COVID-19 cases and deaths

Attila Szuts

28/11/2020

Contents

1 Executive summary **2**

2 Introduction **2**

3 Histogram and summary statistics **2**

4 Variable transformations **3**

5 Model of choice: simple linear regression **3**

6 Hypothesis testing and residual analysis **3**

7 Appendix **4**

7.1 Transforming variables 4

7.2 Estimating different models 6

7.3 Investigating biggest residuals 9

1 Executive summary

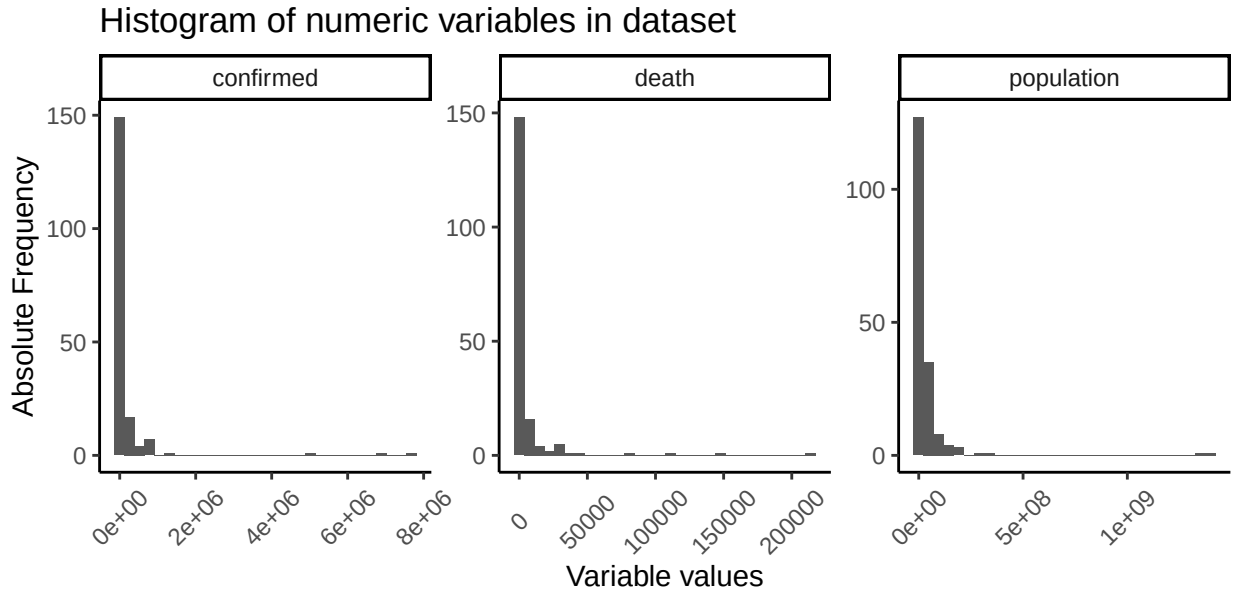
In this report I am going to summarise my main findings on the pattern of association between confirmed COVID-19 cases and deaths on a given day on a country level. I used a simple linear regression with log transformed predictor and outcome variables. I found a clear linear relationship between the log of number of confirmed cases and log of number of deaths: countries that have one percent higher confirmed cases, have roughly one percent higher deaths, on average ($\beta = 1.031$); however causal relationship has not been investigated in any way. The selected model can explain 89% of the variance. ($R^2 = 0.8859$) This suggests that in general as the number of cases increase the number of deaths increase in the same pace, i.e. very broadly speaking high and low number of confirmed cases have the same fraction of deaths. Adding further predictor variables (like population density and response to pandemic situation) could help us better predic the number of cases and explain outliers better.

2 Introduction

My main variables are countries, confirmed cases, deaths and total population of a country. Numbers on confirmed cases and deaths were collected by the Center for Systems Science and Engineering at Johns Hopkins University. Population data for each country was downloaded from the World Bank's site. The population of my analysis are the world's countries during the pandemic and I am analysing a cross-sectional subset on 2020-09-10. Data quality can vary from country to country depending on the countries' healthcare system's level of sophistication.

3 Histogram and summary statistics

In the histograms below we can see all of our variables are skewed to the left, having a long right tail. These extreme values are not measurement errors, but are in fact countries with extremely high numbers of COVID-19 cases. These countries include Brazil, India, and the United states with the top 3 number of confirmed cases; Brazil, India and Mexico with the top 3 number of deaths. These extreme values could be due to a number of factors, including higher than average population, population density and lack of counter-pandemic measures.



