# COVID Cases and Deaths Regression

John Joyce - CEU ID: 2001928

11/29/2020

### Introduction

This report analyzes COVID-19 data and seeks to answer the following research question: is there a pattern of association between **confirmed cases per capita** and **deaths due to COVID per capita**?

COVID data is provided by Johns Hopkins University (source) and population data comes from the World Bank's *World Development Indicators*, which can be accessed through the **WDI** R package. This report focuses on the variables **country**, **population**, **confirmed COVID cases**, and **deaths due to COVID**. The data covers the population of interest (the whole world) but there may potential data quality issues: governments may test their citizens differently and may not report their data accurately. The analysis derives **confirmed cases per million** as an explanatory variable and **deaths per million** as a dependent variable for the models.

### Data Exploration

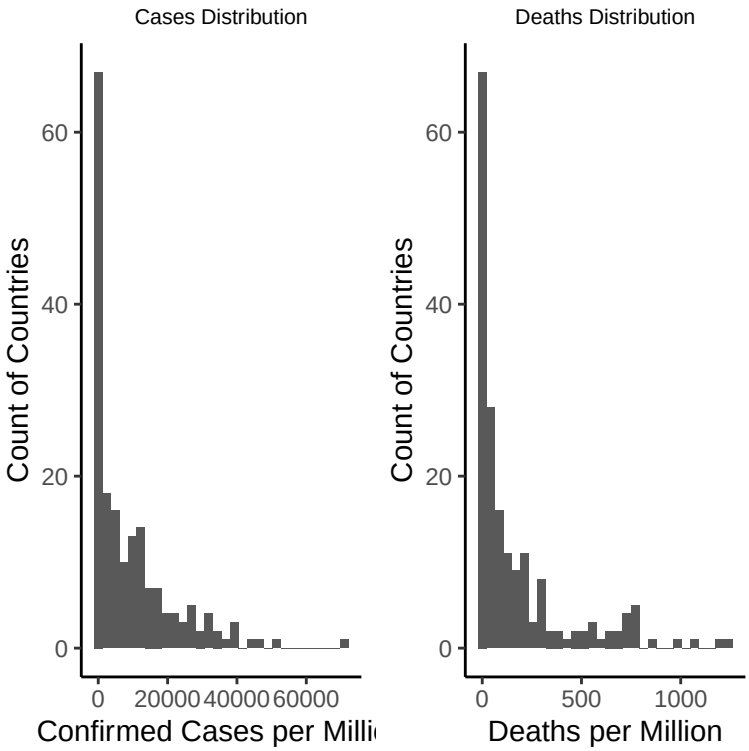
The distribution of all variables is examined with histograms (appendix item 1), and some extreme values on the right are examined more closely. A quick look at some of the values on the right reveals that they are not errors. (This check is done on all variables, but only one is shown in this report as an example.) No observations are dropped at this stage.

```
df %>% arrange(desc(confirmed)) %>% head(3) %>% kable()
```

| country | confirmed | death | recovered | active | population |
|---|---|---|---|---|---|
| United States | 10268446 | 240405 | NA | 6054094 | 328239523 |
| India | 8636011 | 127571 | 8013783 | 494657 | 1366417754 |
| Brazil | 5699005 | 162802 | 5183970 | 352233 | 211049527 |

Variables which represent cases and deaths per capita are created and visualized. To make these variables easier to interpret, both are scaled to represent cases/deaths per million people in the population. The two variables have similar distributions which resemble a power-law distribution: strongly skewed to the left side and with a long right tail. The range of observed values is enormous, with deaths per million as low as zero.

| variable | mean | median | min | max | std_dev |
|---|---|---|---|---|---|
| Confirmed Cases per Million | 9433.6256 | 4796.54867 | 3.334645 | 70998.937 | 12035.8961 |
| Deaths per Million | 174.9143 | 59.97319 | 0.000000 | 1240.402 | 255.1353 |




Both variables are transformed with a log transformation, dropping some values. Substantive reasoning: the variables are affected in multiplicative ways. For example, the maximum value of **confirmed cases per million** is approximately 1800x the minimum value. Statistical reasoning: linear regression will do a better job at estimating the average differences if the dependent variable is normally distributed. Also, when we examine the visualization of the log-log distribution, we can see that it will fit a simple linear regression more easily than any other transformation. (See appendix item 2.)

