Linear Regression Analysis to Predict COVID-19

Amy Nguyen, Saroop Samra, Arely Vasquez

1. Introduction

The COVID-19 pandemic was declared by the WHO on March 11th 2020[1]. In this paper develop a set of linear models of the number of deaths in week 5 as a function of the other variables. Using this model we will predict point estimates and prediction intervals as well as making inferences. Research in statistical methods to understand the underlying impact of COVID-19 is critical to prepare healthcare professionals and reduce loss of life.

The summary of our analysis is a model that used all the parameters in a using multiple linear regression gave the highest correlation factor of 0.98. However, simplifying the model rather than using all the parameters yields the same predictive power due to the issue of multicollinearity. Additionally, using a log transform is recommended to reduce the range of the prediction interval as well as giving real world interval values (which are not negative values that are meaningless). Finally, backward-elimination was able to improve our model's adjusted R^2 score by eliminating total cases, recovery cases , and week 4 deaths. However, after performing Lasso and Ridge regression, our best model is shown to have only removed total cases and week 4 deaths.

2. Data

The data is from 16 countries and have 6 parameters including the number of cases and their type, the case fertility rate (CFR) as well as the week 4 and 5 deaths. The data will be used to predict the week 5 deaths using a combination of the other parameters. Limitations of the data include issues of a lack of the relative amount of data, at only 16 observations, the impact of outliers can be significant on the model. Another limitation is the observations are from a broad range of countries with different climate, geographies and economic conditions that can impact the predictive power for say India, a developing country in a hot climate.

3. Analysis

3.1 Multiple Regression Model: Default (No Transforms)

3.1.1 Descriptive Statistics

In Table 1 shown below, the statistics shows that throughout all of the nations we have a minimum of 69 cases in Indonesia, a maximum number of 74185 cases in China, and a mean number of 8897 cases throughout all nations. With that being said, we can see a very large jump from the number of deaths in week 4 to the number of deaths in week 5. We can numerically see the jump of maximum deaths of 2004 in week 4, to 4825 deaths in week 5. Although we also see a minimal rate of increase in the minimum number of deaths in week 4 of 0 to 6 deaths in week 5. With that being said, we can see that some countries are having fast growing rates of deaths, while some are having a stable rate.

	Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Week 5 Deaths
min	69	60	38	0	0	6
max	74185	57805	65112	2004	6.81	4825
range	74116	57745	65074	2004	6.81	4819
median	2183	2126	495	11	1.38	106
mean	8897.33	7216.27	6936.63	288.93	2.14	757.2
std.dev	18933.38	14731.6	16561.31	605.66	2.08	1353.01

[1] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html

kurtosis	9.46	9.31	10.17	5.19	2.68	5.79
skewness	2.71	2.68	2.85	1.91	0.98	1.94

Table 1: Descriptive Statistics

3.1.2 EDA Visualizations

In Figure 1, the box plots show a large range for each value. For Total Cases, China is an extreme outlier having the max number of total cases of 74185. China also presents the largest outlier in active cases, recovery cases, week 4 deaths, and week 5 deaths. Italy is another nation that is an outlier, but only for week 4 deaths and week 5 deaths. Although we can see that China had a proportional amount of deaths according to their number of total cases, whereas Italy had much larger death rates compared to their number of total cases. With that we can see that not only China and Italy were the two nations that were hit most by Covid-19, but Italy had a dramatically larger death rate with a much higher case fatality rates (CFR) of 6.811, with iran following not so far behind with the second largest CFR of 4.523.

