MATH 7343 Applied Statistics
Course Project:
State Coronavirus Severity Association with Various Statewide Metrics

Northeastern University
Team C:
Josh Galloway (galloway.j@northeastern.edu),
Elal Segev (segev.e@northeastern.edu),
Zachary Larson (larson.z@northeastern.edu),
Louis Carpenter (carpenter.lo@northeastern.edu)

April 2021

1 Introduction

As the last year has been dominated by news surrounding the varied responses and other contributing factors to the severity of novel coronavirus (COVID-19) infections and deaths from state-to-state in the United States, Team C chose to perform a statistical analysis of the differing factors possibly associated with the intensity of infection and/or deaths in each state.

The questions investigated are as follows:

- Do economic factors such as Gini coefficient (a measure of economic inequality), per capita income, GDP and so on correlate to the severity of infection or death?
- Do policy decisions such as having a statewide mask mandate or closing public schools correlate to severity of infection or death?
- Do heath system metrics such as number of ICU beds, per capita health spend, number of physicians and hospitals correlate to COVID-19 death rate in the state?
- Does the composition and lifestyle choices of the population within the state correlate to increased severity of COVID-19 outcomes?

2 Dataset

For the questions regarding economic factors, policy decisions, and population make-up/habits, a publicly available dataset found on Kaggle [Ran20] with data amended from CNN outlining the statewide mask-wearing mandates [Asm20] both from November of 2020, was utilized. Inquiry with respect to health system metrics utilized a multi-source dataset with data for COVID-19 metrics as recent as March 2021 and other sources dating back to 2014. A detailed description of the specifics for these sources is provided in the appendix [Appendix C].

3 Economic Factors

The objective of this study is to determine whether there is a statistical relationship between economic factors, including the Gini coefficient, per capita income and GDP, and the severity of impact of COVID-19 by state. The raw number of infections and deaths in each state was normalized to be the natural log of infections and deaths per 100,000 people, as larger states will inevitably have more infections and deaths. A performed a series of statistical tests were then performed to determine whether these economic factors truly influence the impact of the disease and how.

The first step in determining the relationship between economic factors and impact of disease is formulating the correct hypothesis to test: determining if a higher Gini coefficient, income, or GDP has an association with infections and deaths. The data was then split at the median Gini, income and GDP, and tested the corresponding number of infections and deaths in each group; thus, the hypothesis to be tested is that the mean of the infections and deaths for each group are the same. The Wilcoxon rank sum test, ANOVA, and linear regression were used to test the hypothesis.

The Wilcoxon rank sum test returned a p-value of 0.0409 between number of infected and income, as well as a p-value of less than 0.001 between number of deaths and Gini coefficient. ANOVA returned a p-value of 0.0045 between the number of deaths and Gini coefficient. Linear regression tells us the same, with a death rate and Gini coefficient p-value of less than 0.001. However, linear regression also reveals that the three metrics being tested are quite poor when it comes to explaining the variance of the infected data, with none of them having a significant p-value and the overall fit having an adjusted R^2 of just 0.03 [Figure 5].

Based on the results of the statistical tests, there is strong evidence that economic factors do indeed have an association with the severity of impact of COVID-19. With significant p-values (below 0.05) among these tests, the study has determined with some confidence that there is an association between Gini coefficient and death rate. The association between income and infection rate was only found in a single test and, if the Bonferroni correction is taken into consideration (which would lower the required significance to 0.025), does not reach a high enough significance threshold so there is not enough evidence to dispute the null hypothesis that there is no association. Similarly, there was little evidence to suggest GDP is associated with the impact of COVID-19 either

4 Policy Decisions

States have implemented policies such as face masks mandates and school closures to curb the spread of COVID-19. To measure efficacy of such policies, infection and death rates for states that enforced policy measures are compared to the same measures for states that did not impose mandates. Infection and death rates are recorded as infections/deaths as proportion of population times 100,000 (100K).

States with mask mandates and school closures are marked with 1 in the dataset while those without mandates are marked with 0. 3 of the 51 states (6%) did not close schools and 12 (24%) did not enforce mask mandates. Only 1 (1.9%) had no mandate of either kind (Nebraska) while 37 (73%) had both mandates in effect [Figure 6].

