Analysis of COVID-19 Data - Data Analysis 2 Assignment

Kata Süle

29th November 2020

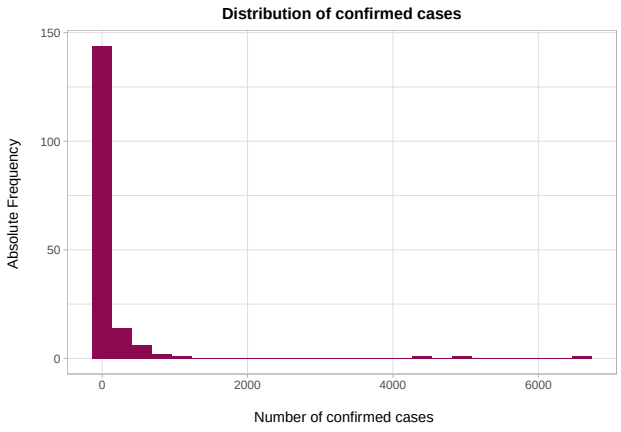
Introduction

The aim of this cross-sectional analysis is to find out whether countries with higher numbers of confirmed COVID-19 cases also experience a higher number of deaths due to COVID-19. In order to investigate this issue I use data on the number of confirmed COVID-19 cases and deaths due to COVID-19 for 170 countries for 15th September 2020. Both variables are quantitative and are measured on a ratio scale. The corresponding population of the sample used in this analysis would be all the countries where there was at least one confirmed positive case, while the sample only contains a certain amount of these. Data quality issues can arise mainly regarding reliability and coverage since it is possible that not all the cases are reported in a country due to communication problems or there are deaths that are not attributed to COVID even though they should be. As for coverage the lack of testing could be a problem.

Exploratory Data Analysis

After creating the workfile I created histograms and summary statistics for both of my variables. Based on these there were four extreme values for the dependent variable - deaths - which belong to Mexico, India, Brazil and the United States. As for the explanatory variable - confirmed cases - there were three namely India, Brazil and the United States. Since these observations ensure higher variance in the variables and are unlikely to be errors I decided to keep them. However, because of the skewed distributions of both variables I used ln transformation, therefore I dropped countries that reported zero deaths. Furthermore, I scaled the confirmed cases variable by dividing it by 1000 to make it easier to interpret.

The histograms show that the distributions of both variables are skewed with a long right tail. The summary statistics also show this feature with the means being much greater than the medians. The standard deviation of confirmed cases is large indicating big differences between countries. Therefore, as mentioned above I decided to use ln transformation for both variables. The substantitve reasons were that percentage changes would be straightforward to interpret and that the variables were likely to be affected in a multiplicative way, while the statistical ones were that the transformation made the relationship closer linear and that it made the modelling problem simpler as well.



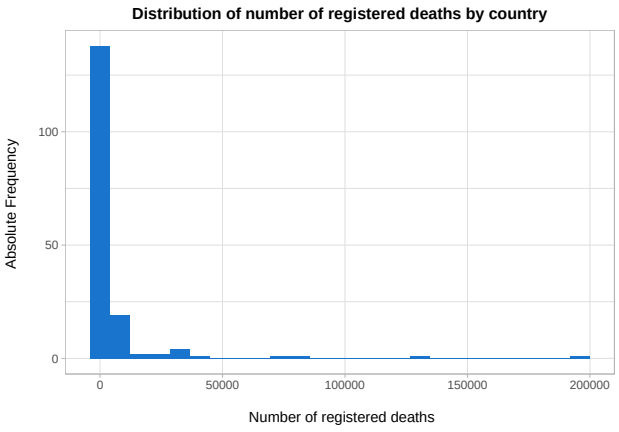


Table 1: Summary statistics for number of confirmed cases and number of deaths

variable	n	Mean	Median	Min	Max	Std
confirmed	170	173.803	13.849	0.032	6596.385	723.104
deaths	170	5496.194	266.000	1.000	195793.000	20578.424

Choice of Model

After running four different models on the data (see Appendix) I opted for using a simple linear regression with the following formula:

```
## [1] "ln_deaths = -4.32 + 1.04 * ln_confirmed"
```

The alpha parameter of the regression shows the average value of the log of deaths when ln_confirmed equals 0 which means that the number of confirmed cases equals 1. This means that the average number of deaths is 0.013 when the number of confirmed cases equals 1. While according to the beta parameter when the number of confirmed cases is greater by 1% then the number of deaths is greater by 1.04% on average.

