Misinformation during COVID-19

Robert Dunn **rdunn@ucsd.edu**
Gavin Tran **gatran@ucsd.edu**
Christopher Ly **chl818@ucsd.edu**

1. Introduction

In this report we will be addressing a separate report which used linear regression to predict deaths from the COVID-19 pandemic. COVID-19 or the Coronavirus disease 2019, is caused by SARS-COV-2, or severe acute respiratory syndrome coronavirus 2, which currently has no form of treatment. In particular, the report was focused on predicting the number of deaths in India at week 6 from day 0 using linear regression. The dataset used in the report comes from Elsevier's COVID-19 resource center which was made freely available to the public in January 2020. The included information revolves around data from the 15 most affected countries (including India) and for each country includes the total amount of cases, active cases, recovery cases, week 4 deaths, CFR (Case Fatality Rates, or the number of reported deaths per reported case), and week 5 deaths. Some limitations of the report they reported includes the use of only the top 15 countries affected, potentially leading to overestimations, as well as most of the data being used in the analyses which could potentially lead to issues like multicollinearity.

In our analysis section, the first thing we did was build a linear model to find a point estimate and prediction interval for the number of deaths in India in week 5. What we found was a p-value of 1.925e-07 and instead of a single prediction value of 196 deaths in week 5 from the report, a confidence interval with a range of 34 to 357 deaths. After this the next thing we did was apply a log transformation to the dataset, fit a linear model with the log scale and again provided a point estimate and prediction interval for the amount of deaths in India in week 5. After dropping Brazil, due to the country having 0 deaths in week 4 and in turn a CFR of 0, which would be heavily affected by the log transformation, we found a p-value of 0.0001 and a confidence interval with a range of 22 to 406 deaths. In the next part of our analysis, we compared the models and predictions from parts 1 and 2 of our analyses, looking at the pros and cons of both sections. What we found for the original linear model was that it was easier to understand and had a linear relationship, but issues were that the data was skewed, had more noticeable outliers and had issues of multicollinearity. In the transformed linear model, we discovered a more symmetrical distribution of clusters and linear relationships, issues though include that the transformations led to ambiguity with outliers, potentially affecting the quality of data with an observed example being the reinforcement of collinearity between variables. In the last part of our analysis, knowing that some variables may be correlated, we tried using model selection methods to possibly help improve the models. Looking at two different models, we discovered that the best model which didn't reveal signs of collinearity was a linear model including the variables 'Recoverycases,' 'Week4deaths,' and 'CFR.'

After these four analyses, we included a brief conclusion of our findings and ended our report with an appendix including all of the figures used in the analysis section.

2. Analysis

2.1. Part 1

Since several countries have experienced their fifth week of the coronavirus, we build a linear model to predict the outcome of India's fifth week. As a result, India is predicted to have 196 deaths in week 5. However, a single value prediction is likely not correct. In this regard, a range of possible number of deaths lie between 34 and 357 in week 5. These results were rounded to the nearest integer for a more meaningful interpretation of the prediction. To determine the correctness of these predictions, we need to evaluate the strength of the linear model.

The linear model was built using all of the available variables to predict the deaths of week 5. In Figure 1.1, the coefficients of the linear model are presented. The coefficients of each variable represent the expected change in the number of deaths in week 5 for a unit increase in that variable given that other variables are held constant. While the exact number of the coefficients are less meaningful to interpret, they represent the relationship between them and the number of deaths in week 5. Additionally, the p-values associated with the variables indicate the meaningfulness of that variable explaining the response variable. The number of deaths in the fourth week seems most meaningful in predicting the number of deaths in the fifth week.

The strength of the model relies on the relationship between the explanatory variables and the response variable. In addition, any case of multicollinearity may reduce the practicality of the linear model. From Figure 1.2, the relationship between the variables and each other can be observed. From this, it can be seen that many points are clustered together which could reduce the linearity of the model as outliers become more influential. Despite this, the model fits well with an adjusted R^2 value of 0.9701 and a F test returning a significant p-value of 1.925e-07. The residuals of the linear model were evaluated for heteroskedasticity using the Breush Pagan test which resulted in insignificant findings with a p-value of 0.71.

2.2. Part 2

From the results of the previous model, we found that by the nature of the variables, clusters of points were commonplace. To reduce this effect, the variables undertook a log transformation. However, as a result, Brazil had to be dropped after the transformation as it had 0 deaths for week 4 and in turn, its case fatality rate. Log transformation on points with a value of 0 would extremely influence the result of the linear model. The linear model based on the transformed variables predicted the number of week 5 deaths to be 94 with a potential range of 22 to 406. The predictions had to be transformed back from the log scale to have an interpretable meaning.