This research utilized the two-sample t-tests, the rank-sum test, and regression analyses to test the difference in means of infection and death rates. The paired t-test revealed only one significant difference in means: the mask mandate with infections per 100K. At 95% confidence, states with a mask mandate had between 460 and 1,563 less infections per 100K than states without. No other means testing revealed any statistically significant results [Figure 7].

Univariate and bivariate logistic regressions were used to measure correlation between mask mandates and school closures (X variables) on infection rates and death rates (Y variables). Among the six regressions employed, only two were correlated at the 0.05 statistically significant threshold. Only mask mandates were negatively correlated to infections per 100K with a slope of -1011. That number drops to -1018 when school closures were included. There is no significant relationship between school closures and infections per 100K nor are there any between masks and school closures with deaths per 100K [Figures 8 and 1].

X Variable 2	Slope (X1, X2)	Intercept	P-Value (X1, X2)	Adjusted R-Squared	Std. Error (X1, X2)
n/a	-1011.8	3564.5	0.0049	0.1333	343.3
School Closure	-1018.8, 218.3	3364.3	0.0051, 0.7289	0.1174	347,625.5

Figure 1: Mask Mandate (X2) on Infections per 100K

This study was able to determine a statistically significant difference in infections per 100K between states with and without mask mandates. Moreover, there is a significant negative correlation between mask mandates and infections per 100K. However, the relationship explains little variation as the R^2 is 0.15. In short, this research can say that enforcing a mask mandates has a correlation of reduced infections per 100K by an estimate of 1011 [Figures 8 and 1].

5 Health System Metrics

The following section studies the possible association between health system metrics, such as the number of ICU beds, hospitals, number of physicians, and per capita health

spending, and deaths that occurred during the COVID-19 pandemic. These examined statistics are quantitative datasets. This section looked at these metrics in a statewide examination, so the sample size is 50. Creating histograms showed the results follow an approximate normal distribution. Thus, there is a possibility for underlying statistical distributions for the dataset and this report will compare COVID-19 deaths per 1000 populations to the health system metrics using both parametric and nonparametric tests. Since the research is observational, only association not causation could be inferred. By grouping the states by various metrics that can be seen in table [Figure 2], the COVID-19 deaths were compared using various health metrics using parametric and nonparametric methods including: Two Way ANOVA, Pairwise Bonferroni, Tukey's HSD, and Wilcox Sum Rank [Figure 3]. Additionally, the relationship between the health system metrics and deaths (two random variables) was examined in terms of Spearman's correlation and Pearson's correlation.

Health System Metrics	ICU Beds (per 1,000 population)	Hospital (per 100,000 population)	Number of Physicians (per 100,000 population)	Per capita health spending ($)
Group 1	≤ 2	≤ 1.3	≤ 260	≤ 8000
Group 2	>2	>1.3	>260	>8000

Figure 2: Grouped Health Metrics and Number of COVID-19 Deaths per 1,000 Populations

Health System Metric	ICU Beds (per 1,000 population)	Hospital (per 100,000 population)	Number of Physicians (per 100,000 population)	Per Capita Health Spending ($)
Two Way ANOVA	0.000107	0.562	0.211	0.8590
Pairwise Bonferroni	0.00011	0.56	0.21	0.86
Tukey's HSD	0.00011	0.562	0.211	0.859
Wilcox Rank-Sum Test	0.0001066	0.467	0.2615	0.711
Pearson's Correlation ρ	0.598	0.201	0.0119	-0.0551
Pearson's Probability p	5.693e-06	0.161	0.935	0.704
Pearson's 95% Confidence Interval	(0.381, 0.753)	(-0.0817, 0.454)	(-0.267, 0.289)	(-0.328, 0.227)
Spearman's Correlation ρ_s	0.620	0.166	-0.0544	-0.0467
Spearman's Probability p	3.43e-06	0.248	0.707	0.747

Figure 3: Test Results for Studying Association and Correlation of Health Metrics and COVID-19 Deaths

