09082020_COVID_analysis

Tamas Stahl

29/11/2020

Subject of the Analysis

My task was to prepare an analysis on the pattern of association between the number of confirmed cases and deaths of COVID-19 on 8th September, 2020. The data of the COVID related deaths and confirmed cases were collected from a publicly availabe github repo of John Hopkins University, where there is data for any specific date since the start of the pandemic.

The data was processed, cleaned and finalized with the provided codes from the teacher. My additional cleaning step was that the zero values for registered death were not included due to the fact that the log of 0 is -Inf, which could not be interpreted in our case. Some people could also argue that excluding these values might alter the outcome and we should replace those values with a really small number, but I found it more meaningful to exclude them than assign a random low number instead of zero.

An additional note, during the cleaning of the data in order to interpret our numbers more easily, as scaling we used the population divided by a million. So when we are interpreting the results the population will be in million people for population. The number of registered deaths and cases remained intact.

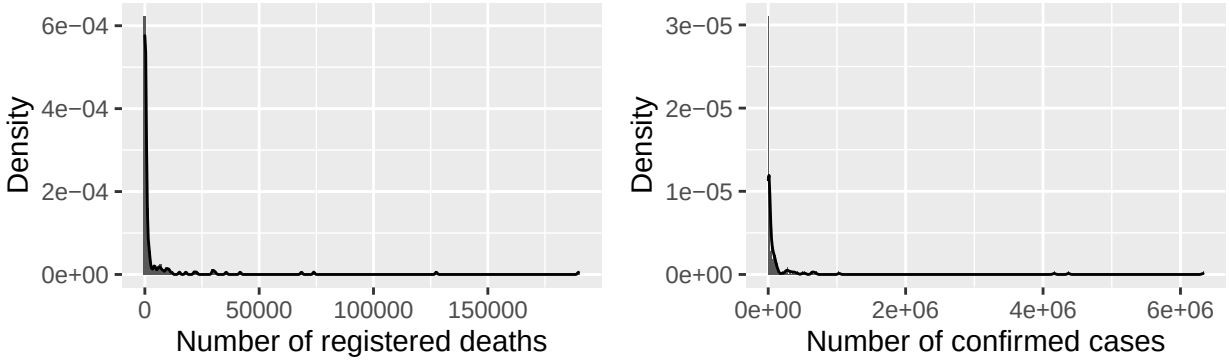

Figure 1: Distribution of deaths and confirmed cases of COVID-19

variable	mean	median	min	max	skew	NumberOfObservations
Number of registered deaths	5275.76	262.00	1.00	189660.00	6.75	170
Number of registered cases	162107.78	13194.50	31.00	6330007.00	7.22	170

Table 1: Summary statistics

We could see from either the histograms as well as the summary statistics that these are skewed with a right tail and that both the deaths and confirmed cases are distributed lognormaly, hence we could argue that log-log transformation will be the best option.

Model chosen

As previously mentioned, I excluded the zero value deaths for the reasons detailed above. Table 1 shows the summary statistics of the confirmed cases and deaths of COVID-19 on 8th September 2020.

After checking the level-level, log-level, level-log and log-log transformations (please refer to the Appendix) my chosen model would be log-log as in our case the percentage change of number of registered death and registered case would be more informative than the other options. Additionally, taking the log of both variables would make the association more linear as seen in the attached figure.

The substantive reasoning for taking log-log transformation would be that it is easier to interpret than a level-level or log-level transformations as both variables will be in percentage change. In case of log transformation the zero values need to be dealt with as those could not be interpreted.

The statistical reasoning for taking the log-log interpretation would be that the R2 is comparatively high and captures the variation well. Furthermore, it makes the pattern closer to linear.

After choosing the log-log transformation we will start to make models, I chose weighted OLS as my model (Model 4 of the Appendix), which was: $ln(death) = \alpha + ln(confirmed), weights = population$, as for scaling it would be useful to see the population as well.