4 Variable transformations

Since we would like to model percentage differences in this analysis (for countries that have higher number of cases, how will the average number of deaths change), it would make sense to transform our x and y variables logarithmically (if it also makes sense from a statistical point of view). To investigate this question, I plotted my data on a scatterplot and used the previous histograms and summary statistics to establish, that indeed the clearest linear pattern of association emerges when both the predictor and outcome variable is transformed. Also, my data is cross-sectional and taking logs of variables we can solve the lack of baseline for comparison.

Because of log transformations, I had to exclude countries, where the number of confirmed cases or deaths were 0. There were no countries that did not have at least a single case, but there were 11 countries for example Mongoloia, Bhutan and Cambodia that all had 0 deaths.

5 Model of choice: simple linear regression

In my analysis I used different models to try to uncover the pattern of association between confirmed cases and deaths and to try to find the model that can explain the most variance in my data while also being easily interpreted. Because of these reasons I chose the simple linear regression that I am going to introduce in the following section.

The formula for this model is:

$$ln(\text{deaths}) = \alpha + \beta \times ln(\text{cases})$$

$$\alpha = -4.272$$

$$\beta = 1.031$$

These parameters can be interpreted as the number of deaths is higher 1.031% on average for observations with 1% higher confirmed cases. The intercept (α) means, that when there is only one confirmed case the average of the log of deaths is -4.272 (which is meaningless in this context).

6 Hypothesis testing and residual analysis

Next, I would like to investigate, if the true β is 0 or not, or said otherwise: if there is a true relationship between our x and y. So for this I set up the following hypothesis test:

$$H_0 : \beta_{true} = 0, H_A : \beta_{true} \neq 0$$

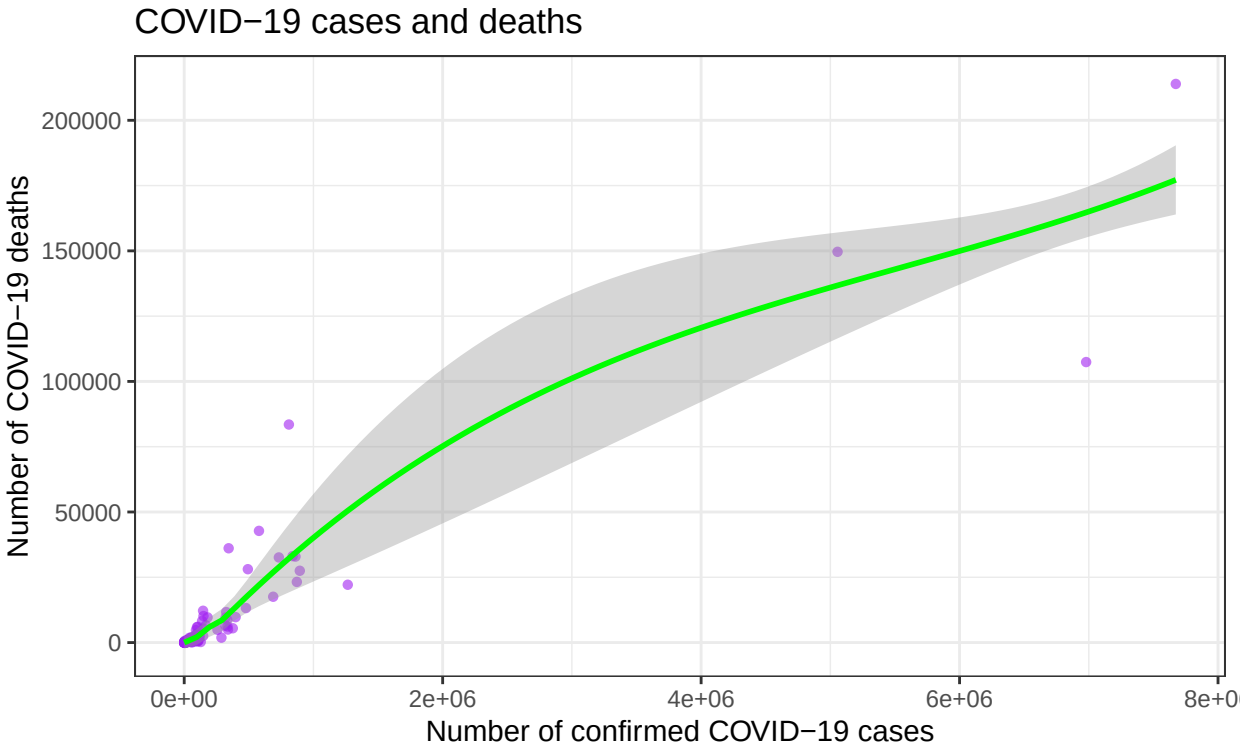
The estimated t-statistics is 35.94, with p-value: 8.63e-81, which means our test is significant at 1%. Thus I reject the H_0, which means the confirmed cases is not uncorrelated with the number of deaths.

