Erick Eaton

Professor Elisa Perrone

Linear Statistics Modeling and Regression

4/29/2020

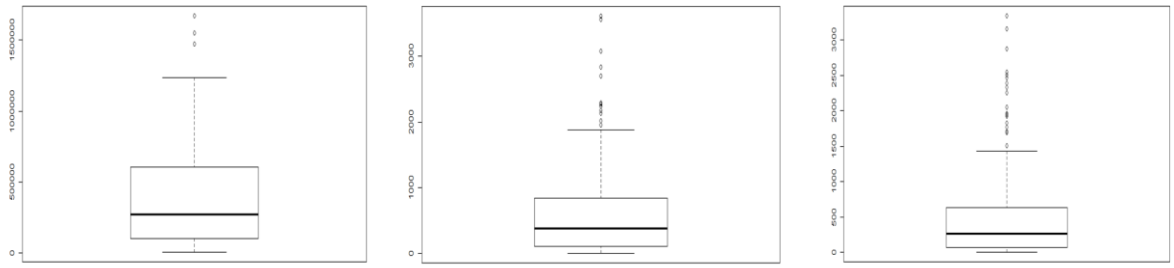
Population Density and the Spread of Covid-19

For my project, I chose to focus on the spread of the novel coronavirus, Covid-19, within the United States for its relevancy and my interest in what affects the spread of it. I began by brainstorming what might cause a virus to reach more people in some communities rather than others. This could be many factors, some more obvious than others. I started with what came most obviously to me, which was how close people are, or the population density. This relationship is most easily viewed in cities like Boston, San Francisco, Seattle, and New York. If we consider population and population density as natural factors, we can also see how important non-natural or governmental factors are in the spread of a virus, factors like the availability of tests or quarantine measures. However, it soon became clear to me that these are much harder to find data on, especially while in the midst of the crisis. This meant that I would unfortunately only be modeling my response using county population and population density data in the United States.

The population data I used was from the U.S. Census Bureau's 2019 population estimates for each county. I also used county land area data from the U.S. Census Bureau's most recent records. I then used R to divide population by land area for each county to get the respective population density in that county. For my response, I chose to express the spread of Covid-19 as the number of new cases in each county x days after the first recorded instance. Choosing x proved to be a challenge of its own because if x were too small, the timeframe would be too short to see the exponential growth. On the other hand, if x were too large, there would not be enough data since not enough counties have had the virus for that long. After

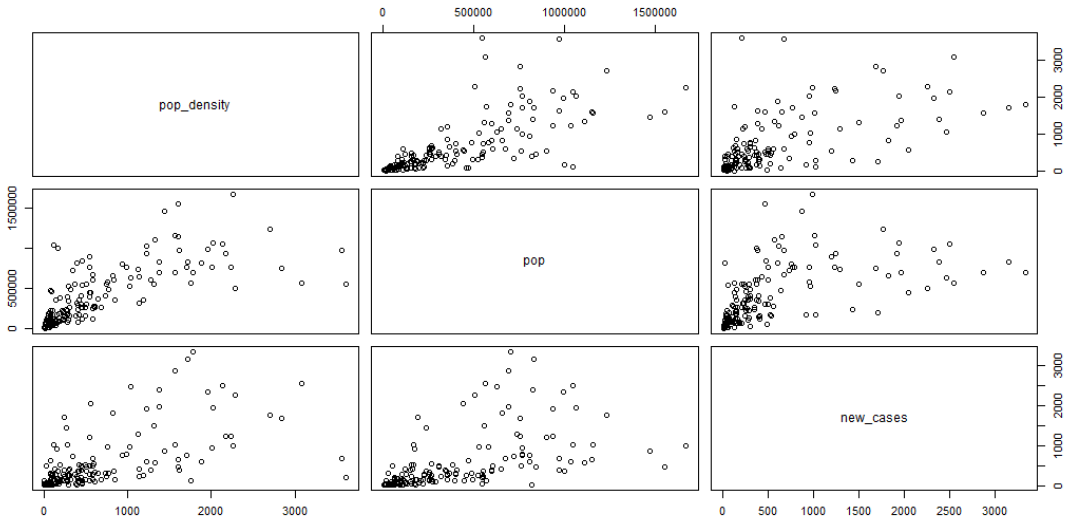
comparing x with the scatterplots, data size, and box plots, I found 45 days to be the optimal length of time for the response.

