the regression model explain approximately half of the variation in the dependent variable. The far lower adjusted R-square value indicates that many of the variables contribute little to the model, which is supported by the low R-squares found in the simple linear regression models.

Compared with the simple regression models, the t-stats and p-values of the explanatory variables changed drastically. The nursing and mask variables remain significant, yet the lockdown variable’s p-value moved from 0.025 in the simple linear regression all the way to 0.952 in the multiple regression. All other variables that were not statistically significant in the simple linear regressions had similar bumps in their p-values. The population density variable, for example, had a p-value of 0.137 in the simple linear regression and moved all the way to 0.777 in the multiple regression.

The increase in p-values across all variables indicates multicollinearity exists among the explanatory variables. To determine the extent that collinearity is an issue with the variables in the model, data for the explanatory variables will be examined both visually and with correlation coefficients.

*Examining Multicollinearity*

After testing for multicollinearity, there is a high likelihood that multiple variables will move in tandem with one another in a linear regression model. The correlation matrix below visualizes the extent to which multicollinearity affects each variable. For a quantitative approach, the table beneath the correlation matrix provides the absolute value of the zero-order correlation coefficients between variables.

**Correlation Matrix**

**Zero-Order Correlation Coefficients (Absolute Value)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mask | Density | Tests | Elderly | Nursing |
| Mask | 1.00 | 0.36 | 0.29 | 0.05 | 0.32 |
| Density | 0.36 | 1.00 | 0.38 | 0.28 | 0.16 |
| Tests | 0.29 | 0.38 | 1.00 | 0.24 | 0.07 |
| Elderly | 0.05 | 0.28 | 0.24 | 1.00 | 0.22 |
| Nursing | 0.32 | 0.16 | 0.07 | 0.22 | 1.00 |

Of the 10 possible combinations of the explanatory variables—excluding the lockdown variable because Pearson correlations require continuous variables—all but three exhibit a negligible degree of multicollinearity (defined as a zero-order correlation coefficient less than |0.3|). To examine the three combinations with a higher degree of multicollinearity, the t-stat between each relationship will be considered. The following table details the statistical significance of the variable combinations with a zero-order correlation greater than |0.3|:

**Examining for Multicollinearity, Variables with Correlation Coefficient Greater than |0.3| (Absolute Values)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 (Y) | X2 | r | R-square | t-stat | p-value |
| Mask | Density | 0.36 | 0.1314 | 2.723 | 0.00894 |
| Mask | Nursing | -0.320 | 0.1017 | -2.356 | 0.022521 |
| Density | Tests | 0.382 | 0.1463 | 2.897 | 0.00561 |

As the table shows, all three variables having a substantial zero-order correlation coefficient also have a statistically significant p-values, meaning measures should be taken to reduce multicollinearity’s effects on the model.

To explore the problem multicollinearity poses to the linear model, collinearity diagnostics and statistics, including condition indices and tolerance levels, were considered. The table below displays the figures calculated from these tests, using the death rate as the dependent variable:

**Collinearity Diagnostics: Variance Proportions**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dimension | Condition Index | B2,  nursing | B3,  tests | B4,  elderly | B5,  mask | B6,  density |
| 1 | 1.00 | 0.00372 | 0.0030278 | 0.0003832 | 0.0001495 | 0.0031 |
| 2 | 2.27 | 0.00612 | 0.0000240 | 0.0002760 | 0.0000326 | 0.4957 |
| 3 | 3.11 | 0.00583 | 0.0011433 | 0.0000588 | 0.0000177 | 0.2622 |
| 4 | 6.76 | **0.68881** | 0.1974967 | 0.0004356 | 0.0013210 | 0.0705 |
| 5 | 8.54 | 0.16361 | **0.6930295** | 0.0229533 | 0.0059013 | 0.0717 |
| 6 | 26.92 | 0.04951 | 0.1052569 | **0.9615187** | 0.1277454 | 0.0655 |

\*Dependent Variable: Death

**Collinearity Statistics**

|  |  |  |
| --- | --- | --- |
| Variables | Tolerance | VIF |
| B2, Nursing | 0.8322 | 1.202 |
| B3, Tests | 0.7850 | 1.274 |
| B4, Elderly | 0.8044 | 1.243 |
| B5, Mask | 0.6110 | 1.637 |
| B6, Density | 0.7241 | 1.381 |

The table above indicates that there are three conditions with a variance proportion greater than the “acceptable” range of 0.5. When observing the variance of inflation factors, all of the variables are close to 1, meaning there is little correlation among the kth predictor and the remaining predictor variables, and hence the variance of B­k is inflated minimally by the presence of other variables.

