Positive COVID Test Results and COVID Mortality in the United States
Authors: Diana Liang and Ryan Duncan

The United States first identified cases of a novel coronavirus (COVID-19) within our country's borders in January of 2020. Since then, COVID has ravaged America with a death toll of almost 300,000 lives. We aim to quantify this impact; more specifically, we will assess the extent that positive COVID test counts affect mortality rate in the U.S.. To accomplish this, we use a hierarchical Poisson generalized linear model (GLM), which allows us to analyze the relationship between death counts with weekly positive test totals as covariates. Results from the selected model tell us that the most influential positive test-based predictor of COVID deaths in America is the positive test total from three weeks prior to the resulting death count.

Our data comes from the COVID Tracking Project (https://covidtracking.com/), which is unaffiliated with any organizations but collects data from multiple sources to better understand national trends of COVID-19. The data goes back to February 2020 and continues to update as the pandemic continues; usually daily. While it's impossible to have complete data on COVID, which would include all cases and full demographic information, enough major news groups trust and rely on this dataset for us to make meaningful inferences from its information. We obtained the most recent data on Wednesday December 11th, 2020 that contained variables describing state, date, different testing results, and mortality. Since we were only interested in daily positive test cases and daily deaths, we isolated the variables and discarded observations where the daily cases and deaths were negative, which are theoretically impossible. Death tends not to come immediately after positive test results, so we had to take into account the possible lag between a positive test case and the resulting death. We created new 3 lagged weekly positive cases to compensate. For each observation, we summed the number of positive test cases for the week prior (0-7 days before) for the first week of cases, and then repeated a similar

process for two weeks prior (7-14 days before) and three weeks prior (14-28 days before). We also wanted to consider possible confounders by controlling for geography and time, the two other types of variables available in the original data set. We used state for geography, and we created a new variable to enumerate days since the first case in that state for time to better match epidemiological behavior.

Since we are counting the number of deaths, we modeled the relationship of positive tests and deaths on a Poisson GLM as mentioned above. Poisson GLMs use a log link between the mean structure and the linear function of the predictors. In turn, the variance structure should be equal to the mean structure:

$$\log(\mu) = E[Y|X=x] = \beta_0 + \beta_1 * x_1 + \ldots + \beta_p * x_p$$

$$Var[Y|X=x] = \Phi * Var[E[Y|X=x]], \text{ where } \Phi = 1$$

The relationship between the predictors and the response, then, would be multiplicative, but we included the log of the weekly positive cases as the predictors rather than just the weekly cases because we expected the positive cases to be additive rather than multiplicative in relation to the number of deaths. If the relationship of the predictor of positive tests, specifically the log of the positive test cases, was 1:1 with the response of deaths, we could conclude that there is an additive relationship between positive test cases and deaths and that we can reasonably predict the coming number of deaths from the previous COVID deaths and current information on positive tests. We created three lagged weekly positive case predictors, though, so the sum of their effects should equal 1 rather than individually for the same conclusion to apply.

Poisson assumes that the variance is equal to the mean, but confounders could still have an effect on the variance that is impossible to distinguish from the actual variance from positive test cases. To limit the influence of confounders, we controlled for geography by using a

hierarchical model and for time by including it into the linear function as a spline. A hierarchical model can group the data so that the observations within the group come from the same distribution, but the observations between groups came from technically different distributions that are defined by a higher distribution. So, all states have the same higher distribution of test cases and deaths, but each state has a single draw of that higher distribution with different errors for each state. Grouping by state would also control for demographics other than geography, such as population. We added days since the first case in the state as a spline to control for time and other confounders that might be related to time, such as availability of supplies.

After building the model under the stated framework, we find that the estimates for each of these weeks sum to 0.909. Since the sum is approximately 1, we can justify the existence of an additive relationship between positive test cases and mortality such that we can predict the coming number of deaths from the current information on positive COVID tests. Among our three predictors of mortality (positive test totals for one week, two weeks, and three weeks prior to death count), positive test totals for the third week before death count effectively account for almost half (0.453) of a death. Positive test totals for the second and first week prior to death count account for 0.366 and 0.09 of a death, respectively. Figure 1 shows the trace plots of the beta estimates, or the sampled values of these estimates throughout the non-discarded iterations of our chains, as well as the deviance. As we can see, the mixing in the first three subplots is not ideal, however, in table 2, which shows our estimates for our four parameters, as well as our effective sample size (n.eff) and R-hat, which is a measure of how well our three Markov chains have mixed, we can see that our R-hat values are not substantially larger than 1. This means that our chains have mixed well enough and that posterior estimates can in fact be trusted. Unfortunately, our effective sample sizes are small across the board, which indicates that

autocorrelation is still present in our saved samples. We suspect that our estimates are not final, but they do seem to stay within a range where their sum would still be 1, justifying a 1:1 relationship. Also, our deviance estimate is much higher than we'd like, which indicates that this model does not fit well to our data. This could come from the assumption that the variance structure is equal to the mean structure, which is likely untrue due to overdispersion in our data and model.