My first approach to the data were boxplots to get an idea of the scale of the data and potential outliers.



Population Population Density New Cases

The boxplots show that most counties have relatively low populations, population densities, and new cases. The outliers marked at the higher parts of the plots are the counties that contain cities. I removed several outliers from the data, though most cities still remained. Next, I created a scatterplot matrix of the two predictors and the response to determine the relationships between them.



There is a positive relationship between population density and the amount of new cases in each county. There is also a positive relationship between total population and the amount of new cases. Additionally, there is a positive relationship between the two predictors, though not perfect. Lastly, for my initial analysis of the data, I created a correlation matrix to view the relationships numerically.

```
> cor(final_x)
                    fips        pop pop_density   new_cases
fips         1.00000000 -0.1750600  -0.1025527 -0.03463267
pop         -0.17505996  1.0000000   0.7226909  0.56043232
pop_density -0.10255266  0.7226909   1.0000000  0.61885057
new_cases   -0.03463267  0.5604323   0.6188506  1.00000000
```

The correlation matrix tells us much of the same information we can find in the scatterplot matrix, but in a more concise manner. Population and new cases are 56% positively correlated, population density and new cases are 61.8% positively correlated, and our two predictors are 72.3% correlated.

To choose a model, I created five models, m0 to m4, beginning with the null model: new_cases ~ 1, and ending with the full model: new_cases ~ pop_density + pop + pop_density:pop. Comparing the adjusted r-squared value in each model, I was able to select the model that best explained the variance of the response. This led me to the conclusion that the full model with both predictors and their interaction term was the best model because it had the highest adjusted r-squared, with a value of 41.3%. I then also ran an anova test on the full model to see the significance of each term.

```
> anova(m3)
Analysis of Variance Table

Response: new_cases
                Df   Sum Sq  Mean Sq  F value  Pr(>F)
pop_density      1 29878924 29878924 102.4037 < 2e-16 ***
pop              1  2092540  2092540   7.1717 0.00821 **
pop_density:pop  1  1112785  1112785   3.8138 0.05264 .
Residuals      154 44933493   291776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both individual predictors tested significant with p-values below 0.01, while the interaction term only had a p-value of 0.0526. Because the interaction term was very close to a significance of 0.05 and also because the inclusion of it in the model improved the adjusted r-squared, I decided to leave it in.

Admittedly, an adjusted r-squared value of 41.3% is not a very strong value. This shows that while population and population density are both important factors in predicting the spread of Covid-19 (and most likely other viral infections), they are not the only important factors. If I were to continue working on this model, I would add several more predictors to see how they could improve my r-squared value. I would include the following factors: number of available tests, average income, average age, days under quarantine, and number of hospitals per 1000 persons. I believe these factors, or their interactions, could potentially improve the accuracy greatly. The model could also be improved with more data, which is only available in limited amounts since the pandemic is still far from over. The sensitivity of the model's r-squared value became apparent when playing with the number of days of the response variable and the inclusion or exclusion of outliers. For instance, when the model is run with the outliers included, the adjusted r-squared value shoots up to 80.1%. In conclusion, there is clearly a strong relationship between population, population density, and the spread of Covid-19, but it is not a perfect relationship and the predictors are insufficient to adequately model the spread of the virus within a community; more predictors and data are necessary.

Sources

- https://github.com/nytimes/covid-19-data
- https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html#
- https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_counties_current.html