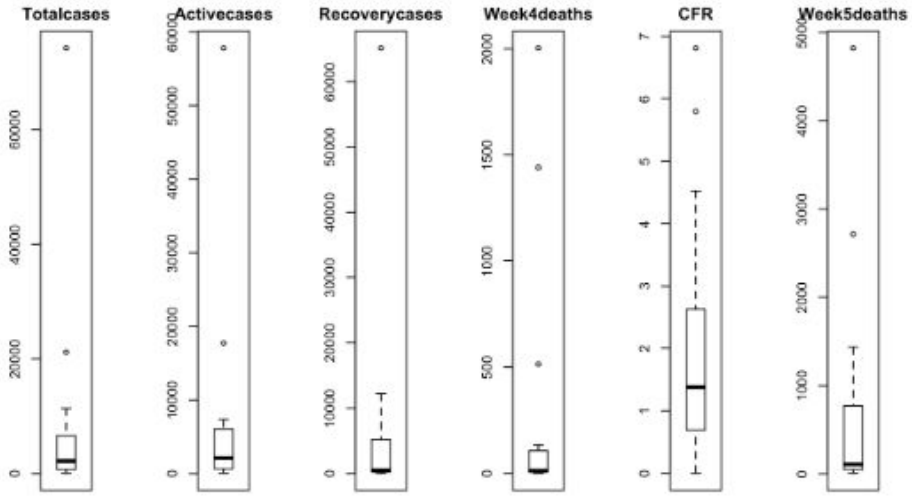

Figure 1: Box-Plot of parameters

In Figure 2 we can see below that our histograms are all right skewed with the exception of CFR. This reiterates the fact that our means in those right skewed graphs are larger than the medians. This positive skewness shows how a majority of the values are all focused on smaller values. On the other hand, although case fatality rates (CFR) is also right skewed, it has a variation of a bimodal distribution as you can see two peaks of about 0 and 3. From these histograms, we can see the presence of outliers in each value that are far from the rest of the data values. These low frequency values you can see in the histograms are from China and Italy as previously stated.

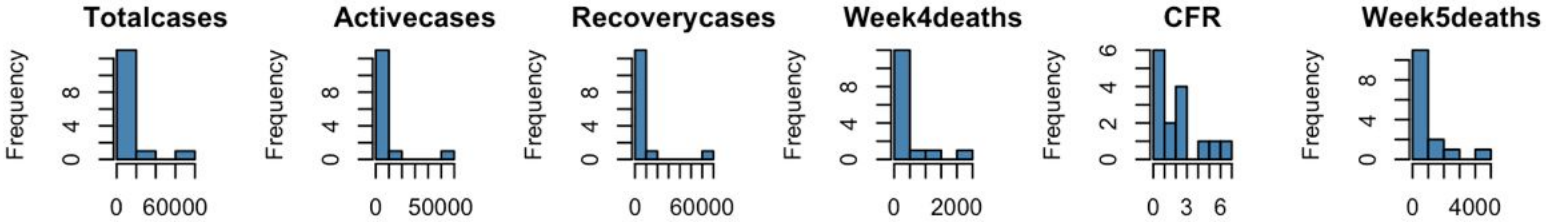

Figure 2: Histogram of parameters

In Figure 3, the QQ plots expose the lack of correlation with the normal distribution with outliers. As you can see, the outliers on the far right cause the data points to skew far from the line. These outliers, China being the furthest right data point in all the plots except got CFR, and Italy Iran and China being the three furthest outliers in CFR. With that being said, it is easy to see how a majority of the countries follow the normal distribution, but China and Italy are two large outliers and influences on the linear regression line. Although those data points are necessary and vital to the Covid-19 studies and predictions, they make it harder to predict using linear regression.

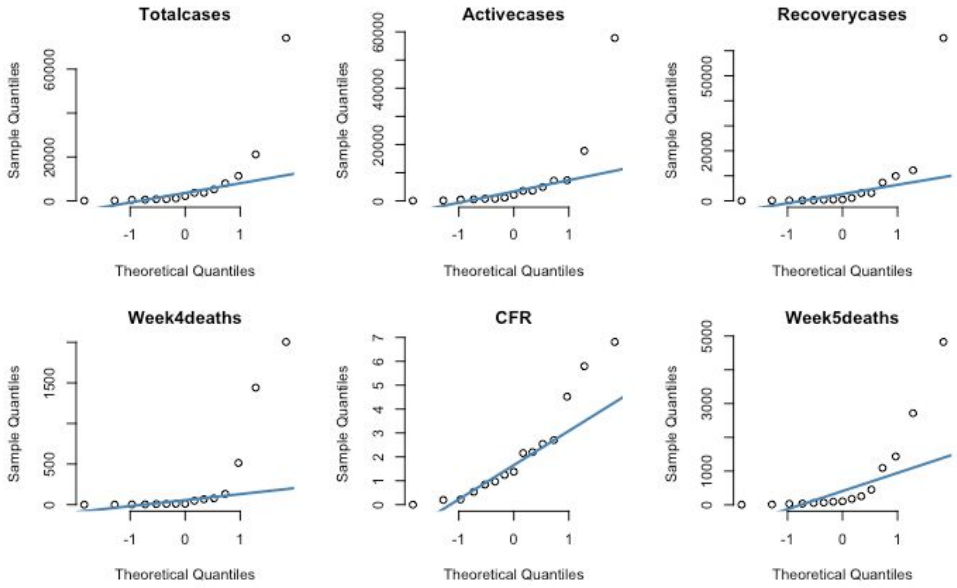

Figure 3: QQ Plot of parameters

Our first model we will use all 5 parameters to predict the week 5 deaths using multiple linear regression.

