The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties

by

Nivedita Mukherji

Department of Economics

Oakland University

Rochester, MI 48309, USA

email: mukherji@oakland.edu

May, 2020

Abstract

This paper uncovers the socioeconomic and health/lifestyle factors that can explain the differential impact of the coronavirus pandemic on different parts of the United States. Using a dynamic panel model with daily reported number of cases for US counties over a 20-day period, the paper develops a Vulnerability Index for each county from an epidemiological model of disease spread. County-level economic, demographic, and health factors are used to explain the differences in the values of this index and thereby the transmission and concentration of the disease across the country. These factors are also used in a zero-inflated negative binomial pooled model to examine the number of reported deaths. The paper finds that counties with high per capita personal income have a high incidence of both reported cases and deaths. The unemployment rate is negative for deaths implying that places with low unemployment rates or higher economic activity have higher reported deaths. Counties with higher income inequality as measured by the Gini coefficient experienced more deaths and reported more cases. There is a remarkable similarity in the distribution of cases across the country and the distribution of distance-weighted international passengers served by the top international airports. Counties with high concentrations of non-Hispanic Blacks, Native Americans, and immigrant populations have a higher incidence of both cases and deaths. The distributions of health risk factors such as obesity, diabetes, smoking are found to be particularly significant factors in explaining the differences in mortality across counties. Counties with higher numbers of primary care physicians have lower deaths and so do places with lower hospital stays for preventable causes. The stay-at-home orders are found to be associated with places of higher cases and deaths implying that they were perhaps imposed far too late to have contained the virus in the places with high-risk populations. It is hoped that research such as these will help policymakers to develop risk factors for each region of the country to better contain the spread of infectious diseases in the future.

1 Introduction

The novel coronavirus, also known as COVID-19, has brought the global economy to a screeching halt. It is sweeping through the United States and the country has now taken a lead in not only the total number of positive cases but in terms of the number of reported deaths as well. New York's total number of cases exceeds the total number reported for any country, including China.

The virus is believed to have originated in Wuhan, China, in late 2019. As the virus spread through Wuhan and the rest of China, it raised alarms across the scientific communities and governments around the world. With every passing day, the virus continued to spread exponentially. The impact of the virus in the United States started to grab public attention from late February, and many states started imposing stay-at-home orders in mid-March, 2020. The dramatic increase in the number of infected patients in a nursing home in Seattle, Washington, stunned a nation and made evident the contagiousness of the virus and its lethality. While the national attention was focused on Seattle, the virus was taking a deadly hold in New York City and its surroundings. With every passing day, one state after another started announcing their first reported cases. No US state has been spared. However, the spread of the virus has been anything but uniform. Figure 1 shows the geographical distribution of reported cases across US counties in April 2020.

This paper attempts to uncover the socioeconomic conditions that are dominant in the areas with the high number of cases and deaths. The literature on the transmission of infectious diseases often finds that highest impact areas have low income, poor sanitary conditions, and poor health care conditions due to their focus on viruses that have significantly impacted developing countries (Campos et al. (2018) for Zika, Redding et al. (2019) for Ebola are recent examples). Moore et al. (2017) used Ebola to develop an Infectious Disease Vulnerability Index for countries in Africa. The literature on the socioeconomic determinants of the spread of infectious diseases in developed countries is not extensive - Adda (2016) is an exception. Using data from France, it offers an extensive analysis of the transmission of three viruses -

influenza, gastroenteritis, and chickenpox. The paper asks the important questions whether the virus spread more rapidly during periods of economic growth and if their spread follows a "gradient determined by economic factors." Using data from France, Adda (2016) finds that the viruses studied indeed propagated faster during times of economic boom due to increased economic activity and contact between people. Qiu (2020) have conducted a similar analysis for Wuhan, China. Both papers find a positive relationship between the spread of the virus and economic activity. Avery et al. (2020) offers a list of resources both in terms of relevant research and data sources for researchers.

Unlike some of the literature cited above that concentrate on the impact of mitigation andor containment strategies along with economic conditions such as GDP, employment, and weather-related factors such as temperatures, and pollution (Wu (2020)), this paper focuses on economic, demographic, and health conditions in explaining the number of cases and deaths in the US. Figure 1 clearly indicates that the spread of the virus has been in regions of high economic activity on the two coasts. The virus has arrived on the US shores through international travel. While the initial spread of the virus is expected to be triggered by international travel and economic activity, it is important to understand whether its continued spread and concentration are restricted to such places. As the lockdown continues and the medical profession is trying to understand the susceptibility of individuals in contracting the disease, this paper attempts to understand the underlying socioeconomic conditions of the geographic regions around the US that make them susceptible to becoming hotspots. This is related to the question about the factors that determine the gradient followed by the virus as it spreads through the country.

Introducing heterogeneity that captures region-specific uniqueness in an epidemiological model of disease spread, the paper develops a Vulnerability Index for the counties included in the study. These indexes capture the underlying factors that impact the vulnerability of a region to the virus. Economic, demographic, and health/lifestyle factors are used to explain the observed differences in the vulnerability index. These factors are also used to explain the differences in the number of deaths reported across the countries. The results indicate

that the underlying demographic and health/lifestyle factors have a more significant impact in explaining deaths than disease spread. This is not a surprising result since the spread of the virus does not depend on a person's ethnicity or education status. Once people contract the disease, however, the health outcome depends on a multitude of factors that go beyond an individual's control.

The paper uses available county-level data to identify economic, demographic, and health/lifestyle risk factors for different parts of the US. The paper finds that people in regions of high economic activity and economic inequality are particularly at elevated risk of both disease spread and mortality. There is a remarkable parallel between the spread of the disease and the distance-weighted distribution of passengers arriving at international airports. The demographic distribution in terms of race shows higher vulnerability both in terms of disease contraction and death for non-Hispanic Blacks, Native Americans, and immigrants. Counties with higher numbers of personal care physicians per 1000 individuals have lower deaths and so do places with fewer preventable cases of hospital stays. Some of the high-risk factors such as obesity, diabetes, are found to have a more mixed result. This can be partly explained by the fact that many of these risk factors have a high degree of concentration in many of the southern states. These states have not reported as many cases or deaths as the regions around New York, Detroit, Chicago, and the western states of California and Washington. The paper includes the number of days since the onset of stay-at-home orders issued by governors at the state level across the country. This variable is not found to be statistically significant in influencing the vulnerability index in the spread of the virus. Regions that have longer stay-at-home orders have experienced a higher number of deaths. These regions would have experienced much higher number of cases and deaths without those orders. It is likely that they were imposed too late to have been successful in containing the virus.

This paper has identified socioeconomic and health/lifestyle factors that have played a critical role in helping the virus to develop a stronghold in certain parts of the country and cause high fatalities. It is true that a single gathering of individuals can lead to a spike in the number of cases and large number of deaths in a region. The members of the Coronavirus Task Force

are monitoring where a sudden spike is occurring. As data on cases and deaths are collected, it is important to be able to better predict if the population of a certain area is particularly vulnerable to the disease. This paper shows that it is possible to develop a vulnerability index both for disease spread and deaths based on the socioeconomic composition of the population and their health/lifestyle choices. Developing such a profile will be particularly important as the various parts of the country contemplate lifting stay-at-home orders before the invention of a therapeutic or a vaccine.