### Model Creation and Evaluation

This report explores four different models: simple linear regression, quadratic linear regression, linear splines, and weighted linear regression (using population as weights). The first three models all have similar R-Squared values of approximately 0.8. There are no dramatic differences in fit among the first three models - in fact, the simple linear model and the spline model are extremely similar based on their slopes and intercepts. There are no general changes in slope for splines to adapt to. Therefore, the more complicated quadratic and spline models do not offer any advantages which justify their use. The weighted linear regression stands out as a model with a higher R-Squared of approximately 0.9. The regression captures more of the variation in y than the other models. (See appendix item 3 for scatterplots.) The weighted linear regression is selected as the preferred model. (See appendix item 4.)

The formula is ln_death_pc_scaled = alpha + beta * ln_confirmed_pc_scaled, weights: population

The intercept is -2.6498, which indicates that the natural log of **deaths per million** is -2.6498 on average when the natural log of **cases per million** is 0. This is meaningless. The beta parameter is 0.8591, and it is easy to interpret: on average, observed values of **deaths per million** are 0.8591% higher for every 1% higher **cases per million**.

### Hypothesis Testing on Beta

A hypothesis test is conducted to determine whether there is a significant linear relationship between **confirmed cases per million** (x) and **deaths per million** (y). The null hypothesis is that the beta parameter is equal to 0, and the alternative hypothesis is that it is not equal to zero. If the p-value is less than 0.05, we will reject the null hypothesis.

```
linearHypothesis(reg4, "ln_confirmed_pc_scaled = 0") %>% kable(digits = 50)
```

| Res.Df | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|
| 165 | NA | NA | NA |
| 164 | 1 | 124.3377 | 7.105973e-29 |

The table shows that the p-value is much less than 0.05, therefore we can confidently reject the null. There is significant linear relationship between the two variables. ### Analysis of the Residuals

The countries with the largest positive errors are Yemen, Mexico, Ecuador, Bolivia, and Iran. The points which represent these countries are above the regression line. These countries performed poorly - they experienced more deaths per million relative to their cases per million than average.

| country | ln_death_pc_scaled | reg4_y_pred | reg4_res |
|---|---|---|---|
| Yemen | 3.032365 | 1.012646 | 2.019719 |
| Mexico | 6.621748 | 5.035328 | 1.586420 |
| Ecuador | 6.606066 | 5.272918 | 1.333148 |
| Bolivia | 6.639930 | 5.447429 | 1.192501 |
| Iran | 6.158680 | 5.121775 | 1.036905 |

The countries with the largest negative errors are Singapore, Sri Lanka, Qatar, Botswana, and the United Arab Emirates. The points which represent these countries are below the regression line. These countries performed well - they experienced fewer deaths per million relative to their cases per million than average.

