Count Data Regression: Modeling Daily Count of New Covid Cases

Jonathan Amdur

12/14/2021

Beginning in January of 2020, the novel corona virus COVID-19 had started to appear in the news and around the world. The WHO and began tracking it to ensure the growth of it did not get out of hand. As the virus spread and more was learned about how it works, grows, and spreads, more and more data collection allowed us to track the daily count of new covid cases. As time has gone on, we have seen new ways of combating the virus take affect from lockdowns to vaccinations and boosters. With this data, we can model the progression of these daily covid cases which can help explain what techniques are working, what is failing, and when the next wave of the virus is coming. With this in hand, front line workers and decision makers can utilize mathematical regression models on count data to ensure the safety and continued success fighting the virus. Now, our goal is to find the best fitting model from the linear regression model, Poisson GLM, Negative Binomial GLM, and COM-Poisson GLM for the count of daily new covid cases in 6 countries studied with data gained from https://ourworldindata.org/coronavirus.

Data for this project was sourced from https://ourworldindata.org/coronavirus. This data is a daily aggregation of key country statistics and tracked factors that are known to affect Covid-19 case rates. This dataset contained 134,015 rows and 67 columns for data from 1/1/2020 through 11/17/2021. This represented tracked features for 237 countries across the globe. The amount of data for each country varied based on variability, so the decision was made to limit to just 6 countries. This was done based on the study design in the *Count regression models for COVID-19* study, which varied the countries based on length of covid presence and geographic locations. The countries of Canada, United States, Great Britain,India,Korea,and Sweden were chosen as our representative countries based on similar criteria to the study, as well as the variety of lockdown approaches and availability of data for the first confirmed cases. For all countries, the new_cases variable represents the daily count of Covid-19 and is our target variable for the proposed count models. Analysis on this variable revealed that while counts should be strictly positive, a few countries did include negative entries. These entries were investigated and while the initial assumption was that they were indicating corrections to previously reported cases, this was not the case and seemed to be an error. To handle these values and ensure a proper count model was performed, the absolute value of these were taken and the negatives assumed to be erroneous.

While performing initial analysis, 3 new variables were also created to ensure dates were usable for the model building process. The days_covid_present variable represents the days between the date field and the first date where the new_count was greater than 0, which represnts the first confirmed case(s). This variable allows us to model new cases based on how far into the pandemic each country is, rather than the specific date which has varied points in the lifecycle. Further, the date columns month and year were preserved as numeric columns to account for seasonality of the virus. It has been seen that the case count does increase in the winter months so these new variables allowed for that to be accounted for. Since all countries are located in the northern hemisphere, the season also will align and means these variables make sense within the context of the model.

With these new variables created, basic data cleaning and null value imputation was performed. It was found that many columns were unecessary since they were just duplicates, smoothed, or scaled versions of the same data, or completely null. These variables were dropped from the data. Further analysis showed that for all total columns, the current days counts were already pre-added. To ensure the total was

representative of the total up to that point, which would remove the new days count from affecting the new covid case count variable, the corresponding new count variables amount was subtracted for the total of that day. Once the variables were cleaned up, null value imputation was performed to ensure there was no missing values in the data. For variables with nulls specific to countries, recent news articles were consulted to fill out total boosters and the mean of all other counties was used to fill out missing days. For all other nulls, 3 approaches were taken. First, for any value that represented a running total, any day before the first day with a non-null value was assumed to mean no affect was present, such as no people fully vaccinated before the first time a fully vaccinated person was counted, and a 0 was used in place of null. Second, for any nulls in a running total that had values the day(s) before or after, such as total_tests, the last valid value from the country was used for the running total. Finally, standard mean imputation was used for anything still remaining.

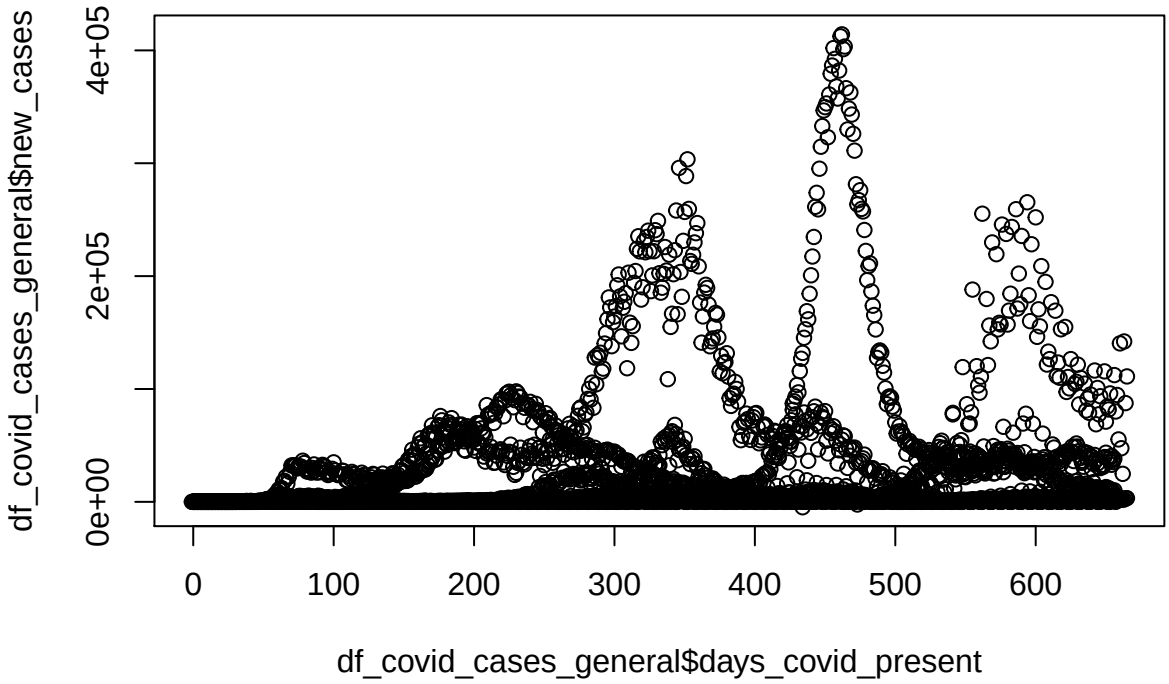
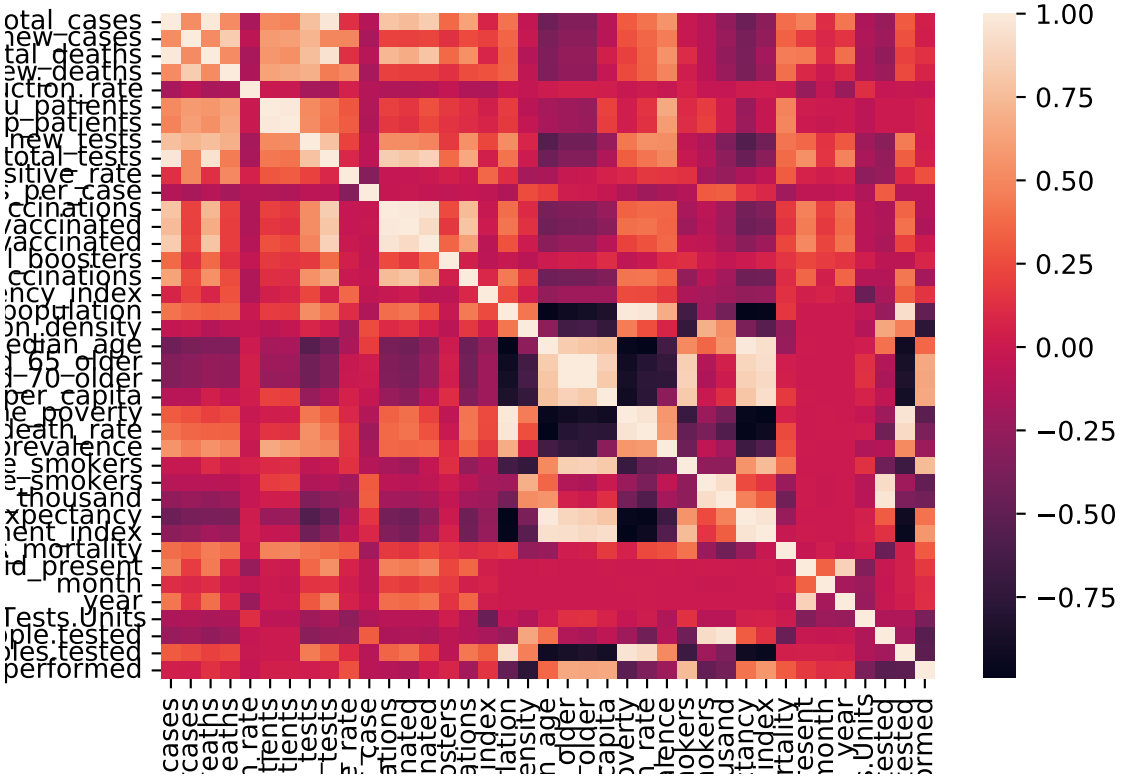