Recent experience suggests that infectious diseases are a major threat to both the health and economic well being of people around the world. In spite of the experience with HINI, SARS, and Ebola, countries such as the United States did not develop a coherent infrastructure or strategy to determine which parts of the country are at a particularly higher risk of disease transmission. This paper shows that it is possible to utilize the economic, demographic, and lifestyle profiles of regions to develop a risk factor for each geographical area so that when the next epidemic arises, public officials are better prepared to anticipate where the hotspots are likely to arise and take the necessary containment steps. The experience with COVID-19 shows how rapidly an infectious disease can bring an economy down. Without advance preparation the next disease will be just as difficult to contain as this. The large differences within state boundaries show the importance of developing more local strategies that take into consideration a multitude of factors.

2 Methodology and Data

The coronavirus pandemic has impacted all 50 states in the United States. The experience of each state, county, and city has been anything but homogeneous. To understand this differential effect across counties in the US, we consider two sets of factors. Epidemiological models explain how an infectious disease evolves in a region based on population and the size of the pool of infected individuals. We will use epidemiological models such as the SIR model to determine the fundamental differences in cases based on population size and the number of

infections. These factors alone cannot explain the entire heterogeneous outcomes across the country. We expect differences in types and amounts of economic activities, living conditions, demographic makeups, and lifestyle choices to determine the vulnerabilities of communities in the spread of a highly contagious virus such as the coronavirus.

We will conduct this analysis in two steps. In the first step, an epidemiological model of disease spread will be used to generate estimates of a vulnerability index for each county once population and infections are accounted for. In the second step, we will use county-level economic, demographic, and health data to explain differences in the vulnerability indexes across counties.

Epidemiological models of the SIR type such as in Blackwood et al. (2018) describe disease spread dynamics based on three main factors - the size of the population, the number of susceptible individuals, and the number of infected individuals. With a population of size N, if I denotes the number of infected individuals, R denotes the number of recovered individuals, the number of individuals susceptible to the disease is given by $S = N - I - R$. At each time t, the number of new infections will depend on the interactions of the susceptible (S) and infected (I) individuals. The infected individuals are non-infectious during the latent period and asymptomatic but infectious from the end of the latent period to the end of the incubation period and infectious with symptoms after the end of the incubation period. If j denotes the number of days it takes to become infectious, at time t, the interactions of susceptible people with people infected $t - j$ days earlier will lead to new cases.

Using daily reports of coronavirus cases for counties across the United States, we generate a panel dataset of US counties over a 20 day period from March 30 to April 18. The panel data approach in estimating the growth of the virus in different parts of the US allows us to introduce county-specific fixed effects in the estimation. The panel estimates the number of cases as a function of the potential pool of susceptible and infected individuals and time and county-specific fixed effects and is given by the following equation:

$$C_{it} = \beta_0 + \beta_1 C_{it-1} + \beta_2 \frac{S_{it} * I_{i,t-j}}{N} + \gamma_i + \delta t + u_{it} \tag{1}$$

where, C_{it} denotes the number of reported cases in county i at time t, γ_i gives the fixed effect parameter for county i, δ is the parameter for the time variable, and u_{it} is the error for county i at time t. The lagged value of the cases shows that the number reported in any day depends on the numbers reported the previous day.

Estimation of the above regression will generate parameter values, γ, for each county. These values will reflect the county-specific fixed effects that influence the vulnerability of each county to the virus. From these fixed effects, we generate a vulnerability index for each county. This approach is similar to the one used by Mukherji and Silberman (2013) in studying patent citations between metro areas in the US. In the second step of the analysis, we use county-level economic, demographic, and health care factors to explain how they influence the vulnerability index for each county. The factors that may explain the county vulnerability index are classified into three groups. The first group of factors relate to the economic conditions and include factors such as: per capita personal income, the unemployment rate, the level of income inequality, poverty, access to housing, and others. The second group of factors relate to a set of demographic factors including the size of the population and its density, the racial profile of the counties, the age distribution of the population, and the percentage of the population that was born outside the United States. The third group of factors considered includes health or lifestyle related factors such as the number of primary care physicians per capita, the percentage of the population with obesity and diabetes, the percentage of the population that smokes and drinks, the percentage of the population with inactive lifestyles.

In addition to the county level economic, demographic, and health data, spatial factors are considered as well. The contagious nature of the disease compels one to consider the spillover effects to neighboring counties. We introduce inverse-distance weighted values of the number of international passengers served by the top 46 international airports in the contiguous US. Since the virus is presumed to have originated in China and then spread to other parts of the world including Europe, before taking hold in the United States, international passenger data is introduced to examine if proximity to international airports is related to the concentration of confirmed cases. While international passengers often arrive at a particular airport and then

use domestic airlines to travel to other parts of the country, the locations of the international airports are closely tied to areas with concentrations of activities that are globally oriented. Consequently, a large number of international passengers served by these airports are expected to interact in the regions around these airports. Using a 300-mile radius around each county where the airports are located, an inverse-distance matrix is used to assign the number of international passengers in the areas surrounding the airports. The bottom part of Figure 1 displays the weighted distribution of international passengers. While this data is unrelated to the number of confirmed COVID-19 cases, the spatial distribution of the passenger data is similar to the spatial distribution of confirmed COVID-19 cases.

The estimation of the impact of these regional factors in explaining differences in vulnerabilities to the disease will be based on Equation (2).

$$V_i = \alpha + \upsilon W I_p + \lambda_k \sum_k e_{ki} + \phi_m \sum_m d_{mi} + \rho_n \sum_n h_{ni} + \varepsilon_i \tag{2}$$

In the above equation, V_i represents the vulnerability index of county i, e_{ki} represents the set of k economic variables that makes a county susceptible to the spread of the disease due to the enhanced interactions between people and working in close proximity. Although the economic activity of a county changes with time, the general distribution of such activities across the country remains relatively stable within short periods of time. d_m represents the demographic factors and h_n represents the health-care factors discussed above. This equation includes a spatially weighted number of international passengers in the region by multiplying an inverse distance-weighted matrix W with the number of international passengers, I, served by an international airport in the neighborhood of county i.

This paper uses county-level data for the United States. The data on COVID-19 cases and deaths is obtained from the COVID tracking data provided by the New York Times and Johns Hopkins University. Figure 1 displays the distribution of cases in the 2512 counties.

