James Yang

DATA 512

Part 1 – Common Analysis

How did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021?

1. **What needs to be cleaned and standardized over the three datasets?**

In order to standardize and clean the three datasets, we need to first link the raw us deaths with the state associated to the New Jersey state and the columns associated with the Bergen county. Another thing we need to clean is the dates associated with each of the states and the FIPS value that links with the mask use by county. Lastly, the number of deaths and the date associated with each date needs to be standardized across each of the datasets so they can be interpretable.

1. **There is a delay between the time of infection and the time a case is confirmed. Many factors may contribute to such delay. People may not show symptoms right away after infection. It may take a few days for the testing results to become available especially during the early period of the pandemic. Should we model the delay?**

To model the delay, it can cause discrepancy between the actual vs. estimated vs. delayed. I believe it would create a more convoluted model that has one more predicting feature that wouldn’t necessarily give a better output. The delay itself could be modeled in the output of the death simply through the drastic increases in recorded deaths in short spans of time. Unless it is attributed to a rapid increase of people getting covid, we can speculate that it is also incorporating the delay already.

1. **Masking may simply make it longer to get infected or it may prevent some percentage of infection. How should we consider the effect of a mask?**
2. **The research question is about how a time series changes. The infection time series is a set of slopes. Therefore the question is about a derivative function. That is, you want to answer a question about the change in slope over time. How can we test the difference in the derivative function?**

One method is the ruptures module or Pruned Exact Linear Time (PELT). The regression line is determined by minimizing the sum of distances of all the data points. The algorithm essentially splits the data into sectors that have change within the data. It can detect when the overall sectional changes occur, which allows for calculating variance as well. (<https://medium.com/dataman-in-ai/finding-the-change-points-in-a-time-series-95a308207012>)

1. **Masking survey data shows probability of compliance in several categories. How can we model different proportions for population compliance?**
2. **Masking policies varied in their implementation (e.g., size of “crowd” required, different situations, restaurants, bars, clubs ...). How should I handle things when my County implemented two different policies at different times?**
3. **The County I was assigned did not implement a masking policy! What is a reasonable way to answer this question? That is, how might I model “voluntary” masking?**
4. **Vaccinations probably impacted the apparent effectiveness of masks. How should we account for different vaccination rates in different populations within the same County?**