Figure 1: Covid Cases By Days Present



Initial analysis was performed on the data. From figure 1 we see that there is a nonliner relationship between days_covid_present and new_cases variable. We also notice that there are several patterns within the data. This indicates that this may be a mixture model or that separate models for each country will be a better representation than one total model. We will consider the multiple model approach. Further, we also investigated the data dispersion to see how well the models might fit. We found that for the new_cases variable the sample mean is 23945.7748361069, the sample variance is 2546551497.35433, and the sample standard deviation is 50463.37. We notice a large standard deviation and a variance greater than the mean. These point to overdispersion of the data, which makes sense as the growth is exponential and cyclical. Now, for the variables at hand we look to figure 2 to see how the pearson correlation between all variables looks. By the heatmap in figure 2, we see there are multiple variables that show a very strong correlation to each other. This indicates that there is collinearity between multiple factors in the variables and so we may need to further remove variables for modeling purposes. This makes sense since some variables are static by country, such as population over 65 or gdp. When one is present, it automatically means there are relationships between it and anything else that is static at that country. Thus, we will be careful to remove any colinear variables within the models.

Now, to model the count of Covid-19 cases, we look to the literature contained in the *Introduction to Linear Regression Analysis*, *Count Regression Models for COVID-19*, *USE OF POISSON REGRESSION MODELS IN ESTIMATING INCIDENCE RATES AND RATIOS* and *A Flexible Regression Model for Count Data*. We want to attempt 4 different models so we can pick the best one for this data.

The first choice is the multvariate linear regression model. This standard model assumes the errors are i.i.d Normal with $\epsilon_i \in N(0, \sigma^2)$ and the ϵ uncorrelated. It also assumes that linear relationship exists between the features and target variable modeled by

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon_i$$

Given these assumptions, we do not expect this model to work best. From *An Intermediate Course in*

Probability, *Introduction to Linear Regression Analysis*, and *A Flexible Regression Model for Count Data*, we note that count data is best modeled typically by a Poisson distribution. Further, these assumptions most likely do not hold based on our analysis. The data showed increasing trends over time that were non linear in nature and it's likely the errors are non-normal or uncorrelated. Now, we compute this regression model in order to confirm these hypothese. To compute this, we utilize R's lm() function. We first build the full model utilizing the cleaned up data from the python program, with the individual countries removed to make this model more general. We will consider rebuilding this with these back in if the model calls for it.The model output for this follows the model

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n$$

The output corresponding to this is below.

```
##                (Intercept)                 total_cases
##              -3.492408e+05               -1.561113e-03
##               total_deaths                   new_deaths
##              -2.631242e-02                3.622020e+01
##          reproduction_rate                icu_patients
##               7.163199e+03               -7.153758e-01
##              hosp_patients                   new_tests
##               6.157914e-01                2.473609e-02
##                total_tests               positive_rate
##               2.453243e-04                2.367475e+05
##             tests_per_case          total_vaccinations
##               1.232193e+01               -1.190409e-03
##          people_vaccinated     people_fully_vaccinated
##               1.054215e-03                1.189259e-03
##             total_boosters           new_vaccinations
##               -3.596015e-04               -1.504732e-03
##           stringency_index                  population
##              -7.779679e+01                1.851255e-05
##         population_density                  median_age
##              -4.663293e+02                3.675843e+04
##              aged_65_older               aged_70_older
##              -2.674192e+05                3.100806e+05
##             gdp_per_capita            extreme_poverty
##                         NA                          NA
##      cardiovasc_death_rate         diabetes_prevalence
##                         NA                          NA
##             female_smokers                male_smokers
##                         NA                          NA
## hospital_beds_per_thousand            life_expectancy
##                         NA                          NA
##    human_development_index           excess_mortality
##                         NA              -3.227508e+02
##         days_covid_present                       month
##              -2.041136e+01                1.507425e+03
##                        year         Missing.Tests.Units
##               3.728985e+03                2.052717e+04
##              people.tested              samples.tested
##               2.038524e+04                7.360169e+03
##            tests.performed
##                         NA
```

We notice that the R^2 value is not terrible at .8, indicating 80% of the total variance in the count of

covid cases is explained by this model. Looking at the variables, we first note that there are multiple NA coefficient estimates. These are due to the collinearity and thus are not included in the model. Further, there are multiple cases where there is no significance to the variable based on the t-test handled by R. Thus, there are multiple variables that may be unnecessary given the rest of the variables contained within the model. Now, to combat this we utilize the olsrr package ols_step_both_p that performs stepwise regression. In this, all variables previously entered into the model are evaluated for significance using the t-test and if the p-value is greater than the removal requirement, it is removed from the model, and if a new variables p-value is less than the entrance requirement, it is added. This continues until no new variables are removed or added and a final model with all relevant variables are chosen. Below is the output of this model. We see the nonsignificant and colinear variables have now been removed. We will use this as our final linear model for evaluation.

```
##                (Intercept)                 new_deaths
##              -2.942160e+04               3.469212e+01
##              hosp_patients                  new_tests
##               4.175459e-01               2.636739e-02
##              positive_rate           excess_mortality
##               2.290939e+05              -3.223922e+02
##          people_vaccinated                total_tests
##               9.920728e-04               2.372005e-04
##               total_deaths                      month
##              -1.224608e-01               1.215455e+03
##        Missing.Tests.Units             total_boosters
##               1.998257e+04              -5.724461e-04
##          reproduction_rate    people_fully_vaccinated
##               6.792313e+03               1.182290e-03
##             tests_per_case         total_vaccinations
##               1.303055e+01              -1.150766e-03
##             female_smokers              aged_65_older
##              -2.506732e+03               2.991364e+03
## hospital_beds_per_thousand             people.tested
##              -3.001303e+03               1.975503e+04
##         days_covid_present           new_vaccinations
##              -1.017915e+01              -1.409566e-03
##      cardiovasc_death_rate
##              -5.752855e+01
```

