Modern Applied Data Analysis Project

Aidan Troha - College of Public Health, University of Georgia

5/7/23

# 1. Abstract

Data is obtained from the Georgia Department of Public Health (GDPH) on deaths attributed to COVID-19 and from the University of Georgia’s Carl Vinson Institute on county-wide economic and political data. LASSO (Least Absolute Shrinkage and Selection Operator) methodology is employed to select the most effective, simplest model to predict county-wide COVID-19 deaths in Georgia. Poisson regression is employed to estimate the total number of COVID-19 related deaths per county. Predictors included in the model were high school graduation rates, hospital bed capacity, and log per capita income and includes interaction terms.

# 2. Introduction

## 2.1 Background

Sars-Co-V-2, most commonly known as COVID-19, is a highly contagious respiratory viral infection from a class of virus known as corona viruses that was first recorded in humans at the end of 2019. This class of virus is particularly dangerous as hundreds of thousands of Americans have died, directly or indirectly, from complications relating to Sars-Co-V-2. Early into the Sars-Co-V-2 pandemic, Southern states were disproportionately affected by the pandemic than some other regions of the United States (Johnston & Chen (2020)).

Particularly, social and political aspects, such as income inequality, political ideology, and education may have an affect the chance of infection with COVID-19 (Liao & Maio (2021)). Due to efforts from the University of Georgia, state-wide data is routinely compiled regarding social and political aspects of the population of each county in the state.

## 2.2 Objectives

This study endeavors to synthesize an understanding of the relationships between death as a complication of COVID-19, and by extension contraction of COVID-19, itself, with other potential factors of disease occurrence, such as social and political factors and demographic factors to help predict which groups are at higher risk of COVID-19 infection in Georgia.

# 3. Methods

## 3.1 Data Acquisition

The data used in this study is derived from an open source initiative from the Georgia Department of Public Health (GDPH) in an effort to provide transparency about the pandemic and provide frequently updated trend data for Georgia residents to stay informed about the current situation. The data can be downloaded from the site dashboard [here](https://ga-covid19.ondemand.sas.com/) (GDPH 2023).

In addition, socio-political data was taken from the University of Georgia’s Carl Vinson Institute of Government database which has been providing data to state agencies for over 30 years. Topics include anything from agriculture to labor statistics to transportation. Various data files were used, including files from: the economics, health, education, government, and population tabs. The dashboard to the Carl Vinson Institute of Government data archive can be found [here](https://georgiadata.org/data/data-tables) (UGA 2023).

Data was synthesized from these two sources in such a way that deaths relating to COVID-19 were assessed at the county-wide level.

## 3.2 Analysis Steps

### 3.2.1 Data Summary

The data from GDPH includes individual-level data with about 35,000 subjects that includes county origin, age, and other demographics data. The data from the Carl Vinson Institute is county level data including the relevant aforementioned demographics topics. Since the data from UGA has a less fine metric (county-level instead of individual-level), the individual-level data was summarized to adhere to the county level metric by way of summing the number of cases for each county. Further analyses used the county identifier to compile predictor metrics.

### 3.2.2 Data Processing

Variables were chosen according to the year 2020 in which the GDPH data was collected, and variable names were altered to better accommodate variable calls in R. Some variables were mutated, via dplyr, to factors to better account for the bivariate nature of certain conditions like party majority and graduation improvement. Notes according to the original data, though not included in the final analyses, were used to make judgements about some observations found in the data. The data was joined according to county name. Following this, Poisson regression was perform to predict the number of deaths as complications from COVID-19 from covariates such as per capita income, population, high school graduation rates, and party majority by county in Georgia.

