Analyzing the Efficacy of COVID-19 Travel Bans: A Regression Analysis Approach

Mallory Kochanek
mkochan@bgsu.edu

Follow this and additional works at: https://scholarworks.bgsu.edu/honorsprojects

Part of the Statistical Models Commons

How does access to this work benefit you? Let us know!

Repository Citation
https://scholarworks.bgsu.edu/honorsprojects/918

This work is brought to you for free and open access by the Student Scholarship at ScholarWorks@BGSU. It has been accepted for inclusion in Honors Projects by an authorized administrator of ScholarWorks@BGSU.
Analyzing the Efficacy of COVID-19 Travel Bans: A Regression Analysis Approach

Mallory Kochanek
Bowling Green State University

Faculty Advisors:
Dr. Riddhi Ghosh, Department of Mathematics and Statistics
Dr. Bradley Fevrier, Department of Public and Allied Health

December 2023

Abstract

Some might associate the term ‘public health’ with the pandemic that occurred in 2020. COVID-19 spread like most have never seen in their lifetime. It is useful to look at the effectiveness of the travel restrictions in mitigating the spread of the global pandemic. Using linear regression and network regression, we obtain parameter estimates to determine the relation of predictors, such as network effect, percentage of urban population and GDP, on the COVID-19 incidence rate for the months January to April of 2020. Linear regression does not account for the correlation structure of the data. Network regression, on the other hand, performs this task effectively, following a community detection algorithm such as the infomap method, and calculates parameter estimates using the communities, or clusters, within the data. Through simulations, the consistency of both linear regression and network regression estimators is evaluated. We compare the network regression estimates to the linear regression estimates. For further research, we determine if the network effect is significant. The results of linear regression estimates and the network regression estimates differed, though still providing similar conclusions. More emphasis is placed on the estimates obtained through network regression given
how it accounts for the networks within the data. Considering results from the estimates, it does not appear that the travel restrictions set around the globe assisted in minimizing the spread of COVID-19.
1 Introduction

In the wake of the unexpected COVID-19 pandemic that swept the globe, governments and health authorities were faced with the tireless job of implementing various measures to curb the rapid spread of the coronavirus. One measure, employed early in the pandemic, was the implementation of travel bans and restrictions. These restrictions, which limited the transit of individuals between countries, were credited as effective tools in mitigating the spread of the virus. However, their long-term impact on controlling the pandemic has been a subject of debate.

The COVID-19 pandemic did not only pose a health crisis to the world but brought attention to the intersection of public health policy and mathematical statistics. Data was collected immediately, research began quickly, and vaccines were administered within the year. Gaining a deeper understanding of the impact of travel bans on the control of the pandemic becomes an important aspect of future public health readiness and response. Mathematical and statistical research has contributed to the discussion surrounding COVID-19 mitigation strategies, providing valuable insights for further research and public health officials.

This research will be employing a network regression model and comparing it to common linear regression. Linear regression does not account for subgroups embedded in the data and assumes independence among observations. It is not suitable for every data set. Network regression, on the other hand, controls for community structure, which establishes a more meaningful correlation assumed by the network connections [9]. By using network regression, more insight about the hidden structure of data provided, and correlation among covariates is identified. Network regression provides a more compelling representation of the data, which will retain the important information about the network and reflect the fact that interactions between observations in complex systems are interdependent [18].

Network regression relies on community detection algorithms, which detect communities based on some similarity measures inherent in a network data. Community detection algorithms, used in a newly proposed network regression developed by [9], account for the structures nested in the data. These algorithms do not assume independence among covariates, and instead use those relations as communities. The communities are densely connected to each other, and have few connections to other individuals in the network [21].
Such structures are found using the infomap method. The infomap method, one of many community detection algorithms, minimizes the map equation, or the description length of a random walker [18]. This method is stochastic and is best used for spreading of information [21]. Community detection algorithms, specifically the infomap method, are able to provide better insight into problems relating to public health due to its ability to find the structures within a network that are significant, with respect to how information flow through that network [17].

The rest of this thesis is organized as follows. The Literature Review provides the necessary information to understand the research, such as what is meant by public health, epidemiology, COVID-19 background and information, and travel restriction statistics. Methodology explains how the data was prepared and the statistical testing used to gather results. The results will answer the overarching question of how effective the travel restrictions were in halting the spread of COVID-19. In the Discussion and Implications for further research, the research is summarized, and avenues for continuation are discussed.

