The Law of Excluded Middle:

COVID-19 Mortality in Rural America

Julian Bernado, Shashank Rammoorthy, Daulet Tuleubayev

Abstract

We apply the Augmented Inverse Propensity Weighted Estimator of the Average Treatment Effect to a dataset considering a rural/urban indicator variable as a treatment and COVID-19 mortality as our outcome. We find the average treatment effect to be statistically significant and negative. This indicates that after controlling for confounders that are present in both rural and urban counties, mortality rates are lower in urban areas. We compare our causal estimate to the observed mean difference in mortality rates between rural and urban counties and suggest some areas for further research.

Introduction

About 60 million people live in America's countryside. Historically, rural areas have been lagging behind the rest of the country in terms of healthcare - hospitals are often understaffed, and resources are often lacking. We explore the effect this reality has had on the way rural America has coped with the pandemic. We note that while rural areas are less population dense, and could thus have less transmission, the COVID-19 mortality may be higher for the above reasons of health infrastructure. Furthermore, we note that the disproportionate impact of COVID-19 on People of Color has been well documented, but there are less claims in the literature comparing Rural vs. Urban areas. We hope to contribute to the understanding of COVID-19 mortality in various subpopulations of America.

Methods

We work within the potential outcomes framework for Causal Inference (Rubin 2005) and focus our attention on COVID-19 mortality as an outcome and Urban/Rural Classification as a treatment. We analyze this relationship using 'Dataset 4' from Vertical 2 which contains Abridged County Level data from the Yu Group at UC Berkeley. This dataset contains demographic and COVID-19 related covariates for $n = 1325$ counties across the United States. One such covariate is the Economic Research Service's 2013 Rural-Urban Continuum Codes. This code runs from 1 (most urban) to 9 (most rural). The ERS designates that a code of 3 or below indicates a metropolitan county, while a 4 or above indicates nonmetropolitan. Following in-suit, our indicator variable for 'Urban' is 1 when a county's Continuum Code is less than or equal to 3. Note that dichotomization is not desired, but is necessary for the method we are to use. With this dichotomy we have $t = 802$ treated (urban) counties and $n - t = 523$ control (rural) counties.

With this in mind, our goal is to estimate the average treatment effect (ATE) of a county being urban on our outcome, COVID-19 mortality. Note that COVID-19 mortality is defined as a county's total number of deaths divided by its total number of cases. To do so, we make use of the Augmented Inverse Propensity Weighted (AIPW) Estimator (Glynn and Quinn 2010). This estimator has the attractive quality of being doubly robust. In other words, we fit two models to get the AIPW estimator and if either of them are correctly specified, our estimator is consistent and unbiased. This works by estimating the main effect using the outcome model and running Inverse Propensity Weighting on the residuals. We deal with the assumptions of this model in the Appendix, but note that there are some potential issues. To fit this model we estimate propensity scores for each county as well as regression models for our outcome. We fit the propensity score using a logistic regression model with the following variables

- Percent Male Population
- Median Age
- Percentage Diagnosed with Diabetes
- Mortality Rate of Heart Disease
- Mortality Rate of Strokes
- Mortality Rate of Respiratory Issues
- Percentage of Smokers in County
- Democrat to Republican Ratio
- Percent of County at least 60 years old
- Percent of County Eligible for Medicare
- Factor Variable indicating when county issued stay at home order
- Factor Variable indicating when county issued limitations on gatherings above 50 people
- Factor Variable indicating when county closed public schools

Similarly, we fit the two outcome models using Ordinary Least Squares Regression using the above variables as well as

- Factor Variable indicating when county issued limitations on gatherings above 500 people
- Factor Variable indicating when county issued limitations on dine-in service at restaurants
- Factor Variable indicating when county issued limitations on entertainment and gym services

The above variables were not included in propensity score fitting as they had very little predictive power.