3.1.3 Diagnostics

In Figure 4, the Scatter Plots show the relationship between the number of deaths in week 5 and the other variables. From these plots we can see outliers in all the graphs. The outliers are also influential points that represent China and Italy. China is the outlier on the furthest right, and Italy is the data point second furthest to the right. These two countries have a high leverage on the linear regression line. In the CFR plot, Italy is the major outlier on the CFR plot that is also high leverage. If those influential points (countries) were to be removed, the linear regression line would change drastically. The trends are not so clearly linear.

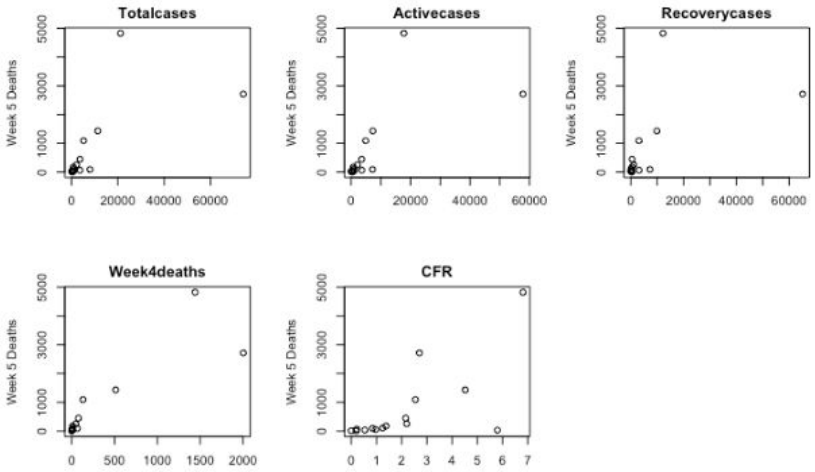

Figure 4: Scatter Plot Weekly 5 Deaths against other parameters

Finally we diagnose the normality of the residuals using a QQ Plot and classic residual plot as shown in Figure 5 and 6 respectively. The plots show outliers as well as a variability of regressions and do not appear random.

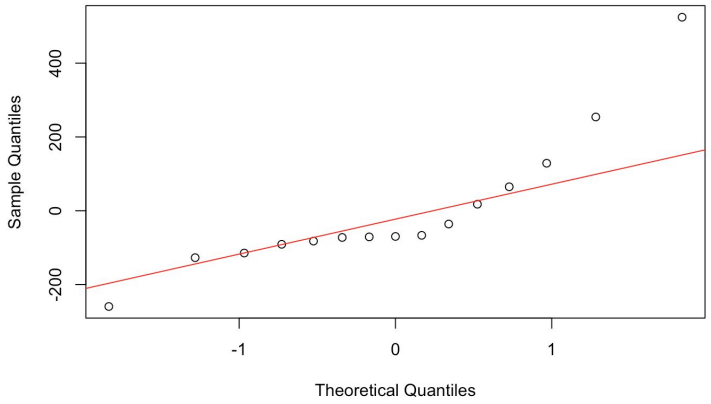

Figure 5: QQ Plot of Residuals

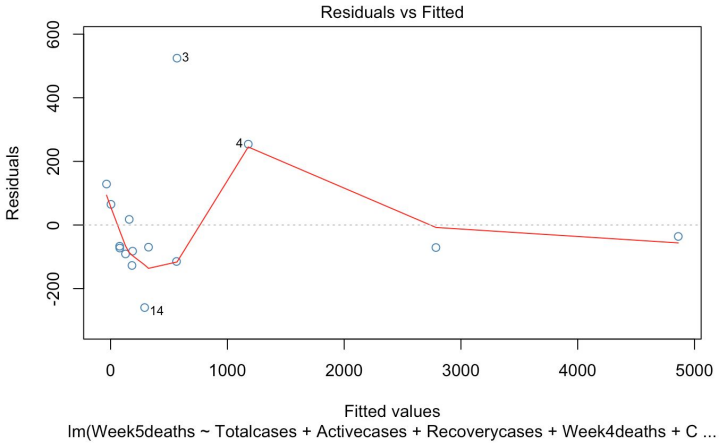

Figure 6: Residual Plot

As there are a few data points we perform a Breusch-Pagan Test (*Advanced Analysis*) to establish if Heteroskedasticity exists. The p-value for the test is 0.7028, indicating that we cannot reject the Null hypothesis that the data follows a Chi-Square distribution with the parameters.