Hypothesis Testing

When testing whether the beta parameter was equal to zero I got the following p-value:

Table 2: P-value (H0: beta = 0)

measure	value
p-value	0

Since my chosen level of significance was 5% I could conclude that the beta parameter was significant since the p-value was very close to zero and smaller than the chosen level of significance.

Residual Analysis

When analysing the residuals I could conclude that the top 5 countries which lost the most people due to COVID-19 are mainly middle-sized countries whereas the ones that lost the least amount of people due to COVID-19 are small states. This could be because making sure that all citizens follow the implemented measures is easier to check in smaller countries or because testing capacities are probably better in these states so as a result less people get infected so less people lose their lives.

Table 3: Top countries with largest positive and largest negative errors

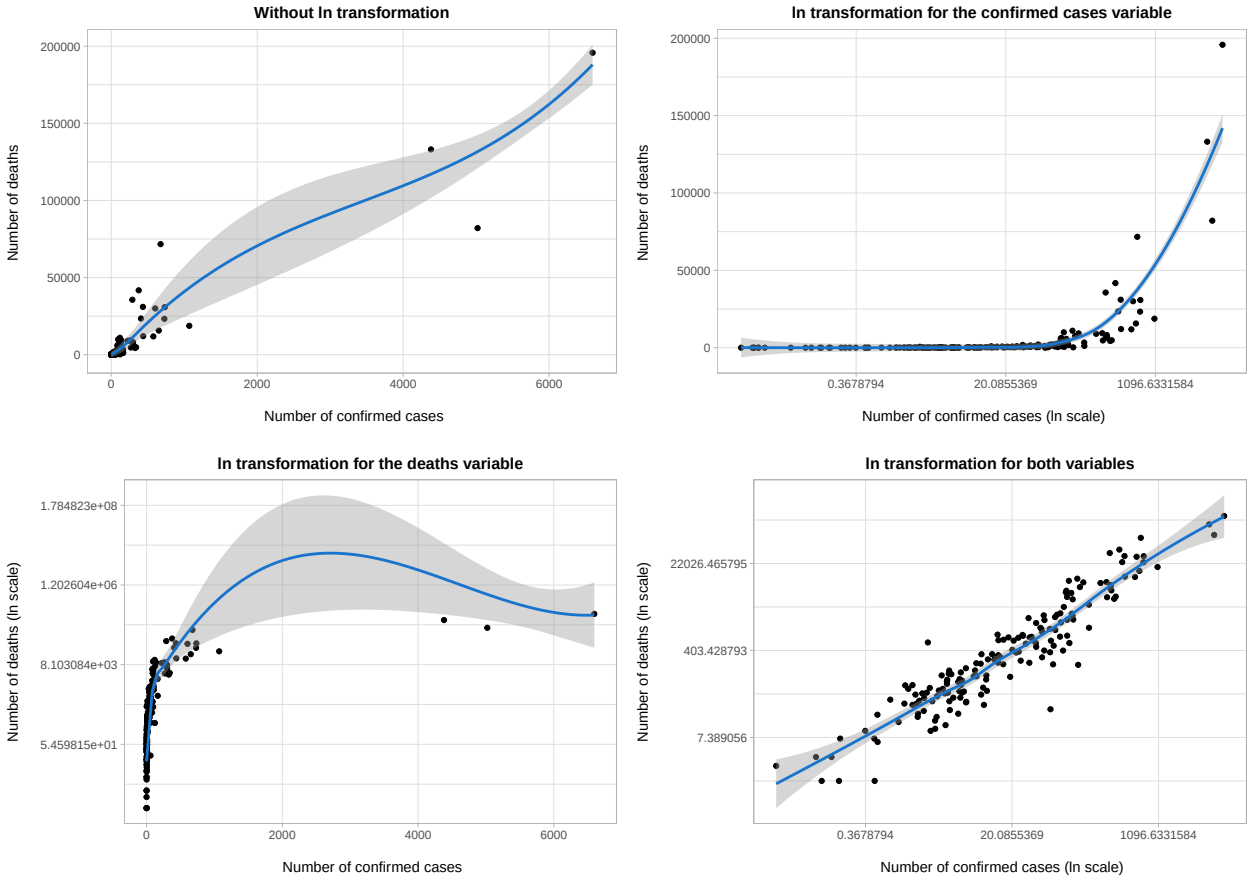
country	ln_deaths	linreg_y_pred	linreg_res
Mexico	11.179939	9.683718	1.496221
United Kingdom	10.639527	9.072931	1.566596
Belgium	9.203316	7.633798	1.569518
Italy	10.481027	8.800133	1.680895
Yemen	6.368187	3.617189	2.750998
Singapore	3.295837	7.112095	-3.816259
Qatar	5.337538	7.898808	-2.561270
Burundi	0.000000	2.102703	-2.102703
Bahrain	5.361292	7.184887	-1.823595
Maldives	3.496508	5.215138	-1.718630