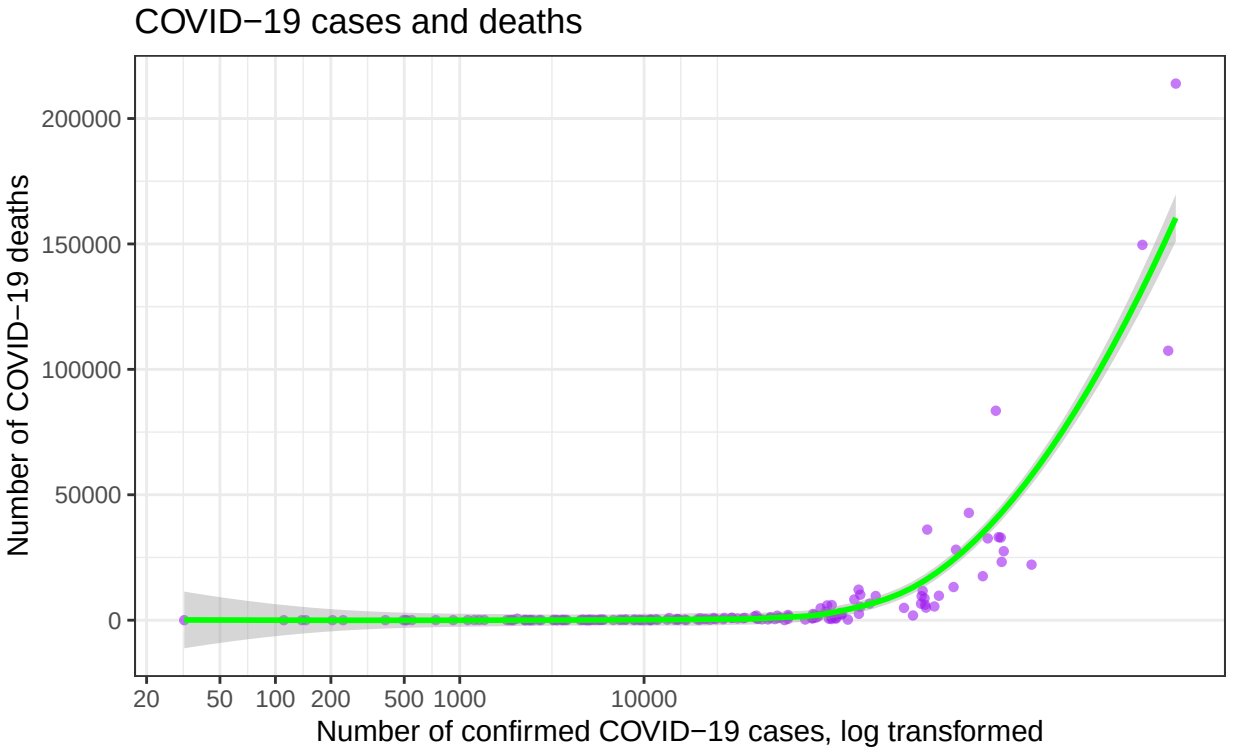
Next, I investigated the residuals:

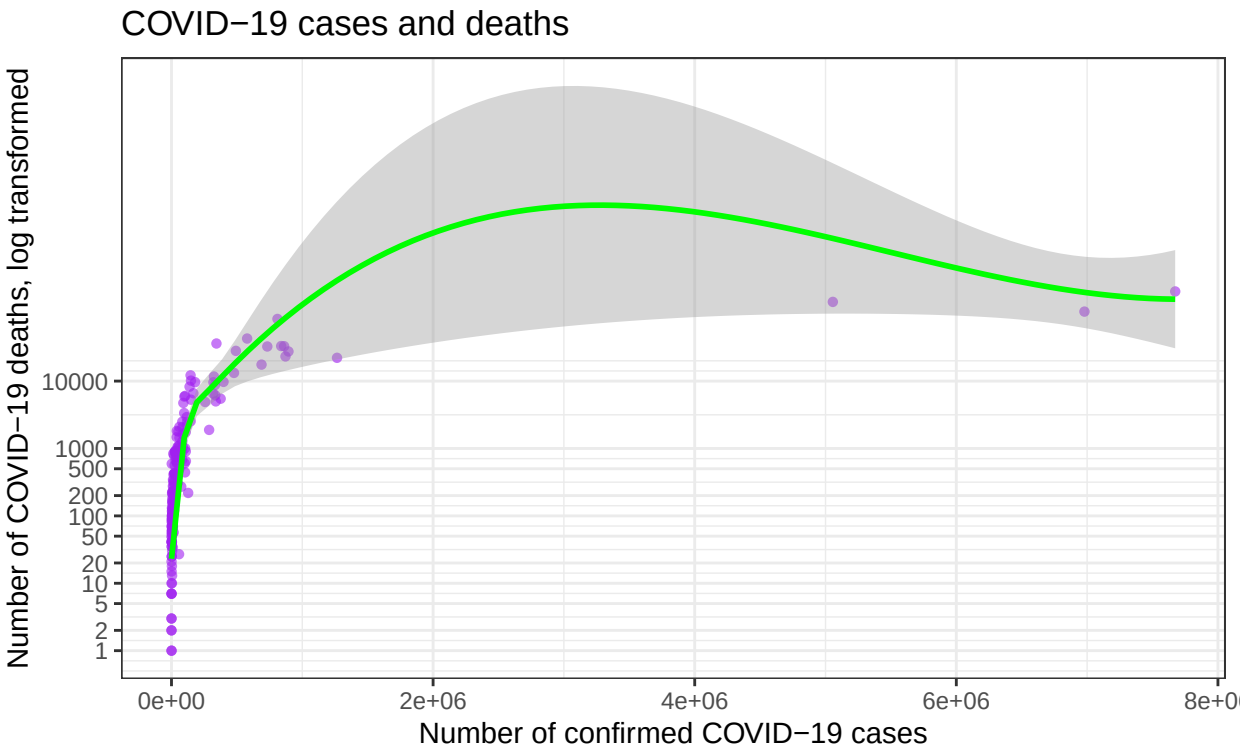
The largest negative deviance from the predicted value is found in Singapore with predicted number of deaths of 1139, but the real value is only 27. This means, that in Singapore the situation is better, than the model would predict based on the number of active cases. The largest positive deviance from the predicted value is found in Yemen with predicted number of deaths 36, but the real value is 593. This means, that in Yemen, the situation is worse, than the model would predict based on the number of active cases.