The coefficients produced from the log transformation require a different interpretation of the values. Instead of an expected amount of change per additional unit in the explanatory variable, the coefficients indicate the percentage change given a 1% change in the explanatory variable. These coefficients can be found in Figure 2.1. In the same manner, the exact values are

less significant rather than the relationship they describe between them and the response variable. Unlike the untransformed model, the transformed model did not find a variable with significant contribution to predicting the number of deaths for week 5.

The log transformation alters the relationship between the variables. The resulting pair plots can be observed in Figure 2.2. The clusters have been decompressed and we see a greater linear relationship between the explanatory variables and the response variable. However, multicollinearity is more observable as well. This explains the lack of a meaning variable in the coefficients mentioned earlier as multicollinearity increases the standard deviation resulting in greater p-values. The transformed model did not fit the data as well with an adjusted R^2 value of 0.9038 but still significant as a F test returns a p-value of 0.0001. The heteroskedasticity of the residuals was tested returning a p-value of 0.94 reinforcing the homoskedasticity of the residuals.

2.3. Part 3

The untransformed model and the transformed model presented a similarly wide range of predictions for the amount of deaths for week 5 while having differing results for a single estimate. The untransformed model has a more interpretable meaning and surprisingly uncovered a linear relationship despite the skewed data as shown in the numerous clusters between the variables. Outliers are more obvious while the data is not transformed. The transformed model obtained a more symmetrical distribution of the clusters and linearized the relationships. However, this is also a potential downfall as this may not be the correct transformation to use. Blindly transforming all the variables has the potential of obfuscating outliers and low quality data. Another downfall of the transformed model is that it reinforces the collinearity between the explanatory variables. A limitation of both models is the presence of multicollinearity which will be addressed in the next section.

2.4. Part 4

The first thing we look at for our model selection is the presence of collinearity among the explanatory variables in the data. We want to avoid having collinearity in our data because it tends to overcomplicate the model. This can have certain effects on the model such as inflating the variance. Therefore, we want to have a Variance inflation factor (VIF) of less than 10 among our explanatory variables to avoid such problems. When looking at these values (Figure 4.1), we can see that ‘Totalcases,’ ‘Activecases,’ and ‘Recoverycases’ variables have a strong correlation between each other. Therefore, we can create a new linear model without a combination of two of these variables. Doing so removing ‘Activecases,’ and ‘Recoverycases’ results with an adjusted R-squared value of 0.9602, without ‘Totalcases’ and ‘Recoverycases’ results in a value of 0.945 and with ‘Totalcases’ and ‘Activecases’ results in a value of 0.9701. As a result, it can be seen that removing ‘Totalcases’ and ‘Activecases’ from our model results in an equal adjusted R-squared value to the model without removing any variables. Therefore, we can conclude that

by removing these two variables, we get a better model through removing collinearity between variables which consequently lowers the variance in our model.

A different approach to picking a better model is to explicitly attempt to increase the adjusted R-squared value by removing one variable at a time and calculating the respective value for the given model. Given that our original model has an adjusted R-squared value of 0.9701, we can then begin the backwards-elimination approach. As seen in Figure 4.2, the model with the highest adjusted R-squared value is the one that removes the 'Totalcases' variable. We can then repeat this step, revealing that removing the variable 'CFR' increases the adjusted R-squared value again. Finally, it can be seen that removing one more variable does not lead to an increase in the adjusted R-squared value. Therefore, we find the best model according to the backwards-elimination approach includes the variable 'Activecases,' 'Recoverycases,' and 'Week4deaths.' However, it should be noted that we have already established that the 'Activecases' and 'Recoverycases' variables are correlated implying the existence of collinearity in this model. Also, it can be seen that removing one of these variables does not increase the effectiveness of the model which can most likely be attributed to removing too many variables from the model.

In the end, both of these models have a relatively close adjusted R-squared value while the first choice does not have the presence of collinearity, which most likely implies that the first choice is a better model. Therefore, we can conclude that an improved version of the original model is a linear model with the variables 'Recoverycases,' 'Week4deaths,' and 'CFR.'