Data sources for the various demographic and economic variables such as population distribution by ethnicity, population density are listed in Table 1. While many of the data listed in the table are obtained from the USDA's Atlas of Rural and Small Town America and the

Federal Communication Commission, the original data sources are the Census Bureau and the American Medical Association. Some of the demographic data such as the distribution of the population by race and education are from the 2010 census. The total population, per capita personal income, unemployment data are from 2018. The percentage of the population with various heath-related factors such as obseity, diabetes, and life-style habits such as smoking and drinking are available from the 2014-15 period. Data on international air passengers were obtained from the Bureau of Transportation Statistics. This source provides the number of international passengers served by the top 50 international airports in the United States. Using airports in the contiguous United States only, 46 of the 50 airport data were used. The total number of passengers on international flights is over 109 million for 2018. In order to account for local spillover effects of the virus in the form of increased susceptibility due to higher prevalence of cases, an inverse distance weighted matrix was created with positive weights assigned upto a 300 mile radius around a county. This radius is just large enough to ensure that each county in the study had at least one other county in the study as a neighbor.

3 Estimation

3.1 Estimation of Cases

The previous section explained that the foundation of the analysis of the socioeconomic factors that can contribute to the spread and concentration of the coronavirus in the various parts of the country lies in the epidemiological model of disease transmission. The first step is to generate county-level vulnerability measures from an estimation of equation (1). The daily coronavirus data is available for over 2500 counties. To manage the computational load of estimating a panel that large, we restrict our analysis to counties that reported an average of 30 cases per day from March 30 through April 19. This generates a panel of 771 counties covering all 50 states. Each of the counties reported at least one confirmed case during the period of analysis resulting in a balanced panel. Equation (1) includes a lagged value of infections in determining the proportion of the population that is susceptible at any time t.

The incubation period for this virus is estimated to be anywhere between 2 to 14 days. People are infections a few days before they develop symptoms and after they develop symptoms. We assume a seven day lag for the results reported in the paper. Sensitivity analysis was conducted for different lag lengths.

Equation (1) shows that cases in period t depend on the number of cases in period $t-1$ and also on the number of susceptible and infected people whose values depend on the number of cases in previous periods. The inclusion of the lagged dependent variable makes this a dynamic panel and requires the use of dynamic panel estimation methods. A model with small T (20) and large N (771) with a lagged dependent variable is expected to have the Nickell's bias Stephen (1981). A difference GMM estimation is found to be the best option for the data. The Allerano-Bond estimation method Arellano and Bond (1991) that uses lagged values as instruments as implemented by Roodman (2006) was used. The results are reported in Table 2 and show that although autocorrelation of the first order exists, there is no second order autocorrelation. The Sargan and Hansen tests of no overidentification of instruments are satisfied and the F statistic shows that the model fits the data well. The table shows that the one period lagged number of cases has a significant impact on the number of cases reported on any day. The interaction of the infected and susceptible population is also significant and positive.

One of the key objectives of this regression is to obtain a set of estimates for the county level fixed effects. The method of dynamic panel estimation that utilizes first differencing removes the impact of time-invariant variables such as the time-invariant fixed effects. These are, however, recoverable from the residuals. It is to be noted that for a dynamic panel model of the form, $y_{it} = \rho y_{it-1} + a_i + e_{it}$, the residual $\hat{e}_{it} = a_i + e_{it} + (\hat{\rho} - \rho) y_{it}$. The average $\bar{e}_i$ can be used as an estimator of the fixed effects to analyze how the underlying conditions in the various counties impact the fixed effects as long as those factors are uncorrelated with the e_{it}. That condition is satisfied with the average e_{it} equalling -7.00e-09 for the results of the regression of equation 1. The plot of the fitted and observed values in Figure 5 shows the distance between the observed values and the fitted line and will be the county-level fixed effects.

3.2 Estimation of the Vulnerability Index

The estimates of the fixed effects derived from the dynamic panel regression of cases are converted to an index by transforming the mean value to 100 and is termed the Vulnerability Index. High values of the index indicate that the counties are more susceptible for the growth of the disease. The value of the index ranges from a low of 63 for Lincoln, Arkansas to a high of 229 for New York City, New York. Table 3 offers a list of the 20 lowest and highest values of the index. The results show that the higher values are in the so-called "hot spots". The table lists the region codes and Urban Influence Codes (UIC) used by the USDA to distinguish between rural and urban areas. Codes 1 and 2 are for metro areas, 11 and 12 are for non-core areas that are not adjacent to any metro area. The table shows the concentration of the high index areas in the northeast and in large metro areas. The bottom values are found in counties mainly outside the northeast. There is a large difference in the population densities of the places with high values of the index than the ones with the smallest values. The table shows that there are differences in both location and type of county that distinguish areas with high values of infections from places with smaller outbreaks. We attempt to introduce additional factors that can shed light on why some places experienced significantly higher infection rates than others after controlling for the pool of susceptible individuals.

The values of the vulnerability index[1] are used to estimate equation (2). Descriptive statistics of the variables are reported in Table 2 while the results are reported in Table 4. The differences in the three sets of results are based on the inclusion of population and population density in the regression. These two variables have a correlation of 0.76. As discussed in the previous section, the independent variables are classified into three broad groups - economic, demographic, and health/lifestyle. The results show that in the economic group, per capita income has a positive and significant effect showing that places of high income have higher vulnerability. The Gini coefficient measuring the degree of income inequality and severe

[1]Different specifications used in estimating the index led to similar rankings of the index even if the actual values of the coefficients varied. Since it is the relative rankings that matter for the analysis of the underlying socioeconomic factors, the results of the following analysis are largely insensitive to the exact specifications used to derive the index values.

housing problems are positive and significant if only population density is included. Another measure of economic hardship measured by the degree of food insecurity has a significant and negative effect. This is consistent with the result on income. Figure 2 shows that the largest concentration of counties with the highest levels of food insecurity is in the southern states of Georgia, Mississippi, Arkansas, Alabama - places that have not reported as many cases as some of the hot spot counties in the northeast and west. The unemployment rate and the indicator of deep poverty are not found to be significant variables. The results also show that places of severe housing shortage have higher vulnerability indexes only when population is not included. Together these results show that counties with higher vulnerability have higher economic activity. The measures of income inequality and severe housing problems have a positive impact on the vulnerability index but they are only significant when population is not included.

Figure 1 showed that the locations of the international airports are close to the regions of high infection and the results show that the distance-weighted number of international passengers served by these airports is positive and significant. Since the source of the virus is traced outside the United States and is expected to have spread here through people traveling from outside the United States, this result is not surprising. The results for the number of international passengers served by the airports measure the impact of the passengers in the counties in which the airports are located[2] and this variable is positive and significant in most models.

The most significant variables in the demographic and health-related groups relate to the racial profiles of the counties and some lifestyle choices such as the percentage of the population that drives alone to work and are physically active. The results show that counties with higher concentrations of non-Hispanic Blacks, Native Americans, and immigrants have higher infections. The foreign-born or immigrant variable is significant in only the model with no population. It is not surprising that the other factors such as the age distribution or health

[2]The diagonal values of the weight matrix used for the calculation of the weighted international passengers are zeros. Consequently, the weighted values measure the impact in the surrounding areas only.

indicators are not significant since anyone regardless of age and other health conditions can get infected. The economic indicators are significant because they determine the type of interactions people have that make them vulnerable in getting in contact with other carriers of the disease. The variable about driving alone to work is negative and highly significant. This is consistent with the notion that driving alone causes less exposure to others and can serve as a protection against getting infected. Population size is a highly significant indicator and so is population density as long as population is not included.