The second choice is the poisson regression model. This standard model assumes the errors are not i.i.d Normal, but that the model can be made linear utilizing a link function. It also assumes that mean and variance are equivalent. From the analysis, we note that the assumption of equivalent variance and expectation is violated. From *A Flexible Regression Model for Count Data* and *Count Regression Models for COVID-19*, we note that overdispersed data will be best served by COM-Poisson or Negative Binomial. To compute this model, we utilize R's glm() function with a log-linked poisson family. We first build the full model. The coefficient output for this is

$$\hat{\mu_i} = \hat{y_i} = e^{\hat{\beta_0} + \hat{\beta_1} x_1 + \cdots + \hat{\beta_n} x_n}$$

The output corresponding to this is below.

```
##                (Intercept)                 total_cases
##               3.182007e+02               -1.604180e-07
##               total_deaths                   new_deaths
##               1.059724e-05                1.340473e-04
##          reproduction_rate                icu_patients
##               1.256551e-01                3.966790e-05
```

```
##              hosp_patients                 new_tests
##              -5.465880e-06              4.430204e-07
##                total_tests              positive_rate
##              -2.322322e-09              9.248804e+00
##             tests_per_case         total_vaccinations
##              -1.130161e-02              2.205185e-08
##           people_vaccinated    people_fully_vaccinated
##              -2.070312e-08             -2.712665e-08
##             total_boosters           new_vaccinations
##              -1.594750e-08              1.886416e-08
##            stringency_index                 population
##               8.056023e-03             -8.027796e-08
##         population_density                 median_age
##               6.814446e-02             -1.058743e+01
##              aged_65_older              aged_70_older
##               3.450861e+01             -4.247778e+01
##             gdp_per_capita           extreme_poverty
##                         NA                        NA
##      cardiovasc_death_rate        diabetes_prevalence
##                         NA                        NA
##             female_smokers              male_smokers
##                         NA                        NA
## hospital_beds_per_thousand           life_expectancy
##                         NA                        NA
##    human_development_index           excess_mortality
##                         NA             -5.414707e-04
##         days_covid_present                     month
##               6.991085e-03              8.151396e-03
##                       year        Missing.Tests.Units
##              -2.193905e-01             -5.267167e-02
##              people.tested             samples.tested
##              -1.247894e+00              7.500025e+00
##            tests.performed
##                         NA
```

We notice that the null deviance is significantly larger than the residual deviance, suggesting good model fit. We again perform stepwise regression, but this time using the Akaike Information Criterion(AIC) and the stepAIC() function from R. This works as before but replaces the t-test with the lowering of the AIC, which represents the estimated prediction error. The output model is below.

```
##             (Intercept)              total_cases             total_deaths
##            3.182007e+02            -1.604180e-07             1.059724e-05
##              new_deaths        reproduction_rate             icu_patients
##            1.340473e-04             1.256551e-01             3.966790e-05
##            hosp_patients                new_tests              total_tests
##            -5.465880e-06             4.430204e-07            -2.322322e-09
##            positive_rate           tests_per_case       total_vaccinations
##             9.248804e+00            -1.130161e-02             2.205185e-08
##       people_vaccinated people_fully_vaccinated           total_boosters
##            -2.070312e-08            -2.712665e-08            -1.594750e-08
##         new_vaccinations         stringency_index                population
##             1.886416e-08             8.056023e-03            -8.027796e-08
##       population_density               median_age            aged_65_older
##             6.814446e-02            -1.058743e+01             3.450861e+01
```

```
##           aged_70_older        excess_mortality      days_covid_present
##           -4.247778e+01           -5.414707e-04            6.991085e-03
##                   month                    year     Missing.Tests.Units
##            8.151396e-03           -2.193905e-01           -5.267167e-02
##           people.tested          samples.tested
##           -1.247894e+00            7.500025e+00
```

Our third choice is the negative binomial regression model. This model is similar to the poissoin, but it does not assume that mean and variance are equivalent and allows for under/overdispersion. From the analysis, we note that the assumption of equivalent variance and expectation is violated so this should fit well. By https://data.library.virginia.edu/getting-started-with-negative-binomial-regression-modeling/ we note that we can relate estimate the theta parameter usingthe sample variance and means. By the site we have that $\sigma^2 = \mu + \mu^2/\theta$ so we use the sample estimates to create an estimate of theta as $\hat{\theta} = \bar{X}^2/(S_x^2 - \bar{X})$. To compute this model, we utilize R's glm() function with a log-linked negative binomial family and the estimated theta of 0.2252. We first build the full model. The coefficient output for this is

$$\hat{\mu_i} = \exp(\ln(\theta_i) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

The output corresponding to this is below.

```
##                (Intercept)                 total_cases
##                8.060446e+01               -3.941520e-07
##                total_deaths                   new_deaths
##                1.790390e-05                4.104398e-04
##          reproduction_rate                 icu_patients
##                7.088203e-01               -5.292848e-05
##              hosp_patients                    new_tests
##                9.434712e-06                5.235801e-07
##                 total_tests                positive_rate
##                1.147834e-08                1.167926e+01
##             tests_per_case           total_vaccinations
##                -3.622745e-03               -9.753241e-09
##           people_vaccinated    people_fully_vaccinated
##                9.341503e-09                3.904752e-09
##             total_boosters            new_vaccinations
##                -9.820679e-09               -1.094509e-08
##           stringency_index                  population
##                5.732640e-02               -2.045191e-08
##         population_density                  median_age
##                2.607892e-02               -3.326065e+00
##               aged_65_older               aged_70_older
##                1.462086e+01               -1.764921e+01
##              gdp_per_capita            extreme_poverty
##                          NA                          NA
##      cardiovasc_death_rate        diabetes_prevalence
##                          NA                          NA
##             female_smokers                male_smokers
##                          NA                          NA
## hospital_beds_per_thousand             life_expectancy
##                          NA                          NA
##    human_development_index            excess_mortality
##                          NA                9.526547e-03
##         days_covid_present                       month
##                3.737006e-03                1.093099e-01
```

```
##                        year        Missing.Tests.Units
##               2.240954e-01               6.607808e-01
##              people.tested             samples.tested
##              -4.892933e-01               4.266760e+00
##            tests.performed
##                         NA
```

We notice that the null deviance is significantly larger than the residual deviance, suggesting good model fit. Similar issues arrise in the model as did in the linear model. We again perform stepwise regression using the Akaike Information Criterion(AIC) and the stepAIC() function from R. T The output model is below.

```
##         (Intercept)         total_cases        total_deaths          new_deaths
##        8.043165e+01       -3.917537e-07        1.785271e-05        4.100819e-04
##   reproduction_rate        icu_patients       hosp_patients           new_tests
##        7.068048e-01       -5.146709e-05        9.198674e-06        5.197070e-07
##         total_tests       positive_rate      tests_per_case  total_vaccinations
##        1.132679e-08        1.164642e+01       -3.621349e-03       -5.570678e-09
##   people_vaccinated      total_boosters    stringency_index          population
##        5.007150e-09       -1.347397e-08        5.700284e-02       -2.032606e-08
##  population_density          median_age       aged_65_older       aged_70_older
##        2.599791e-02       -3.313092e+00        1.457705e+01       -1.758979e+01
##    excess_mortality  days_covid_present               month Missing.Tests.Units
##        9.576609e-03        4.378743e-03        9.081553e-02        6.488333e-01
##       people.tested      samples.tested
##       -4.910258e-01        4.268953e+00
```