3.1.3 Inference

We report the summary of the multiple linear regression with the full 5 feature parameters in Table 2. The R^2 (goodness of fit) and adjusted R^2 values of 0.9807 and 0.9701 with a F-statistic of 91.7 with a p-value of $1.9\ 10^{-7}$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	84.4247	115.0089	0.7341	0.4816
Totalcases	-0.0700	0.2182	-0.3208	0.7557
Activecases	0.1216	0.1554	0.7823	0.4541
Recoverycases	-0.0957	0.1097	-0.8728	0.4055
Week4deaths	3.4975	0.7039	4.9686	0.0008
CFR	33.5133	46.3383	0.7232	0.4879

Table 2: Inference summary of model with 5 feature parameters

The interpretation of coefficients is as follows: 84.4247 intercept is the value of week 5 deaths when all feature parameters are zero. The week 5 deaths increase by -0.0700 when one additional Total case is reported and all other feature parameters are constant. The week 5 deaths increase by 0.1216 when one additional Active case is reported and all other feature parameters are constant. The week 5 deaths increase by -0.0957 when one additional Recovery case is reported and all other feature parameters are constant. The week 5 deaths increase by 3.4975 when one additional Week 4 death case is reported and all other feature parameters are constant. The week 5 deaths increase by 33.5133 when CFR increases by one and all other feature parameters are constant. The regression formula is as follows:

$$Week5Deaths = 84.42 - 0.07*Totalcases + 0.12*Activecases - 0.1*Recoverycases + -3.5*Week4deaths + 33.51*CFR$$

We also notice that the t-test has p-values that we cannot reject the null hypothesis other than the Week4deaths which shows that we can make a stronger inference about week 4 deaths but not strong with the other features. Finally we also plot the regression pair plot in Figure 7 that shows a high degree of outliers.

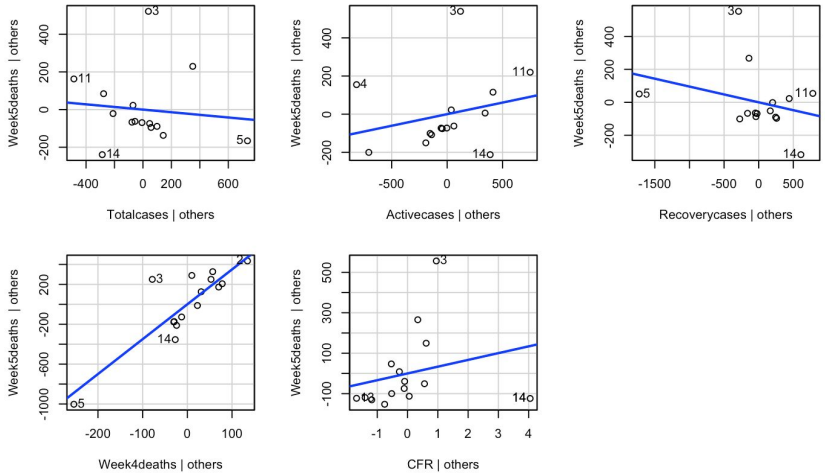

Figure 7: Pair Plots

Finally, we can perform a point estimate and create a prediction interval for India using our model with all 5 feature parameters. The resulting point estimate is 196.60 and has a prediction interval of -358.04 - 749.24.

3.2 Multiple Regression Model: Log Transformation

We develop a linear regression model that applies a log transform to the parameters as the scatter plots indicate an exponential relationship, This is expected as pathogens and infectious disease will spread exponentially. However, we will need to remove Brazil from our observations as it has zero week 4 deaths and a CFR of zero as well that would result in NaN values that cannot be used in the model.

3.2.1 Diagnostics

In Figure 8, the scatter plots show the transformation of log compared to Figure 4. Although there are still visible outliers on the far right and far left, the outliers of China and Italy are no longer having a high leverage on the linear regression line, therefore they are no longer influential points. There are fewer outliers, and they are not as extreme and don't have as much of a high leverage as the outliers in the original scatter plot. With that being said, we can say that it is easier to see the linear trend between the two variables with a log transformation. If those influential points were to be removed, the linear regression line wouldn't change much.

