Impact of Population on Coronavirus Testing and Deaths

Max Tokman

June 2020

Github Repository: `https://github.com/maximetokman/CS-349-Final`

Introduction

As the Covid-19 pandemic leads to over 100,000 deaths in the United States and many more across the world, many countries have put in place testing and various restrictions. We will focus on the United States in this series of modeling. The majority of states put in place stay-at-home orders of varying levels. Different states had different testing capabilities, and population density among states is quite different too. As we continue to search for ways to tackle this virus, some states are more ready than others to begin opening up safely. I examined various data sets containing the following information about the virus in the United States: the population of each state, the number of individuals tested in each state, and the number of deaths per state. I wanted to better understand how population as well as population density affect the quantity of tests conducted in a state, and how that might affect the level of deaths in a state.

Modeling

I created two different models, one using population as the input and the other using population density per square mile. It makes sense that population density would give a better representation of how testing and deaths are affected because perhaps a higher density of people means that the virus is spread more easily, but it could also mean that more tests can be administered if people are more closely located. I built the models to visualize this. In order to predict the number of tested individuals in a given state given the population or population density, I built a polynomial regression using data from mid-May. It is important to note that I split the population density models: one includes Washington D.C in the data and the other does not. D.C was a major outlier in terms of population density because of its small area, skewing the regression and making the other data points illegible. I played around with various degrees of polynomials for the regression, and decided to use degree 6 for the regression involving raw population, degree 2 for the population density regression including D.C, and degree 15 for the population density regression not including D.C.

After the number of tested individuals is predicted based on the population/population density using regression, we can then determine the death level given the population and testing information. I defined 3 levels of death, denoted on the plots by green, orange, and red points. Each level corresponds to a certain threshold proportion of the entire population that died due to the virus. I used the following proportions: green: 0.0001 (0.01%), orange: 0.0005 (0.05%), red: 1 (100%). I selected these thresholds by trying a few different sets of values out and seeing how spread the death proportions were. These thresholds gave a reasonable amount of each color, and while a single death due to this virus is significant, it is important to look at the number of deaths relative to the population of each state.

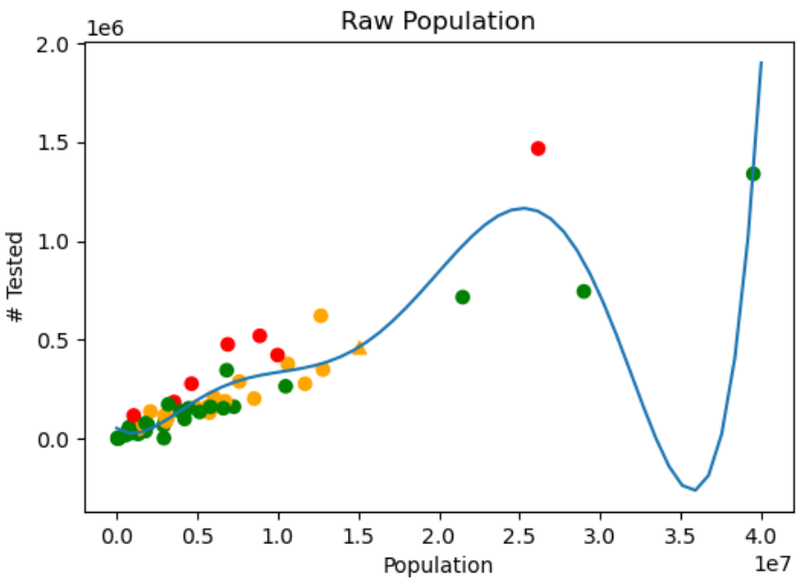
The next part of each model is determining the death level given the population/population density and the number of individuals tested. Predicting the death level requires having a value for population or population density and a value for the number of individuals tested, whether predicted from regression or an actual amount. To predict the death level, I used the k-nearest neighbors algorithm. First, I assigned a numerical value to each of the 3 color labels for death: green = 1, orange = 3, red = 6. The reasoning behind this is to add weight to the more severe death levels. In other words, if a certain point is close to other points that are red or orange, I argue that it is important to weigh the red points more, since a point near the red point may be in a red zone, and more precautions should be taken for the safety of that region. Thus, I add weight to more severe death rates so that the model overestimates the death level. It is better to overestimate than underestimate when making decisions about how to slow the spread of the virus in a certain region.