Our final choice is the Conway-Maxwell Poisson (COM-Poisson) regression model coming from the *A Flexible Regression Model for Count Data* paper. This model is similar to the poisson, but it does not assume that mean and variance are equivalent, allowing for under/overdispersion, and it accounts for this utilizing a dispersion parameter ν and a normalization factor. From the analysis, we know that the assumption of equivalent variance and expectation is violated so this should fit well. To compute this model, we utilize R's glm() function with a COM-Poisson family, from the spaMM package, and the estimated ν of 0.7. We first build the full model. The coefficient output for this is

$$\hat{y_i}|x_i = \hat{\lambda}^{a/\hat{\nu}} - \frac{\hat{\nu}-1}{2\hat{\nu}}, \hat{\lambda} = \exp(x_i'\beta)$$

The output corresponding to this is below.

```
## Warning in .COMP_maxn(lambda, nu): maxn truncated to 10000 for (lambda,nu)=
## (2.22044604925031e-16,0.7) and possibly other values.
```

```
##                (Intercept)                 total_cases
##               2.236580e+02               -1.122936e-07
##               total_deaths                   new_deaths
##               7.418396e-06                9.382681e-05
##          reproduction_rate                icu_patients
##               8.795495e-02                2.776966e-05
##              hosp_patients                   new_tests
##              -3.826414e-06                3.101098e-07
##                total_tests                positive_rate
##              -1.626447e-09                6.474220e+00
##             tests_per_case          total_vaccinations
##              -7.912082e-03                1.543821e-08
```

```
##          people_vaccinated    people_fully_vaccinated
##              -1.449383e-08              -1.899068e-08
##             total_boosters           new_vaccinations
##              -1.116357e-08               1.320657e-08
##           stringency_index                 population
##               5.639521e-03              -5.646977e-08
##         population_density                 median_age
##               4.807145e-02              -7.454772e+00
##              aged_65_older              aged_70_older
##               2.437535e+01              -2.999657e+01
##             gdp_per_capita           extreme_poverty
##                         NA                        NA
##      cardiovasc_death_rate       diabetes_prevalence
##                         NA                        NA
##             female_smokers               male_smokers
##                         NA                        NA
## hospital_beds_per_thousand           life_expectancy
##                         NA                        NA
##    human_development_index           excess_mortality
##                         NA              -3.789545e-04
##         days_covid_present                     month
##               4.894319e-03               5.706037e-03
##                        year       Missing.Tests.Units
##              -1.535661e-01              -3.692830e-02
##              people.tested             samples.tested
##              -8.738105e-01               5.359003e+00
##            tests.performed
##                         NA
```

We notice that the null deviance is significantly larger than the residual deviance, suggesting good model fit. Similar issues arrise in the model as did in the linear model. We again perform stepwise regression using the Akaike Information Criterion(AIC) and the stepAIC() function from R. Due to compute power the stepwise did not work so the variables from previous models were used to compute this. The output model is below.

```
##         (Intercept)         total_cases        total_deaths          new_deaths
##        8.043165e+01       -3.917537e-07        1.785271e-05        4.100819e-04
##   reproduction_rate        icu_patients       hosp_patients           new_tests
##        7.068048e-01       -5.146709e-05        9.198674e-06        5.197070e-07
##         total_tests       positive_rate      tests_per_case  total_vaccinations
##        1.132679e-08        1.164642e+01       -3.621349e-03       -5.570678e-09
##   people_vaccinated      total_boosters    stringency_index          population
##        5.007150e-09       -1.347397e-08        5.700284e-02       -2.032606e-08
##  population_density          median_age       aged_65_older       aged_70_older
##        2.599791e-02       -3.313092e+00        1.457705e+01       -1.758979e+01
##    excess_mortality  days_covid_present               month Missing.Tests.Units
##        9.576609e-03        4.378743e-03        9.081553e-02        6.488333e-01
##       people.tested      samples.tested
##       -4.910258e-01        4.268953e+00
```

Now, we must asses these models for adequacy. Note, all plots were placed in the appendix to conserve discussion space. We performed regression analysis on each of the models. For the linear model, this involved utilizing the residuals to compute a normal q-q plot and to plot the residuals vs. predicted response. For the generalized linear models, the residuals is replaced with the deviance residuals as per the *Introduction*

to Linear Regression Analysis. Further, for the COM-Poisson model, Sellers and Shmueli reccommend using bootstrapped deviance residuals for the q-q plot pulling from the empiric distribution. This was completed utilizing the boot package in r and running the diagnostic plots. We see that for the linear model, generally it fits well for larger predicted responses, but struggles with the smaller responses. Further, the normal assumption is slightly violated with a heavy left and right tail. A transformation could improve the fit of this model. Overall, this doesn't fit the data great but it does fit well enough. Looking at our glm models, we see that the Poisson and COM-Poisson have the best fits according to our adequacy checks according to the q-q plot. The negative binomial has the best residual vs. predicted response plot indicating it has consistent variance unlike the other 2 models. Further, from the COM-Poisson and Poisson diagnosis plots we see that the variance is not consistent and varies by the outward facing cone shape in the predicted vs. residual plot. This makes sense as the pandemic has changed as new variants have come about through mutations and lockdowns have come and gone in some places, but returned in others. Thus, a better fit for COM-Poisson may be achieved threw a model that can handle non-consistent error and variance. Overall, our chosen models do adequately fit our data.

Finally, to compare our models we utilize the mean squared error (MSE), the AIC of the model, and the predicted outputs vs. days covid present plots in the appendix. the output below shows our evaluation criteria.

```
## [1] "Linear Model Evaluation: "

## [1] "AIC = "           "90728.0174668775" "MSE ="            "498181131.540502"

## [1] "Poisson Model Evaluation: "

## [1] "AIC = "           "12003694.3102554" "MSE = "           "496343019746.511"

## [1] "Negative Binomial Model Evaluation: "

## [1] "AIC = "           "74992.6715061154" "MSE = "           "1.79455867877802"

## [1] "COM-Poisson Model Evaluation: "

## [1] "AIC = "          "8415769.2658995" "MSE = "          "14865437.094045"
```