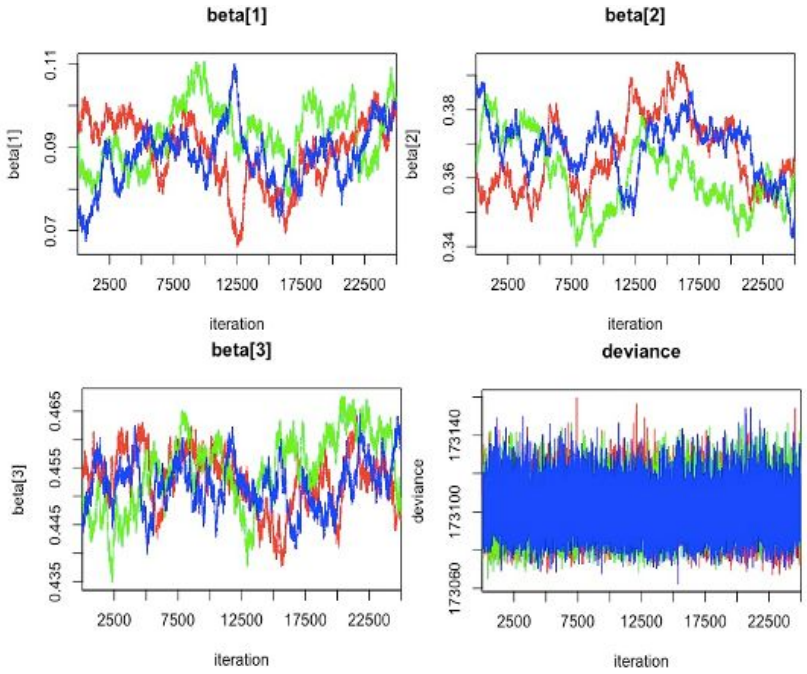

Figure 1. Traceplots of the sampled values of our parameters over 25,000 iterations.

	Estimate	**n.eff**	**R-hat**
beta[1]	0.09	25	1.1
beta[2]	0.366	19	1.146
beta[3]	0.453	48	1.075
deviance	173100.706	1	1

Table 2. Summary of jags output for chosen model parameters.

As discussed above, the limitations to our analysis include the possibility of remaining autocorrelation from running too few iterations. The size of our data set and the complexity of the model demanded more iterations than our computational and time restraints would allow. At 50,000 iterations and 3 chains, the model ran for over 30 hours. The weekly positive test estimates seemed to have stayed within a range that confirmed a 1:1 mean structure, but many of the estimates for states were not stabilized. Another limitation is the large deviance estimate pointing toward a model-misspecification. Our count data may not be exactly Poisson-like due to additional sources of heterogeneity, including administrative errors in reporting and demographic data that we did not control for. News about COVID reporting suggests there were days when information compounded over time was entered on a single day instead of on the days that they occurred; and we had to remove several observations of negative deaths or positive tests. Even if our results did not fully support a Poisson-like model for the data, our findings do not require the variance model being correct. Our queries only need to verify that a 1:1 relationship for the mean structure is correct. Our best model has suggested that it does exist, but there is a chance that the covariate estimates may change once autocorrelation no longer remains.

Assuming that a 1:1 relationship between the number of positive test cases and the number of deaths still exists after proper mixing for all covariates is achieved, we can justify using a percent change in positive tests to predict a similar percent change in the upcoming deaths. The overdispersion likely present in our analysis suggests that the variance is too high or complicated for the exact number of deaths to be predicted, only that an increase in positive tests foretells an increase in deaths. In future studies, we would like to find the resources to run the model for a sufficient number of iterations for proper mixing and to include other demographic confounders such as sex, race, income, and type of living area (rural, urban, suburban, etc.).