For the correlation statistics, there are no significant relationships between the number of physicians and per capita health spending and COVID-19 deaths. For the hospitals, there is the possibility of a weak positive relationship seen through the Pearson and Spearman correlations; however, the null hypothesis could not be rejected. The Spearman's and Pearson's methods for hospital beds, however, show a moderate positive relationship between the number of ICU Beds and COVID-19 deaths and reject the null hypothesis of no relation between these variables. This is further supported by the scatter plot showing the visual relationship between ICU Beds and COVID-19 deaths [Figure 10]. The relation between deaths and beds has a p of 5.69e-06, rejecting the null hypothesis of no relation between them at the 0.05 significant level. The linear relationship between deaths and beds has an R^2 of 0.3576 and is given by the equation below.

$$\text{Relationship: COVID-19 Deaths per 1000} = -0.006336 + 0.668340 \times \text{Beds per 1000} \quad (1)$$

This study was unable to significantly determine the possibility of correlation or association between the COVID-19 deaths and health metrics of hospitals, physicians, and per capita health spending. However, there is an association and moderate correlation between hospital beds and COVID-19 deaths that is statistically significant to the 0.05 level. This association is modelled by the linear relationship in equation [Equation 1].

6 Population Composition and Lifestyle Choices

In order to investigate the question—Does the population composition (Age, Sex-Ratio) and lifestyle choices (Percent Urbanization, Smoking Rate) effect the severity of COVID-19 outcomes within a state?—the natural log of deaths per 100,000 people (LoDR) was calculated for each state. This metric was used to decouple the effects of population differences from state to state and the natural log was utilized to transform the data to effect a more normal distribution [Figure 11]. All hypotheses in investigation of this question were tested with a significance level of 0.05.

The selected variables for investigation were then broken into groups. States were grouped by age-range fractions following the rubric: "Older" states have brackets 26-54 and 55+ both greater than 0.33, "Younger" states have brackets 0-25 and 26-54 both greater than 0.33, and all other states are grouped as "Middle" aged. The remaining variables were all grouped as lower quartile, interquartile (IQR), and upper quartile. The resulting groups are assumed to be independent.

To test the resultant group means/medians for equality, first, normality, and homoscedasticity were checked via the Shapiro-Wilk and Breusch-Pagan tests, respectively. The resultant p-values were adjusted by way of the Bonferroni technique, and conclusions drawn in regards to the shape of the underlying distributions. Then, an appropriate inner-variable multiple group mean/median test for the null hypothesis of equality was selected and conducted with false discovery rate adjustment. As a result, Sex-Ratio and Urbanization were found to have at least one inner-variable group mean/median differing from the others indicating that a significant association to LoDR between one of the variable's groupings may be present [Figure 12].

After discovering that the Sex-Ratio and Urbanization inner-variable groups have at least one differing population mean/median LoDR, testing of all individual groups within the two variables was conducted to identify where the specific differences lie. Testing by way of pairwise Wilcoxon rank sum test with Bonferroni multiple testing adjustments demonstrated the specific differences, and from this, a final null hypothesis was developed and tested.

For Sex-ratio, the "More Male" states' median varied from the "More Female" states. Further testing with the null hypothesis, "More Male" median LoDR > "More Female" was conducted. Here, the null hypothesis was rejected; thus, "More Male" states have a lower median LoDR than "More Female" states.

For Urbanization, the "More Urban" states' median LoDR varied from the "IQR" and "More Rural" states', and here the null hypothesis "More Rural" median LoDR > "More

Urban" was rejected as well. Thus, "More Rural" states have a lower median LoDR than "More Rural" states [Figure 4].

	Sex Ratio				Urbanization		
	IQR	More Female	More Male > More Female		IQR	More Rural	More Urban > More Rural
More Female	0.14	--	1.34E-05	More Rural	0.788	--	0.0036
More Male	0.16	8.00E-05	Reject Null	More Urban	0.027	2.20E-02	Reject Null

Figure 4: Sex-Ratio and Urbanization Inner-variable Specific Median Differences

As such, an association between both sex-ratio and urbanization within a state and severity of COVID-19 outcomes measured by log of death rate was identified and confirmed. States with a more female (lower quartile) sex-ratio are more likely to have a more severe COVID-19 outcomes than states with a sex-ratio leaning toward male (upper quartile). Additionally, states in the upper quartile in urbanization suffered more severe COVID-19 outcomes than other states.