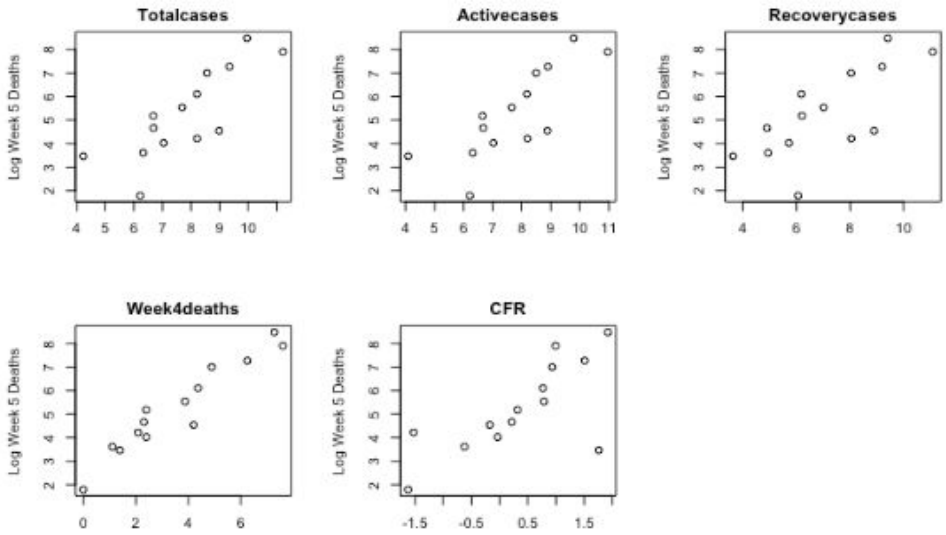

Figure 8: Scatter plot of Log Transformed: Week 5 deaths vs Features

Finally we diagnose the normality of the residuals using a QQ Plot and classic residual plot as shown in Figure 9 and 10 respectively. The plots show outliers as well as a variability of regressions and do not appear random.

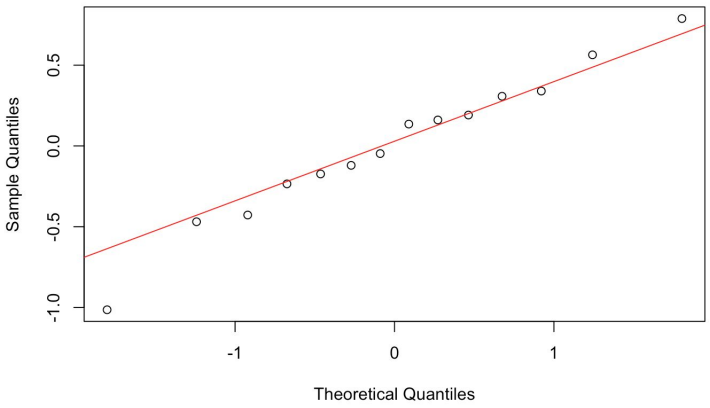

Figure 9: QQ Plot of Residuals (Log Transform)

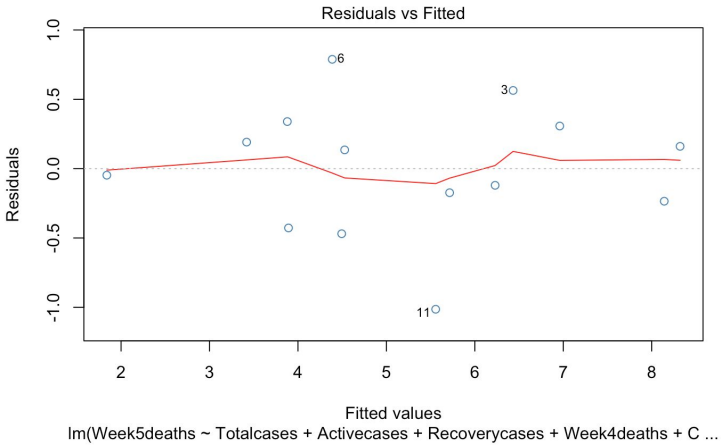

Figure 10: Residual Plot (Log Transform)

As there are a few data points we perform a Breusch-Pagan Test (*Advanced Analysis*) to establish if Heteroskedasticity exists. The p-value for the test is 0.6937, indicating that we cannot reject the Null hypothesis that the data follows a Chi-Square distribution with the parameters.

3.2.1 Inference

We report the summary of the multiple linear regression with the full 5 feature parameters in Table 3. The R^2 (goodness of fit) and adjusted R^2 values of 0.9408 and 0.9038 with a F-statistic of 25.43 with a p-value of 0.000103.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-581.1014	1067.4115	-0.5444	0.6010
Totalcases	125.2276	231.8770	0.5401	0.6039
Activecases	1.6159	2.1969	0.7355	0.4830
Recoverycases	-0.1072	0.3223	-0.3324	0.7481
Week4deaths	-125.9814	231.8386	-0.5434	0.6017
CFR	127.0035	231.8375	0.5478	0.5988

Table 3: Inference summary of model with 5 feature parameters (Log Transform)

The interpretation of coefficients is as follows is different than previously as we have to take into account the log scale as follows:

*log(**Week5Deaths**) = -581.10 + 125.23*log(**Totalcases**) + 1.62*log(Activecases) -0.11*log(**Recoverycases**) + -125.98*log(**Week4deaths**) + 127.00*log(**CFR**)*

We also notice that the t-test has p-values that we cannot reject the null hypothesis for any feature parameter and limits. Finally we also plot the regression pair plot in Figure 11 that shows a high degree of outliers.