In our case in Model 4, $\alpha = -3.32$ and $\beta = 0.98$, where α is hard to interpret. However, $\beta = 0.98$ means that if the registered cases are 1 percent higher then the average number of registered deaths is 0.98 percent higher as well.

As my task was to analyze the number of registered cases and deaths of COVID-19 on 8th September 2020, I did not use the per capita value of these. Therefore, I used the population as weight in my chosen model.

Hypothesis testing

The hypothesis testing was conducted based on whether the estimated β parameter is significant at 5%.

$$H_0 : \beta = 0 \quad and \quad H_A : \beta \neq 0$$

Below we could see the result of the hypothesis test, that resulted in a a 95% CI of [0.85 1.11]. This means that H_0 could be rejected with 95% confidence level and that β is significant at that level, hence, there is a positive pattern of association between the registered deaths and confirmed cases of COVID-19. From the two-sided hypothesis the p-value is almost 0, which is lower than 0.05, so we could reject H_0.

variable	estimate	std.error	pvalue	conf.low	conf.high
ln_confirmed	0.98	0.06	0.00	0.85	1.11

Table 2: Hypothesis test

Analysis of residuals

Table 3 shows the 5 best performing countries, where less people died than projected. In order to find out what might have been the reasons for this "over-achievement", people must be informed regarding those countries. In our list of 5 there are really wealthy countries like Singapore, Qatar, etc. where the health care system might be more developed and more people could be saved. Another reason could also be that the data is not entirely correct, there are regions where not all the COVID-19 related deaths are reported, therefore, there might be missing values.

country	ln_death	reg4_y_pred	reg4_res
Bahrain	5.31	7.38	-2.08
Burundi	0.00	2.69	-2.69
Maldives	3.37	5.55	-2.19
Qatar	5.32	8.12	-2.80
Singapore	3.30	7.39	-4.09

Table 3: The 5 best best performing countries

Table 4 shows the 5 worst performing countries with the largest positive errors, including Belgium, Italy, Mexico, UK. Keep in mind that these numbers are weighted with the population. The reason behind the "poor performance" of these countries might be that COVID-19 infected people in these regions first. Therefore, there was no best practice, doctors and people did not know how to treat this virus. Another reason for this performance could be that the health care system is not that developed or even overcrowded, like in Italy and Spain during the spring.

country	ln_death	reg4_y_pred	reg4_res
Belgium	9.20	7.82	1.38
Italy	10.48	8.94	1.53
Mexico	11.13	9.76	1.38
United Kingdom	10.64	9.18	1.46
Yemen	6.36	4.11	2.25

Table 4: The 5 worst best performing countries

Executive summary

In this assignment my task was to analyze the pattern of association between the number of confirmed cases and registered deaths of COVID-19 on the 8th of September 2020. In order to analyze the data I used weighted regression, where the weights were the population of each country in million people. During my analysis I calculated that there is a positive pattern of association between the number of registered deaths and confirmed cases of COVID-19.

The quality of the data is questionable as all these data are collected from the authorities of each country, and as I mentioned earlier, some countries my report more cases or deaths than there actually is or even less cases or deaths. The model could be strengthened by more variables like the location (urban or rural areas), age of infected, numbers of days required to recover and whether there were symptoms.

To summarize my findings in Model 4 (weighted-OLS), we could say that if the confirmed cases are 1 percent higher than the average number of registered deaths is 0.98 percent higher.