3.3 Estimation of Deaths

While the age distribution and health indicators are not significant in explaining the differences in the number of cases across the counties, it is well established at the individual patient level those are important factors. The daily data provided by the New York Times and Johns Hopkins University report the number of deaths as well. Table 5 displays the results of an estimation of the deaths based on variables similar to the one for the cases reported in Table 4. Unlike the regression related to the analysis of the number of confirmed cases, there are many instances of zero values of the dependent variable for the regression on deaths. The number of zero values reported for this sample is 1545 out of a total number of 9984 observations. The zero-inflated negative binomial distribution is preferred to the negative binomial when excessive zeros are present. Comparison of the model fit in terms of AIC and BIC values shows that the zero-inflated negative binomial better fits the data. Due to the very large number of observations, county-level fixed effects and a panel approach are computationally difficult. A pooled model with indicator variables for the days for which the data is analyzed is used for the analysis.

The coefficients of the regressors are reported as incidence rate ratios to help in the interpretation of the values. Unlike the estimation of cases reported in Table 3, a window of 14 days is used from infection to death. The results of the pooled zero-inflated negative binomial model show that the number of deaths is positively related to the number of cases reported 14 days prior and the size of the population. The increase of reported cases by 1 increases the

death rate by 0.023%. The indicator variables for the days of the analysis show that relative to the 20th day, days 15 through 17 had significantly fewer deaths. This is to be expected since the death counts have been rising during this period. The lack of significance in the values for days 18 and 19 relative to day 20 may indicate some slowing of the rise in the death counts after April 16. The results of this regression as they relate to the economic variables are very similar to what was reported for the cases. Counties with higher personal income and higher inequality in terms of income distrbution (Gini coefficient), severity of housing shortage have higher numbers of reported deaths even after controlling for the number of cases. Consistent with the income result, the unemployment coefficient is negative.

While the demographic and health-related factors were largely not significant except for the racial distribution of the population, in this regression of the number of reported deaths, the demographic factors are more impactful. The results show that the counties with a higher percentage of the population with less than a college education have higher deaths and so do counties with a higher percentage of females. Counties with higher percentages of non-Hispanic Blacks, Native Americans, and immigrants have higher deaths while populations with higher Hispanics, Asian Americans, and multi-racial populations have lower values relative to the excluded category of non-Hispanic Whites.

On the health-related factors, counties with more primary care physicians have fewer deaths reportedly. The remaining results related to health indicators are as follows - places with more preventable hospital stays, a higher percentage of the population that has diabetes, HIV, and are physically inactive have higher reported deaths. These are not surprising since people with underlying health risks are expected to experience more severe reactions to the infection. Counties with higher obesity and the percentage of the population that engages in excessive drinking have fewer deaths.

Obesity is a personal medical risk factor for morbidity. The county-level result reported in Table 6 is inconsistent with that. This is also true about the results for gender and age. Numerous variations of choices of variables and regression techniques show that when a large

number of variables are considered, the signs and significance of all variables are not consistent with what are known as risk factors at the individual level. Using principal component analysis as an alternative method to address the correlations between variables, the results related to factor loadings are consistent with the results reported in Table 6. This suggests that when a region's vulnerability to an infectious disease such as the coronavirus is concerned and multiple factors need to be taken into account, aggregated regional statistics that mask patient-level data may not be fully consistent with patient-level risk factors. In preparation for future epidemics and pandemics this is an issue that needs more attention.

The coefficients of the region codes 2-4 are less than 1, indicating that relative to the excluded region, the northeast, the other regions had a smaller incidence of death.

The results show that the economic factors are important for explaining the differential impacts experienced by counties across the country both in terms of confirmed cases and deaths reported. The demographic and health-related factors are more pronounced in the estimation of deaths than reported cases. This is not surprising since the virus does not discriminate based on any factor other than immunity but the severity of the disease that can lead to a fatal outcome depends on underlying health and demographic factors.

4 Conclusion

This paper has examined the differential experience of infections and deaths across the United States due to the COVID-19 pandemic. Daily reported cases of confirmed cases and deaths were examined over a 20 day period from March 30 through April 19, 2020. Although data is available for over 2700 counties, this paper focused on 771 counties that reported an average of 30 cases over the 20 day period. The counties that are not included in the study had far fewer cases and reported deaths. The counties that remain in the sample includes a vastly diverse set of counties. The excluded counties are largely similar in the small number of cases and reported deaths and added significant costs in terms of computational complexity without adding much in terms of added value.

The analysis of the number of cases is based on an epidemiological model in which we included a county fixed effect. This is a novel way to introduce heterogeneity in such a model. A dynamic panel regression of the number of cases included the potential number of interactions between susceptible and infected individuals as a proportion of the population along with county fixed effects. The results of the model were used to construct a Vulnerability Index for each county. Economic, demographic, and health/lifestyle factors were used to explain the differences in the Vulnerability Index across the counties. The results showed that counties with higher economic activity have higher vulnerability. The results show that regions around international airports experienced higher numbers of cases than ones that are over 300 miles away. This is consistent with the fact that the virus has arrived on the US shores through travelers coming to the US from abroad. The results also show that places with higher vulnerability also have a higher proportion of the population that does not use public transportation to go to work. Counties with more non-Hispanic Black, Native American, and immigrants are more vulnerable. The remaining demographic and health variables were largely insignificant.

Due to many counties reporting zero deaths during many of the days used in the sample, a zero-inflated negative binomial pooled regression was used to analyze how the economic, demographic, and health conditions impact the severity of the infection experienced by the counties. The results show that the economic factors have a similar impact on deaths. That is, counties with higher income and cases also experienced higher deaths. Counties with higher income inequality and housing shortage also experienced more deaths. In contrast to the results of the reported cases, this regression showed that not only are counties with higher percentages of non-Hispanic Blacks, Native Americans, and immigrants more likely to die relative to counties with non-Hispanic Whites, so are counties with a higher concentration of people with less than a college education. Counties with more personal care physicians per capita experienced lower deaths, and so did counties with a lower percentage of the population with diabetes, inactive lifestyles, and preventable hospital stays. Counties with higher obesity, HIV, and drinking are associated with lower deaths. It is to be noted that results here are based on reported deaths at the county level and do not include any patient-level information.

The coronavirus pandemic has demonstrated how quickly a highly contagious respiratory illness can bring the global economy to a standstill. There have been several such infections in the last ten years, although none of them had the virulence or lethality of this virus. Most of them spread to a few countries and then disappeared. The developed world remained largely unaffected by most of them, and the experience of this pandemic has laid bare the lack of infrastructure to respond to such an incident. The economics literature is not extensive in the area of pandemics and epidemics in developed countries. The contribution of this study is to understand the various socioeconomic conditions that can make a county or region more vulnerable to both disease spread and severity of cases. A national strategy to prepare the infrastructure for controlling the spread of infectious diseases should consider these factors and develop Vulnerability Indexes for regions across the country.