Despite the promising results of the collinearity statistics, steps to minimize multicollinearity in the regression model will be pursued, first by removing a variable deemed nonessential to the investigation. If removing variables does not remove multicollinearity enough to make a meaningful difference, a proxy variable will be introduced to attempt to resolve this issue. Finally, if neither of those fix multicollinearity, transformations of the variables in the initial model will be attempted using log-lin, lin-log, and log-log methods to examine the relationship between changes in explanatory variables to describe changes in the dependent variable.

*Resolving Multicollinearity*

Method 1: Removing nonessential variables

The best candidate to remove from the model is the population density variable. It has an extreme outlier due to the District of Columbia having a far higher population density than that of entire states and has the two strongest correlations among variables in the model. The District of Columbia was not removed from the model because its observation in the population density variable was the only variable in the entire model containing an outlier more than three sample standard deviations away from the mean. Even after removing the District of Columbia from the dataset, changes in the significance were minimal. Regardless, removing the population density variable will also remove the only outlier in the dataset. Further, population density does not serve as an ideal proxy for how frequently people come into contact with one another (and hence increase the spread of the coronavirus); it is deeply flawed, as people are not dispersed evenly within the state’s borders—many (if not most) live in cities while the rest of the state remains rural. The following table shows the results from removing population density variable from the initial regression model:

**Regression Results after Removal of Population Density**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | .708449 | .5019 | .4466 | 1.161048 |

\*Predictors: Percent Wearing Masks, Total COVID Tests Per 100k, Percent of Population over 65, Nursing Home Residents Per 100k Adults

**Collinearity Statistics**

|  |  |  |
| --- | --- | --- |
| Variable | Tolerance | VIF |
| B2, Nursing | 0.8323 | 1.201 |
| B3, Tests | 0.8308 | 1.204 |
| B4, Elderly | 0.8619 | 1.160 |
| B5, Mask | 0.6504 | 1.537 |
| B7, Lockdown | 0.7319 | 1.366 |

**Collinearity Diagnostics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dimension | Eigenvalue | Condition Index | B2, nursing | B3, tests | B4, elderly | B5, mask |
| 1 | 5.13925 | 1.00 | 0.00395 | 0.0033095 | 0.0004333 | 0.000167 |
| 2 | 0.64916 | 2.81 | 0.01363 | 0.0000561 | 0.0003347 | 0.000041 |
| 3 | 0.12475 | 6.42 | **0.54098** | 0.2852141 | 0.0000204 | 0.000822 |
| 4 | 0.07685 | 8.18 | 0.29852 | **0.5340663** | 0.0261404 | 0.006319 |
| 5 | 0.00776 | 25.74 | 0.06136 | 0.1763269 | **0.9359389** | 0.141982 |
| 6 | 0.00224 | 47.86 | 0.08156 | 0.0010271 | 0.0371323 | **0.850669** |

\*Dependent Variable: Death

**Removal of B6, Zero-Order Correlation Coefficients (Absolute Values)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Control Variable | Variable | Mask | Tests | Elderly | Nursing |
| Death | Mask | 1.00 | 0.29 | 0.05 | 0.32 |
|  | Tests | 0.29 | 1.00 | 0.24 | 0.07 |
|  | Elderly | 0.05 | 0.24 | 1.00 | 0.22 |
|  | Nursing | 0.32 | 0.07 | 0.22 | 1.00 |

**Multivariate Regression Model Coefficients, without Population Density**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 9.903402924 | 2.635102699 |  | 3.758 | 0.00049 |
| B2, Nursing | 2.675760061 | 0.789680941 | 0.39074385 | 3.388 | 0.00147 |
| B3 Tests | -0.000002601 | 0.000013344 | -0.02250076 | -0.195 | 0.84631 |
| B4, Elderly | -0.019991412 | 0.086555518 | -0.02617393 | -0.231 | 0.81839 |
| B5, Mask | -0.105800304 | 0.028700670 | -0.48087911 | -3.686 | 0.00061 |
| B7, Lockdown | 0.035340441 | 0.404992481 | 0.01073120 | 0.087 | 0.93085 |

Although removing the density variable had negligible effects on R-square, the p-values lowered for all values except B3 [Tests], which was expected. Since collinearity between masks and nursing is still above the |0.3| benchmark, further steps should be taken to soften this issue.