Appendix

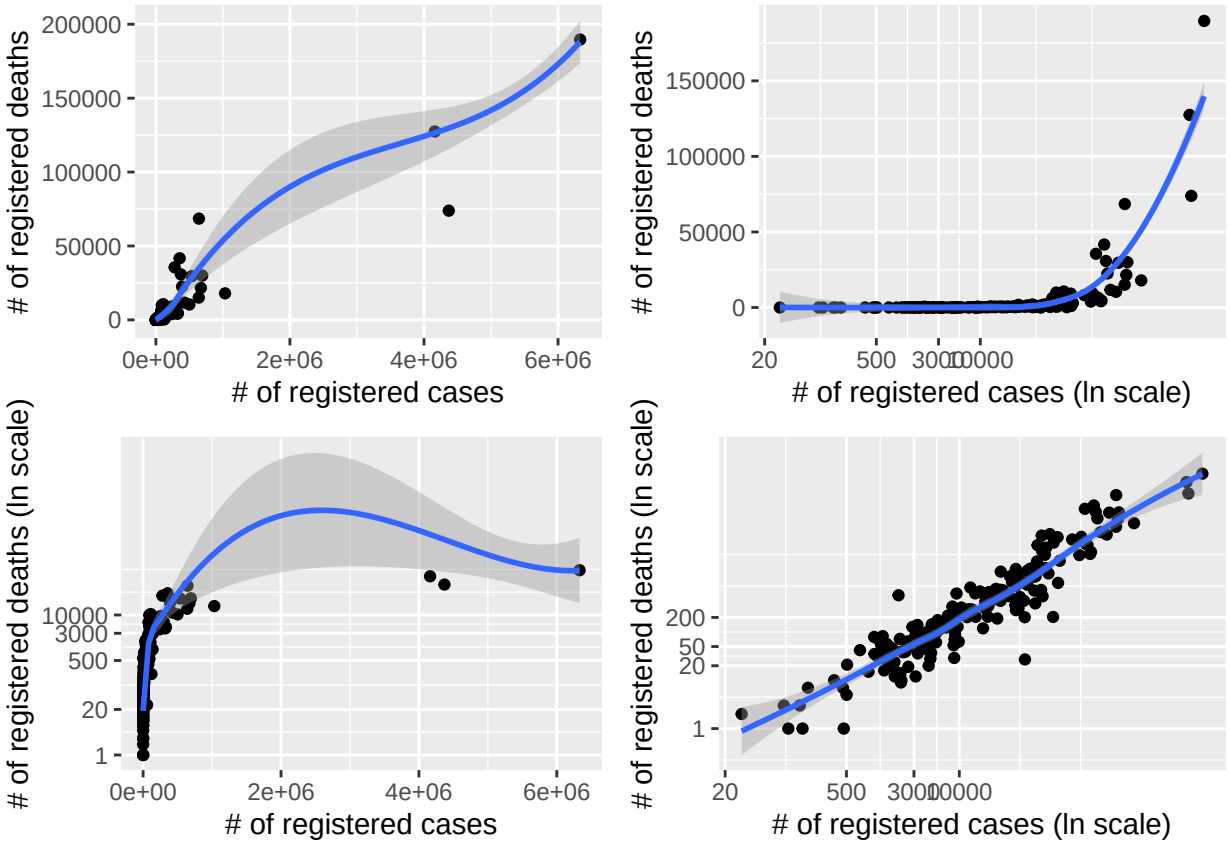

Figure 2: Scatter plot for the level and log transformations

Models description

Below you could find scatter plots for each model (1 through 4) and a summary table of each model with intercept, beta, R2, etc. Our 4 models were the following: Model 1: Simple linear regression, Model 2: Quadratic, Model 3: PLS, Model 4: Weighted OLS.

We could say that all the models capture the positive pattern of association between the variables, although I chose Model 4 as the R2 was the highest with 0.92 and the population weight is important to give depth for our analysis as we used only the number of registered deaths and confirmed cases. From a statistical point of view this model is also beneficial as the standard error and CI is smaller, resulting in more precise prediction.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−4.32*	−2.63*	1.49	−3.32*
	[−4.90; −3.73]	[−4.37; −0.90]	[−4.46; 7.44]	[−4.81; −1.84]
ln_confirmed	1.05*	0.68*		0.98*
	[0.99; 1.10]	[0.33; 1.03]		[0.85; 1.11]
ln_confirmed_sq		0.02*		
		[0.00; 0.04]		
lspline(ln_confirmed, cutoff_ln)1			−0.22	
			[−1.58; 1.14]	
lspline(ln_confirmed, cutoff_ln)2			1.04*	
			[0.70; 1.38]	
lspline(ln_confirmed, cutoff_ln)3			1.06*	
			[0.99; 1.12]	
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	0.87	0.86	0.87	4.53

