Exploring the Global relationship between COVID-19 Positive Tests and Deaths

Since positive COVID test cases are some of the most direct information organizations around the world can provide for the 2020 coronavirus pandemic, our objective is to analyze the relationship between these positive cases and deaths from COVID to collect more information about the effects of this virus. We used the COVID data from the European Centre for Disease Prevention and Control (ECDC) starting from the beginning of the year 2020 until September 20, 2020 to model the number of deaths on the number of positive test cases per country per day. Using information available from this data set and current information available about this virus, we reasoned that the relationship between cases and deaths would be additive (or a 1:1 relationship), where an increase in cases would match an equivalent increase in deaths, and the majority of our analysis focused on whether this claim is reasonable. If our analysis supported this 1:1 relationship, the results would also bolster the effectiveness of positive COVID tests as a measure of COVID exposure and severity.

The data already included daily numbers of positive test cases and deaths per country, but we did additional data cleaning. To manage the lag between testing and death, we created four more covariates by taking the sum of the cases for each week for the previous 4 weeks before that specific day (e.g. the number of cases for week 0 is the sum of cases in the 7 prior days). We converted the date information into days since the beginning of the year and days since the first death in that country. We also removed all observations with negative deaths and/or cases since these are likely outliers due to administrative issues.

With the adjusted data, we modeled deaths on the 4 lagging case covariates based on the Quasi-Poisson family because of the count structure of the data and the flexibility for both the mean and variance structure. We used the log of those covariates along with the inherent log link of the Poisson mean structure so that the coefficients adding to 1 would be a check for a 1:1 relationship. Fixing the effects of country and time seemed the best way to lower variance and best fit our data compared to other combinations of effects and interactions. Adding country as a categorical variable, we fit separate models to decide if rdays, the days since the first COVID death in that country, or days, the days since the beginning of the 2020 year, would result in a better fit.

The coefficients of the log-lagged positive cases summed close enough to 1 for both the rdays model (0.30, 0.33, 0.27, 0.02) and the days model (0.31, 0.26, 0.21, 0.10) that both seemed to appropriately fit the Poisson mean structure of a 1:1 relationship with approximately 10-15% standard error and all relevant covariates being significant at the 0.05 level. The two differed more in the variance structure. The rdays model had a dispersion scale parameter of 35.9 while the days model had 41.9, suggesting that rdays could explain more of the variance in deaths rather than days. We noticed many possible outliers in the number of deaths and compensated with robust Huber scale parameters, which came to 1.49 for rdays and 1.45 for days. The decrease by a factor of 10 in dispersion scale parameters supported our theory that our models were relying heavily on outliers, but the resulting parameters were so similar that we could not say that the two models differ after taking outliers into account. We then modeled using Tweedie with a power of 1.5, which has the same mean structure as Poisson but has a variance structure that is the power of the mean rather than a scale of the mean, to fit on a different distribution. These Tweedie models provided coefficients that matched the expected mean structure (0.47, 0.21, 0.17, 0.11 for rdays and 0.47, 0.19, 0.14, 0.14 for days) with similar standard errors and significance to the Quasi-Poisson models and lower scale parameters for the variance structure (7.05 for rdays and 6.78 for days). These results implied a more complicated variance structure than Quasi-Poisson rather than a heavy reliance on outliers.

The mean structure of all our models supported a 1:1 relationship between positive test cases and deaths. All had a sum close to 1 of the log lagged number of cases, suggesting that there is an additive relationship between cases and deaths, and that knowing the number of positive test cases will give us information about the severity of COVID in that country that is proportion to those cases. While the mean structure was consistent with our claim, the variance structure demonstrated overdispersion that is difficult to account for. The robust Huber scale parameters of Quasi-Poisson models suggested that heavy reliance led to the overdispersion, while the scale parameters Tweedie models suggested that a more complicated variance structure explained most but not as much of the overdispersion. There may have been an unaccounted for pattern in the outliers, so the Tweedie models would be the safer option for inference without any additional data or outside knowledge.

We considered the possibility that the overdispersion could decrease with fewer countries of lower populations or with a different combination of covariates, but the resulting models were more overdispersed. There may be natural overdispersion because of the global nature of the data or unknown dependencies and inhomogeneity that were not captured in this ECDC data set.