Method 2: Substituting in proxies

As noted in the beginning of this section, one of the greatest problems with the state population density variable is that it does not take into account where in a state most people are clustered. In an attempt to remedy the lingering multicollinearity, a proxy variable containing urban population density was substituted in the place of the density by state population and landmass variable. States with more urban land and more people living in it have greater urban densities than states with only a few large cities but extensive landmass. The new regression model is shown in the table below:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | .711476 | .5063 | .439 | 1.156 |

\*Predictors: Percent Wearing Masks, Urban Population Density, Total COVID Tests Per 100k, Percent of Population over 65, Nursing Home Residents Per 100k Adults

**Multivariate Regression Model Coefficients, with Urban Population Density as B6**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 9.784102024 | 2.660101905 |  | 3.678 | 0.000637 |
| B2, Nursing | 2.725932124 | 0.799172208 | 0.398070524 | 3.411 | 0.001398 |
| B3, Tests | -0.000003727 | 0.000013557 | -0.032239053 | -0.275 | 0.784650 |
| B4, Elderly | 0.003151193 | 0.094735798 | 0.004125726 | 0.033 | 0.973615 |
| B5, Mask | -0.111609853 | 0.030364956 | -0.507284428 | -3.676 | 0.000641 |
| B6, Urban | 0.000101597 | 0.000163068 | 0.083235062 | 0.623 | 0.536478 |
| B7, Lockdown | 0.007179095 | 0.410271548 | 0.002179948 | 0.017 | 0.986118 |

The following tables were created after checking for multicollinearity in the new model:

**Zero-Order Correlation Coefficients (Absolute Values), with Urban Population Density Replacing B6**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mask | Urban | Tests | Elderly | Nursing |
| Mask | 1.00 | 0.41 | 0.29 | 0.05 | 0.32 |
| Urban | 0.41 | 1.00 | 0.33 | 0.40 | 0.30 |
| Tests | 0.29 | 0.33 | 1.00 | 0.24 | 0.07 |
| Elderly | 0.05 | 0.40 | 0.24 | 1.00 | 0.22 |
| Nursing | 0.32 | 0.30 | 0.07 | 0.22 | 1.00 |

**Collinearity Statistics**

|  |  |  |
| --- | --- | --- |
| Variables | Tolerance | VIF |
| B2, Nursing | 0.8239 | 1.214 |
| B3, Tests | 0.8161 | 1.225 |
| B4, Elderly | 0.7294 | 1.371 |
| B5, Mask | 0.5891 | 1.698 |
| B6, Urban | 0.6287 | 1.591 |
| B7, Lockdown | 0.7230 | 1.383 |

**Collinearity Diagnostics**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dimension | Eigenvalue | Condition Index | B2, Nursing | B3, Tests | B4, Elderly | B5, Mask | B6, Urban |
| 1 | 5.94913 | 1.00 | 0.00284 | 0.002439 | 0.000270 | 0.00011259 | 0.00340 |
| 2 | 0.65479 | 3.01 | 0.01610 | 0.000146 | 0.000373 | 0.00005330 | 0.00226 |
| 3 | 0.22183 | 5.18 | 0.13393 | 0.010452 | 0.001293 | 0.00000162 | 0.39826 |
| 4 | 0.08968 | 8.14 | 0.32190 | **0.548081** | 0.000378 | 0.00035833 | 0.33993 |
| 5 | 0.07626 | 8.83 | 0.42903 | 0.350619 | 0.025540 | 0.00645210 | 0.02264 |
| 6 | 0.00608 | 31.27 | 0.01966 | 0.087572 | **0.958441** | 0.16099759 | 0.22289 |
| 7 | 0.00222 | 51.74 | 0.07654 | 0.000691 | 0.013703 | **0.83202448** | 0.01062 |

\*Dependent Variable: Death

As the charts above illustrate, substituting urban density in for state population density did not resolve multicollinearity in the regression model. Therefore, the use of urban population density is not justified.

Method 3: Changing the form of the initial regression equation

Log-Lin Method

The “log-lin” method of correcting multicollinearity takes the natural logarithm of the explanatory variables and repeats the regression using the new, transformed explanatory variables. The dependent variable remains unchanged. It did not improve the p-values of the variables compared to the initial regression.

Lin-Log

The “lin-log” method of correcting multicollinearity takes the natural log of the dependent variable and performs the regression using the original explanatory variables on the transformed dependent variable. There is a problem with this method in particular because Death (Y) contains values of zero, which is outside the domain of the natural logarithm. For the period sampled for this study, neither Maine nor Vermont had any new deaths due to the coronavirus. To allow this transformation to take place, the zeroes were replaced with values arbitrarily close to zero. The results of the new regression are provided in the tables below:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.696563 | .4852 | .415 | 3.947 |

\*Predictors: Percent Wearing Masks, State Population Density, Total COVID Tests Per 100k, Percent of Population over 65, Nursing Home Residents Per 100k Adults