2 Literature Review

The topic of the global pandemic is a public health matter. A field of multidisciplinary nature, public health authorities interact with governments, health agencies, researchers, and the community. [26], a leading figure in the field of public health in the early 20th century, defined public health as “The science and the art of preventing disease, prolonging life, and promoting physical health and efficiency through organized community efforts for the . . . control of the community infections, . . . and preventative treatment of the disease.” As a society, we take certain measures to initiate and maintain health improvement.

Epidemiology is one of the branches of public health. It is central to public health because of its population focus and quantitative methods [1]. Epidemiology is the intersection of mathematical statistics and public health. Epidemiological methods, used to investigate causes of diseases and to identify trends in disease occurrence that may influence the need for medical and public health services, are relevant to this research [20]. When epidemiologists determine information like how an epidemic is caused, what the spread of the contagion is like, and measures for control, it provides public officials
with knowledge for future policies. For example, the emergence of COVID-19 is still not exactly known, but this information would benefit epidemiologists and public health officials, to help earlier determine to rate of spread and methods for prevention.

Public health has increased drastically since the 19th century. People are living longer, getting effective medicine and treatments, and enjoying a better quality of life. Due to the improvement of technology and medicine, the research is like nothing before. The invention of the microscope led society to be able to investigate microorganisms [23]. Social, economic, and political reforms have also greatly impacted the improvement of public health. In 1948, after World War II, the World Health Organization (WHO) was created [23].

This is not to say that the system is perfect and unflawed. Society has had its many struggles with public health. There are debates and discussions about health care reform. Health care reform, according to [1], focuses on cost containment, rather than the health of the entire population. As history has shown, the health of the entire population should be a concern. The bubonic plague, smallpox, influenza, cholera, and HIV/AIDS are all examples of deadly pandemics that encourage more public health policies [19]. Most recently, the coronavirus disease in 2019, or COVID-19.

In December 2019, in Wuhan, China, the first case of COVID-19 was reported. By April 30th, 2020, more than 3,200,000 cases were reported, with at least 233,000 deaths worldwide. In the United States, as of April 30th, 2020, 1,069,534 cases were reported with at least 63,000 deaths [27]. COVID-19 is one several strains of coronavirus caused by severe acute respiratory syndrome coronavirus-2, or SARS-CoV-2 [20]. All strains have been deadly and caused outbreaks, but none like COVID-19. We are still learning about this virus, as data is still being collected and research is still being conducted.

Measures were put in place to stop the spread of the pandemic. On January 31st, 2020, President Trump issued a proclamation that imposed travel restrictions on passengers coming into the United States from China in the past 14 days (about 2 weeks) [4];[22]. The CDC began screening at prominent international airports. In February, another proclamation was issued with the same intention, but for Iran. March proved to be one of the biggest months for travel restrictions. After the declaration of the global pandemic on March 11, 2020, the Trump administration issued a travel restriction for non-Americans entering the United States from 26 European countries [22]. A few days later, the United Kingdom and Ireland were added to this list.
According to Section 2 of the March Proclamation, this ban did not apply to “legal permanent residents and most immediate family members of U.S. citizens” [22];[2]. Soon after, the United States closed its borders.

The United States was behind in imposing the travel restriction. Around the globe, other countries were imposing serious travel restrictions as early as January. Before the end of January 2020, 11 countries had restrictions in place for visitors from China. In February, 62 counties had travel restrictions in place, in early March, 81 countries had travel restrictions, and in late March, 181 countries had travel restrictions [24]. As mentioned, the WHO declared COVID-19 a global pandemic on March 11. By April 6, 2020, 96% of countries had implemented travel restrictions, whether it be closing borders, destination travel restrictions, suspension of flights, or other measures such as quarantine. Out of the 5 regions defined by the WHO, three had 100% destinations imposing restrictions, and two had 92% and 93% [24]. The United States was one of the countries that had destination specific travel restriction, limiting passengers who have recently been to one country not to enter the U.S. The airline industry suffered heavily, with 43 airlines declaring bankruptcy [11].