# 4. Results

## 4.1 Exploratory Analysis

|  |
| --- |
| Figure 1. Number of deaths from COVID-19 by county in Georgia by the projected population for 2020. Trend line represents the estimates association between projected population and number of deaths attributed to COVID-19; grey area represents the 95% confidence interval for the trend line. |

Based on previous studies which looked at incidence fluctuations based on population size and ambient temperature changes, relative population size may be a relevant predictor for COVID-19 attributed deaths (Jahangiri et al. (2020)). Figure 1 shows the relationship between population size and deaths attributed to COVID-19 in Georgia counties, ultimately showing a positive correlation between population size and COVID-19 incidence.

|  |
| --- |
| Figure 2. Distribution of deaths attibuted to COVID-19 by county in Georgia separated by party majority in 2020 (>50%) for each county. |

Figure 2 shows the number of deaths attributed to COVID-19 separated by counties that are Democrat controlled versus Republican controlled. There is a clear difference in distribution between the median number of COVID deaths in each category. This indicates that modeling efforts may be more fruitful if political prominence in a county is included in the model.

|  |
| --- |
| Figure 3. Number of deaths from COVID-19 by county in Georgia by hospital bed capacity in 2020. Trend line represents the estimates association between hospital bed capacity and number of deaths attributed to COVID-19; grey area represents the 95% confidence interval for the trend line. |

Cavallo et al. discuss the implications of hospital capacity and COVID-19 complications in their study in an effort to predict the most effective course of action in overly serviced hospitals (Cavallo et al. (2020)). Based on the results seen in Figure 3, it may be fair to conclude an association between hospital capacity and deaths that have been attributed to COVID-19.

|  |
| --- |
| Figure 4. Number of deaths from COVID-19 by county in Georgia by log per capita income in 2020. Trend line represents the estimated association between log per capita income and number of deaths attributed to COVID-19; grey area represents the 95% confidence interval for the trend line. |

Figure 4 demonstrates the clear exponential relationship between log per capita income and COVID-19 deaths, albeit allowing for wide variation. Nevertheless, per capita income might help to explain COVID-related deaths in Georgia as available income is known to influence health prospects and healthcare options (Liao & Maio (2021)).

|  |
| --- |
| Figure 5. (Left) Number of deaths from COVID-19 by county in Gerogia by high school graduation rate in 2020. Trend line represents the estimated association between graduation rate and number of deaths attributed to COVID-19; grey area represents the 95% confidence interval for the trend line. (Right) Distribution of deaths attributed to COVID-19 in the State of Georgia organized by county graduation status (‘Improved’ or ‘Not Improved’). |

Figure 5 shows the relationship between high school graduation rates and COVID-19 attributable deaths in Georgia as parameterized by county. There appears to be a positive association between the number of graduates and the number od deaths which demonstrates the opposite direction of the expected association. Additionally, whether graduation rates are increasing or not seems to have very little to no effect on COVID-19 deaths. These associations could be influenced by other confounders as many studies have found that education level have a significant influence on COVID-19 disease incidence (Wake (2020)).

## 4.2 Regression Modeling Outcomes

| RMSE.for.Main.Model | RMSE.for.Main.Model.and.Interactions |
| --- | --- |
| 72.07711 | 52.88645 |

Table 1. RMSE metrics for each of the resulting LASSO cross-validation methods previously described.

Poisson regression was selected to model for count data in the form of predicted number of deaths attributed to COVID-19. Model workflows were generated via the tidymodels workflow, and LASSO and cross-validation methodology was employed to select the best model for the data. Two models, a main effects model and a saturated model including all main predictors and their interaction terms, were generated using LASSO. RMSE is used to compare across the two resulting models.

According to Table 1, the saturated model (main effects + interactions) fits the data better than the main effects model. For future analyses, the saturated model will be used to present trends across the predictor variables.

## 4.3 Model Performance

|  |
| --- |
| Figure 6. (Left) Main effects model predictions compared to observed number of deaths by county. (Right) Saturated model predictions compared to observed values by county. Dashed line represents a 1-to-1 interaction between observed and expected values. |

Figure 6 compares the model fit between the main effects model and the saturated model using an observed versus predicted plot format. Counties are categorized by party majority and relative population size. According to the figure, the saturated model fits slightly better than the main effects model alone as more points are closer to the dashed line.

|  |
| --- |
| Figure 7. (Left) Main effects model residuals compared to the observed number of deaths by county. (Right) Saturated model residuals compared to the observed number of deaths by county. Red line indicates a residual of 0. Blue dashed lines represent 2 RMSE values away from 0. |

Figure 7 shows a similar comparison between the two models wherein the saturated model is a better fit for the data. However, this plot shows that the saturated model typically accurately predicts large, Democrat-driven cities but less accurately predicts moderately sized Republican-driven counties. All in all, the saturated model is preferable as it accounts for a smaller amount of variation between predictions.