**Multivariate Regression Model Coefficients, “Log-Lin” Transformation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 3.480183738 | 0.911753549 |  | 3.817 | 0.000419 |
| B2, Nursing | 0.530960809 | 0.271503293 | 0.2318768510 | 1.956 | 0.056877 |
| B3, Tests | 0.000008938 | 0.000004720 | 0.2312030776 | 1.894 | 0.064840 |
| B4, Elderly | 0.000033734 | 0.030801910 | 0.0001320831 | 0.001 | 0.999131 |
| B5, Mask | -0.040085012 | 0.010180852 | -0.5448550142 | -3.937 | 0.000290 |
| B6, Density | 0.000119857 | 0.000039822 | 0.3825898500 | 3.010 | 0.004316 |
| B7, Lockdown | -0.121050924 | 0.139772839 | -0.1099245189 | -0.866 | 0.391158 |

\*Dependent Variable: Death (natural logarithm)

With the transformation of the dependent variable, both Density (B6) and Masks (B5) are significant at the 0.05 level of significance, while Nursing (B2) is also significant at the 0.05 level of significance using a one-tailed test. Had the sign on the coefficient estimate for Tests (B3) remained negative—as it had in all previous regressions—it would be reasonable to consider using a one-tailed test to make the variable significant at the 0.05 level of significance. The sign on the coefficient became positive after the variable transformation, however, so that would mean that the more coronavirus tests a state has per 100,00 persons, the more people per capita who die due to the virus. Conventional wisdom suggests that having more tests will *decrease* transmission of the virus and therefore lower the chances that someone will die to the virus, so a positive sign on the Tests variable—in tandem with previous regressions having a negative sign (though estimates for all regressions using Tests were all very close to zero)—rules out the possibility that a one-tailed test can be used to determine significance. The p-values improved drastically for all values with the one exception being Elderly (B4), and although the R-square dipped slightly from .5028 to .4852, the difference between the R-square and adjusted R-square in this and the initial regression remains roughly equal (~.07).

Log-Log

The “log-log” method of correcting multicollinearity involves taking the natural logarithm of both the dependent variable and the explanatory variables. As with the “lin-log” method, the values for Maine and Vermont were given arbitrarily small values to allow the natural logarithm to remain within its domain. This method did not improve the statistical significance of the variables in the regression compared to the original regression.

*An Improved Model*

The degree of multicollinearity within the original model mandated that measures be taken to alter the equation. Removing density from the model made modest improvements in the significance of variables, but the gains were merely incremental and changed little in the regression results. Substituting in a proxy variable for the state population density did not improve the results on the variables in their original form. Variable transformation proved very effective at improving the results of the regressions. Although not shown in the regressions above, the p-value in a simple linear regression of new COVID deaths against urban density was significant at the 0.1 level of significance. This made me curious to see whether a “lin-log” transformation of the urban variable would improve the results of the regression. Intuitively, the urban density variable is a better proxy for how frequently people come into contact with one another than the state population density and should therefore make a more accurate predictor in the model. Despite the fact that multicollinearity in the model increased after replacing state population density with urban density, the “lin-log” method should have eased the problems posed by the greater level of multicollinearity. The results of the regression using urban density and the “lin-log” method are the following:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.671044 | 0.4503 | 0.3754 | 0.4079 |

\*Predictors: Percent Wearing Masks, Urban Population Density, Total COVID Tests Per 100k, Percent of Population over 65, Nursing Home Residents Per 100k Adults

**Multivariate Regression Model Coefficients, Variable Substitution (B6, Urban Density) and “Log-Lin” Transformation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 3.009948283 | 0.938560913 |  | 3.207 | 0.002501 |
| B2, Nursing | 0.605513002 | 0.281971076 | 0.26443467 | 2.147 | 0.037307 |
| B3, Tests | 0.000010754 | 0.000004783 | 0.27815632 | 2.248 | 0.029626 |
| B4, Elderly | 0.007374845 | 0.033425530 | 0.02887541 | 0.221 | 0.826397 |
| B5, Mask | -0.040389822 | 0.010713635 | -0.54899814 | -3.770 | 0.000483 |
| B6, Urban | 0.000137287 | 0.000057535 | 0.33636000 | 2.386 | 0.021394 |
| B7, Lockdown | -0.122313761 | 0.144755672 | -0.11107128 | -0.845 | 0.402701 |

\*Dependent Variable: Death (natural logarithm)

Further, removing the elderly variable from the regression is an excellent choice to increase the degrees of freedom while also reducing the level of multicollinearity among the variables. Keeping both a rate of the number of people in nursing facilities—people who are typically very old—and the percent of the population over age 65 seems redundant. The reason the elderly population variable was included was to have a proxy for the population of people who are most susceptible to the virus. The nursing facility variable already fills this role, so there is little justification for keeping the Elderly variable in the model. The results after this removal are as follows:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.6705967 | 0.4497 | 0.3886 | 0.4035 |