Summary

In this analysis the aim was to model whether if a country has more confirmed COVID-19 cases it experiences a larger number of deaths as well. For the estimations I used data on confirmed cases and number of deaths for 170 countries for 15th September 2020. After transforming both variables with ln transformation and estimating four different models I chose a simple linear regression to model the pattern of association. The results showed that there is a positive, linear relationship between the number of confirmed cases and the number of deaths meaning that if a country has more confirmed cases then it has more deaths as well.

Appendix

Checking Ln Transformations Based on the histograms and the summary statistics both the dependent and the explanatory variable showed a skewed distribution with a long right tail therefore ln transformation might be necessary. Therefore I checked scatter plots with lowess and different ln transformations. Based on which I decided to transform both of my variables.

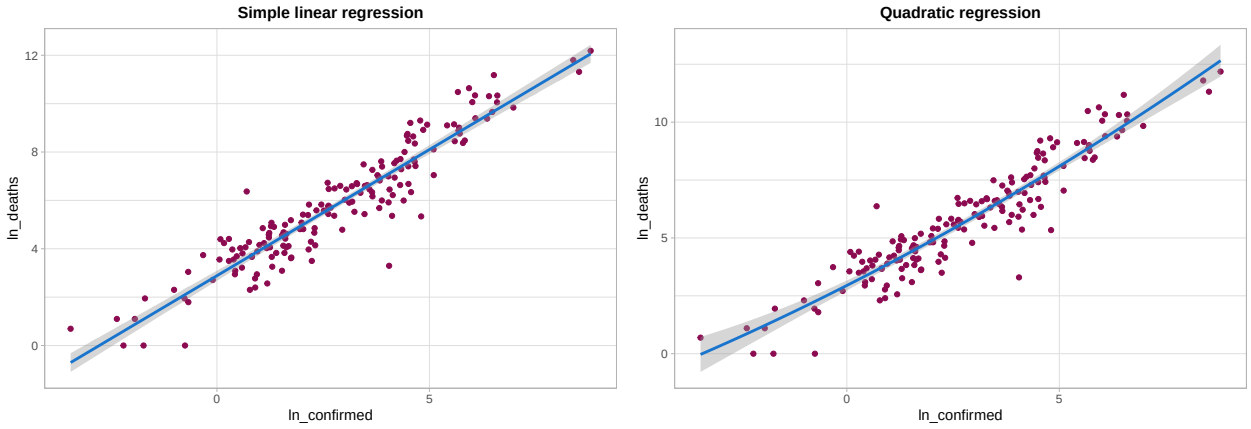


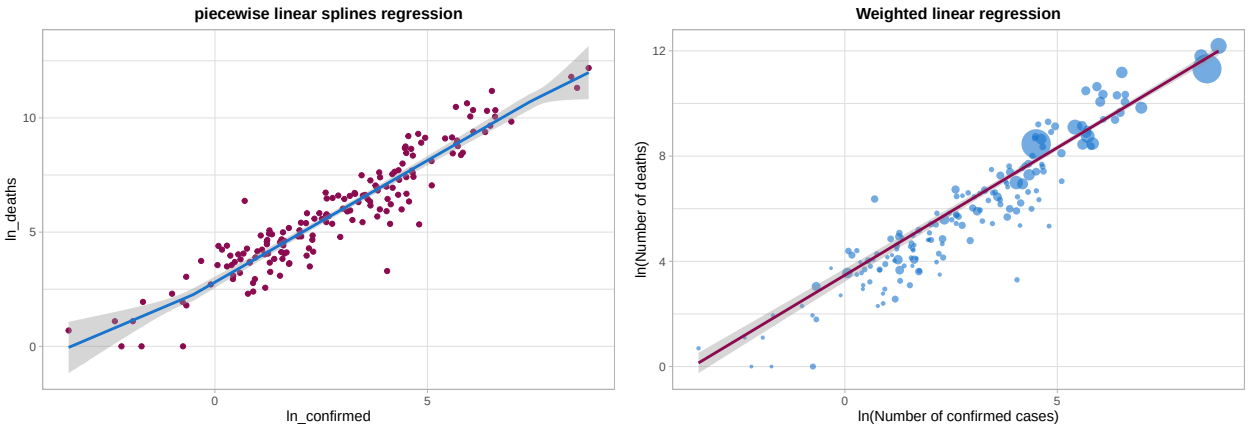
Estimating Different Models In total I estimated four models which were the following: simple linear regression, quadratic regression, regression with piecewise linear spline and weighted linear regression where I used the total population of countries in 2019 as weights. The comparison of their features can be seen in the table below. While their scatter plots are included below the comparison table.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.89*	2.95*	2.66*	3.47*
	[2.68; 3.09]	[2.73; 3.16]	[2.26; 3.07]	[2.81; 4.14]
ln_confirmed	1.04*	0.93*		0.97*
	[0.98; 1.10]	[0.81; 1.05]		[0.84; 1.10]
ln_confirmed_sq		0.02*		
		[0.00; 0.04]		