References

Adda, Jérôme, "Economic Activity And The Spread Of Viral Disease," *The Quarterly Journal Of Economics*, 2016, *131* (2), 891–941.

Arellano, M. and S. Bond, "Some tests of specification for panel data: monte carlo evidence and an application to employment equations," *Review of Economic Studies*, 1991, *58*, 277–297.

Avery, Christopher, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison, "Policy Implications of Models of the Spread of Coronavirus: Perspectives and Opportunities for Economists," Working Paper 27007, National Bureau of Economic Research April 2020.

Blackwood, Julie C., Lauren M. Childs, and Lauren M. Childs, "An Introduction To Compartmental Modeling For The Budding Infectious Disease Modeler," *Letters In Biomathematics*, Dec 14, 2018, *5* (1), 195–221.

Campos, Monica C., Jamille G. Dombrowski, Jody Phelan, Claudio R. F. Marinho, Martin Hibberd, Taane G. Clark, and Susana Campino, "Zika Might Not Be Acting Alone: Using An Ecological Study Approach To Investigate Potential Co-Acting Risk Factors For An Unusual Pattern Of Microcephaly In Brazil," *Plos One*, 2018, *13* (8), E0201452.

Moore, Melinda, Bill Gelfeld, Adeyemi Okunogbe, and Paul Christopher, *Identifying Future Disease Hot Spots*, Santa Monica: Rand Corporation, The, 2017.

Mukherji, Nivedita and Jonathan Silberman, "Absorptive Capacity, Knowledge Flows, And Innovation In U.S. Metropolitan Areas," *Journal Of Regional Science*, Aug 2013, *53* (3), 392–417.

Redding, David W., Peter M. Atkinson, Andrew A. Cunningham, Gianni Lo Iacono, Lina M. Moses, James L. N. Wood, and Kate E. Jones, "Impacts Of Environmental And Socio-Economic Factors On Emergence And Epidemic Potential Of Ebola In Africa," *Nature Communications*, Oct 15, 2019, *10* (1), 4531–11.

Roodman, David, *How to do xtabond2: An introduction to "difference" and "system" GMM in Stata*, Washington, D.C.: Center for Global Development, 2006.

Stephen, Nickell, "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 1981, *49* (4), 1417–1426.

Xiao, Nethery Rachel Et.Al. Wu, "Exposure To Air Pollution And Covid-19 Mortality In The United States: A Nationwide Cross-Sectional Study," April, 2020.

Yun, Chen Xi Shi Wei Qiu, "Impacts Of Social And Economic Factors On The Transmission Of Coronavirus Disease 2019 (Covid-19) In China," 2020.

Table 1: Data Sources

Variable	Data Source
Cases and Deaths	New York Times COVID-19 Project
	Johns Hopkins University
Per Capita Personal Income	USDA -Atlas of Rural and Small Town America
Unemployment Rate	USDA -Atlas of Rural and Small Town America
Deep Poverty	USDA -Atlas of Rural and Small Town America
Gini	US Census Bureau
International Passengers	US Department of Transportation
Population	USDA -Atlas of Rural and Small Town America
Population Density	USDA -Atlas of Rural and Small Town America
Education Statistics	USDA -Atlas of Rural and Small Town America
Age Distribution	USDA -Atlas of Rural and Small Town America
% Female	USDA -Atlas of Rural and Small Town America
Population by Race	USDA -Atlas of Rural and Small Town America
Severe Housing Problems	Connect2HealthFCC
Food Insecurity	Connect2HealthFCC
% Poor to Fair Health	Connect2HealthFCC
% Adult Obesity	Connect2HealthFCC
% Diabetes	Connect2HealthFCC
% Smoking	Connect2HealthFCC
% Drinking	Connect2HealthFCC
HIV Per 100000	Connect2HealthFCC
Preventable Hospital Stays	Connect2HealthFCC
Physical Inactivity	Connect2HealthFCC
Long Commute Driving Alone	Connect2HealthFCC
Driving Alone to Work	Connect2HealthFCC
PCP Per Capita	Connect2HealthFCC
Stay at Home	Kaiser Family Foundation

Table 2: Descriptive Statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
ln Vulnerability Index	769	0.849	0.885	-0.429	5.213
Economic Variables					
Weighted International Air Passe	769	61.079	122.864	0	1842.562
ln Per Capita Income	769	10.778	0.267	9.91	12.175
Gini	769	0.45	0.035	0.356	0.624
Unemployment Rate	769	4.023	1.211	1.7	18.1
Severe Housing Problems	769	16.821	4.53	6.8	35.7
Food Insecurity	769	14.847	3.87	5	33
Demographic Variables					
ln Population	769	12.044	1.139	8.757	16.129
ln Population Density	769	5.567	1.313	1.475	11.149
% Edu Less Than HS	769	12.093	5.377	2.098	36.311
% Edu HS Diploma Only	769	29.54	7.306	7.982	52.182
% Edu Some College	769	21.124	3.487	8.362	31.076
% Edu Assoc Degree	769	8.543	1.955	3.056	14.781
% Edu College Plus	769	28.7	11.086	7.937	74.133
% Age Less 18	769	23.204	3.149	7.7	35
% Age 18 to 65	769	62.201	3.385	40.7	76.2
% Age over 65	769	14.595	3.976	7	51.6
% Female	769	50.687	1.468	34.3	53.9
% Black Non Hispanic	769	13.521	15.358	0.108	82.951
% Asian Non Hispanic	769	2.65	3.812	0.01	43.015
% Native American Non Hispanic 2	769	0.905	4.546	0.038	73.298
% Hispanic	769	10.295	12.739	0.415	95.745
% Multiple Race	769	2.057	2.065	0.161	35.008
% Foreign Born	769	8.036	7.185	0	52.945
Heath & Lifestyle Variables					
% Poor to Fair Health	769	15.44	4.94	0	38.5
% Adult Obesity	769	29.129	5.044	12	45
% Diabetes	769	10.274	2.346	3.9	18.6
% Smoking	769	18.601	5.237	0	33.1
% Drinking	769	15.67	5.006	0	32.9
HIV Per 100000	769	240.83	268.853	0	2704.3
Preventable Hospital Stays Per 1	769	59.714	18.177	0	142.43
Physical Inactivity	769	24.22	5.499	9.2	39.4
Long Commute Driving Alone	769	31.322	11.135	7.7	64.2
Driving Alone to Work	769	79.84	6.926	6.4	88.8
ln PCP Per 1000	768	-7.375	0.473	-9.545	-5.426
Stay at Home	769	18.966	7.679	0	29

Region Codes R1 - R4 indicate the US regions Northeast, Midwest, South, and West, respectively. The analysis includes 132 counties from R1, 165 from R2, 360 from R3, and 112 from R4.