Similar studies have been done regarding the effectiveness of the travel ban as a means to slow the pandemic. [11], examined lifting travel restrictions while still protecting health. Their results show that globally coordinated travel policies are not only necessary for resuming international travel, but that it is also possible to accomplish this goal with minimal effect on public health relative to full border closure. [3] concluded additional travel limitations have only a modest effect unless paired with public health interventions and behavioral changes, such as isolation and use of face masks. [5] looked at the effects of a pandemic influenza and showed that under most scenarios restrictions on air travel are likely to be of surprisingly little value in delaying epidemics, unless almost all travel ceases very soon after epidemics are detected. [25] and [9] had similar findings: travel restrictions are primarily effective at the early stage of a pandemic, but not as long-term effective as social distancing and mask wearing.

The field of public health is expanding, especially in the wake of COVID-19. With the help of methods in epidemiology, policymakers can make informed decisions about what to do next amid a pandemic, so that, as a society, we are better equipped for the next outbreak. As this research relates closely to epidemiology, this research, and the others listed above, should provide insight into whether or not travel restrictions are effective in mitigating
an epidemic, so that new policies may be set in place.

3 Methodology

We describe the methods used in the research to obtain results. The methods are organized into two sections: preparing the data and the statistical tests used. Preparing the data will discussed how the data was gathered, steps needed to clean the data, and the functions created to make the data compatible to merge into one data set. Statistical testing will discuss how methods such as consistency, linear regression, and network regression are used with the data.

3.1 Preparing the Data

Here, we describe the steps for prepping the data. Individual data sets must be gathered from separate sources. Each data set must be cleaned and be prepared to merge with other data sets. Lastly, we create functions so that data may displayed effectively as one master data set. Features of interest for the master data are the COVID-19 incidence rate, network effect, percentage of urban population, and GDP. All data sets are recent, published in the last 5 years.

3.1.1 Gathering the Data

Data was collected from a variety of sources using an internet search. Only the months January to April of 2020 were analyzed. This is approximately two months before and after the surge of the pandemic. These four months were chosen for ease and data control. Given the amount of data per month from all data sets, it was best to keep concise.

From [15], data on flights from one country to another during 2020 was found. The monthly flight data illustrates the development air traffic and spans all flights seen from January 2019 to the end of the pandemic [15]. The monthly flight data includes features (all of which were not useful) such as callsign, number, aircraft UID, type code, origin, destination, first seen, last seen, day, starting latitude, longitude, and altitude, and ending latitude, longitude, and altitude.
The other important data to search for was the reported COVID-19 cases for every country. Obtained from popular the data site, GitHub, [12] provided data for the number of cases that were reported from each country, daily.

Data on the predictor variables also needed to be obtained. Because we are interested in how GDP, percentage of urban population, and population changed with the restrictions on travel, data was collected on each. Collected from [28];[29];[30], each data set had 3 features: country, year, and the statistic of interest (GDP, % urban population, or population). Because we are only interested in the 2020, only data from 2020 was relevant.

Because the monthly flight data provided the airport codes for origin and destination, while other data provided country name, more data was needed so that all data could be joined. First, data on airport codes was gathered from [7]. This data set includes, among other features, the IATA airport code and alpha-2 country codes. Then, data on country codes was gathered from [8]. The country code data had features such as the country name, alpha-2, and alpha-3. All data is now gathered and ready for preparation to create a master data set.

All data was loaded and read into RStudio. Known for its statistical capabilities, RStudio offers specific packages for data analysis, cleaning, and has a well-organized, user-friendly interface. RStudio is efficient in producing easy to read plots and charts to clearly display findings. Because of its ability to present, manipulate, and statistically analyze data, RStudio is the ideal choice.

3.1.2 Cleaning the Data

After carefully inspecting each dataset, all features, or columns, that were not important to the analysis were removed. Removing this unnecessary data not only consolidates the data, but can improve runtime when the datasets are called. This procedure is done for all data sets.

Recall that the COVID-19 case data set listed the reported number of cases for each country every day. In order to make this data compatible with our monthly format, it is simple to convert the ‘date’ column to a date type. Since the data is of date type, the system can recognize what month the cases were reported. It is now possible to subset the COVID-19 cases data by month. Each monthly subset is formatted as a single column, displaying the number of cases reported per country.

Because GDP, percentage of urban population, and population are from
the same the source and contain the same features, it is simple to join them now. Additionally, the airport code data is joined with each month’s flight data. This step must be done early so that the data can later be joined with the country code data, and later the COVID-19 cases data.

