COVID-19: Influence of Income Inequality on Mortality in the United States

Walter Shen

Imperial College London

Introduction

In mere months, the COVID-19 epidemic has entrenched itself in American society at all levels, sparking conversations over its intersection with socioeconomic issues. Specifically, in this project we take a look at COVID-19 mortality statistics in the United States of America, and determine whether it has a relationship with income inequality metrics.

Exploratory Analysis

We used the `tidycensus` R package to look at the United States American Community Survey (ACS) results from 2018 [1]; we retrieved US county-level data for median inequality, Gini Index, and poverty rate—all relevant metrics in considering income inequality. We used the `covdata` R package to retrieve US county-level aggregated COVID-19 case data collected by The New York Times [2], as of May 28, 2020. For the COVID-19 data, we calculated the death-to-case ratio (DCR) for each county [3, 4]:

$$DCR = \frac{\#\text{ confirmed deaths due to COVID-19}}{\#\text{ confirmed cases of COVID-19}} \quad (1)$$

It is very important to note that we only calculated the DCR for counties with at least 30 confirmed cases, to avoid skewed ratios with a small number of isolated cases. We then created US county-level map visualizations of our four variables of interest, as well as a correlation matrix.

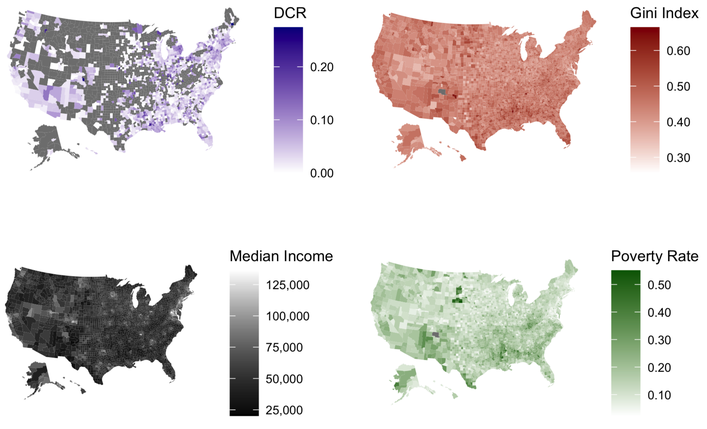

Fig. 1: Clockwise from top left: map visualizations of death-to-case ratio, Gini Index, median income, and poverty rate. COVID-19 data is from the New York Times, income measures from the American Community Survey.

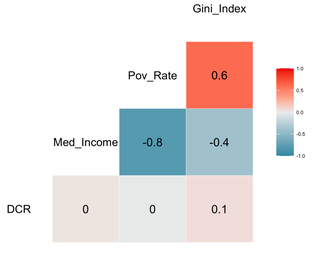

Fig. 2: Correlation matrix of the four variables.

Methodology

We looked at 4 multiple linear regression models, as illustrated in Table 1, and we calculated the Akaike information criterion (AIC) to choose a model for further analysis.

Table 1: AIC for multiple different models

#	Model	AIC
1	DCR $\sim$ Gini	-6133.476
2	DCR $\sim$ Gini + Median Income	-6137.486
3	DCR $\sim$ Gini + Poverty Rate	-6137.206
4	DCR $\sim$ Gini + Poverty Rate + Median Income	-6136.108

Our null hypothesis is that there is no relationship between the DCR versus the income variables. We use $\alpha = 0.05$.
Because Model 2 has the minimum AIC of -6137.486, we mainly investigate this model.

Results

In our R analysis [5], we retrieve the coefficients of the multiple linear regression model of Model 2, and calculate confidence intervals for each parameter.

Table 2: Summary report for Coefficients of Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-9.557e-03	1.538e-02	-0.622	0.53434
Gini Index	9.623e-02	3.002e-02	3.205	0.00137
Median Income	1.683e-07	6.865e-08	2.452	0.01432

Table 3: Confidence intervals for Coefficients of Model 2

	2.5%	97.5%
Intercept	-0.03969	0.02058
Gini Index	0.03739	0.1551
Median Income	3.375e-08	3.029e-07