Table 3: Dynamic Panel Regression of Disease Spread

Variable	Coefficient	St.Err.	t-value	p-value	[95% Conf	Interval]
$lnCases_{t-1}$	0.883***	0.051	17.280	0.000	0.783	0.983
$\frac{lnSI_{t-7}}{N}$	0.039*	0.022	1.790	0.075	-0.004	0.083
D1-D8	Omitted					
D9	-0.019***	0.008	-2.470	0.014	-0.034	-0.004
D10	-0.016	0.010	-1.640	0.102	-0.034	0.003
D11	-0.019	0.012	-1.580	0.115	-0.042	0.005
D12	-0.030**	0.014	-2.180	0.030	-0.057	-0.003
D13	-0.037**	0.016	-2.330	0.020	-0.068	-0.006
D14	-0.030*	0.016	-1.820	0.070	-0.062	0.002
D15	-0.033*	0.019	-1.740	0.082	-0.070	0.004
D16	-0.037*	0.020	-1.800	0.073	-0.077	0.003
D17	-0.026	0.022	-1.220	0.224	-0.069	0.016
D18	-0.024	0.023	-1.030	0.304	-0.070	0.022
D19	-0.029	0.025	-1.150	0.250	-0.078	0.020
D20	-0.034	0.026	-1.280	0.201	-0.085	0.018

t statistics in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
Arellano-Bond test for AR(1) in first differences: z = -8.10 Pr > z = 0.000
Arellano-Bond test for AR(2) in first differences: z = 1.56 Pr > z = 0.119

Sargan test of overid. restrictions: chi2(18) = 15.15 Prob > chi2 = 0.652
Hansen test of overid. restrictions: chi2(18) = 28.18 Prob > chi2 = 0.059

Group variable: Fips Number of Obs = 10794
Time variable : Day Number of Groups = 771
Number of instruments = 44 Obs per Group: Min = 14
F(22, 771) = 6006.10 Avg = 14.00
Prob > F = 0.000 Max = 14

Table 4: Vulnerability Index of Top and Bottom 20 Counties

County	State	Fips	Vulnerability Index	Population Density	Region Code	UIC
Top 20						
New York City	New York	36061	229	69468	1	1
Nassau	New York	36059	200	4705	1	1
Suffolk	New York	36103	197	1637	1	1
Westchester	New York	36119	196	2205	1	1
Cook	Illinois	17031	190	5495	2	1
Wayne	Michigan	26163	185	2974	2	1
Bergen	New Jersey	34003	182	3884	1	1
Los Angeles	California	6037	181	2420	4	1
Rockland	New York	36087	178	1796	1	1
Essex	New Jersey	34013	176	6212	1	1
Hudson	New Jersey	34017	176	13731	1	1
Miami-Dade	Florida	12086	176	1315	3	1
Union	New Jersey	34039	173	5216	1	1
Philadelphia	Pennsylvania	42101	173	11379	1	1
Passaic	New Jersey	34031	172	2715	1	1
Fairfield	Connecticut	9001	171	1467	1	2
Orleans	Louisiana	22071	171	2029	3	1
Middlesex	New Jersey	34023	171	2622	1	1
Middlesex	Massachusetts	25017	170	1838	1	1
Suffolk	Massachusetts	25025	169	12416	1	1
Bottom 20						
Northumberland	Pennsylvania	42097	70	206	1	5
Cass	Missouri	29037	70	143	2	1
Putnam	Florida	12107	70	102	3	3
Nevada	California	6057	70	103	4	3
Washington	Texas	48477	70	56	3	3
Delaware	Oklahoma	40041	70	56	3	6
Washington	Utah	49053	70	57	4	2
DeKalb	Illinois	17037	70	167	2	1
Allen	Ohio	39003	70	264	2	2
Richland	Ohio	39139	70	251	2	2
Perry	Missouri	29157	70	40	2	6
Otsego	Michigan	26137	70	47	2	11
Napa	California	6055	70	182	4	2
Dubuque	Iowa	19061	69	154	2	2
Grant	Indiana	18053	69	169	2	3
Decatur	Georgia	13087	69	47	3	5
Madera	California	6039	69	71	4	2
Muhlenberg	Kentucky	21177	68	67	3	6
Marshall	Iowa	19127	67	71	2	5
Lincoln	Arkansas	5079	63	25	3	2

Table 5: Fixed Effects Regression of Cases

Variable	Model 1	Model 2	Model 3
Economic Variables			
Wtd Intnl Passengers	0.000181**	0.000269***	0.000287***
	(2.42)	(3.71)	(3.75)
International Passengers	0.0000144***	0.00000575	0.00000630*
	(2.64)	(1.59)	(1.71)
Gini	0.837***	0.218	0.207
	(2.78)	(0.79)	(0.75)
ln Per Capita Income	0.112**	0.144***	0.145***
	(2.45)	(3.46)	(3.50)
Unemployment Rate	0.0114	0.00523	0.00431
	(1.52)	(0.96)	(0.79)
Severe Housing Problems	0.00602**	0.00291	0.00259
	(2.02)	(1.29)	(1.13)
Food Insecurity	-0.0103***	-0.0136***	-0.0135***
	(-2.99)	(-4.44)	(-4.38)
Deep Poverty All	-0.00358	0.00463	0.00424
	(-0.90)	(1.28)	(1.16)
Demographic Variables			
ln Pop Density	0.0737***		-0.0114
	(9.42)		(-1.46)
ln Population		0.143***	0.151***
		(18.92)	(17.13)
Age less than 18	0	0	0
	(.)	(.)	(.)
Age 18 to 65	-0.0113***	-0.00763***	-0.00673**
	(-3.26)	(-2.70)	(-2.35)
Age over 65	-0.00837***	-0.00589***	-0.00585***
	(-3.09)	(-2.69)	(-2.65)
% Female	-0.000639	-0.00449	-0.00291
	(-0.15)	(-1.27)	(-0.81)
% Black	0.00373***	0.00427***	0.00435***
	(5.43)	(6.83)	(6.92)
% Asian	0.000294	-0.00303	-0.00309
	(0.09)	(-1.33)	(-1.36)
% Native American	0.00632***	0.00343***	0.00308**
	(3.67)	(2.82)	(2.54)
% Hispanic	-0.000419	-0.00145*	-0.00155*
	(-0.41)	(-1.75)	(-1.85)
% Multiple Race	-0.00469	-0.00601**	-0.00601**
	(-1.45)	(-2.19)	(-2.16)
% Foreign Born	0.00540**	0.00173	0.00213
	(2.06)	(0.87)	(1.06)
Health & Lifestyle Variables			
% Poor to Fair Health	-0.00128	-0.00133	-0.00161
	(-0.55)	(-0.62)	(-0.75)
% Adult Obesity	-0.00295	-0.00313	-0.00295
	(-1.10)	(-1.27)	(-1.20)
% Diabetes	-0.00713	-0.00352	-0.00251
	(-1.18)	(-0.66)	(-0.46)