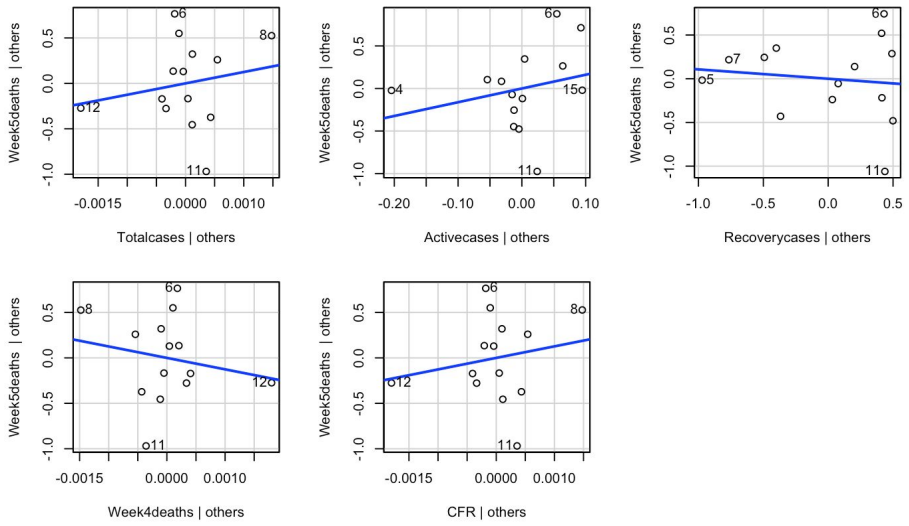

Figure 11: Pair Plots

Finally, we can perform a point estimate and create a prediction interval for India using our model with all 5 feature parameters. Note, we have to apply the transform to reverse the log to evaluate the predicted value as follows:

$$\textbf{Week5Deaths} = e^{-581.10 + 125.23*log(\textbf{Totalcases}) + 1.62*log(\textbf{Activecases}) -0.11*log(\textbf{Recoverycases}) + -125.98*log(\textbf{Week4deaths}) + 127.00*log(\textbf{CFR})}$$

The resulting point estimate is 93.88 and has a prediction interval of 12.76 - 690.48.

3.3 Comparison of Multiple Regression Models

We can now compare the models (original not transformed, and log transformed). Table 4 shows the prediction summary of the models including the point estimate and prediction intervals. One advantage of the log transforms is that the prediction interval is significantly narrower (61% smaller). Another advantage of the log transform model is that the lower intervals are positive and represent a real world scenario compared to the original transform which has lower intervals that are negative and are not representing nature.

Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Transform	Point Estimate	Interval Lower	Interval Upper
✔	✔	✔	✔	✔	None	195.60	-358.04	749.24
✔	✔	✔	✔	✔	Log	93.88	12.76	690.48

Table 4: Prediction Summary of Models

Table 5 shows the inference summary of the models including the pearson correlation coefficients and the f-statistic. One advantage of the non transformed model is that the correlation factor is higher, for example 0.9807 compared to 0.9408 for the model using all the feature parameters. We can also see that the models that included all 5 parameters, have a lower degree of freedom, indicating that there is more freedom to vary and the model is more resistant to the values.

Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Transform	R^2 Coef	Adj R^2 Coef	Degrees Freedom	F-Statistic
✔	✔	✔	✔	✔	None	0.9807	0.9701	9.0000	91.7027
✔	✔	✔	✔	✔	Log	0.9408	0.9038	8.0000	25.4276

Table 5: Inference Summary of Model

3.4 Model Selection

3.4.1 Multicollinearity

We diagnose the potential multicollinearity of the multiple linear regression model (no transform) by plotting a pearson correlation heat map shown in Figure 12. We see that there exists a high correlation between all the parameters and week 5 deaths with the highest correlation of week 4 deaths with week 5 deaths. However, we also see high correlation between feature parameters; for example, total cases and active cases have a 0.99 correlation which indicates issues of multicollinearity will exist.