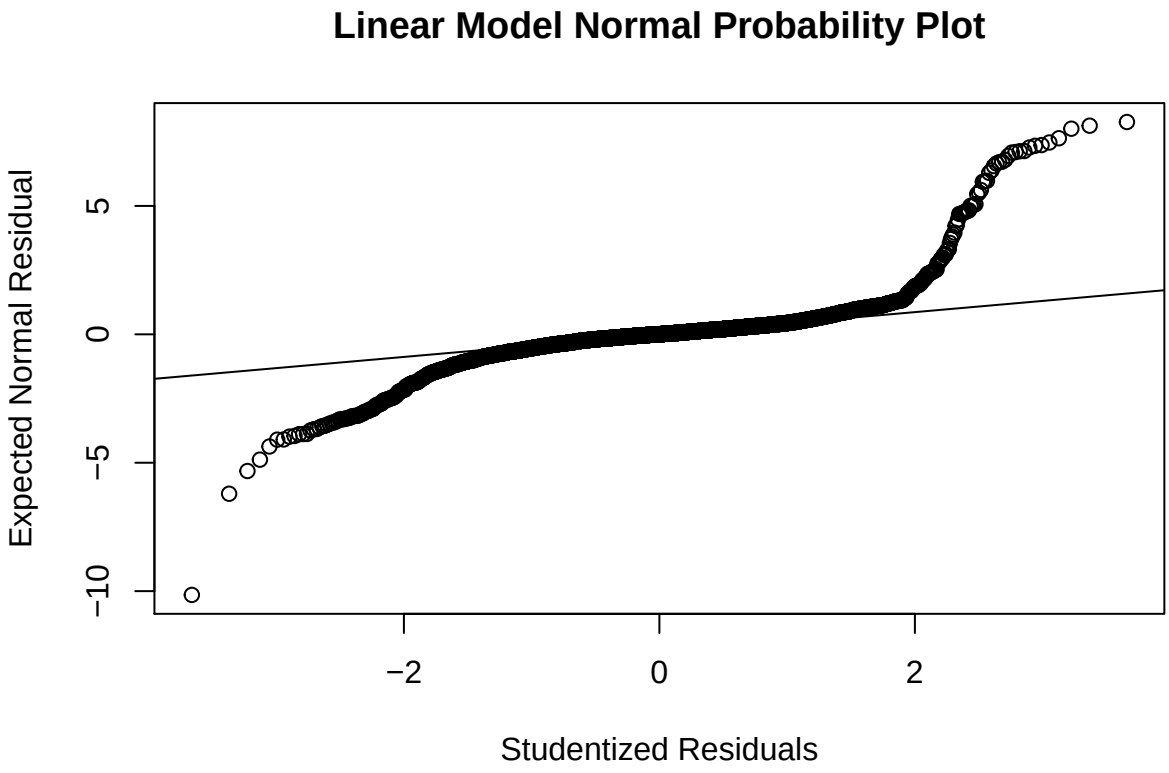
We see that the AIC of the Negative Binomial GLM is the lowest, followed by the Linear Model, COM-Poisson, and then Poisson GLM. This indicates that the Negative Binomial model will have the lowest prediction error compared to the rest of the models. However, the interpretability of the linear model does make a case for its use as well. Further, the Negative Binomial model has significantly lower mean squared error when compared to the rest of the models. Turning to the plots in the appendix of the predicted vs. days since covid present, we first note that red represents the predicted values and black the true values. We see that partly why the negative binomial fits so well is because it is mainly fit for the countries that have little volume. It misses the major spikes in the model. However, it seems this works well for the majority of the countries. Looking at the other plots, we see that the COM-Poisson model did not fit this data very well. The plot shows points scattered about and not much fitting to the data. Now, we see for Poisson and Linear that these do fit the data well but the spikes are located at different points than expected. This is likely due to other features interaction and so we can account it for that. These also account for the large spikes, unlike the negative binomial, making them decent candidate models. Taking these all into account, based on the study done the model choice that best generalizes from our data is the Negative Binomial GLM.

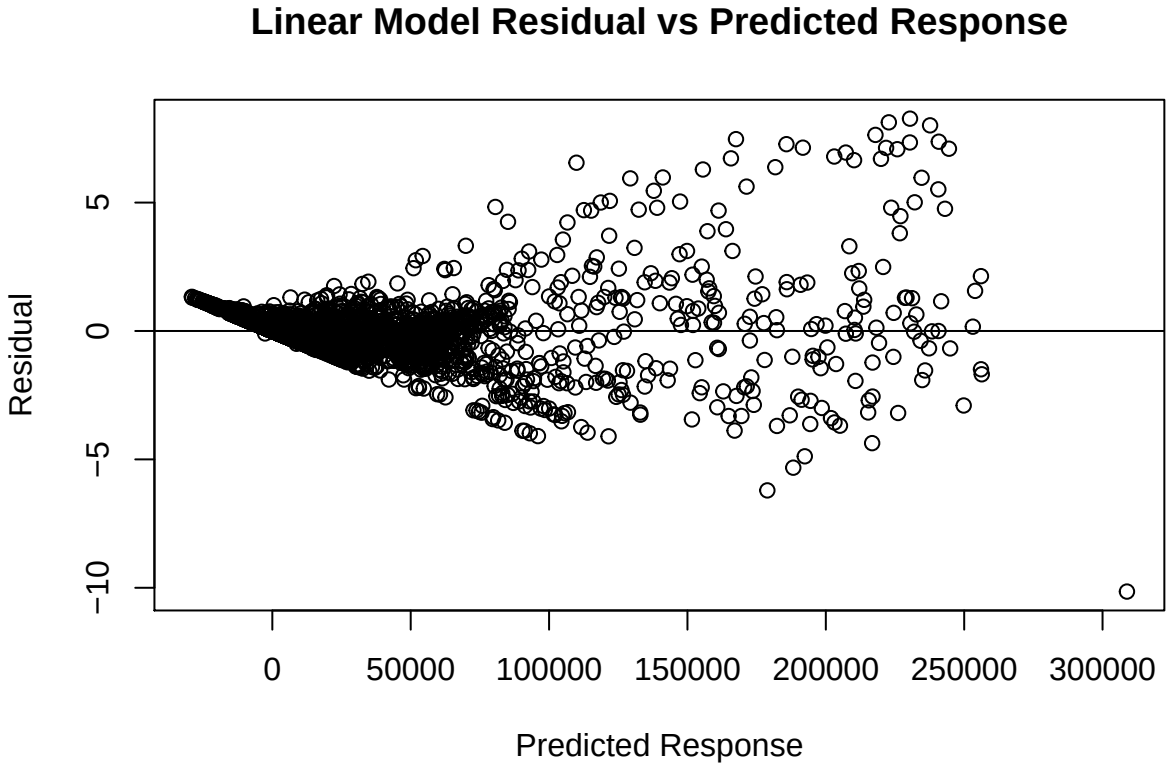
Through this study, we have found that for the count of daily new covid cases in the 6 countries studied, that the Negative Binomial GLM is the model that best fits our count data. It produces the