Once the colors were assigned numerical weights, I created a nearest neighbor learner which would use 3 neighbors and the 'manhattan' metric to compute distance. These settings were used for all 3 scenarios. I then inputted the population/population density and testing value into the nearest neighbor model to get the k (3) nearest neighbors to that point. I computed the average label among those 3 neighbors, which accounts for the increased weight of more severe death levels. Then I found the label that had the smallest absolute

value difference when compared to the computed average label. If there were multiple labels that had the same absolute value difference, I took the greater label, again to avoid underestimation. The corresponding color would then be the predicted death level for a point consisting of a specific population/population density and number of individuals tested.

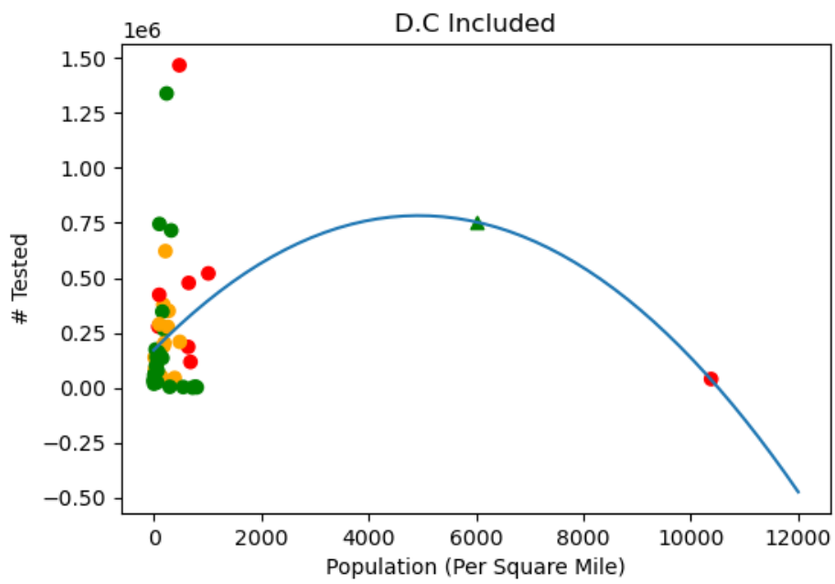
Analysis

I will begin by discussing the model of raw population. I created a degree 6 polynomial regression to relate the number of tested individuals to the raw population of a state. I then tested the model on a population of 1.5×10^7. The regression predicted that there would be roughly 461,354 people tested. Then I inputted these two values into the k-nearest neighbors model to predict that the death level would be orange, or moderate. Below is the graph of all points, as well as an orange triangle denoting the prediction.



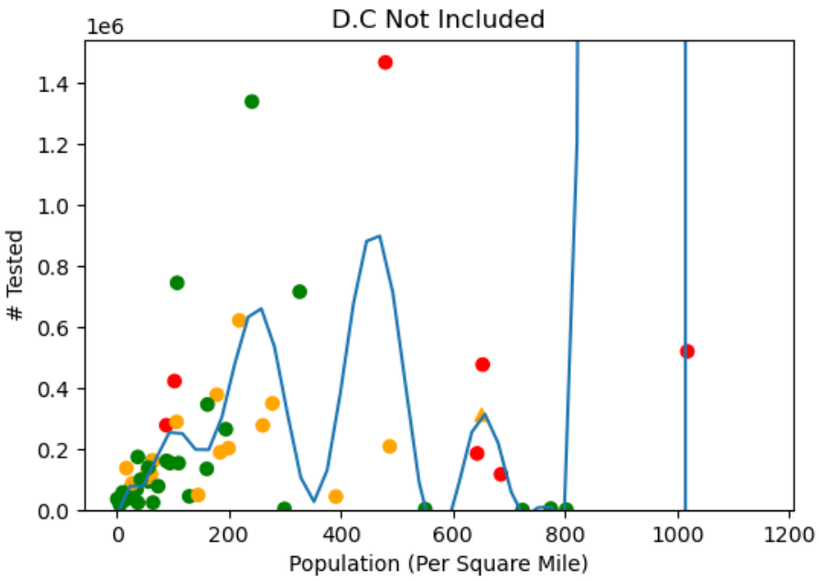
We can notice that there were many more green points at a smaller population than at a higher population. With that said, it also seems that more severe deaths occurred at lower populations and higher testing capacities, which doesn't completely make sense. We would assume that more testing allows for increased prevention of further complications and may prevent death. Simply modeling on raw population does not give us sufficient information to make meaningful conclusions, which is why I proceeded to build 2 more models using population density, with the hope that they might give us some important insights on where officials should focus their virus prevention tactics the most.