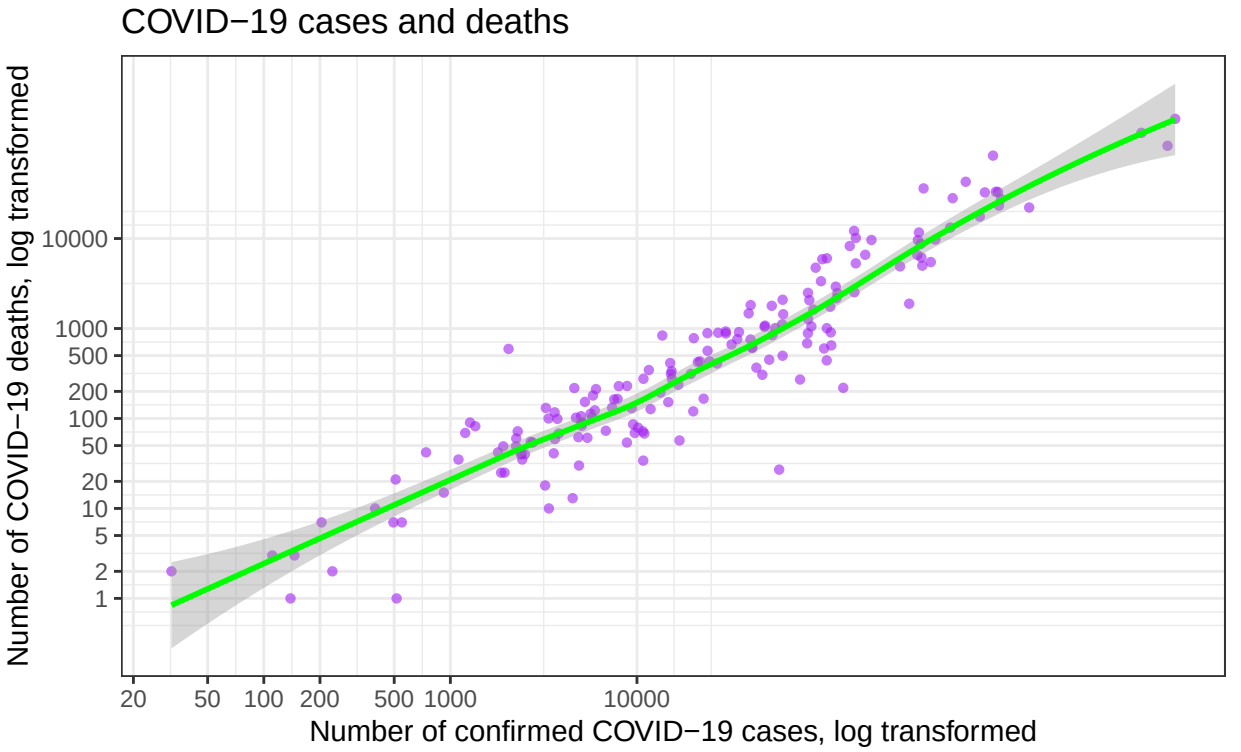
7 Appendix

7.1 Transforming variables






COVID−19 cases and deaths
Number of COVID−19 deaths, log transformed
10000
1000
500
200
100
50
20
10
5
2
1
0e+00
2e+06
4e+06
6e+06
8e+0
Number of confirmed COVID−19 cases


COVID−19 cases and deaths
Number of COVID−19 deaths, log transformed
10000
1000
500
200
100
50
20
10
5
2
1
20
50
100
200
500
1000
10000
Number of confirmed COVID−19 cases, log transformed

7.2 Estimating different models

You can see the different models, that I used in this table. All models use log transformed variables. First, I built a simple linear regression. Then I created a quadratic linear regression, but it did not yield better results at all. Creating a Piecewise Linear Splines model did not prove to be better in terms of explanatory power either. Finally, I built an OLS weighted by population which did yield better R^2, however not much more.

In the table we can see that β-s are all very similar. For the quadratic linear model it is slightly more difficult to interpret the results, as you need to get the first derivative to be able to quantify and interpret the relationship between confirmed cases and deaths. We can see from β_2 that the parabola is convex ($\beta_2 > 0$). Countries which are one unit larger than the average of x (confirmed cases) are higher by $\gamma = \beta_1 + 2\beta_2\bar{x}$ in y (deaths) on average.

The PLS model can be interpreted as there is a 1% increase in deaths on average for countries that have 1% confirmed cases if the total number of cases is less than 3.390621×10^5 and there is a 1.21% increase when the total number of cases is larger than or equal to 3.390621×10^5

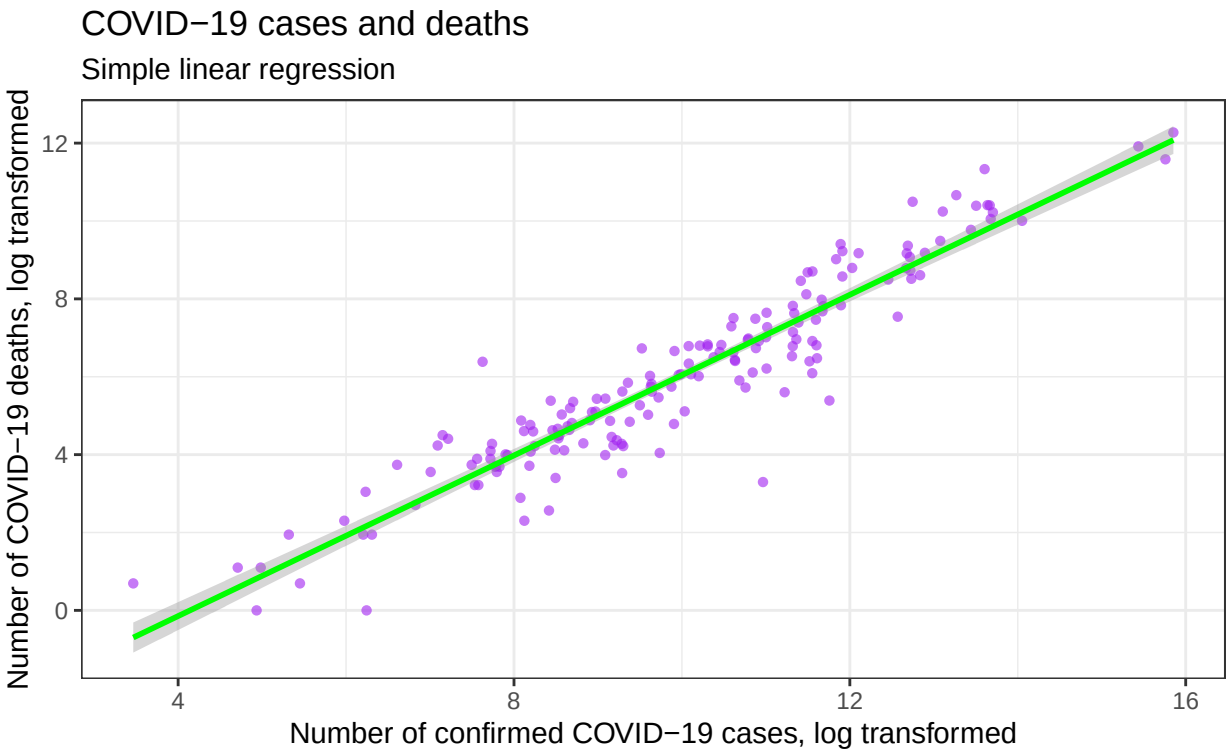
The weighted OLS model can be interpreted exactly the same as the simple linear regression, except that observations are weighted by their population. Results are also very similar, except for having twice as big SE for β.