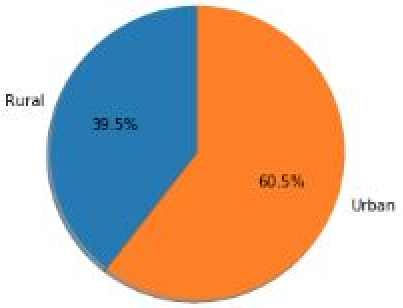
Results

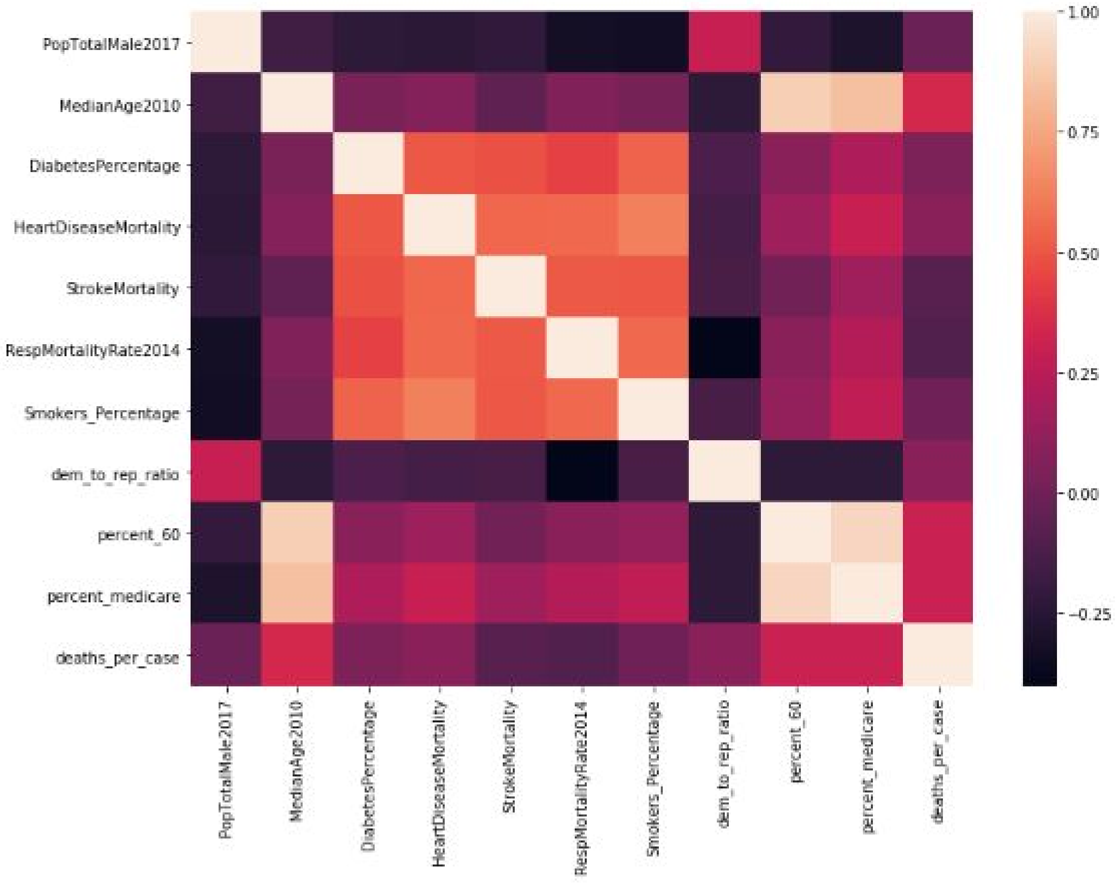
All this in mind, we get an estimated ATE of -4.99%. So, the mortality rate of COVID-19, all else identical, should be about 5% less in an urban county. Furthermore, the standard error of our estimator is 0.40%. So, as our estimator is asymptotically normal, we can use the standard error to build a 95% confidence interval of (-4.18%, -5.74%) for the true average treatment effect. With our bound not containing zero one can statistically significantly claim that the ATE is negative.

We make no statistically sound claim about the reason for this relationship, but expect health infrastructure to be at play.