lspline(ln_confirmed, cutoff_ln)1			0.79*	
			[0.37; 1.20]	
lspline(ln_confirmed, cutoff_ln)2			1.07*	
			[0.99; 1.14]	
lspline(ln_confirmed, cutoff_ln)3			0.92*	
			[0.45; 1.40]	
R^2	0.88	0.88	0.88	0.92
Adj. R^2	0.88	0.88	0.88	0.92
Num. obs.	170	170	170	170
RMSE	0.86	0.85	0.86	4443.24

* Null hypothesis value outside the confidence interval.

Based on the comparison table it can be concluded that the different models grasp the pattern of association between the variables almost equally well since the R^2 values are the same for the first three models and last one is also very close to these. As for the confidence intervals of the parameters only one - the beta parameter of the squared variable in the quadratic regression - contains zero which means that the rest of the coefficients are significantly different from zero. Another important finding is that all the coefficients are positive indicating a positive relationship between the variables which proves the intuition that if there are more confirmed cases in a country then there are probably more deaths as well. As for the piecewise linear splines regression we can see that the coefficients are not very different from each other indicating that the slope of the regression line does not have large breaks.





When looking at the scatter plots of the estimated models we can see that they capture the pattern of association well. The reason I chose the simple linear regression was that it is very straightforward to interpret and that it performed very well compared to the other models. Furthermore, the relationship between the dependent and the explanatory variable is more of a linear one therefore choosing the quadratic or the piecewise linear splines model would have added an unnecessary degree of complexity.