### 4.3.1 Predictors

|  |
| --- |
| Figure 8. (Top Left) Association between hospital bed capacity and the number of COVID-19 related deaths by county in Georgia. (Top Right) Association between log per capita income and the number of COVID-19 related deaths by county in Georgia. (Bottom) Association between high school graduation rate and the number of COVID-19 related deaths by county in Georgia. Trend lines represent predictions from the model with grey areas representing 95% confidence interval values. |

The generated model depends largely on 3 main effect parameters: hospital bed capacity, log per capita income, and high school graduation rate. Figure 8 above demonstrates the relationships each of the three predictors has with deaths from COVID-19 in accordance with their interaction terms. Deaths attributed to COVID-19 appear to increase with increasing hospital bed capacity. Log per capita income varies relatively little with respect to COVID deaths. High school graduation rates and COVID deaths appear to be positively correlated.

# 5. Discussion

From the predictors originally chosen, only a few appeared to effectively detect a strong relationship with deaths attributable to COVID-19. Based on the LASSO cross-validation methods used in the analyses, the interaction terms are integral to understanding the multifacetted interactions between these predictors that influence the rate of COVID deaths at the county level in Georgia.

Based on these findings, an in-depth analysis of demographics not included in this study should be obtained. Many of the predictors discussed in this study influence or are influenced by demographics such as race or religion. These could account for the wide variation seen in the model and could help to better tune future models since other studies have shown that racial disparities have a strong influence on COVID death rates (Karmakar et al. (2021)).

In addition, other cross-validation methodologies may be employed to better tune the model such as a the use of decision trees or random forest analysis. An issue with LASSO is that it often forces out predictors without regard to model hierarchy. This occurs when main effects removed from the model (their coefficient(s) are reduced to zero) while interaction terms with that predictor remain in the model. Other methodologies do not have this problem and could better account for these types of situations.

# 6. Conclusion

Of the predictors included in the models produced from these analyses, hospital bed capacity, log per capita income, and high school graduation rates appeared to be important predictors in the saturated model which out performed the main effects model. These predictors show positive correlation with COVID deaths in Georgia, though further analyses should include demographics data at the individual level to better refine the model.

# 7. References

Cavallo, J. J., Donoho, D. A., & Forman, H. P. (2020). Hospital capacity and operations in the coronavirus disease 2019 (COVID-19) pandemic - planning for the nth patient. In *JAMA Health Forum* (Vol. 1, p. E200345). American Medical Association. <https://doi.org/10.1001/jamahealthforum.2020.0345>

Jahangiri, M., Jahangiri, M., & Najafgholipour, M. (2020). The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of iran. *Science of the Total Environment*, *728*. <https://doi.org/10.1016/j.scitotenv.2020.138872>

Johnston, C. D., & Chen, R. (2020). The COVID-19 pandemic and its impact on the southern united states. *Source: Journal of Comparative Family Studies*, *51*, 314–323. <https://doi.org/10.2307/26976653>

Karmakar, M., Lantz, P. M., & Tipirneni, R. (2021). Association of social and demographic factors with COVID-19 incidence and death rates in the US. *JAMA Network Open*, *4*. <https://doi.org/10.1001/jamanetworkopen.2020.36462>

Liao, T. F., & Maio, F. D. (2021). Association of social and economic inequality with coronavirus disease 2019 incidence and mortality across US counties. *JAMA Network Open*, *4*. <https://doi.org/10.1001/jamanetworkopen.2020.34578>

Wake, A. D. (2020). Knowledge, attitude, practice, and associated factors regarding the novel coronavirus disease 2019 (COVID-19) pandemic. In *Infection and Drug Resistance* (Vol. 13, pp. 3817–3832). Dove Medical Press Ltd. <https://doi.org/10.2147/IDR.S275689>