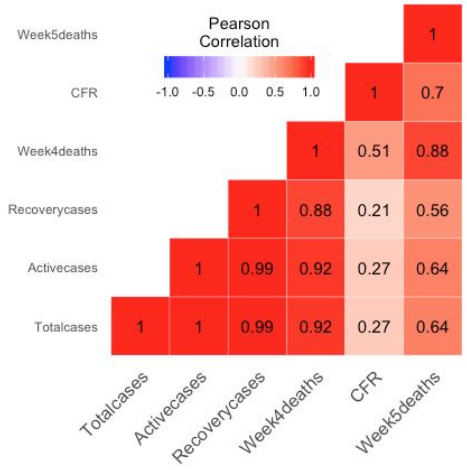

Figure 12: Pearson Correlation Heat Map

To further establish multicollinearity we report the Variance Inflation Factor (VIF) (*Advanced Analysis*) for the model in Table 6. This has VIF values which are substantially larger than 10; for example, total cases have a VIF of 4357 which indicates the multicollinearity present.

	Totalcases	Activecases	Recoverycases	Week4deaths	CFR
VIF	4357.00	1338.14	842.39	46.42	2.37

Table 6: VIF for Model using all 5 features as parameters

In Table 7 we correct this by reducing the parameters to three: recovery cases, week 4 deaths and CFR. The resultant VIF values are below 10. This simpler model will also be used for comparison later.

	Recoverycases	Week4deaths	CFR
VIF	7.21	9.34	2.13

Table 7: VIF for Model using 3 features as parameters

As mentioned earlier we found large VIF values with this model and we now compare the model with a simpler model which has only 3 feature parameters (recovery cases, weekly 4 deaths and CFR). The result of ANOVA shows a p-value of 0.79 meaning that we cannot reject the null hypothesis and the simpler model can be leveraged instead of using the more complex model.

We also diagnose the potential multicollinearity for the log transform of a linear regression model by plotting a pearson correlation heat map shown in Figure 13. We see that there exists a high correlation between all the parameters and week 5 deaths with the highest correlation of week 4 deaths with week 5 deaths. However, we also see high correlation between feature parameters; for example, total cases and active cases have a 1.0 correlation which indicates issues of multicollinearity will exist.

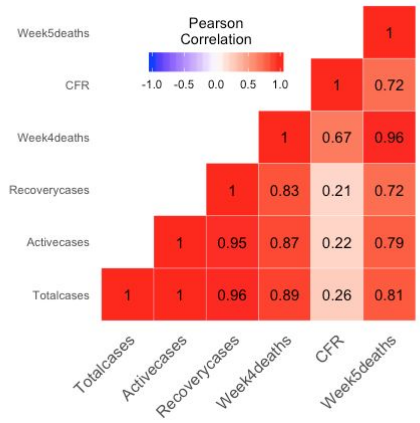

Figure 13: Pearson Correlation Heat Map (Log Transform)

To further establish multicollinearity we report the Variance Inflation Factor (VIF) (*Advanced Analysis*) for the model in Table 8. This has VIF values which are substantially larger than 10; for example, total cases have a VIF of 6452851 which indicates the multicollinearity present.

	Totalcases	Activecases	Recoverycases	Week4deaths	CFR
VIF	6452851.06	540.74	17.08	10922041.51	2438601.29

Table 8: VIF for Model using all 5 features as parameters

In Table 9 we correct this by reducing the parameters to three: recovery cases, week 4 deaths and CFR. The resultant VIF values are below 10. This simpler model will also be used for comparison later.

	Week4deaths	CFR
VIF	1.812540394	1.812540394

Table 9: VIF for Model using all 2 features as parameters

As we found large VIF values with this model and we now compare the model with a simpler model which has only 2 feature parameters (weekly 4 deaths and CFR). The result of ANOVA shows a p-value of 0.61 meaning that we cannot reject the null hypothesis and the simpler model can be leveraged instead of using the more complex model.

We can now compare the models (original not transformed, and log transformed) on the simplified versions of these two models. Table 10 shows the prediction summary of the models including the point estimate and prediction intervals. Again the advantage of the log transforms is that the prediction interval is significantly narrower.

Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Transform	Point Estimate	Interval Lower	Interval Upper
✖	✖	✔	✔	✔	None	223.02	-312.47	758.51
✖	✖	✔	✔	✔	Log	81.34	22.07	299.75

Table 10: Prediction Summary of Models

Table 11 shows the inference summary of the models including the pearson correlation coefficients and the f-statistic. The models that only used 3 parameters have a higher degree of freedom meaning they are less resistant to the values which make sense since there are less parameters to be used to determine the prediction.

Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Transform	R^2 Coef	Adj R^2 Coef	Degrees Freedom	F-Statistic
✖	✖	✔	✔	✔	None	0.9765	0.9701	11.0000	152.4732
✔	✔	✔	✔	✔	Log	0.9408	0.9038	8.0000	25.4276

Table 11: Inference Summary of Model

3.4.2 Step-wise

To see if we can improve our model, we perform backward-elimination on our multiple linear regression model to remove variables that do not lower its AIC score. After performing this model selection, our improved model differs from our multiple linear regression model by its removal of CFR. Comparing the adjusted R^2 of the original multiple linear regression model and the model produced from backward elimination, we note that the adjusted R^2 for the backwards elimination model is lower by 0.02, indicating that the new model is better able to explain the variability in the data.

Total Cases	Active Cases	Recovery Cases	Week 4 Deaths	CFR	Transform	R^2 Coef	Adj R^2 Coef	Degrees Freedom	F-Statistic
✔	✔	✔	✔	✔	None	0.9408	0.9038	8	25.43
✖	✔	✖	✖	✔	Backward Elimination	0.9324	0.9201	11	75.9

Table 12: Inference Summary of Backward-Elimination Model

3.4.3 Regularization

Model	RMSE	R^2
Multiple Linear Regression from Backward Elimination	0.5065948	0.988511
Lasso Regression	0.3833645	0.9984967
Ridge Regression	0.3630199	0.9967985

Table 13: Inference Summary of Lasso and Ridge Regression

To see if we can improve our model, we performed lasso and ridge regression with all of the original predictors and compared their RMSE and R^2 to that of our improved multiple linear regression model. The R^2 differs in Table 13 differs from the one in Table 12 because it was created using the same training data and tested using the same testing data as the models for lasso and ridge regression in order to get a fairer evaluation.

After performing lasso and ridge regression, we note that in both cases the R^2 is larger and the RMSE is smaller than the multiple regression model. The smaller RMSE indicates that performing ridge and lasso regression yield smaller residuals. The better performance is likely because regularization addresses the multicollinearity issue in the data by introducing bias through a penalty.

Since the variables are able to be minimized to 0 in lasso regression, we can look at the coefficients to determine which variables the model dropped. In lasso regression, only *Totalcases* and *Week4deaths* were dropped whereas in backward elimination, those two in addition to *Recoverycases* were dropped. This indicates that *Recoverycases* may be an important predictor that the backward elimination model overlooked.

As shown, we can improve our model further by choosing the model generated from lasso or ridge regression over our model from backward-elimination.

4. Conclusion

We developed multiple different models to estimate the number of week 5 deaths for India. A model that used all the parameters in a using multiple linear regression gave the highest correlation factor of 0.98. However, simplifying the model rather than using all the parameters yields the same predictive power due to the issue of multicollinearity. Additionally, using a log transform is recommended to reduce the range of the prediction interval as well as giving real world interval values (which are not negative values that are meaningless). Furthermore, after performing model selection using backward-elimination, we were able to improve

our model's adjusted R^2 score by eliminating 3 variables. However, our best model is shown to be one that was computed using Lasso or Ridge regression.

Finally, it is recommended to improve the amount of data by including more observations, i.e. data from different countries. With such a small data set it is impractical to do techniques such as test-train split or removing outliers as the number of observations is so small and these techniques could introduce bias.

6. Appendix

6.1 Breusch-Pagan Test

The Breusch–Pagan test is used to test for heteroskedasticity in a linear regression model [1]. The test can establish heteroskedasticity if the variance of the errors from a regression is dependent on the values of the independent variables.

[1] Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation". Econometrica. 47 (5): 1287–1294.

6.2 Akaike's Information Criterion (AIC)

Akaike's Information Criterion is used to compare statistical models to each other in which a lower AIC score indicates a better model[2].

[2] Stephanie Glen. "Akaike's Information Criterion: Definition, Formulas" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/akaikes-information-criterion/

7. Contributions

Saroop Samra worked on the descriptive statistics and EDA visualizations as well as developing the analysis for the multiple regression models including diagnostics, inference and prediction. For advanced analysis, Saroop worked on the multicollinearity section including VIF and ANOVA analysis. Saroop also co-write the model comparison as well as the introduction, data and conclusion.

Arely Vasquez worked on writing the explanations for the tables, charts, and graphs. She also co-wrote the model comparison as well as the conclusion.

Amy Nguyen worked on the backward-elimination and model regularization portion, which consists of the analysis and the code.