\*Predictors: Percent Wearing Masks, Urban Population Density, Total COVID Tests Per 100k, Nursing Home Residents Per 100k Adults

**Multivariate Regression Model Coefficients, Revised Model (without Elderly [B4])**

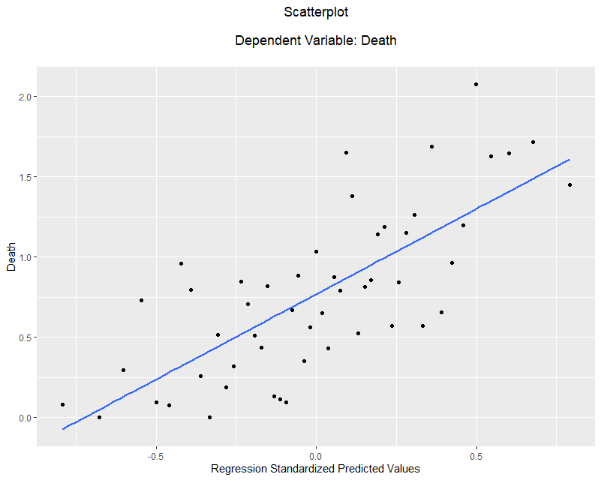
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 3.093254692 | 0.850132117 |  | 3.639 | 0.000704 |
| B2, Nursing | 0.618211907 | 0.273101048 | 0.2699804 | 2.264 | 0.028464 |
| B3, Tests | 0.000010538 | 0.000004633 | 0.2725935 | 2.275 | 0.027747 |
| B5, Mask | -0.039745745 | 0.010198721 | -0.5402435 | -3.897 | 0.000321 |
| B6, Urban | 0.000132309 | 0.000052366 | 0.3241654 | 2.527 | 0.015101 |
| B7, Lockdown | -0.120116952 | 0.142878208 | -0.1090764 | -0.841 | 0.404964 |

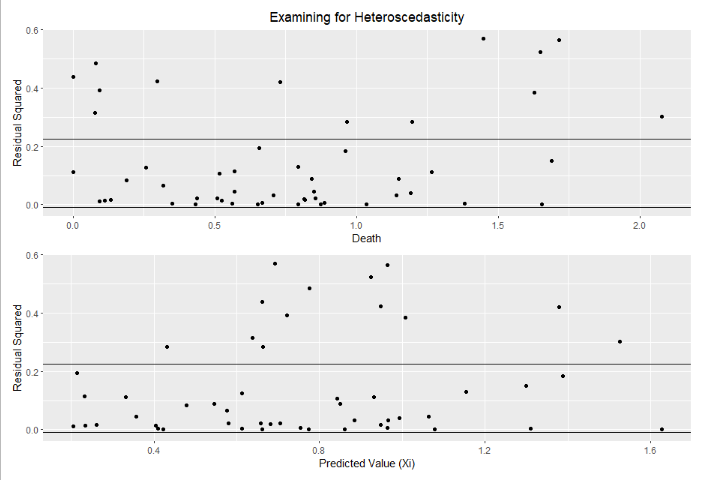
This revised model has reduced multicollinearity to a more acceptable level and improved the significance of the predictors in the process. The multicollinearity which remains is likely minimal and should not dramatically affect the conclusions drawn from this analysis. The measures taken to reduce multicollinearity also lowered the p-values dramatically and doubled the number of significant variables from two to four. Although the R-square dropped by .05 in this revised model, it is still preferable to the initial model because the results obtained in the revised model are considerably more reliable with respect to the significance of the variables.

*Identifying Heteroscedasticity*

Step 1: Visual Analysis of Data

Homoscedasticity, perhaps known better as “equal spread”, of the data in the regression is necessary for the assumptions of OLS to hold. If the standard error at a given area of the sample regression function (SRF) is not the same as all the other areas of the SRF, the model’s results may be misleading. After visual analysis, the data appears to have homoscedasticity.

Step 2: Visual Analysis of Residuals

 In addition to the visual analysis of the regression, the form of the predicted values of both the independent (Yi-hat) and dependent variables (Xi-hat) when graphed against residual squared (ui-hat2) values should be examined. A “parallel band” in the data is the only desired pattern in the output, as it suggests homoscedasticity. Horizontal lines have been added to the plots for ease of reading:

When graphed against the residual squared values, Yi and Xi exhibit moderate levels of heteroscedasticity. Empirical investigation into the presence of heteroscedasticity of the data is necessary to draw more decisive conclusions.