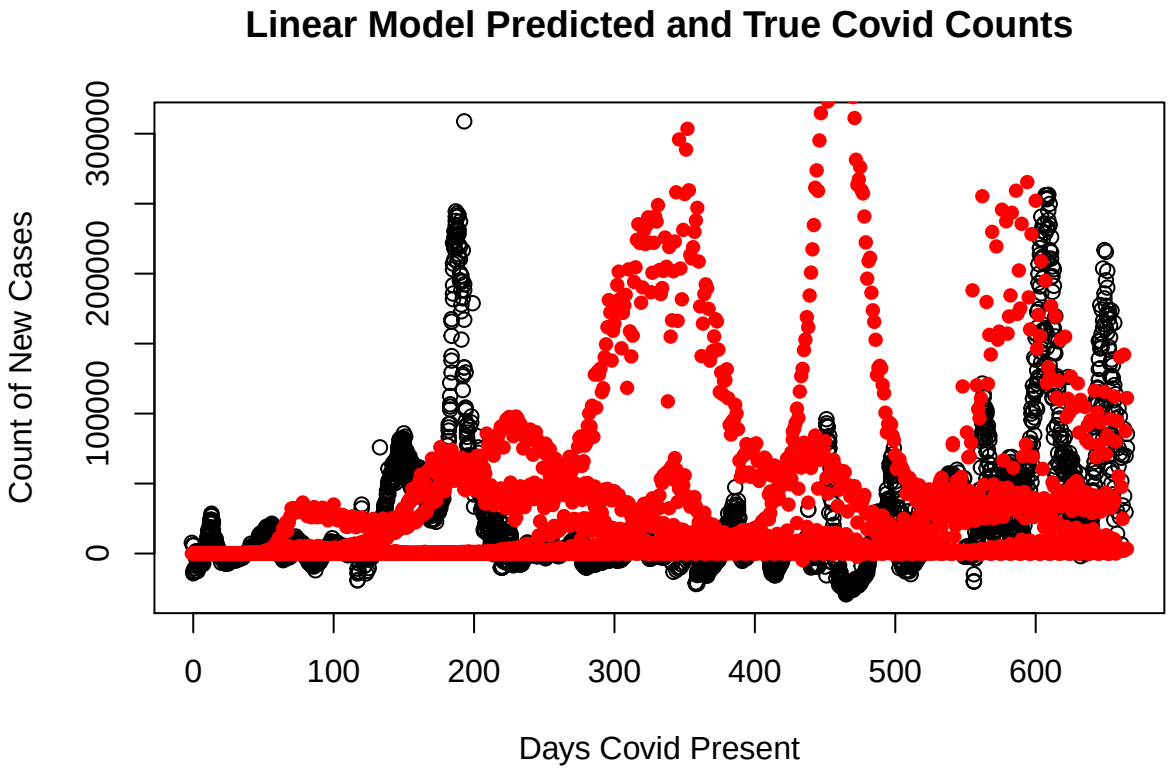
least error estimate and has a good fit to the less extreme counts. For more extreme counts of cases, the Poisson and Linear models seemed to have the best fits and were able to account for the large spikes unlike the Negative Binomial. Further, the interpretability of the linear model makes for a decent second choice. Finding relationships in the coefficients allows for easier study of the change in counts that certain new developments have had over time. Looking at the coefficients of the linear model, we can say that vaccination and boosters have a negative relationship with the count of cases, indicating that for each new patient vaccinated we expect a decrease in the number of total new cases each day. Utilizing these modeling techniques, we can learn more about a new and rapidly changing pandemic. Future work on these models, and potentially including more epidemiological models or branching processes, and greater fit can unlock more knowledge about a novel virus, aiding those fighting it on the front line.