3.1.3 Creating Functions

It is necessary to create a flight count matrix, $C$, to organize the flights per month. Functions are used to create such matrix. Each month’s flight data is used as an argument. Using for loops, the function outputs a matrix, $A$, where the $j^{th}$ row is the country of flight origin and the $i^{th}$ column is the country of flight destination. Thus, the entry $A_{ji}$ is the number of flights that occurred from country $i$ to country $j$. These matrices are now referred to individually as $C_{MONTH}$.

A separate function took in each $C_{MONTH}$ and output a data frame that gave the total number of flights to each destination country, along with the alpha-2 code. All months are represented in this data frame, displayed as columns, where the values in the columns are the number of flights. Note, again, that the country associated with the total number of flights is the country in which the flight arrived, the destination country.

After creating functions to obtain sub-data frames, the master data set is formed. In the first column, all country names are listed. Columns 2-5 list the reported number of COVID-19 cases for January (2), February (3), March (4), and April (5). Columns 6 and 10 list the alpha-3 and alpha-2 code for each country. Columns 7-9 are the GDP, percentage of urban population and population for each country. Columns 11-14 are the monthly flight counts. Creating this master data frame will provide ease when performing our statistical testing. All the necessary information is now in one data frame that can be repeatedly called, rather than calling multiple data frames that are not compatible with one another. Though tedious, creating the master data set is arguably one of the most important parts of the process to make sure all data is organized and synced for countries and months.

3.2 Statistical Tests

Here we describe the statistical tests used to gather results from the master data set.
3.2.1 Linear Regression

Linear regression is one of the methods used to determine the parameter estimates. Linear regression is one of the first taught statistical models in any mathematical statistics course. The model is simple to use, and provides easy to interpret estimates. [16] defined the multivariate linear regression model as

\[ Y = a + \sum_{i=1}^{k} b_i X_i + u, \]

where \( Y \) is the dependent, or response variable, \( X_i \) are the independent, or predictor, variables, \( a \) and \( b_i \) are the regression coefficients, the intercept and the parameter estimates, and \( u \) is random noise. For every month, we will use the model:

\[ Y_i = \alpha_0 + \alpha_1 x_{UP_i} + \alpha_2 x_{GDP_i} + \beta \sum_{j \neq i}^{n} A_{ji} + \varepsilon_i, \]

where \( Y_i \) is the COVID-19 incidence rate for country \( i \), \( x_{UP_i} \) is the percentage of urban population for country \( i \), \( x_{GDP_i} \) is the gross domestic product for country \( i \), \( \sum_{j \neq i}^{n} A_{ji} \) is the number of flights (network effect) from country \( j \) to country \( i \). \( \alpha_0, \alpha_1, \alpha_2 \), and \( \beta \) are the parameter estimates, and \( \varepsilon_i \) is random noise generated from a normal distribution with mean 0 and standard deviation 0.01, \( \varepsilon_i \sim N(0, 0.01) \).

Consider the incidence rate, the number of new cases of a disease divided by the number of persons at risk for the disease, for the COVID-19 cases reported [14]. This is the number of cases reported over the population per million. To make sure the rest of the data is normalized, the network effect is also scaled. GDP is scaled by \( 10^{12} \).

R comes with a pre-loaded stats package that includes the \( \text{lm}() \) function. The \( \text{lm}() \) function is for fitting linear models and takes in a formula of the response and predictor variables as an argument. The \( \text{lm}() \) function is used because it outputs a linear model in a way that is easy to interpret. The function returns components such as the coefficients, residuals, and p-values.

For each month, percentage of urban population, GDP, network effect, and the COVID-19 incidence rate are fed into the \( \text{lm}() \) function. In return, the intercept and coefficient estimates are output. Estimates are compared to the network regression estimates.
3.2.2 Network Regression

As a way to model our data using a different type of regression, consider a type of community structure detection algorithm. Community detection identifies subgroups of highly connected ‘individuals’ [21]. Found in the network are groups with a high density of connections within community and a low density of links between communities. These communities showcase which node interacts with another one by use of connections, or edges. Edges and nodes are connected by optimizing some criteria [6]. Depending on what form of community detection is used (Louvain, Leiden, and Walktrap, for example) there are different methods to identify the hidden structure of the data.

Consider the infomap method, a type of community structure detection. First introduced by [18], the infomap method minimizes the description length of a random walk, where a random walker serves as a proxy for real flow. Random walks on a network tend to get trapped into densely connected parts corresponding to communities [17]. Shorter random walks imply nodes in same community. The goal is to form clusters in which the random walker stays as long as possible, or where the length of a random walk is the shortest. This is referred to as minimizing the map equation, and this is criteria through which the communities are optimized. [17]. The infomap method is best suited for questions about transmission of information or infectious disease, which is useful for the research regarding public health [21]. To use the infomap method, it is necessary to create adjacency matrices for each month.