7 Conclusion

Overall, several significant conclusions were able to be drawn from the analysis of the COVID-19 datasets. It was found among economic factors that the statewide Gini coefficient has a significant association to death rate. For policy decisions, it was found that statewide mask mandates have a significant association with reduced infection rate. The number of hospital beds available in a state was found to have significant association with death rate. And finally, the states in the upper quartile of percent urbanization and lower quartile of sex-ratio (more female) were shown to have a higher association with increased death rate.

Appendices

A Economic Factors

	Wilcoxon		ANOVA		Linear Regression (F-Test)	
	Infection Rate	Death Rate	Infection Rate	Death Rate	Infection Rate	Death Rate
Gini Coefficient	0.617	2.40E-06	0.903	0.0045	0.266	5.00E-08
Income	0.0409	0.564	0.256	0.981	0.108	0.123
GDP	0.239	0.407	0.143	0.571	0.445	0.0284

Figure 5: Economic Factors Hypothesis Testing P-Values and Results (Highlighted is Rejected Null)

B Policy Decisions

Mandate	Requirement	Avg Infections per 100k	Avg Deaths per 100k
Mask	Yes	2553	60
Mask	No	3564	57
School Closure	Yes	2797	61
School Closure	No	2685	33

Figure 6: Grouped Infection Rate and Death Rate per 100K for States by Policy Mandate

Summary		Two Sample T-Test (Unequal Var)			Wilcox Rank-Sum	
Mandate	Measure	95% Lower	95% Upper	P-Value	W	P-Value
Mask	Infections per 100K	-1563	-460	0.0008	98	0.0019
Mask	Deaths per 100K	-18	24	0.7818	224	0.8351
School Closure	Infections per 100K	-4456	4681	0.9279	61	0.6910
School Closure	Deaths per 100K	-18	73	0.1456	108	0.1640

Figure 7: T-Tests and Wilcox Rank-Sum Tests for Infections per 100K and Deaths per 100K by Policy Mandate

X Variable 1	X Variable 2	Y Variable	P-Value	R-Squared
Mask Mandate	n/a	Infections per 100K	0.0049	0.1506
Mask Mandate	n/a	Deaths per 100K	0.8184	0.0011
School Closure	n/a	Infections per 100K	0.8679	0.0006
School Closure	n/a	Deaths per 100K	0.2212	0.0304
Mask Mandate	School Closure	Infections per 100K	0.0187	0.1528
Mask Mandate	School Closure	Deaths per 100K	0.4707	0.0309

Figure 8: Univariate and Bivariate Regressions for Infections per 100K and Deaths per 100K by Policy Mandate

C Heath Metrics

C.1 Dataset

To examine the COVID-19 deaths, this report used the COVID-19 Tracking Project dataset, which was used in the form of a CSV file [API]. The data examined are cumulative state COVID-19 deaths from the start of the pandemic till March 7, 2021. To normalize the health system metrics, this report used the Nations Online "Population of the US States and Principal US territories", which provided estimates for the state populations as of December 2019 [Kla19]. To examine the health metrics for ICU Beds and hospitals, this report used the American Hospital Directory's Hospital Statistics by State data [ELA20]. This dataset is based on each hospital's Medicare cost report updated May 29, 2020. The health system metrics were normalized by state population. To examine possible relations between the number of physicians per 100,000 and COVID-19 deaths,

this report examined the 2019 State Physician Workforce Data Report published by the Association of American Medical Colleges [Goo19]. To study per capita health spending by state, a 2014 study conducted by the Centers for Medicare & Medicaid Services was used [ELA17]. The information published seven years ago is dated, but this report is assuming the information is still relevant.

C.2 Supporting Figures

Statistic	Min	25th Percentile	Median	Mean	75th Percentile	Max
Deaths per 1,000 populations	0.3143	1.1036	1.5114	1.4702	1.8731	2.654
Bedspaces per 1,000	1.338	1.881	2.171	2.369	2.618	9.719
Hospitals per 100,000	0.8011	1.007	1.183	1.338	1.502	2.825
Physicians per 100,000	191	230.2	263	271.2	300.5	450
Per Capita Health Spending ($)	5982	7381	8092	8260	8918	11064