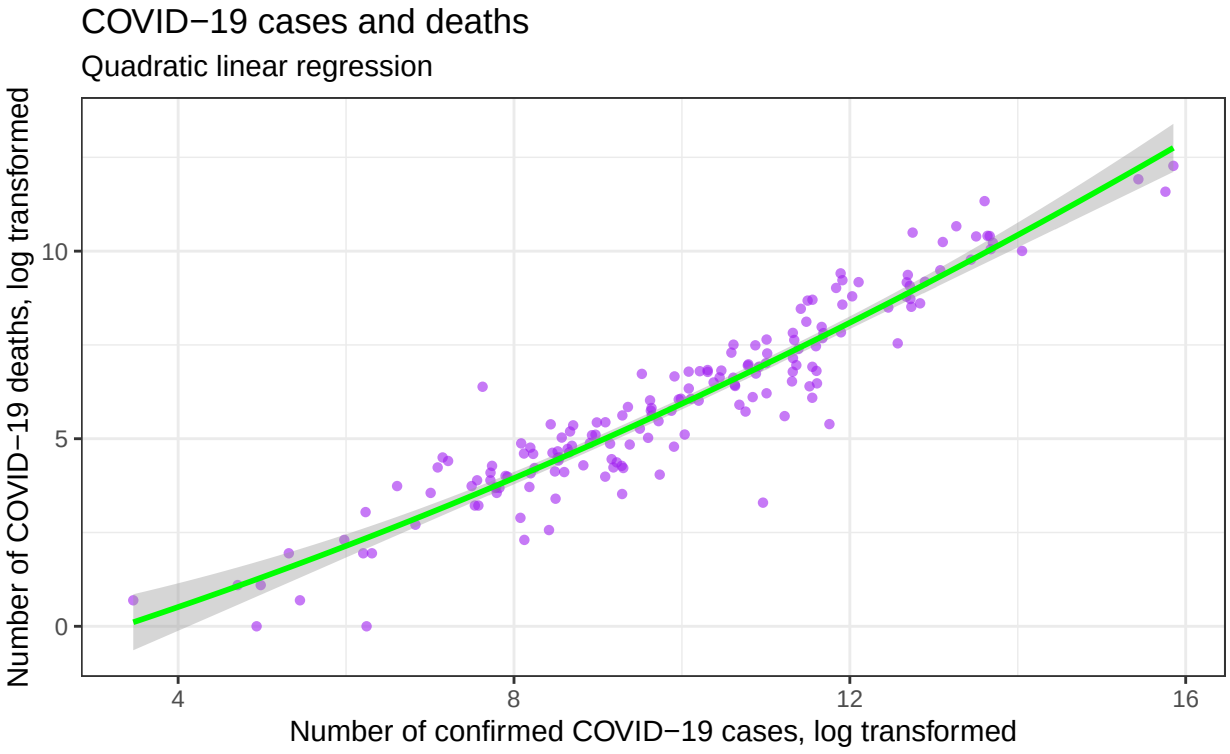
I chose the simple linear regression because it is very similar in terms of R^2, β and $SE(\beta)$ to the other three models, but it is the simplest. Following the law of parsimony, it makes sense to choose the simplest model, that can explain the most of the variance, hence my choice for the simple linear regression. Altough it is worth mentioning, that the PLS shows an interesting insight that could be worth further exploration. Namely, after the cutoff, there was a slight increase in β, altough in $SE(\beta)$ as well. This could be interesting to investigate, if countries with higher cases, are vulnerable to more deaths. Which could make sense, considering the consequences of higher number of cases: higher risk of further transmission, less capacity in healthcare facilities, etc. The WOLS model also could be useful in further analysis, since we can clearly see, that population can be a useful asset in our model, which again makes sense. Countries with more people tend to have more crowded cities which can lead to higher infection rates. However, it would probably be more straightforward to account for these kinds of transmission trends instead of weighing countries simply based on population.