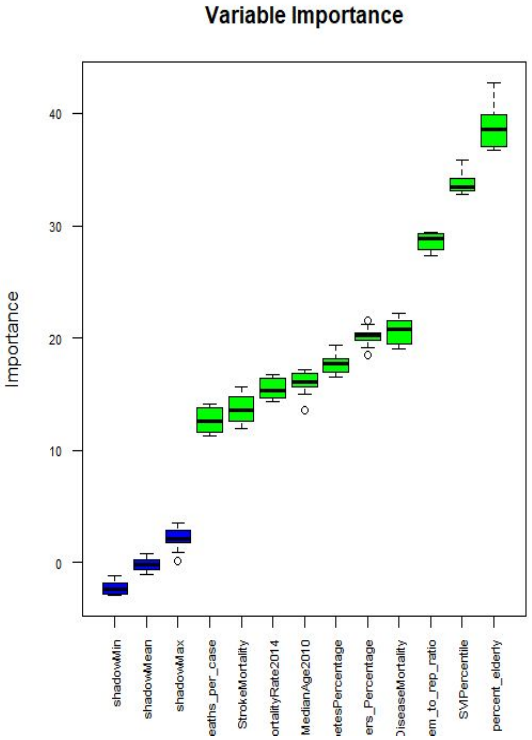
Discussion

We defined a rural county as one having a Rural-Urban Continuum Code of 4 or higher - this is the metropolitan, non-metropolitan divide as per the 2013 Rural-Urban Continuum definitions. With this definition, about 39.5% of counties in the country qualified as rural.

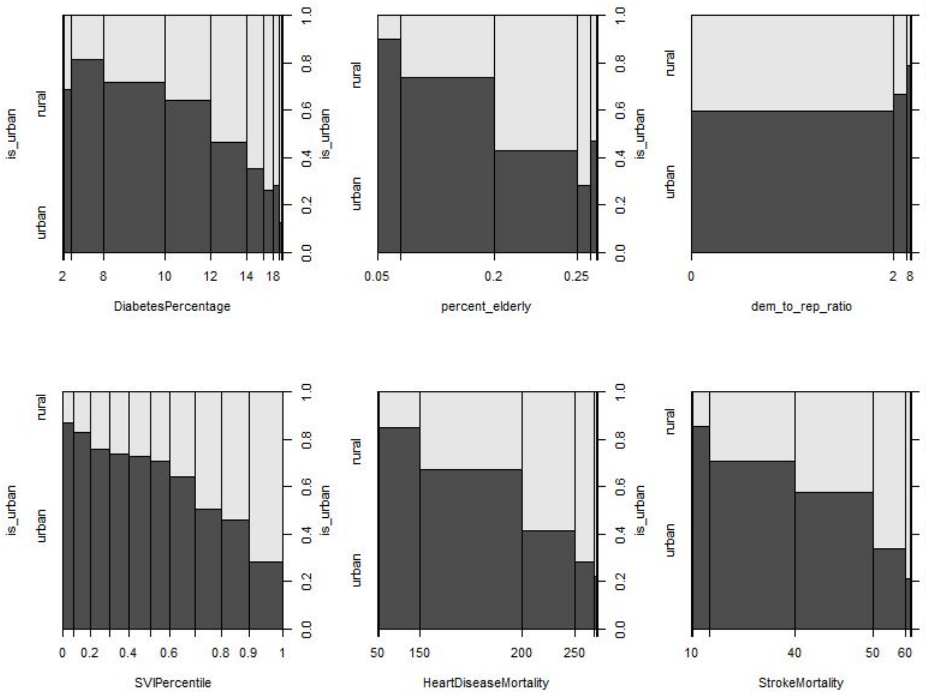




The correlation matrix shows that we have reasonable correlation coefficients for the vast majority of covariates. The high correlation between the percent_medicare and MedianAge2010 variables is explained by the fact that ‘any individual 65 years of age or older who is a United States citizen and paid into the Medicare system through their payroll taxes is Medicare eligible’ (‘Original Medicare’, 2020). The same is trivially true for the relationship between high median age and the percentage of the population above the age of 60.