Appendix - Model Evaluation

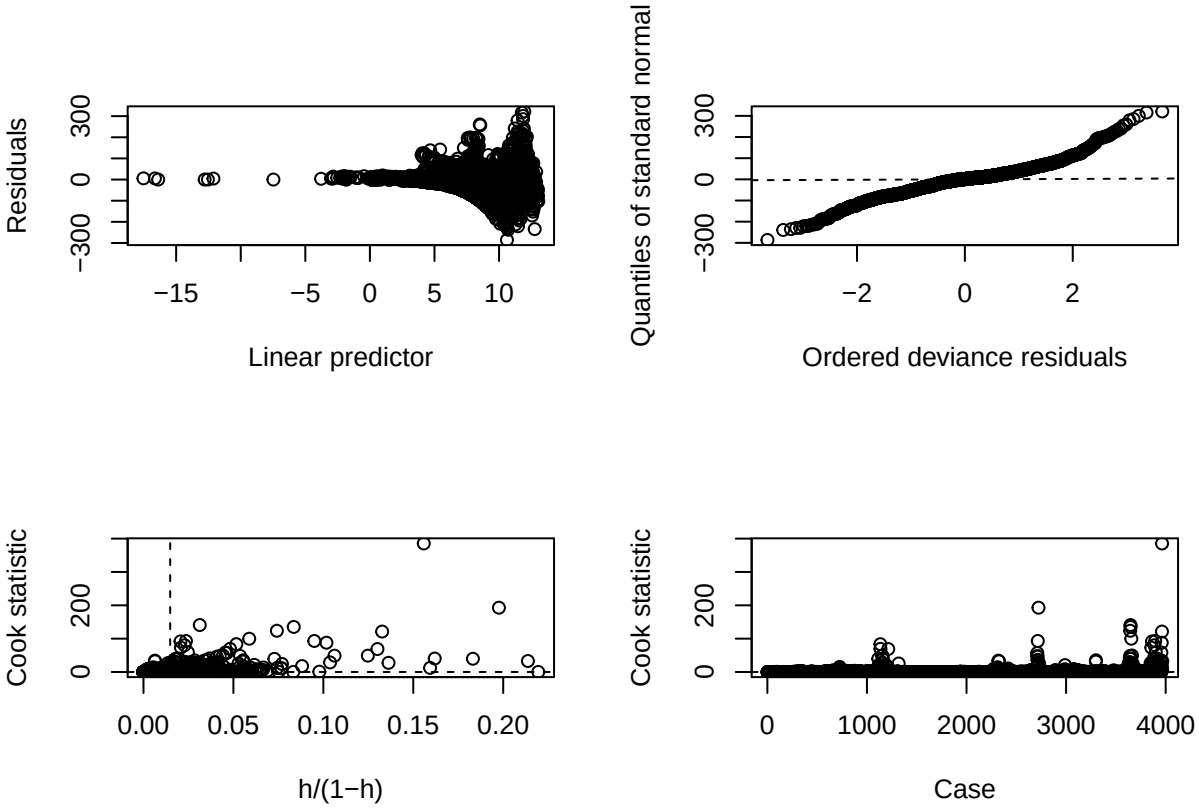
Linear Model Evaluation

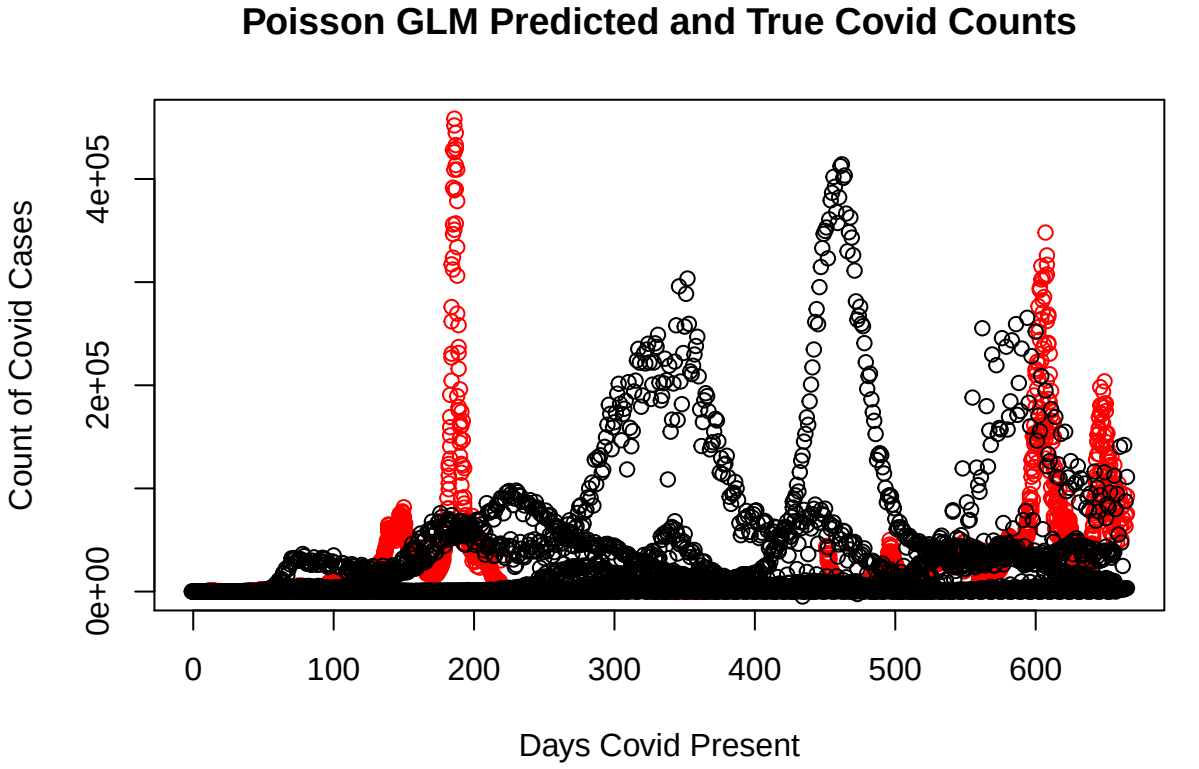




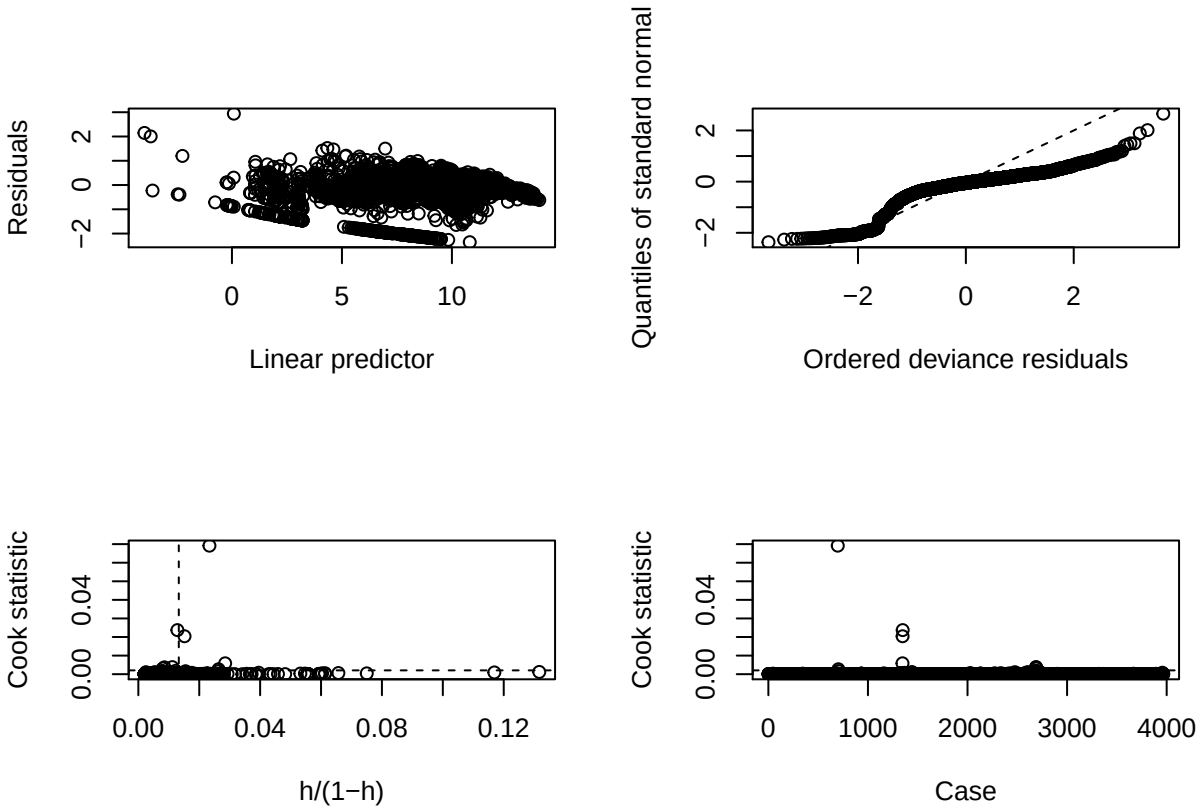

Linear Model Predicted and True Covid Counts
Count of New Cases
0
100000
200000
300000
0
100
200
300
400
500
600
Days Covid Present