I created one such model including Washington D.C, and one without. Below is the model with D.C, and as we can see, D.C is a major outlier. The rest of the points are vertically but closely clustered and are difficult to distinguish because the scale is very skewed. D.C has a density of nearly 11,000 people per square mile.



I ran the model using a population density of 6,000. The resulting death level was green, denoted by the green triangle. It is difficult to generalize much from this because we do not have enough data points around the 6,000 population density mark to make meaningful predictions. We can deduce, however, that a population density nearly as high or higher than that of D.C is very likely going to have a severe death rate because people are generally closer together and the virus can spread more easily. These areas may require more strict distancing measures. Next we will look at a significantly more meaningful model that does not include D.C.

I left out Washington D.C because it is an outlier in terms of population density and it is important to focus on the other states to draw more general conclusions. I generated a degree 15 regression and ran the model on a population density of 650 people per square mile. The predicted number of tested individuals for this population density is roughly 313,198 people, and the k-nearest neighbors death level prediction is orange. This is shown below by an orange triangle.



First, let's look at the relation between population density and the number of individuals tested. In general, a higher population density is related to a higher number of tested individuals. This can be seen by the increased variance of the regression curve. This makes sense because when more people are closer together, or the area of the state is smaller, it might be easier to distribute tests to more individuals. On the contrary, it will be more difficult to get tests across a larger state.

Next, we can see that below a population density of about 400 people per square mile, as more individuals are tested, the death level tends to remain green. This becomes less true as we go beyond 400 people per square mile. Beyond that density we see fewer green points and more orange and red points. Testing capabilities also seem to decrease. This result is inconsistent with the earlier trend below population density 400. The model indicates that these regions should take greater measures to counter the spread of the virus due to their high population density. Furthermore, there are other factors that were unaccounted for in these models that could explain the different trend after the 400 population density mark. Such factors could be political or economic, but if accounted for, they could potentially eliminate more outliers that skew the results. For example, if a certain state has leadership that is unable to distribute tests or makes harmful decisions, that state is not representative of the rest, and like D.C, poses as an outlier. By eliminating such outliers, we might see a similar trend throughout all population densities and not just up to a certain density.

Conclusion

In summary, I took data including the population, number of people tested, and number of deaths due to Covid-19 for every state and territory in the US. I then created 3 models—one using raw population, one using population density including D.C, and one using population density not including D.C. I generated polynomial regressions of varying degrees to optimize fit, and used the regressions to predict the number of tested individuals given a population or population density. Finally, using both of those values, I used k-nearest neighbors to predict the death level, making sure to weigh higher death levels more significantly to ensure an overestimate rather than an underestimate.

Of the 3 models, the population density model not including D.C is the most meaningful and accurate. The raw population model does not provide too much insight because it does not account for the size of the state, which affects things like how efficiently tests can be distributed and how closely people live to one another. The model of population density including D.C is difficult to read because D.C is an outlier in terms of population density, and is only useful if predicting for a population density near that of D.C.

However, such a density is unlikely and furthermore the model is not very accurate because D.C is the only point with such a large population density.

From the model not including D.C, we can see that with increased population density (up to 400 people per square mile), there are more individuals being tested and lower death levels. This type of trend makes logical sense because we would expect that as the density increases, the state is probably smaller, making it easier to distribute tests and test more people. By testing more people, we can treat the virus earlier for many and avoid death. We can also identify more cases and isolate those people from others to prevent spreading the virus. Yet, as population density goes above 400 people per square mile, this trend vanishes and there are mostly red points scattered with some green. The positioning of such points does not seem to provide much of a trend and can be attributed to political, economic, or other factors not accounted for by the models. However, they do show that higher population densities above 400 people per square mile tend to lead to higher death levels, which makes sense because there is less room for distancing. The models would imply that regions with high population density should have stricter distancing regulations. However, if the models accounted for other factors mentioned above, perhaps these points would be outliers and not generally applicable. On the other hand, having these outliers might be causing an overestimate of deaths, which is safer than an underestimate, especially since many regulations involving distancing and staying at home are not fully enforced.

Certainly these models could be further improved by accounting for more factors that affect how testing capabilities relate to population and deaths, but we have seen from these models that a trend does exist between these criteria. Causality cannot be implied from the models, but we can use these models to predict deaths and testing capabilities from population or population density. Such predictions can help make decisions about distancing and safety regulations to help slow the spread and danger of the virus.