We were also looking for factors that correlate with living in a rural environment. Population metrics and the associated factors, such as number of hospitals, ICU beds, and doctors in a given area were excluded due to their direct association with population size. The graph to the left shows the importance of various features in determining whether a county is rural or urban. We can see that the number of elderly (people above 60 years old) tend to more heavily populate rural areas, which leads rural areas to experience higher ratios of people with diabetes, and increased heart disease and stroke mortality. Most importantly, as seen below, there is a clear relationship between rural areas and high scores on the CDC's Social Vulnerability Index (SVIPercentile), demonstrating that rural areas are ill equipped to handle and recover from external stressors on human health, such as a pandemic. Thus rural areas are already predisposed to experience higher COVID-19 mortality rate. Given that our model accounts for such differences, we identify these as factors that potentially further increase mortality rates.



Conclusion

Having shown the results and explored the dataset, we feel confident in claiming that there is a relationship between whether a county is rural or urban and its COVID-19 mortality rate. Our causal estimates place this at about 5% lower for urban counties. Curious readers will notice that a 5% drop seems abnormally high when the average mortality rate by county within the dataset is about 3%. Note that this ATE that is estimated is looking at, in vacuum, what the difference is between rural and urban counties. In practice, each of these counties will have other covariates that add noise to its mortality rate, and within the data if one is to simply subtract the mean outcome for the control from the mean outcome for the treated, we get only about a -0.73% difference.

Given constraints of the datathon and our own expertise, we were not able to explore all paths that were interesting to us. Researchers with more time and skill should consider (1) what are the factors of rural counties that make mortality rate higher? (2) how can we improve upon the AIPW estimate perhaps using more sophisticated models at the intermediate level? (3) Are treatment effects heterogeneous? Researchers with access to data that includes more counties and more covariates should also consider the interactions of rurality with poverty, race, and other demographic variables. Furthermore, our paper makes claims about the County-Level COVID-19 mortality rate but makes no claims about individuals who **live** in rural areas. Potentially patterns of migration or other human patterns could make such findings deviate from our own.

References

Donald B Rubin (2005) Causal Inference Using Potential Outcomes, Journal of the American Statistical Association, 100:469, 322-331, DOI: 10.1198/016214504000001880
Glynn, Adam N., and Kevin M. Quinn. "An introduction to the augmented inverse propensity weighted estimator." Political analysis (2010): 36-56.
Original Medicare (Part a and B) eligibility and enrollment. (2020, July 8). Centers for Medicare & Medicaid Services | CMS.
https://www.cms.gov/Medicare/Eligibility-and-Enrollment/OrigMedicarePartABEligEnrol

Appendix

Assumptions necessary for AIPW:

AIPW relies on three assumptions about the underlying data. First, is the Similarity Under Treatment Value (SUTVA) assumption. One interpretation of this is that a unit's outcome depends only on their own treatment assignment, not on the treatment assignment of others. With a possibility of geographic confounding, we cannot be sure that SUTVA is satisfied. Future study with geospatial regression could say more about this assumption. Next is the overlap assumption. What this essentially states is that our propensity scores are bounded away from 0 and 1. If there are certain values in the support of our data that guarantee assignment to either treatment or control, then we cannot make the same statement. We use the CausalGAM package in R released along with the Glynn and Quinn paper, which checks automatically for this assumption, and find that it is satisfied. The third assumption is strong ignorability, or the statement that conditional on our covariates, the outcome is independent of the treatment. While we cannot eliminate the possibility of a strong confounder, we believe that there is not one hiding within our analysis dataset.

Schema for date-based factor variables:

Original data was in the proleptic Gregorian ordinal format, was converted to MM-DD-YY, and then bucketed into 5 date intervals.

Before 2020-03-16 ⇒ 1
Between 2020-03-16 and 2020-03-19 ⇒ 2
Between 2020-03-19 and 2020-03-23 ⇒ 3
Between 2020-03-23 and 2020-03-26 ⇒ 4
After 2020-03-26 ⇒ 5