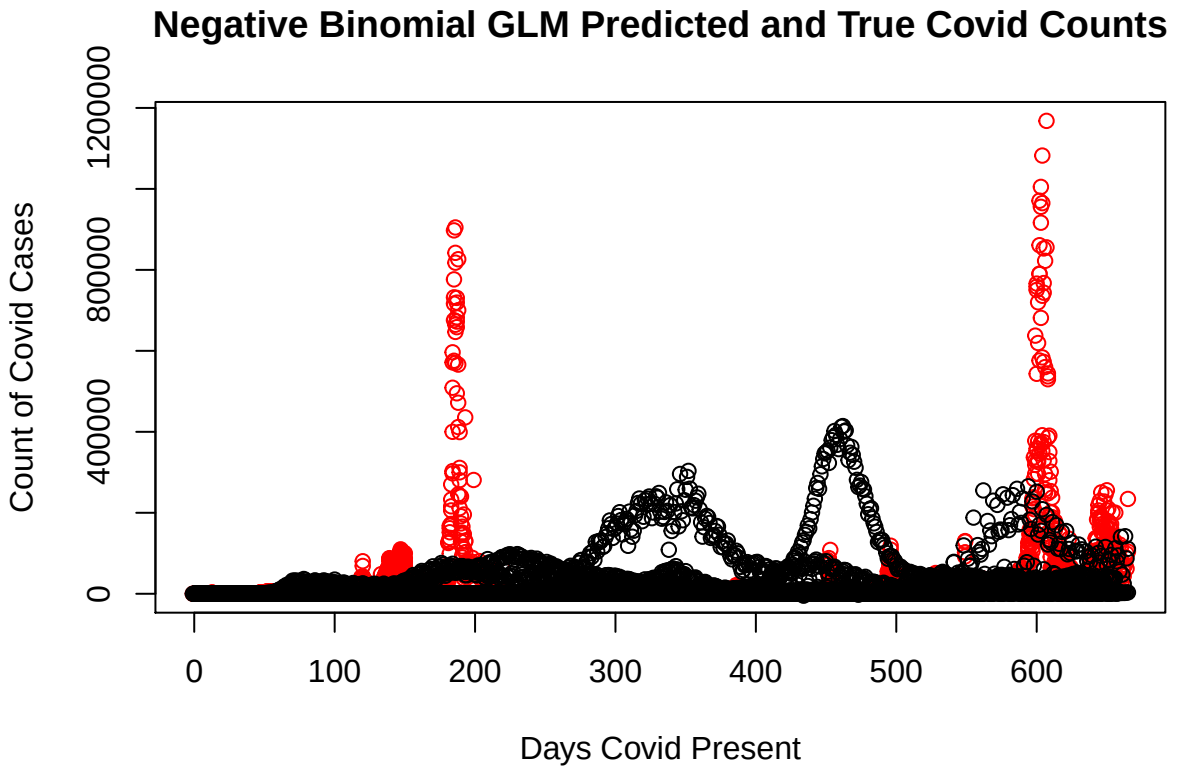
Poisson Model Evaluation



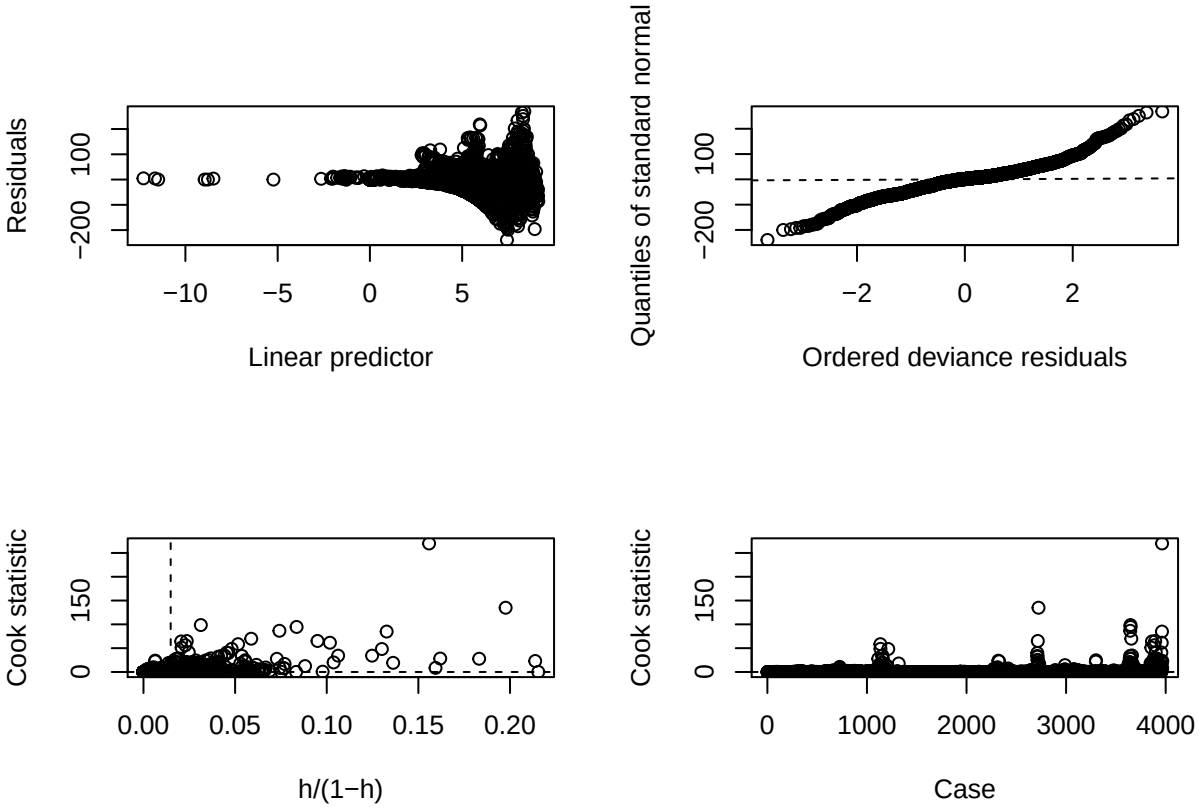


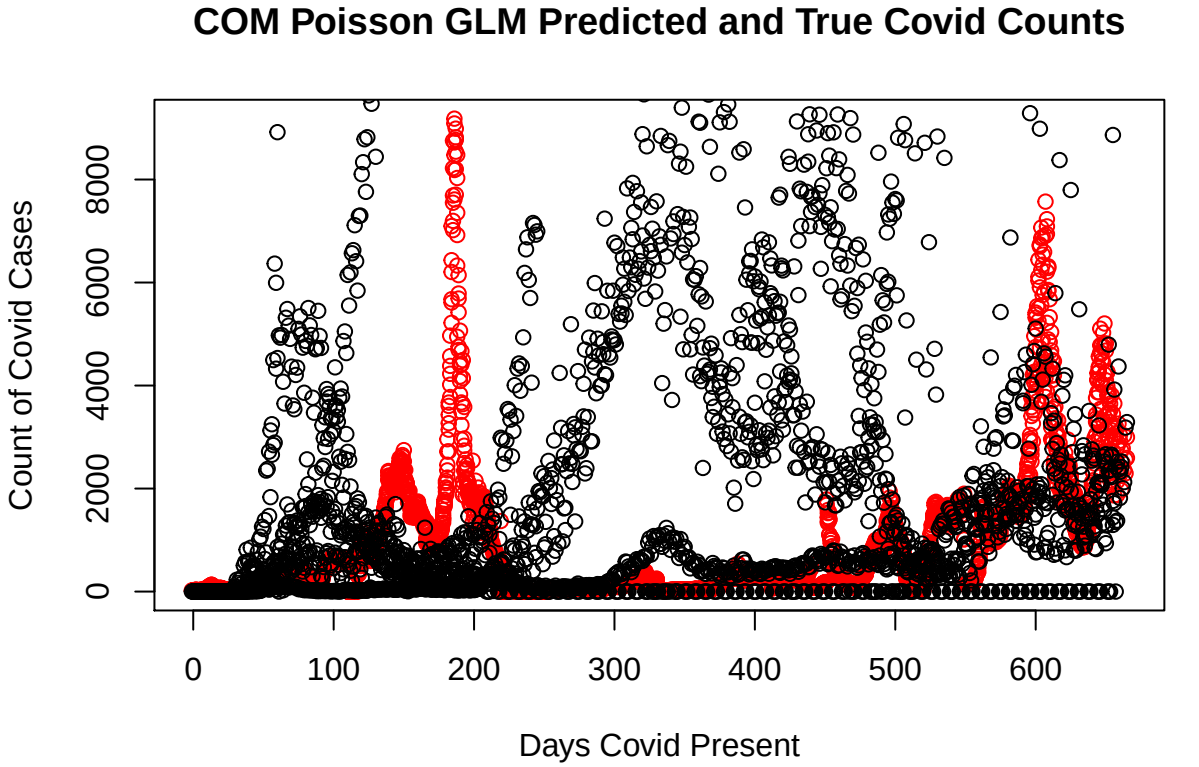
Negative Binomial Model Evaluation





COM-Poisson Model Evaluation





Bibliography

1. Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020)\ - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]\

2. David Inouye, Eunho Yang, Genevera Allen, and Pradeep Ravikumar (2017) – A review of multivariate distributions for count data derived from the Poisson distribution\

3. https://www.rdocumentation.org/packages/boot/versions/1.3-28/topics/glm.diag.plots//

4. Douglas Montgomery, Elizabeth Peck, G. Geoffrey Vining (2012) - Introduction to Linear Regression analysis, 5th edition\

5. Chan, Stephen et al. "Count regression models for COVID-19." Physica A vol. 563 (2021): 125460. doi:10.1016/j.physa.2020.125460//

6. https://data.library.virginia.edu/getting-started-with-negative-binomial-regression-modeling/.//

7. EDWARD L. FROME, HARVEY CHECKOWAY, USE OF POISSON REGRESSION MODELS IN ESTIMATING INCIDENCE RATES AND RATIOS, American Journal of Epidemiology, Volume 121,Issue2,February1985,Pages309-323, https://doi.org/10.1093/oxfordjournals.aje.a114001//

8. Allan Gut (2009) - An Intermediate Course in Probability\

9. Karlis, Dimitris. "Multivariate Poisson Models." Athens University of Economics, 2002, http://www2.stat-athens.aueb.gr/~karlis/multivariate%20Poisson%20models.pdf.//

10. Inouye, David et al. "A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution." Wiley interdisciplinary reviews. Computational statistics vol. 9,3 (2017): e1398. doi: 10.1002/wics.1398//