Identifying Factors Contributing to COVID-19 Outbreaks in the US Prison System

1. Introduction

In order to combat the Covid-19 pandemic, researchers and public health officials have worked to collect data to better understand the virus. This data has been used to inform national guidelines and policies intended to prevent the spread of the virus and reduce infection and mortality rates. Understanding which factors may lead to increased infection and mortality rates is essential in making these decisions.

While the general population has certainly suffered as a result of Covid-19, those in prisons and detention centers across the United States have been impacted even more severely. In the US, 9 in 100 people are known to have had the coronavirus, but in US prisons, this number jumps to 34 in 100 people. Additionally, the virus has also killed incarcerated individuals at a higher rate than the general population [1]. This increased susceptible to infection among the incarcerated population was largely overlooked by government officials particularly in the beginning of the pandemic. As a result, correctional facilities have received fewer resources necessary for mitigating the negative impact of the pandemic.

Using data that has been collected on Covid-19 outbreaks in US prison facilities, this project aims to explore factors which have contributed to this increased infection rate among inmates. To do this, I considered multiple generalized linear regression models to predict the number of total inmate deaths within a state’s prison system. The model that was determined to predict the total number of inmate deaths with the greatest accuracy was then used to identify potential factors which have contributed to the increased vulnerability individuals in prison have experienced during the pandemic.

1. Data

The data used for this project comes from a group of datasets containing information on various statistics related to the Covid-19 outbreak in the United States. This data was compiled by the New York Times and made available to the public through GitHub [2]. Two datasets form this group were used in this project: one containing Covid-19 case and death counts for each state and the other containing this information for each state’s prison system.

The dataset on the state prison systems contained eight variables: system name, number of inmate case, number of inmate deaths, number of officer case, number officer deaths, most recent number of inmates in the system, the maximum number of inmates in the system between May 2020 and March 2021, and the number of P.C.R. tests conducted on inmates. The maximum number of inmates variable was dropped from the dataset as the data was not available for several of the systems in the dataset. All the counts reflect the cumulative number of these variables form the beginning of the pandemic until March 2021. This dataset only includes totals in state prisons and does not include local jails or federal facilities.

The statewide case dataset included the cumulative total of reported cases and deaths for each state for each day since the pandemic began. This dataset was filtered down to only include the totals for each state on March 31, 2021 so that these totals were cumulative up to the same date as the totals for the prison systems data. These statewide case and death totals were then joined with the corresponding state in the prison systems dataset.

As with most data related to Covid-19, accurate case and death counts are difficult to obtain due to major discrepancies in reporting between different areas. Some local and state governments chose to only report cases of Covid-19 that were confirmed by a laboratory test while other governments reported both confirmed cases and probable cases as well. Probable cases and deaths were defined as instances where a positive laboratory test was not observed, but the individual was identified by a public health official to have met diagnostic criteria for Covid-19 based on available medical records. For this reason, the counts in this dataset likely underrepresent the actual number of Covid-19 infections, particularly due to the shortage of available tests in the beginning of the pandemic. These differences in reporting or lack of reporting in some jurisdictions makes analysis of individual facilities or counties difficult. For this reason, statewide data was selected over the county-wide data also available in this group of datasets.

All the predictive variables were scaled so the data points had a mean of 0 and a variance of 1. Several of the predictive variable in the model appeared to show positive correlation with one another based on plots shown in Figure 1. This does imply that the model will feature interaction terms between the covariates. The response variable in the model, total number of inmate deaths, was found to be right-skewed, with a much greater variance than mean (Figure 2).

Chart, scatter chart

Description automatically generated

Figure 1: Plots of several covariates

Chart, line chart, histogram

Description automatically generated

Figure 2: Density plot of the response variable, total number of inmate deaths due to Covid-19

1. Model

As the response variable is a count variable, the best fit regression model was a negative binomial generalized linear model. Poisson and normal models were also considered and determined to be less effective in modeling the inmate death counts. The accuracy of each proposed model was assessed using Pareto smoothed importance-sampling leave-one-out cross validation (PSIS-LOO) via the *loo* package in R [3]. The negative binomial model worked better in comparison to the Poisson model due to the large variance found in the distribution of the response variable in comparison to the mean of the response variable. This would make a Poisson model ineffective since a Poisson distribution requires mean and variance of the model to be approximately equal. Fitting a normal model produced a better fit than the Poisson model and was only slightly worse in predictive power than the normal model when comparing each model’s PSIS-LOO.