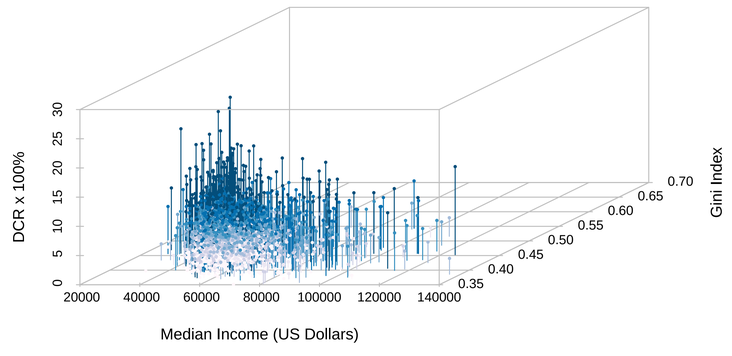

Fig. 3: Plot of DCR versus Median Income and Gini Index

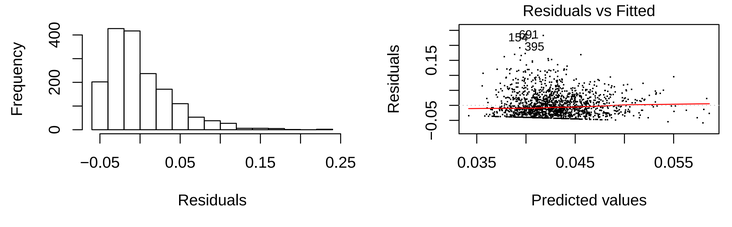

Fig. 4: Residual Analysis of Model 2

The influence of median income on the model outcome seems to be comparatively less compared to the Gini Index, as the coefficient for the Gini Index in Mode 2 was shown to be several magnitudes larger than that for median income. As such, for simplicity, we also looked at Model 1.

Table 4: Summary report for Model 1

Adjusted R-squared	0.002885
p-value	0.01506

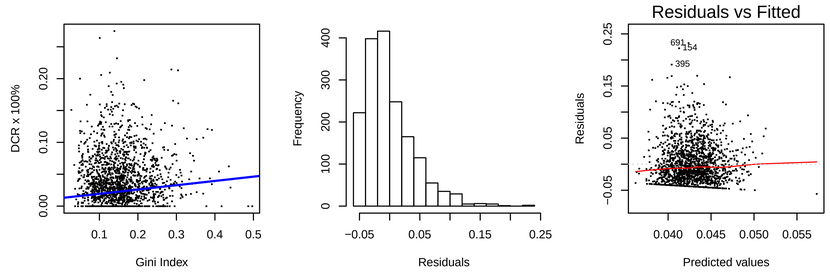

Fig. 5: Plot of Model 1 (left) and Residual Analysis of Model 1 (middle, right)

Discussion

In our multiple linear regression model, Model 2, we observed that the 95% confidence intervals for the Gini Index and median income coefficients did not include 0; the p-values were less than $\alpha = 0.05$. Therefore we reject our null hypothesis; we have shown that there is evidence suggesting that there is a linear relationship between the DCR versus the Gini and income variables.
Even when we looked at Model 1, which only had the Gini Index as the independent variable, there was sigifigant data rejecting the null hypothesis (with p-value = 0.01506 less than α). However, the adjusted R-squared value for Model 1 is low, so the linear relationship explains very little of the variance in the DCR; this possibly suggests that other variables need to be factored in.
We check our assumptions for our models [6]. They (1) follow linear relationships, as explained by the above analysis. (2) The residuals are roughly normal (histograms follow a truncated normal). (3) Homoscedasticity is satisfied, as the variance of the residuals is consistent across all predicted values. Finally, (4) the correlation between Gini Index and median income, from Fig. 2, is fairly low. We feel our assumptions have been sufficiently satisfied.
We may possibly attempt to explain these results. A higher Gini Index means more [7] income inequality; in the USA, this translates to a larger disparity of hospital access, treatment quality, and healthcare infrastructure. As such, the DCR would be higher for counties with worse income equality.

Conclusion

We have shown that there is a relationship between income inequality and COVID-19 mortality in the United States. We used aggregated county-level data; as individual COVID-19 case data did not include individual patient economic status. If there were higher-quality data available, we could have more accurate values for DCR for different economic statuses in the USA.
Further analysis may involve seeing how if counties with worse economic inequality fare with regards to testing access; we could also see how this affects the quality of DCR precision.

References

[1] Walker K. *tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*, 2020. URL https://CRAN.R-project.org/package=tidycensus. R package version 0.9.9.2.
[2] Healy K. *covdata: COVID-19 Case and Mortality Time Series*, 2020. URL http://kjhealy.github.io/covdata. R package version 0.1.0.
[3] Centers for Disease Control and Prevention. Principles of epidemiology - lesson 3: Measures of risk. 2006. URL https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section3.html.
[4] Whitney D. Jendoubi T. Regression analysis of covid-19 data m1r project overview, part ii. 2020. URL https://bb.imperial.ac.uk/bbcswebdav/pid-1768748-dt-content-rid-6294693_1/courses/13439.201910/Regression%20topics%203%264.pdf.
[5] Shen W. Math40008 poster project. 2020. URL https://github.com/shenwalter/MATH40008.
[6] National Centre for Research Methods. Assumptions for multiple regression. 2011. URL http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod3/3/index.html.
[7] Farris FA. The gini index and measures of inequality. *The American Mathematical Monthly*, 117(10):851–864, 2010. doi: 10.4169/000298910X523344. URL https://www.tandfonline.com/doi/abs/10.4169/000298910X523344.