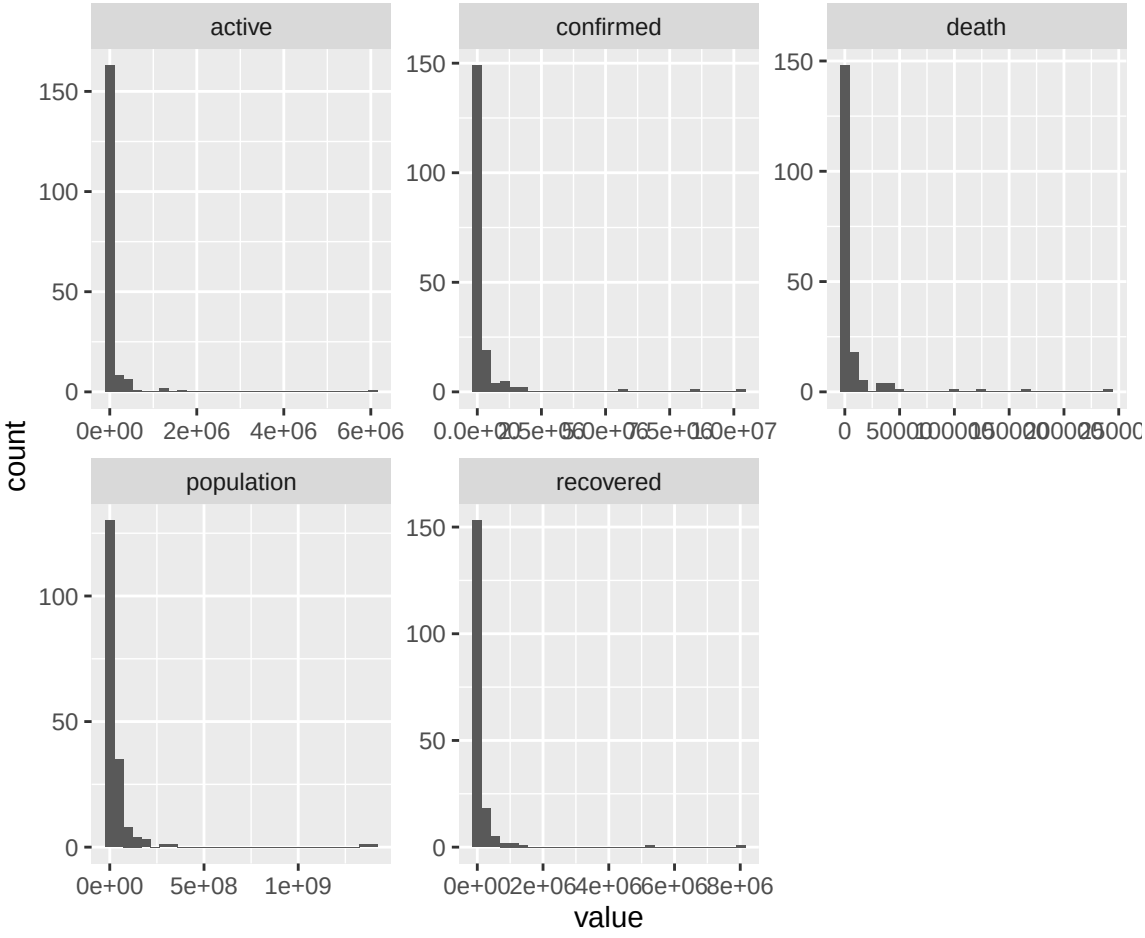
| country | ln_death_pc_scaled | reg4_y_pred | reg4_res |
|---|---|---|---|
| Singapore | 1.5911124 | 5.278694 | -3.687582 |
| Sri Lanka | 0.6315245 | 2.947148 | -2.315623 |
| Qatar | 4.4100316 | 6.602778 | -2.192747 |
| Botswana | 2.4613216 | 4.336619 | -1.875297 |
| United Arab Emirates | 3.9706046 | 5.598730 | -1.628125 |

### Executive Summary

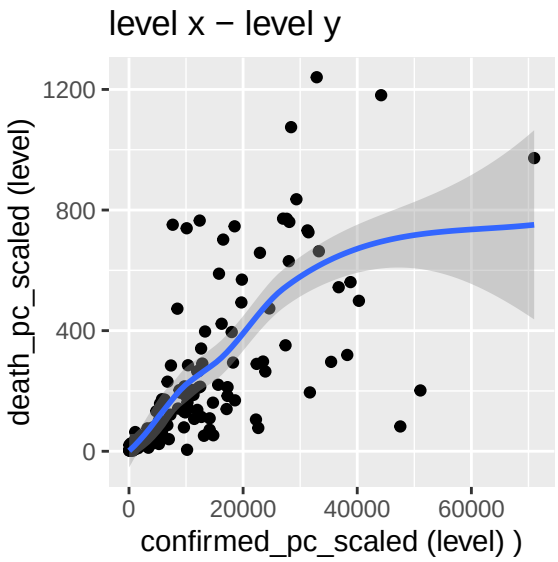
This report found a strong pattern of association between **confirmed cases per capita** and **deaths due to COVID per capita** (using scaled **per million** variables). The **weighted linear regression** model had an R-Squared of approximately 0.9 - the great majority of variation in **deaths due to COVID per capita** could be explained by regression with **confirmed cases per capita**. The results could be weakened by the revelation that some of the data was low-quality (for example, governments fabricated case numbers). The results could be strengthened by adding additional data (for example, updated numbers several months after the release of the vaccine).