Using the monthly count matrix, \( C_{MONTH} \), if an entry in a cell is greater than the calculated total mean of the count matrix, the adjacency matrix will hold a one. If an entry in the count matrix is less than or equal to the calculated total mean, then the adjacency matrix will hold a zero. A one implies that there is a connection between the associated row and column, while a zero implies that there is no connection. After creating the adjacency matrices for every month, we use the `graph_from_adjacency_matrix()` function from the `igraph` package. The `igraph` package houses all the functions necessary to use infomap method. This function provides the visual clusters of the data, displayed as an undirected graph shown in Figure 1. From here, information about the clusters can be accessed. For each node, or data point (country \( i \)), we can see what cluster it belongs to. For example, countries “AM” and “RO” might both belong to cluster 5. Countries in the same cluster have
For each cluster, linear regression is performed, as described above, and parameter estimates are obtained. For however many clusters there are, there are that many copies of parameter estimates. Using all the parameter estimates, the weighted parameter estimate is calculated, a single value for each parameter. This is similar to a weighted least squares. We use the formula,

\[
\hat{\beta}_{MONTH} = \frac{1}{n} \sum_{k=1}^{K} n_k \hat{\beta}^{(k)},
\]

where \(n\) is the total number of countries (nodes), \(n_k\) is the number of nodes in cluster \(k\), and \(\hat{\beta}^{(k)}\) is the parameter estimate for that cluster. Note, this process is done for each month.
### 3.2.3 Simulation: Linear Regression

Consider the behavior of the parameter estimates as the sample size, \( n \), tends to infinity. If the asymptotic behavior of the estimator converges in probability to the true value, then the estimator is considered to be consistent [10]. Instead of using the textbook formula for consistency, squared-error consistency is used. It is acceptable to look at the squared error consistency of an estimator because, if an estimator is squared error consistent, then the estimator is asymptotically unbiased and consistent [10]. To be squared error consistent, an estimator \( \hat{\beta} \) must satisfy:

\[
E\|\hat{\beta} - \beta_0\|^2 \to 0 \quad \text{as} \quad n \to \infty
\]

As the sample size increases, we expect to see the estimator tend to the true value of the parameter. In other words, as \( n \) tends to infinity, the error should tend to zero, indicating that the estimator is accurate.

However, squared error consistency requires that we know the true value. If we had the true value, obtaining regression estimates would not be necessary. Instead, a simulation is conducted to show that the estimator generated from a linear model is squared error consistent. The simulation is conducted by defining the ground truth of the parameter estimators. Using different sample sizes for \( n \) that monotonically increase, random, uniformly distributed data is generated for predictor variables, and random, normal distributed data for noise. Let

\[
y = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + \beta C + \varepsilon
\]

where \( \varepsilon \sim N(0, 0.01) \). For different values of \( n \), we use the `lm()` function to obtain the simulated parameter estimates. It is expected that the parameter estimates converge to the true value as \( n \) increases, as described above.

### 3.2.4 Simulation: Network Regression

A simulation is ran similar to the above method to check the consistency of the network regression parameter estimates. Ground truth is defined for \( \alpha_0, \alpha_1, \alpha_2, \) and \( \beta \). Random samples for \( x_{GDP} \) and \( x_{UP} \) are generated.

To mimic the master set, define \( n = 60 \). Also define \( k = 5 \) for the number of clusters, so there are 12 nodes in each cluster. Create a 5x5 symmetric probability matrix, \( P \), with 0.5 on the diagonal and 0.01 elsewhere. Within a cluster, the probability \( p \) that an edge is present is 0.5. Out of community,
the probability \((q)\) that an edge is present is 0.01. Using this probability matrix, \(n\), and \(k\), we simulate a data generated adjacency matrix using a stochastic block model. Shown in Figure 2 are the communities that were detected from the adjacency matrix, using the simulated probability matrix. Then, for the \(i^{th}\) country, we can now generate \(Y\) as

\[
Y_i \sim N(\alpha^T x_i + \beta \sum_{j \neq i} A_{ji}, \ 0.01)
\]  

\(Y_i\) as

Using the same method as described for network regression, parameter estimates are obtained for the clustered data. Replicate this procedure \(m\) times, where \(m = 50, 100, 300\). Consider the squared error consistency for each replication. The squared error consistency for replications of size \(m\) is

\[
\frac{1}{m} \sum_{r=1}^{m} (\hat{\beta}^{(r)} - \beta)^2
\]

Because \(\beta\) is the only parameter of real interest, the network effect on the COVID-19 incidence rate, we only perform squared error consistency for \(\beta\). As the number of replications increases, we expect to see the loss function decrease.