Step 3: Empirical Analysis

*Glejser Method*

The Glejser method for determining heteroscedasticity provides various means for evaluating the degree of heteroscedasticity within the data by running regression models on the error terms. The Glejser tests used in this analysis take the following form:

1. |u-hati| = B1 + B2Xi + . . . + BkXk
2. |u-hati2| = B1 + B2Xi + . . . + BkXk

The null hypothesis states that the data is homoscedastic, so if the regression statistics have a high p-value and a low R-square, the data has homoscedastic data. If the Glejser test’s null hypothesis is rejected, the model has heteroscedastic data. The results of the first Glejser test for the data used in this model are as follows:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.3521363 | 0.124 | 0.02668 | 0.2284 |

\*Predictors: Percent Wearing Masks, Urban Population Density, Total COVID Tests Per 100k, Nursing Home Residents Per 100k Adults

\*Dependent Variable: Absolute Value of Residuals

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 0.131727297 | 0.481247972 |  | 0.274 | 0.7856 |
| B2, Nursing | -0.041591132 | 0.154598706 | -0.04048281 | -0.269 | 0.7891 |
| B3, Tests | 0.000005235 | 0.000002623 | 0.30182808 | 1.996 | 0.0520 |
| B5, Mask | 0.000667812 | 0.005773354 | 0.02023149 | 0.116 | 0.9084 |
| B6, Urban | -0.000005181 | 0.000029643 | -0.02829155 | -0.175 | 0.8620 |
| B7, Lockdown | -0.138963268 | 0.080881367 | -0.28125579 | -1.718 | 0.0927 |

\*Dependent Variable: Absolute Value of Residuals

The Results from the second Glejser test are as follows:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.4091455 | 0.1674 | 0.07485 | 0.1678 |

\*Predictors: Percent Wearing Masks, Urban Population Density, Total COVID Tests Per 100k, Nursing Home Residents Per 100k Adults

\*Dependent Variable: Residuals-squared

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 0.027900391 | 0.353484455 |  | 0.079 | 0.937 |
| B2, Nursing | -0.048948678 | 0.113555262 | -0.06323955 | -0.431 | 0.668 |
| B3, Tests | 0.000004163 | 0.000001926 | 0.31857956 | 2.161 | 0.036 |
| B4, Mask | 0.000456061 | 0.004240622 | 0.01833897 | 0.108 | 0.915 |
| B6, Urban | -0.000003054 | 0.000021774 | -0.02213692 | -0.140 | 0.889 |
| B7, Lockdown | -0.133101001 | 0.059408678 | -0.35756964 | -2.240 | 0.030 |

\*Dependent Variable: Residuals-squared

As shown by the results of the second regression, the data has heteroscedasticity for the Tests (B3) and Lockdown (B7) variables. Another variable transformation should be tried to correct the heteroscedasticity with these variables.

Step 4: Addressing Heteroscedasticity

One way to adjust a model for heteroscedasticity is performing a “ratio transformation” on the variables. This transformation involves dividing each quantitative variable by the explanatory variable with the worst case of heteroscedasticity. The only variable quantitative variable with heteroscedastic data is the Tests (B3) data, so the it will be the basis for the transformation; Tests (B3) will be divided into each quantitative variable and Tests will be replaced with its inverse. The new regression results, along with the new results from a Glejser test, are displayed in the tables below. First, the new regression results:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.7609205 | 0.579 | 0.5322 | 0.00001154 |

\*Predictors: Ratio of Percent Wearing Masks to Tests Per 100k, Ratio of Urban Population Density to Tests Per 100k, Inverse of Total COVID Tests Per 100k, Ratio of Nursing Home Residents Per 100k Adults to Total Tests Per 100k

\*Dependent Variable: Ratio of New Deaths from COVID (natural logarithm) to Tests Per 100k

**Multivariate Regression Model Coefficients, Final Revision (Ratio Transformation)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 0.00000152031488159 | 0.00000706071492276 |  | 0.215 | 0.830491 |
| B2, Nursing\* | 0.88785326137692577 | 0.23689187645530579 | 0.4741690 | 3.748 | 0.000506 |
| B3, Tests\* | 2.38252259951173029 | 0.72288529641014920 | 1.3944244 | 3.296 | 0.001920 |
| B5, Mask\* | -0.02683572474101781 | 0.00834450892131768 | -1.2679388 | -3.216 | 0.002409 |
| B6, Urban\* | 0.00000000000003545 | 0.00000000000002041 | 0.2004041 | 1.737 | 0.089266 |
| B7, Lockdown | -0.00000466256049465 | 0.00000402153819317 | -0.1294417 | -1.159 | 0.252409 |