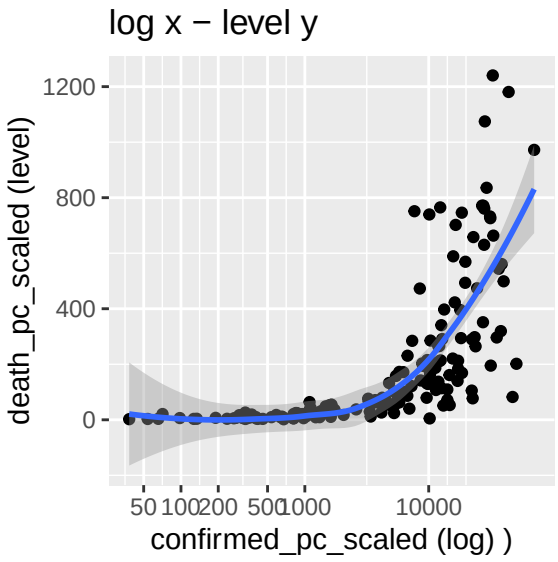
**Appendix**

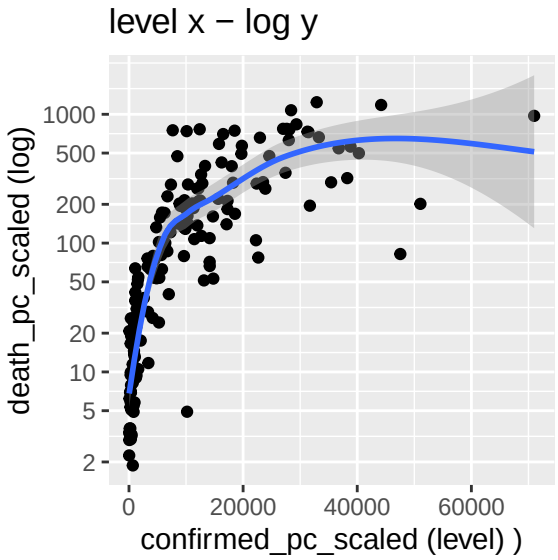
**Item 1 - Distribution of All Variables During Initial Exploration**

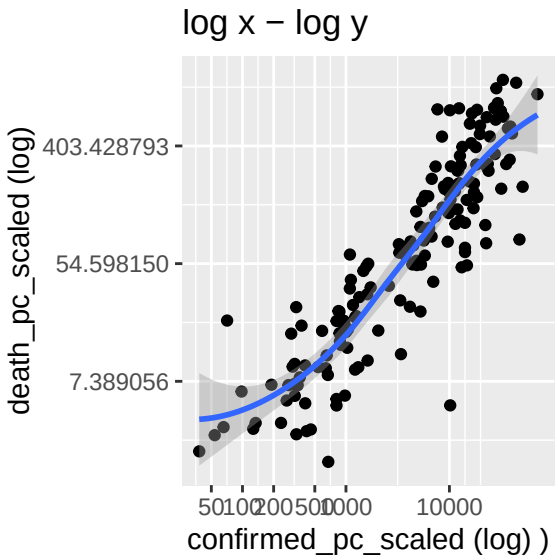



**Item 2 - Variable Transformations**

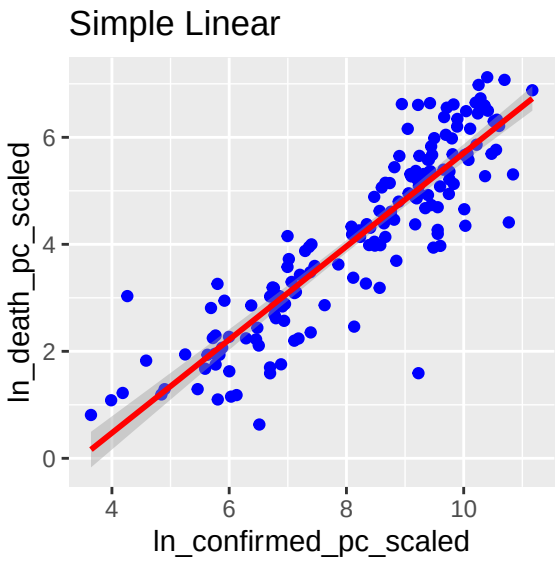

level x − level y
death_pc_scaled (level)
1200
800
400
0
0
20000
40000
60000
confirmed_pc_scaled (level) )