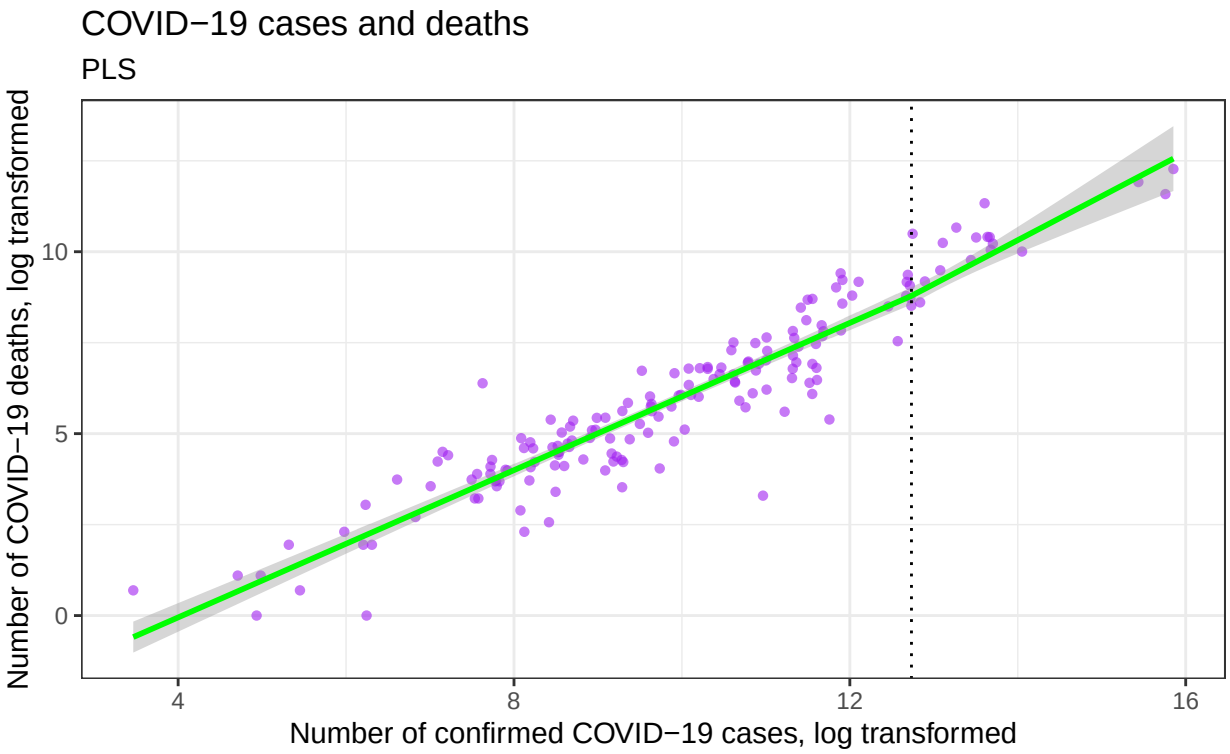
	Simple linear	Quadratic linear	PLS	Weighted OLS
Intercept	-4.27^{***}	-2.22^{*}	-4.10^{***}	-3.08^{***}
	(0.30)	(0.85)	(0.35)	(0.78)
ln(cases)	1.03^{***}	0.59^{***}		0.95^{***}
	(0.03)	(0.17)		(0.06)
ln(cases)^2		0.02^{**}		
		(0.01)		
ln(cases<339062.1)			1.01^{***}	
			(0.04)	
ln(cases>=339062.1)			1.21^{***}	
			(0.13)	
R^2	0.89	0.89	0.89	0.93
Adj. R^2	0.89	0.89	0.89	0.93
Num. obs.	170	170	170	170
RMSE	0.83	0.82	0.83	4315.62