Table 5 – continued from previous page

Variable	Model 1	Model 2	Model 3
% Smoking	0.00323*	-0.000676	-0.000616
	(1.68)	(-0.40)	(-0.37)
% Drinking	0.00210	-0.000378	-0.000383
	(1.34)	(-0.27)	(-0.28)
Preventable Hospitals Per 1000	0.000572	0.000523	0.000527
	(1.21)	(1.25)	(1.26)
% Physical Inactivity	-0.000953	0.00635**	0.00651***
	(-0.36)	(2.57)	(2.63)
% Driving Alone to Work	0.00164	-0.00317***	-0.00328***
	(1.44)	(-3.15)	(-3.26)
ln PCP Per Capita	-0.00407	-0.0184	-0.0175
	(-0.25)	(-1.32)	(-1.26)
Stay at Home	0.00190**	0.00217**	0.00226***
	(2.01)	(2.50)	(2.59)
Cons	3.250***	2.095***	1.914***
	(4.63)	(3.72)	(3.35)
N	768	768	768
R^2	0.595	0.709	0.710
adj. R^2	0.580	0.698	0.698

t statistics in parentheses
*$p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Regression to Explain Distribution of Deaths

Deaths	Coefficient	St.Err.	t-value	p-value	[95% Conf	Interval]
$Cases_{t-14}$	1.00023**	0	2.52	0.012	1	1
Economic Variables						
ln Per Capita Income	2.149***	0.312	5.26	0	1.616	2.858
Gini	7.652**	7.345	2.12	0.034	1.166	50.217
Unempl Rate	0.921***	0.018	-4.28	0	0.887	0.956
Severe Housing Problems	0.998	0.009	-0.21	0.836	0.98	1.016
Food Insecurity	0.923***	0.011	-6.5	0	0.9	0.945
% Deep Poverty All	1.015	0.013	1.13	0.258	0.989	1.041
Demographic Factors						
ln Population	2.017***	0.07	20.3	0	1.885	2.158
% Edu Less Than HS	1.025***	0.008	3.02	0.003	1.009	1.041
% Edu HS Diploma Only	1.02***	0.005	3.92	0	1.01	1.03
% Edu Some College	1.035***	0.009	3.8	0	1.017	1.053
% Edu Assoc Degree	0.921***	0.01	-7.38	0	0.901	0.941
% Edu College Plus	1	.	.	.	.	.
% Age Less 18	0.992	0.008	-0.94	0.345	0.977	1.008
% Age 18 to 65	0.994	0.007	-0.83	0.404	0.98	1.008
%Age over 65	1	.	.	.	.	.
% Female	1.079***	0.02	4.14	0	1.041	1.119
% Black Non Hisp	1.037***	0.003	11.4	0	1.031	1.044
% Asian Non Hisp	0.974***	0.007	-3.65	0	0.96	0.988
% Native American	1.029***	0.007	4.5	0	1.016	1.042
% Hispanic	0.988***	0.004	-3.5	0	0.981	0.995
% Multiple Race	0.961***	0.012	-3.28	0.001	0.939	0.984
% Foreign Born	1.037***	0.008	5.05	0	1.023	1.052
Health & Lifestyle Factors						
ln PCP Per Capita	0.813***	0.041	-4.11	0	0.736	0.897
% Adult Obesity	0.968***	0.008	-3.89	0	0.952	0.984
% Diabetes	1.032*	0.019	1.67	0.095	0.995	1.071
% Smoking	1.004	0.006	0.66	0.508	0.993	1.015
% Drinking	0.986***	0.005	-2.89	0.004	0.977	0.996
HIV Per 100000	1**	0	-2.53	0.011	1	1
Preventable Hospita1 Stays	1.007***	0.001	5.08	0	1.004	1.01
% Poor to Fair Health	0.979***	0.006	-3.56	0	0.967	0.99
Physical Inactivity	1.039***	0.009	4.59	0	1.022	1.057
Stay at Home	1.02***	0.003	6.41	0	1.014	1.026
D1 - D14	Omitted	.	.	.	.	.
D15	0.802***	0.047	-3.75	0	0.715	0.9
D16	0.852***	0.047	-2.89	0.004	0.765	0.95

Table 6 – continued from previous page

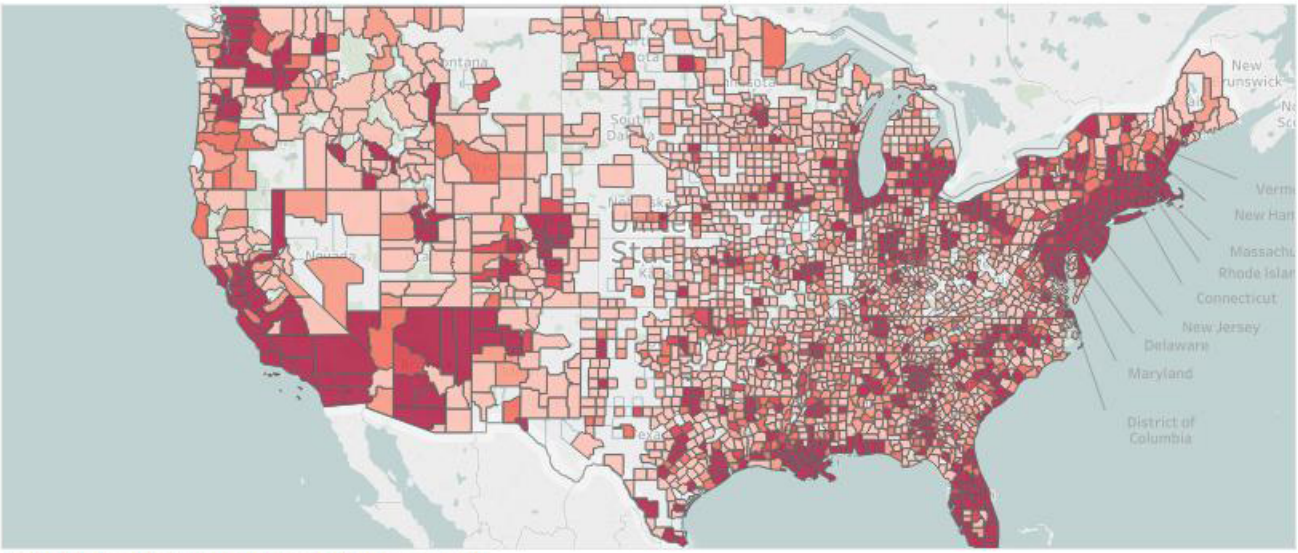
Deaths	Coefficient	St.Err.	t-value	p-value	[95% Conf	Interval]
D17	0.894**	0.048	-2.09	0.036	0.804	0.993
D18	0.935	0.049	-1.28	0.201	0.844	1.036
D19	0.968	0.05	-0.63	0.529	0.874	1.072
D20	1	.	.	.	.	.
R2	0.736***	0.045	-4.96	0	0.653	0.831
R3	0.372***	0.029	-12.48	0	0.319	0.435
R4	0.94	0.088	-0.66	0.507	0.783	1.129
Constant	0***	0	-10.34	0	0	0
Inflate						
$Cases_{t-14}$	-0.068***	0.006	-11.05	0	-0.08	-0.056
ln Per Capita Income	-0.08	0.656	-0.12	0.903	-1.366	1.206
Unemployment Rate	-0.983***	0.16	-6.14	0	-1.297	-0.669
ln Population	-0.187*	0.111	-1.68	0.093	-0.405	0.031
Constant	6.094	7.805	0.78	0.435	-9.202	21.391
$ln\alpha$	-0.322***	0.031	-10.41	0	-0.382	-0.261
α	0.7250048	0.0223961			0.6824117	0.7702563