3. Conclusions

From our report, we discovered that issues from the original report was that the provided number of deaths was misleading and that multicollinearity was an unaddressed issue. To deal with the first issue, using a linear model we discovered that instead of 196 deaths that 34 and 357 deaths was the predicted range of deaths, which is also when we discovered issues with multicollinearity as well as a common cluster of points. To deal with these clusters, we tried using a log transformation which also inevitably revealed that some variables had issues with multicollinearity. After comparing the original linear model to the log transformed one, we discovered that although it helped deal with the clusters of data, the transformation also made collinearity more relevant. To finally deal with the problem of multicollinearity, we discovered that the best variables were 'Recoverycases,' 'Week4deaths,' and 'CFR.' Some takeaways from our findings is that plotting the data is a useful tool for finding patterns of collinearity, as well as to be skeptical of receiving a single number from a model, since the range of a confidence interval is much more accurate than a single estimation of a confidence interval.

Appendix

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.42474  115.00886   0.734 0.481590
Totalcases   -0.06999    0.21816  -0.321 0.755657
Activecases   0.12155    0.15538   0.782 0.454134
Recoverycases -0.09571    0.10966  -0.873 0.405463
Week4deaths   3.49750    0.70392   4.969 0.000771 ***
CFR          33.51329   46.33829   0.723 0.487907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 234.1 on 9 degrees of freedom
```

Figure 1.1 Coefficients of linear model (untransformed)

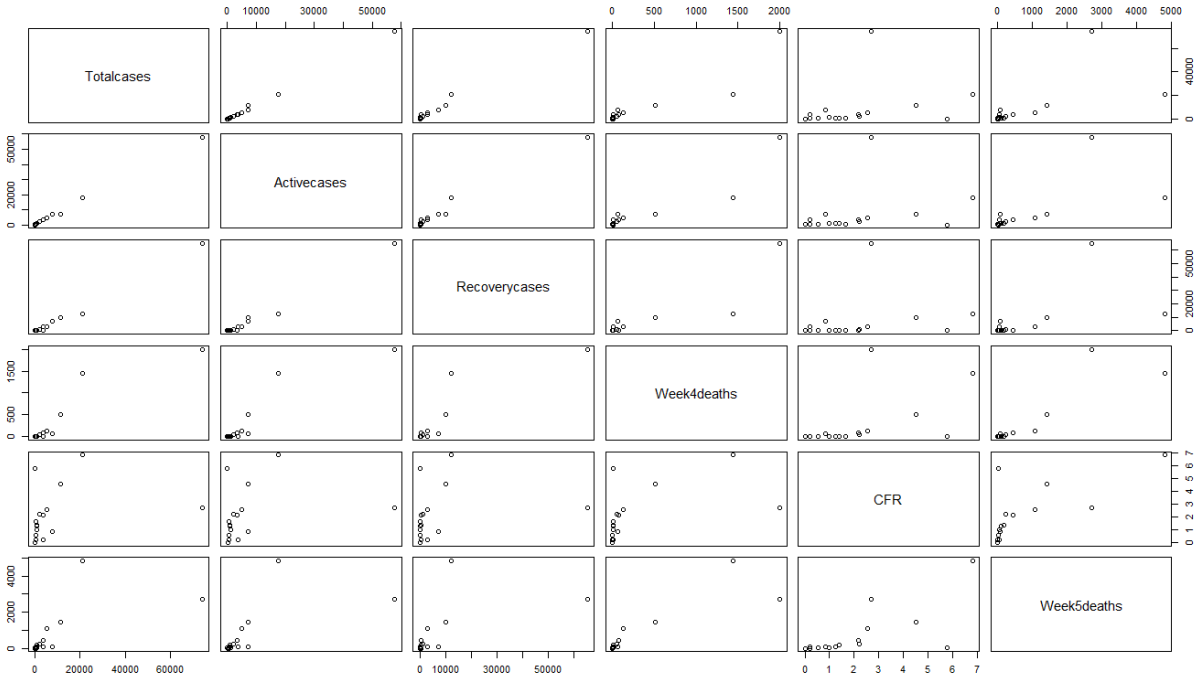

Figure 1.2 Pair plots of variables as is

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -581.1014  1067.4115  -0.544    0.601
Totalcases    125.2276   231.8770   0.540    0.604
Activecases     1.6159     2.1969   0.736    0.483
Recoverycases  -0.1072     0.3223  -0.332    0.748
Week4deaths  -125.9814   231.8386  -0.543    0.602
CFR           127.0035   231.8375   0.548    0.599

Residual standard error: 0.5877 on 8 degrees of freedom
```

Figure 2.1 Coefficients of linear model (log transformed)