Figure 9: Descriptive Statistics of State Health System Metrics

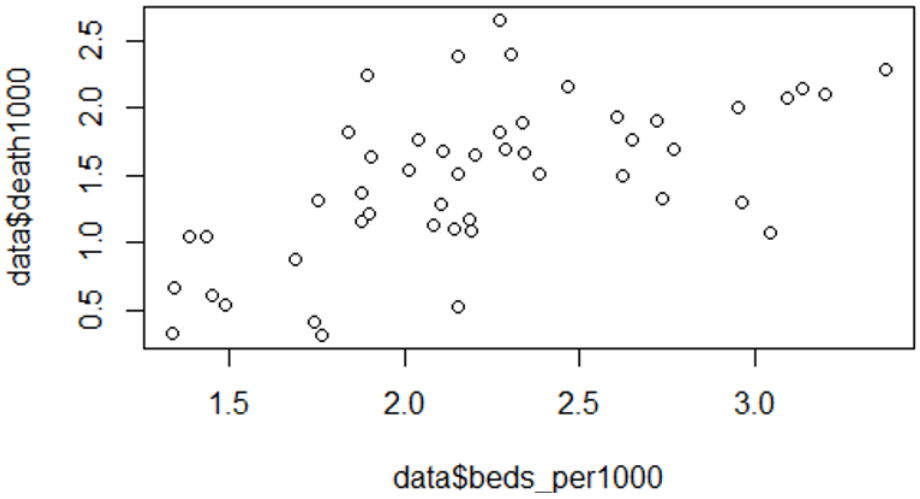

Figure 10: Relationship between COVID-19 Deaths and Hospital Beds per 1000

D Population Composition and Lifestyle Choices

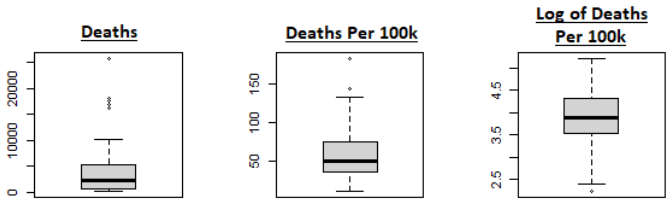

Figure 11: Severity Metric Boxplots Showing Normality of Log of Deaths Per 100k Population Versus Other Metrics

	Shapiro-Wilk Test		Breusch-Pagan		Results	Group Means/Medians Test for Equality			
Grouping Variable	**P-Value**	**Bonferroni**	**P-Value**	**Bonferroni**	**Conservative Conclusion**	**Appr. Test**	**P-Value**	**FDR**	**Result**
Age	0.078	0.391	0.156	0.781	Normal, Equal Variances	ANOVA	0.101	0.126	Fail to Reject
Sex Ratio	0.144	0.720	0.033	0.163	Normal, Unequal Variances	ANOVA Unequal Variances	0.000	0.001	Reject Null
Urbanization	0.011	0.053	0.247	1.000	Not Normal, Equal Variances	Kruskal-Wallis	0.009	0.014	Reject Null
Smoking Rate	0.061	0.306	0.486	1.000	Normal, Equal Variances	ANOVA	0.463	0.463	Fail to Reject

Figure 12: Assumption Testing Methods and Results for Population Groupings

References

[API] Data API. The covid tracking project. The Atlantic Monthly Group. www.covidtracking.com/data/api. Accessed 24 Mar. 2021.

[Asm20] McNabb N. Watts A. Asmelash, L. Most states now require face masks to reduce the spread of covid-19. these are the ones that don't., 12 2020. Retrieved March 03, 2021, from https://www.cnn.com/2020/11/09/us/biden-mask-mandate-nationwide-trnd/index.html.

[ELA17] Health care expenditures per capita by state of residence. KFF, 6 2017. www.kff.org/other/state-indicator/health-spending-per-capita/.

[ELA20] Hospital statistics by state. American Hospital Directory, 5 2020. www.ahd.com/state statistics.html.

[Goo19] Kelly Gooch. 50 states ranked by most active physicians per 100,000 population. Becker's Hospital Review, 11 2019. www.beckershospitalreview.com/workforce/50-states-ranked-by-most-active-physicians-per-100-000-population.html.

[Kla19] Kästle Klaus. Number of inhabitants of the us states ranked by population. One World - Nations Online, 12 2019. www.nationsonline.org/oneworld/US-states-population.htm.

[Ran20] Username: Night Ranger. Covid-19 state data., 11 2020. Retrieved from www.kaggle.com/nightranger77/covid19-state-data.