Figure 2: **Simulated Adjacency Matrix.** The figure shows 5 clusters with 12 nodes each where \((p, q) = (0.5, 0.01)\).
4 Results

4.1 Simulated Data Analysis

We look at the simulated results of the linear regression and the network regression. The network regression, with increased replications, had a smaller squared error than the linear regression with increased sample sizes, according to Table 1. Though, the difference between the values are not significantly different and lead to the same conclusion that both squared errors tend to zero. Because the squared errors tend to zero, we can confirm the estimators for both network regression and linear regression are consistent. Knowing these estimates are consistent, we can proceed with the real data analysis.

<table>
<thead>
<tr>
<th>Linear Regression</th>
<th>n Squared Error</th>
<th>10 $1.93 \times 10^{-3}$</th>
<th>50 $1.17 \times 10^{-5}$</th>
<th>100 $8.88 \times 10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Regression</td>
<td>(m,k) Squared Error</td>
<td>(50,5) $5.82 \times 10^{-7}$</td>
<td>(100,5) $8.94 \times 10^{-7}$</td>
<td>(300,5) $8.35 \times 10^{-7}$</td>
</tr>
</tbody>
</table>

Table 1: Squared Error Consistency. The first row shows the simulated squared error consistency for the linear regression model for sample sizes (n) of 10, 50 and 100. The second row shows the squared error consistency for the network regression model for replication and cluster sizes (m,k) of (50,5), (100, 5), and (300, 5).

4.2 Real Data Analysis

The values of the estimated network effect are shown in the regression table, Table 2. In parenthesis are the linear regression coefficients. The table displays the parameter estimates as columns, and the months as rows. We can now compare the community detection estimates to the linear regression estimates.

The estimates using linear regression are much larger in magnitude than those of the network regression. The linear regression estimates show a monotonic increase for the network effect. In January, the predicted network effect is small, near 0. In April, however, the estimated network effect is about 28, telling us that the relation between network effect and the incidence rate increases. The attention should focus on the estimates for the network effect. The estimates for the network effect using network regression are small. In
March and April, the estimates are 2.6 and -1.4, roughly. Notice that the network effect estimate in April is negative; there is a decrease in the incidence rate as the network effect increases. These estimates are not large, indicating that there is not a strong correlation between the network effect and the incidence rate. What is interesting are the estimates for percentage of urban population and GDP. We notice that as these predictors increase, the incidence rate increases, and increases by a significant amount, especially as we get closer to April. The increased estimates tell us that there is a strong correlation between percentage of urban population and the incidence rate, and between GDP and the incidence rate.

<table>
<thead>
<tr>
<th>Month of 2020</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-0.224 (-0.393)</td>
<td>-0.0006 (0.004)</td>
<td>0.309 (1.447)</td>
<td>0.0003 (0.0002)</td>
</tr>
<tr>
<td>February</td>
<td>-8.956 (-15.108)</td>
<td>0.267 (-0.038)</td>
<td>-0.657 (40.55)</td>
<td>-0.001 (0.005)</td>
</tr>
<tr>
<td>March</td>
<td>-1732.08 (-2393.06)</td>
<td>28.21 (53.74)</td>
<td>619.55 (58.1)</td>
<td>2.61 (2.8)</td>
</tr>
<tr>
<td>April</td>
<td>-5229.41 (-7860.03)</td>
<td>161.45 (224.68)</td>
<td>659.22 (-298.04)</td>
<td>-1.36 (29.07)</td>
</tr>
</tbody>
</table>

Table 2: **Network Regression (Linear Regression)** estimates. The numbers reported in the parenthesis correspond to Linear Regression while the others correspond to Network Regression. $\alpha_0$, $\alpha_1$, $\alpha_2$, $\beta$ correspond to the coefficients of the intercept, percent of urban population, GDP, and network effect.