For the model’s covariates, a normal prior with *mu* = 0 and *variance* = 2.5was placed on each coefficient. Since the covariates were scaled, a normal prior was appropriate for all covariates to model the prior distribution of their coefficients.

Since the predictive variables were heavily interconnected for each observation, all possible pairwise interaction terms were assessed in fitting the final model. Posterior predictive intervals were then assessed for each covariate’s coefficient at a 90% confidence level. Coefficients that were found to be insignificant were then dropped from the model, given the exclusion of these coefficient did not lower the model’s LOOIC.

1. Results

The final model had only somewhat decent accuracy largely due to the issues with accuracy used to build the model. As mentioned in the data section, major discrepancies in reporting information related to Covid-19 greatly impacts the predictive abilities of this model. The model also incorporated relatively few covariates due to the lack of currently available information related to the virus. This points to the idea that a lot of the factors which significantly influence infection and mortality rates of Covid-19 are either difficult to quantify or are not readily available in data that we currently have.

Several covariates were identified as having a significant effect on the response variable based on the predictive model. A lot of these covariates showed the results that were expected. Based on the posterior predictive intervals of the significant coefficients in the model, an increase in the number of inmate cases, an increase in the latest inmate population within the prison, or an increase in the number of officer deaths, the number of inmate deaths was expected to increase as well. The interaction term between the number of statewide deaths and the latest inmate population was found to be significant as well. This indicates that the effects of both the latest inmate population and the number of statewide deaths impact one another, and in this case, when the inmate population is lower than the mean inmate population, a decrease in the number of statewide deaths, we would expect the number of inmate deaths to decrease. Other variables which were include in the model based on their positive effect on LOOIC, but whose coefficients were no longer significant included the total number of officer cases, the number of inmate tests, and the interaction term between the total number of officer cases and the number of state deaths.

One consideration in improving this model was to remove potential outliers. Due to the extreme variation between the data for the different states, identifying which points could be considered outliers was difficult. Based on the PSIS graph for the final model, observations with a Pareto k diagnostic model of greater than 0.7 were identified as potential outliers. Four observations were identified using this approach. Three of the four states identified as potential outliers were California, Florida and Texas which all had a much higher total of inmate deaths compared to other states. This may imply that the model that was fit did not account for factors that these 3 states share which may have caused them to have significantly higher inmate death counts such as larger state populations or conditions which are specific to facilities within these states. The fourth potential outlier identified was Nevada which did not have a high number of inmate deaths. However, it is noted in the dataset that Nevada did not provide data on the number of inmate tests provided past summer 2020. This discrepancy in the data on Nevada likely impacts the ability of the model to make a prediction for this state. Removing Nevada from the dataset significantly improves the fit of the model based on the resulting decrease in the LOOIC of the model the graph shown in Figure 3 which shows the comparison between the observed data and the posterior predictive distributions.

A picture containing text

Description automatically generatedChart

Description automatically generated

Figure 3: Posterior density plots of the model with Nevada excluded (top) compared to the model with all terms included (bottom)

1. Conclusion

While the fit model itself did not offer many significant insights into which factors have influenced the number of inmate deaths due to Covid-19 within US state prisons, it does offer a few potential factors to be analyzed in more detail. Hopefully, as researchers begin to collect more data on Covid-19, there will be more consistent datasets with a more comprehensive set of variables which can be used to provide more meaningful information on factors which have severely influenced the way the pandemic has impacted individuals across the country.

Works Cited

1. <https://www.nytimes.com/interactive/2021/04/10/us/covid-prison-outbreak.html>
2. <https://github.com/nytimes/covid-19-data>
3. <https://cran.r-project.org/web/packages/loo/vignettes/loo2-example.html>