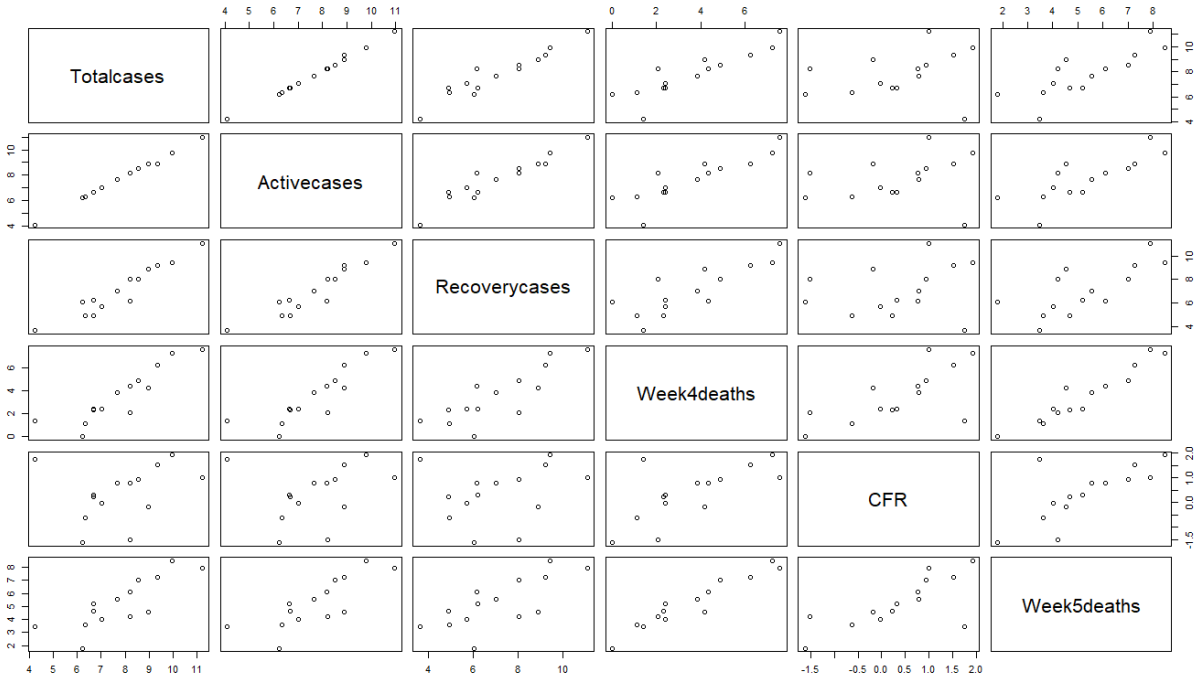

Figure 2.2 Pair plots of variables under log transformation

	X1	X2	R-Squared	Tolerance	VIF
1	Totalcases	Activecases	0.99809812	0.001901875	525.796800
2	Totalcases	Recoverycases	0.98953543	0.010464570	95.560542
3	Totalcases	Week4deaths	0.85123398	0.148766019	6.721965
4	Totalcases	CFR	0.07191492	0.928085084	1.077487
5	Activecases	Recoverycases	0.98301580	0.016984202	58.878244
6	Activecases	Week4deaths	0.85474381	0.145256190	6.884388
7	Activecases	CFR	0.07076285	0.929237154	1.076152
8	Recoverycases	Week4deaths	0.78145053	0.218549465	4.575623
9	Recoverycases	CFR	0.04381820	0.956181803	1.045826
10	Week4deaths	CFR	0.26160210	0.738397898	1.354283

Figure 4.1 Tests for collinearity among explanatory variables

```
"Linear model without Totalcases:"
0.9727406
"Linear model without Activecases:"
0.9712163
"Linear model without Recoverycases:"
0.9707678
"Linear model without Week4deaths:"
0.8991211
"Linear model without CFR:"
0.9714825

"Linear model without Totalcases and Activecases:"
0.9701123
"Linear model without Totalcases and Recoverycases:"
0.9449635
"Linear model without Totalcases and Week4deaths:"
0.8389326
"Linear model without Totalcases and CFR:"
0.9739899

"Linear model without Totalcases, CFR, and Activecases:"
0.972298
"Linear model without Totalcases, CFR, and Recoverycases:"
0.9491396
"Linear model without Totalcases, CFR, and Week4deaths:"
0.7352957
```

Figure 4.2 Adjusted R-squared values for linear models in a backwards-elimination approach

4. Contributions

- Introduction - Gavin Tran
- Part 1 - Christopher Ly
- Part 2 - Christopher Ly
- Part 3 - Christopher Ly
- Part 4 - Robert Dunn
- Conclusion - Gavin Tran