## 5 Discussion

This research has investigated the methods of linear and network regression. Using community structure, network regression captures more efficient estimators. Real COVID-19 data was used, with predictors such as percentage of urban population, GDP, and network effect. The response variable was the COVID-19 incidence rate, or the number of cases per population per million of each country. Simulations were conducted to establish consistency of the estimators. For both linear and network regression, estimators were consistent. The simulations mimicked the real data, using the same amount of observations.

We use network regression for its ability to detect hidden group structures. Network regression employs methods of community detection to identify the communities. Communities can explain the network behavior by assuming
that the covariates are interdependent. The infomap method is chosen to put the data into communities by exploiting the duality between finding community structure in networks and minimizing the description length of a random walk across the network [17]. By providing good representations of a network, the underlying structure is showcased and the relationships between observations is simplified.

The sizes of the clusters should be noted for network regression using real data. For each month, some the nodes (countries) belonged to their own cluster. This is linear regression: independence among observations. No relation. However, those that did belong to clusters were in a community ranging from 10 to 30 other nodes. The community detection observed independence in some countries, and large communities with the others. This indicates a need for the network regression. Therefore, while we will still look at the linear regression estimates, emphasis should be placed on the network regression estimates. The community structure that was detected will tell us more about what is being observed.

There is not enough evidence to conclude that the travel restrictions put in place during the 2020 pandemic helped mitigate the spread of the pandemic. Shown in Table 2, the estimate for network effect in April using network regression is -1.362. The incidence rate does not decrease by much. The estimate for March explains an increase in the incidence rate. There is not a consistent decrease of the incidence rate of substantial value. Similarly, the linear regression estimates do not show a decrease of the incidence rate. In fact, they show a monotonic increase from January to April. We can conclude that there is not a strong correlation of the network effect on the COVID-19 incidence rate. The incidence rate is explained better by predictors such as percentage of Urban Population and GDP. We see that if a country has a higher percentage of urban population, they have a higher COVID-19 incidence rate. This makes sense. In highly populated areas, where a disease is more likely to spread, we do see a spread of virus. The same conclusion can be made for GDP. This tells us that countries with densely populated areas and countries with high economic health have more COVID-19 cases per population per million. This research ends with the conclusion that the travel restrictions did not help mitigate the spread of COVID-19, and the incidence rate is better explained by the percentage of urban population and GDP for individual countries.
6 Implications for further research

We have interpreted and determined the correlation of the predictors on the COVID-19 incidence rate. We should determine whether the coefficient estimate for network effect, \( \beta \), has significant correlation. By determining the significance of the correlation, we learn if our estimates are important enough to include in the research. We include this as an implication for further research because although confidence intervals were constructed for linear regression, they were not constructed for network regression.

We should produce a confidence interval for \( \beta \) given that we do not know the true distribution of the coefficient. Because the sample is small, it is necessary to bootstrap the data. Bootstrapping is a way to resample with replacement using the data available. With bootstrapped samples, it is popular to produce confidence intervals. If the confidence interval does not contain the null hypothesis value, in our case \( \beta = 0 \), then the results are statistically significant [13]. Using the quantiles of the bootstrapped estimates of \( \beta \), monthly confidence intervals are created, and it is established whether or not 0 is in the confidence interval.

Notice that, in Table 3, the bootstrapped estimates for linear regression for January and February include zero in the 95% confidence interval, indicating that there is no significant correlation between the network effect and COVID-19 incidence rate. However, the estimates for March and April do not include zero in the interval. For the months where travel restrictions were in place, there was significant correlation between the network effect and COVID-19 incidence rate.

<table>
<thead>
<tr>
<th>Monthly Quantiles</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-0.000115</td>
<td>0.0006618</td>
</tr>
<tr>
<td>February</td>
<td>-0.005426</td>
<td>0.013231</td>
</tr>
<tr>
<td>March</td>
<td>0.09882</td>
<td>6.877694</td>
</tr>
<tr>
<td>April</td>
<td>1.820.605</td>
<td>63.5365.572</td>
</tr>
</tbody>
</table>

Table 3: Monthly 95% Confidence Intervals. Sample quantiles for 2.5% and 97.5% for each month.

Though we discuss the network regression estimates, it is not determined if they are important estimators. Confidence intervals should be constructed for the estimates to determine the significance of the correlation.
References


as immigrants and nonimmigrants of certain additional persons who pose a risk of.