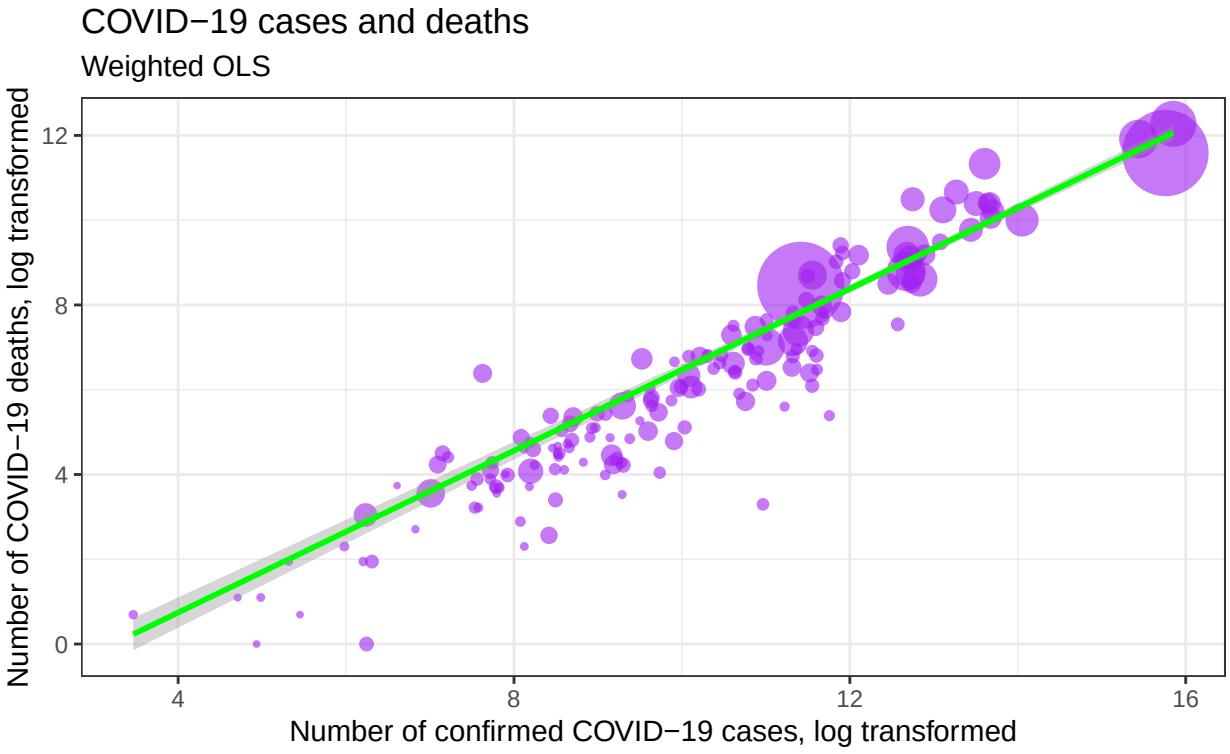
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Modelling case fatality and confirmed COVID-19 cases


COVID−19 cases and deaths
Simple linear regression
Number of COVID−19 deaths, log transformed
0
4
8
12
4
8
12
16
Number of confirmed COVID−19 cases, log transformed


COVID−19 cases and deaths
Quadratic linear regression
Number of COVID−19 deaths, log transformed
0
5
10
4
8
12
16
Number of confirmed COVID−19 cases, log transformed


COVID−19 cases and deaths
PLS
Number of COVID−19 deaths, log transformed
0
5
10
4
8
12
16
Number of confirmed COVID−19 cases, log transformed


COVID−19 cases and deaths
Weighted OLS
Number of COVID−19 deaths, log transformed
0
4
8
12
4
8
12
16
Number of confirmed COVID−19 cases, log transformed

7.3 Investigating biggest residuals

Table 2: Countries with higher predicted than actual deaths

Country	Confirmed cases	Actual deaths	Predicted deaths	Residual
Singapore	57859	27	1139	-3.742357
Qatar	127600	219	2576	-2.464849
Burundi	515	1	9	-2.168331
Sri Lanka	4523	13	82	-1.844374
Iceland	3373	10	61	-1.804153

Table 3: Countries with lower predicted than actual deaths

Country	Confirmed cases	Actual deaths	Predicted deaths	Residual
Yemen	2051	593	36	2.791553
Italy	343770	36111	7159	1.618238
Mexico	810020	83497	17329	1.572451
Ecuador	145848	12175	2957	1.415357
Chad	1274	90	22	1.397286