* Null hypothesis value outside the confidence interval.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

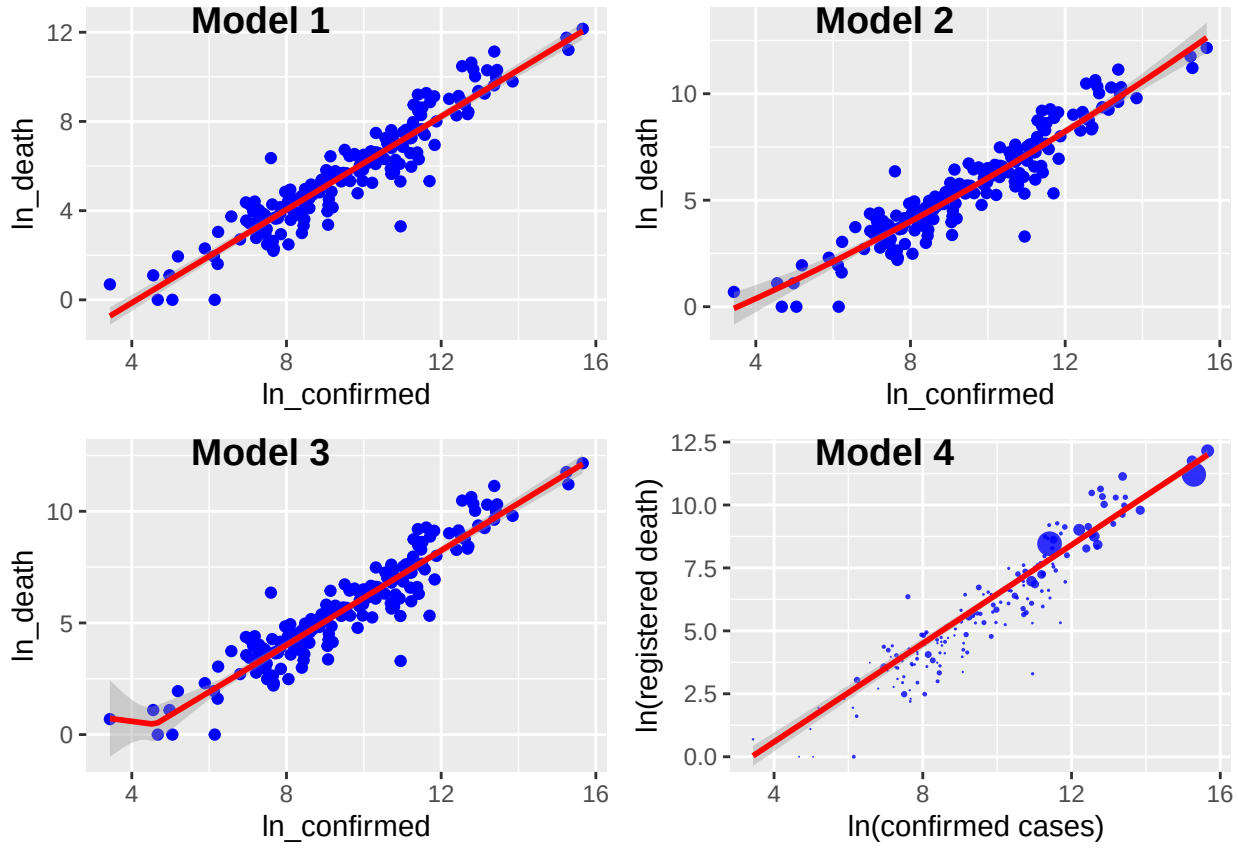

Figure 3: Scatter plots for visualization of models