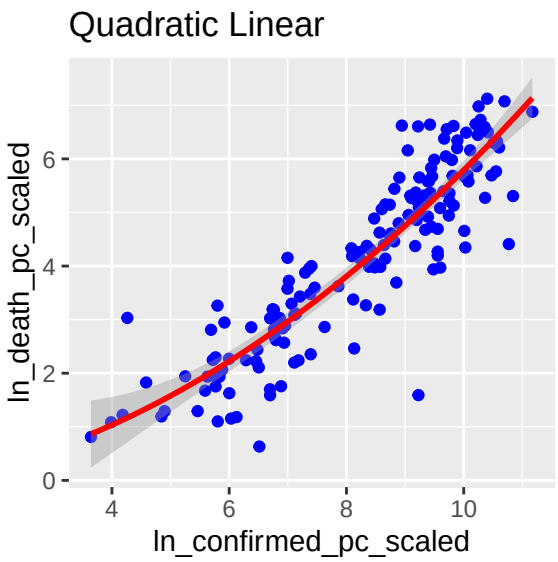

log x − level y
death_pc_scaled (level)
1200
800
400
0
50
100
200
500
1000
10000
confirmed_pc_scaled (log) )

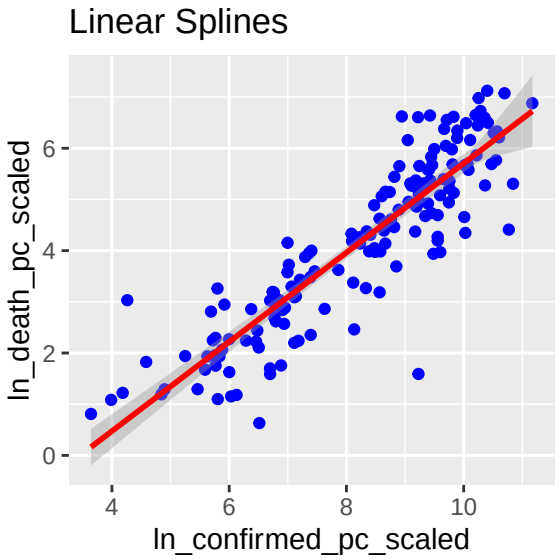


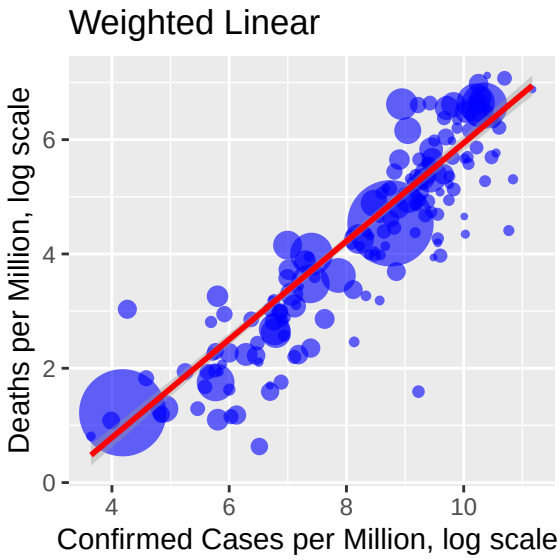




**Item 3 - Model Scatterplots**


Simple Linear
ln_death_pc_scaled
ln_confirmed_pc_scaled
0
2
4
6
4
6
8
10


Quadratic Linear
ln_death_pc_scaled
ln_confirmed_pc_scaled
0
2
4
6
4
6
8
10




**Item 4 - Why weighted linear regression?** Visual inspection of the scatterplot for the weighted linear regression model makes it clear that the model captures the pattern fairly well. It has the highest R-Squared value of any model tested. Additionally, is easy to interpret the coefficients mathematically, and easy to interpret the trend visually.