Mean Dependent Var	39	SD Dependent Var	324
Number of Obs	4608	Chi-square	7458.128
Prob > chi2	0	Akaike crit. (AIC)	30570.965

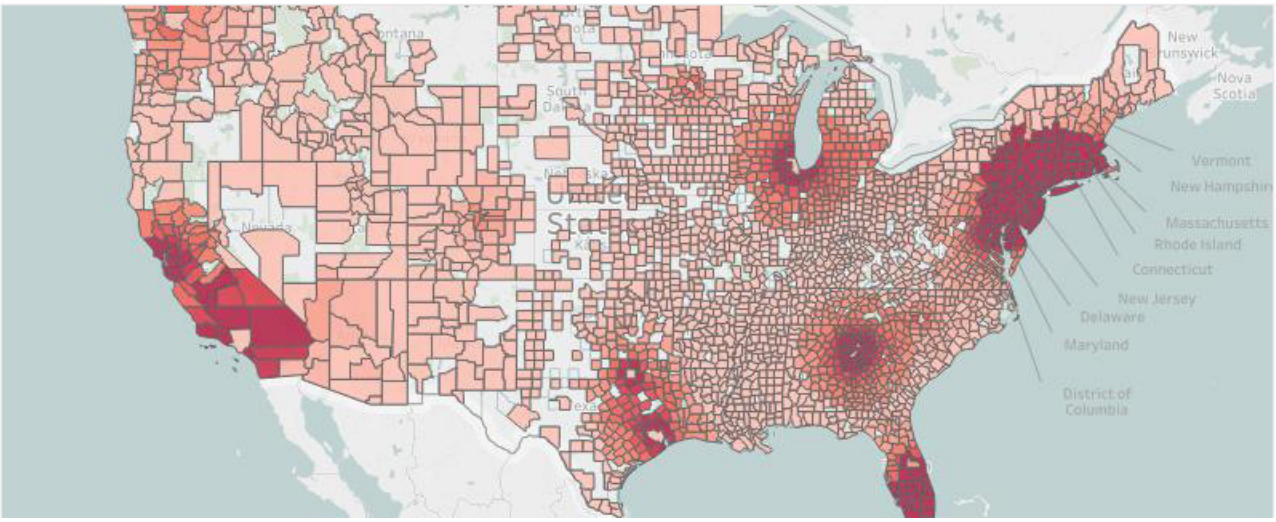
The coefficients reported in this table are incidence rate ratios.

Figure 1: Distribution of Cases and International Passengers

Number of Cases



Weighted International Passengers



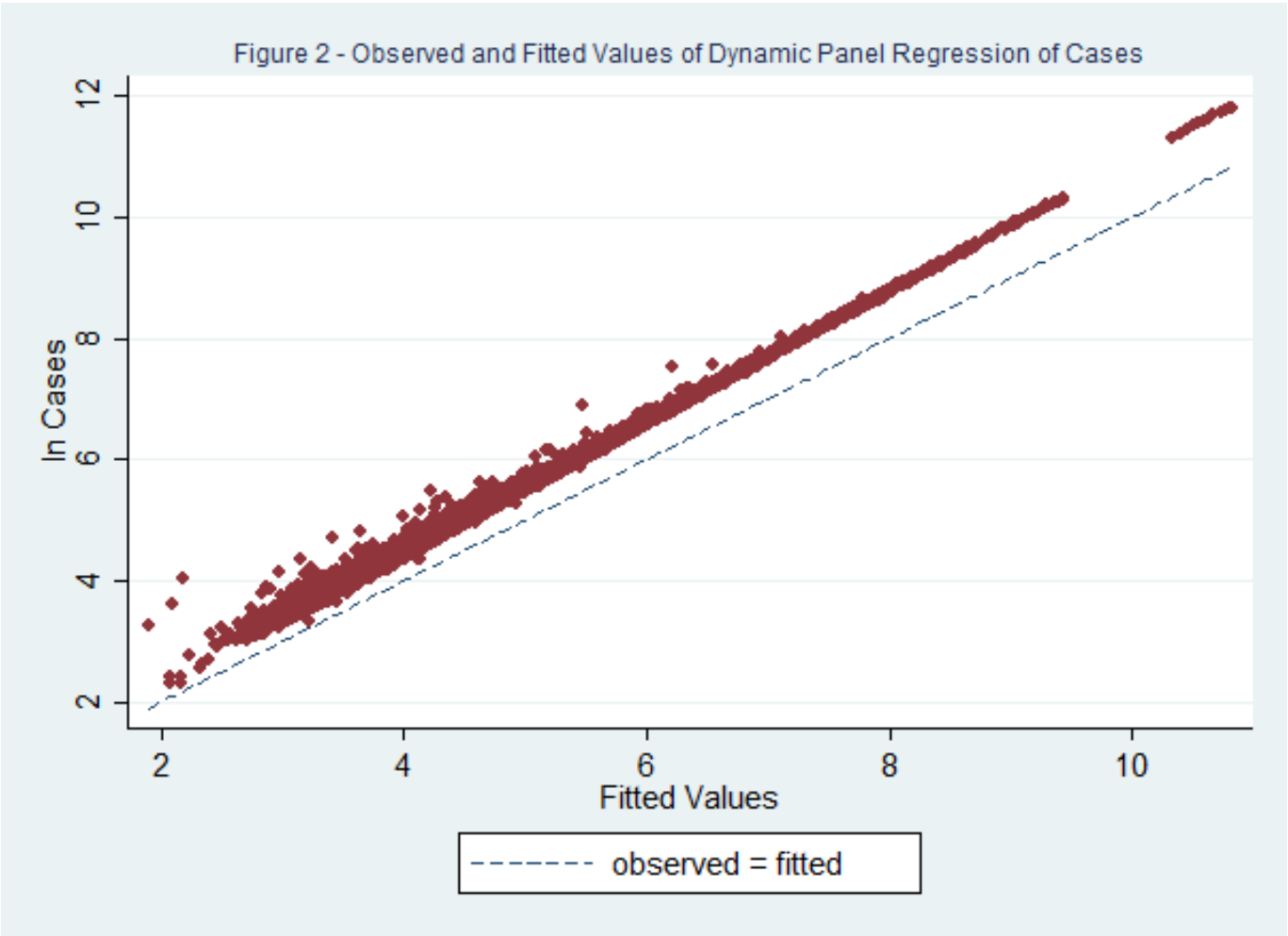


Figure 2 - Observed and Fitted Values of Dynamic Panel Regression of Cases
ln Cases
2
4
6
8
10
12
2
4
6
8
10
Fitted Values
observed = fitted