Jordan Roberts

July 24, 2020

BAN 530

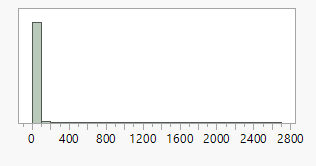
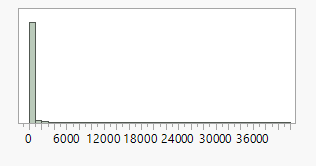
Cross-Validation/Sampling

I am using JMP Pro 15 as my main model building application. After importing all six datasets into JMP, I begin by creating validation columns for cross-validation purposes. Two of the datasets (country\_wise\_latest and worldometer) are not time-series sets, but merely reports for a specific day.

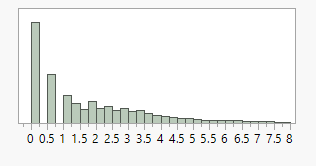
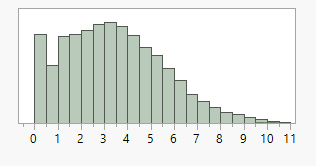
The other four datasets contain time-series data from January 22 to June 23. To the time-series datasets I have added a validation column that splits data into one of three groups (Training, Validation, and Test) based on time elapsed from the beginning. The remaining two datasets receive a random assignment; both versions of the validation column follow a 60-20-20 split. This will ensure that an adequate number of data points from each set can be used to create a model to predict confirmed COVID-19 cases as well as deaths.

**State/County Data**

I removed Cook, Los Angeles, and New York City counties for my US analysis in JMP, as they are outliers that skew the data quite a bit. Also removed are Unassigned counties and cases “Out of” particular states, as they do not help with the work of my project scope, which is trying to find specific regional statistics.



The left graph shows distribution for Confirmed Cases, while the right shows Deaths distribution. We can easily see that most data points are at or near zero, which is backed up by previous visualizations as well. Using the log function in JMP, we can create a more even distribution for both:



Again, left is Confirmed Cases and right is Deaths. On a log scale, we can see a more even distribution. Deaths still skew toward zero, but Confirmed Cases spreads out nicely.

The first months of the pandemic saw very low activity in the US, with fewer than ten counties affected within the first month. Cases exploded toward the middle of March, however. Because of this, I created three validation groups: starting from January 22, one month in, and two months in. With fewer zeroes skewing predictions low, I am hoping that models will be more accurate with more recent data. By June 23, 145 counties still had no confirmed cases present. March 26 was the first day that over half of all US counties had at least one infection, which closely lines up with validation group 3.

I added several new variables to the county data, including change from the previous day as well as lagged changes from two days ago. This is to see if past changes are having an extended effect on cases and deaths later on. In addition, I have added a categorical variable that tells me if a county is above or below the average for confirmed cases or deaths for that day of the year. This will be very helpful as it will show me which areas are struggling compared to others. Also, it serves as a useful response variable that I can use my models to predict. I will also include this variable in country data.

**Country Data**

Five of the six datasets deal with worldwide statistics instead of US areas. They are interconnected, but each has its own uses. For the time-series datasets, I have added the same categorical variables - Above Average Active Cases and Above Average Deaths - that I added to the county dataset. A simple YES or NO response tells me if a country is struggling to keep new cases or deaths down below the average. As almost every country has been infected by the end of the timeline, this seems like an appropriate metric by which to gauge each country’s success in containment.

The “worldometer” dataset is not very useful for my predictive methods in terms of my scope. It gives me the data by country for the last day of the timeline, June 23. This is also present in both “covid\_19\_complete” and “full\_grouped”. However, it does have information about tests and populations, which may come in handy during the prescriptive portion of this project.

The “country\_wise\_latest” dataset actually seems to contradict other datasets within its records. However, it has statistics of deaths and recoveries per case. This may come in useful later on, but for now I will probably stick to the time-series datasets for predictive modeling. These sets have much more data for me to work with, too.

Ultimately, for worldwide COVID-19 data, I will be creating models using “covid\_19\_clean\_complete” and “full\_grouped”. I perform the same process to create a validation column for these sets as with the county data. Using a 60-20-20 split, I create a group each for Training, Validation, and Test data for my predictive modeling going from the first date toward the last. I am looking for the best predictors of Above Average Cases and Deaths, and I will then use them to predict future struggles for countries and counties of interest.