The results from the Glejser test are as follows:

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R-Square | Adjusted R Square | Standard Error of the Estimate |
| 1 | 0.7609205 | 0.579 | 0.5322 | 0.00001154 |

\*Predictors: Ratio of Percent Wearing Masks to Tests Per 100k, Ratio of Urban Population Density to Tests Per 100k, Inverse of Total COVID Tests Per 100k, Ratio of Nursing Home Residents Per 100k Adults to Total Tests Per 100k

\*Dependent Variable: Absolute Value of Residuals

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unstandardized Coefficients | | Standardized Coefficients | t-value | p-value |
| Estimate | Std. Error | Estimate |
| (Intercept) | 0.00000368497916192 | 0.00000437890000015 |  | 0.842 | 0.405 |
| B2, Nursing\* | 0.24571272645288789 | 0.14691512816949182 | 0.30494211 | 1.672 | 0.101 |
| B3, Tests\* | -0.31697531808158635 | 0.44831755129423079 | -0.43110309 | -0.707 | 0.483 |
| B5, Mask\* | 0.00429248422993774 | 0.00517508078381975 | 0.47129345 | 0.829 | 0.411 |
| B6, Urban\* | -0.00000000000000137 | 0.00000000000001266 | -0.01799506 | -0.108 | 0.914 |
| B7, Lockdown | -0.00000214611858927 | 0.00000249406947984 | -0.13845264 | -0.860 | 0.394 |

To simplify the analysis, there are three versions of the regression model which will be discussed for the remainder of the paper. The first is the initial regression, which was the very first multiple regression in this paper. The second is the revised model, which involved taking the natural logarithm of the dependent variable (Deaths) and replaced state population density with urban population density (B6). The third, most recent regression (with the ratio transformation) will be referred to as the final regression.

As the tables above show, the ratio transformation removed heteroscedasticity from the variables in the model, in addition to bumping the R-square up by about 0.13. Nursing\* (B2), Tests\* (B3), and Mask\* (B5) are all significant at the 0.05 level of significance using a two-tailed test. Urban\* (B6) is also significant at the 0.05 level of significance when using a one-tailed test. A one-tailed test is justified because conventional wisdom suggests that the closer people live to one another—the more densely populated the state’s urban landmass—the more likely someone is to come into contact with an infected individual. When the virus spreads faster, more people are at risk of getting sick and thereby dying from the virus. The positive signs on the coefficients in both the revised model and final regression suggest this is the case, and it seems unreasonable to assume that as urban population density increases, new deaths due to COVID decrease. The Glejser tests on the final regression all showed that none of the explanatory variables contained heteroscedasticity, judging by the fact that none of the variables had significant p-values in the two tests performed.

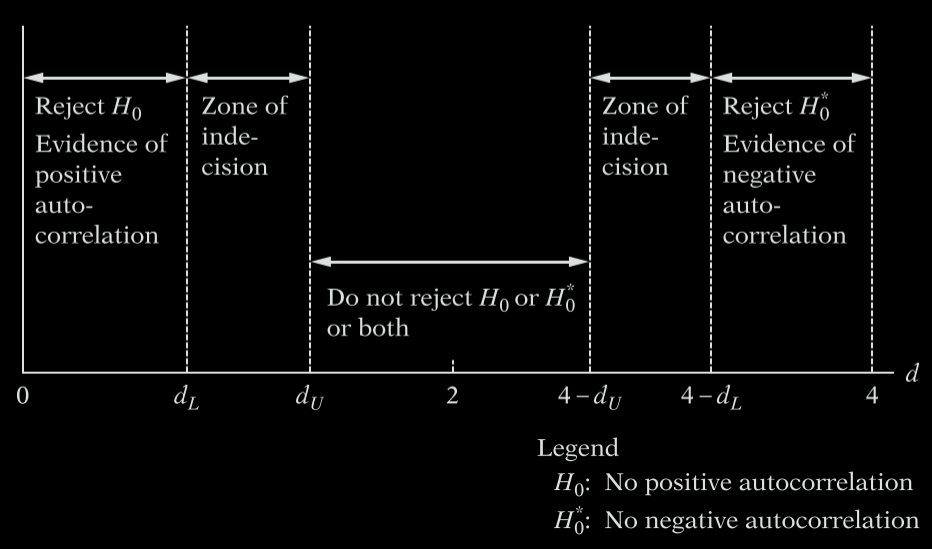
*Testing for Autocorrelation*

Autocorrelation refers to the degree of correlation between the values of the same variable across different observations in the data. Autocorrelation occurs most often in time series and panel data, and can be caused by a multitude of factors, though it is quite rare in cross-sectional data. To test for autocorrelation, the Durbin-Watson test will be performed on the data. If the D-W statistic is close to 2.0, the data are unlikely to be autocorrelated. The results from the Durbin-Watson test using the data from the final regression are shown below:

**Durbin-Watson Test**

|  |  |  |  |
| --- | --- | --- | --- |
| lag | Autocorrelation | D-W Statistic | p-value |
| 1 | 0.05859691 | 1.833173 | 0.486 |

**Durbin-Watson *d* Statistic**



**D-W Statistic: 1.83**

Since the D-W statistic is close to 2.0, it is reasonable to dismiss the possibility that the data are autocorrelated. This result is not surprising because the data used in this study is cross-sectional.

**Conclusion**

The data used in the regressions provide a unique look into the various factors determining the number of people who die from the coronavirus. One conclusion which was very interesting was that the elderly variable (B4) was not significant in any of the regressions performed in the study. This disputes the narrative claiming states with older populations, such as Florida, are more likely to have higher death rates solely due to the fact that there are more people over the age 65 as a percent of the population. The data in this study provides mixed results on the role lockdowns play on the death rate. It was significant in the simple regression model, though lingering multicollinearity may have raised its p-value. Regardless, the significance of the lockdown variable should not be entirely trusted because of the time disparity noted in the section on hesitations within this model.

The final regression’s testing variable disputes claims that increased test capacity will lower new deaths to the virus. The sign on the coefficient for Tests (B3) was positive with a value of 2.38, which means that an increase of one in the inverse of tests per 100k people means the ratio of the natural logarithm of new COVID deaths per 100k to the number of tests per 100k increases by 2-3. This defied expectations from the beginning of the study stating that more tests would reduce the number of deaths. Further research should be undertaken to confirm the results of this study. Even if the number of tests per 100k implies that there are more deaths due to the coronavirus, perhaps testing capacity aids other purposes, such as slowing the spread of the virus. Alternatively, though, the results from testing could be a case study on opportunity cost—it is possible that states increased spending on the production of tests at the expense of other needed equipment. Again, this is mere speculation and further research is required to determine the efficacy of tests to reduce the number of COVID deaths.

The Mask (B5) variable was highly significant in almost all regressions done in this study. In each revision of the model, however, the masks variable had a coefficient close to zero, although it was negative. This suggests that masks only have a modest impact on the increase in deaths. Put differently, masks only reduce the deaths due to COVID modestly. Despite low practical significance, masks do what was predicted at the start of beginning of this paper. The more people wear masks, the fewer the deaths due to COVID—controlled for population.

Another surprising result from the final regression was that the Urban (B6) variable, though statistically significant using a one-tailed test, was not practically significant. This upsets the predictions from the beginning of the paper expecting moderate practical significance. The sign of the coefficient is positive—as urban densities increase, the number of deaths due to COVID increases, but they increase by a negligible amount. Urban densities seem to have little bearing on the number of new deaths due to the virus, which is surprising because one expects that as urban densities increase, people are more likely to come into contact with one another. Further exploration into that expectation would also make for a compelling follow-up study.

The lockdown variable had promising results in the simple regression model but continued to disappoint the prediction of moderate significance made at the beginning of the paper. Judging by the results of the final regression, the lockdown variable was not significant enough to rule out the possibility that the true value of B7 was zero. This suggests strict lockdowns do not have one of their most important intended effects—reducing the number of deaths due to COVID. To reiterate from the Hesitations section of the paper, however, this variable should be recalculated with more accurate data; simply because a state had a lockdown in mid-September does not necessarily mean they have a strict lockdown in mid-October—the policy could have since reversed for states with or without a strict lockdown.

In all regressions, the variable of interest—Nursing (B2)—was highly significant. What is worth mentioning about this variable is that it says nothing about the conditions of the facilities in which the residents live, staffing shortages, or a whole host of other factors relevant to the spread of the virus. Simply the number of patients per the adult population was enough to make the variable very significant. Future research taking discrepancies between facilities into account may tell a very interesting story. This outcome is consistent with the publicity surrounding deaths coming from nursing homes throughout the country. The nursing home variable supports the notion that the coronavirus has hit these facilities especially hard, claiming the lives of many. In a practical sense, these results support what people have thought all along: those at greatest risk of developing terminal illness—those in nursing facilities—should take the utmost care to avoid close contact with carriers of the virus. I leave for the reader the decision on whether the Governor of New York State has any culpability in the death of Granny.

1. Connor Krenzer is an undergraduate student at Siena College (cr14kren@siena.edu). Department of Economics, 102 Kiernan Hall, Siena College, Loudonville, NY 12211. [↑](#footnote-ref-1)