DA2_Assignment1

Bruno Helmeczy

29/11/2020

Executive Summary

This report analyses the pattern of association between confirmed Covid-19 cases, and the respective number of Covid-19 related deaths countries experienced, as of 20th October, 2020. Visual inspection (available in the appendices) highlighted a log-log linear association pattern between deaths and confirmed cases, leading to estimation and comparison of 4 log-log linear regression models.

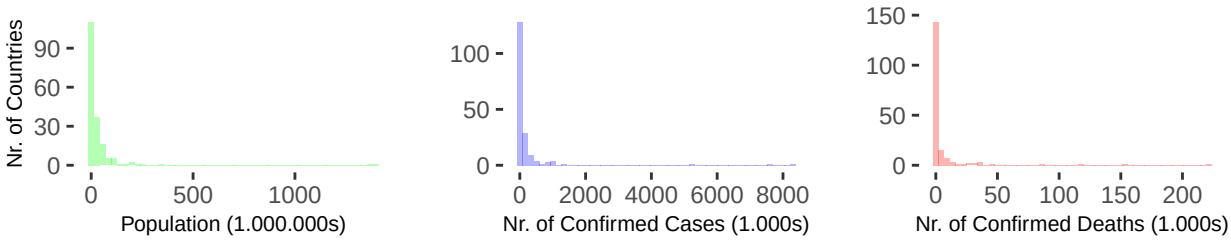
The log-log population-weighted linear regression model was chosen, boasting statistically significant regression parameters, and explaining 92.9% of variation in Covid-19 related deaths among countries with variation in confirmed Covid-19 cases (adjusted R-squared). According to this model, people observing 10% larger number of confirmed cases in a country, on average can expect to observe 9.5% larger number of deaths. The core message is two-fold: 1st, the association pattern is close to 1 for all tested models, where 1 implies a constant ratio between case & death figures i.e. virus deadliness, and 2nd the 0.05 deviation in the slope parameter, points to how other factors could also be useful in predicting covid-19 related deaths.

Indeed, age and health were shown to influence covid-19 infected individuals' probability of dying, therefore including such demographic variables at the national level could improve the chosen models' predictive power. On the other hand, some national governments may choose to "massage" statistics to avoid losing face, which if proven, undermines the validity of results & conclusions derived in this report.

1) Introduction: This report investigates how well confirmed covid-19 cases explain covid-19 related death figures. Thus, the research question is: "How many deaths can be expected, knowing how many confirmed cases there are in a country?" The population is all Covid infected people worldwide, from which Covid-related statistics where sampled in cross-sectional format as of 20th October, 2020, aggregated to country-level. Data Quality issues arise in terms of Validity & Reliability. It takes 7 days to show symptoms once infected, then once the decision arises to test (which may take additional time, depending on nations' size and healthcare infrastructure), people may shrug it off, and if they decides not to test, he/she would not be registered.

Indeed, researchers claimed observed covid-19 figures are rather an estimate of the situation 10-14 days earlier. Additionally, a country's size may be a source of bias: 1st due to faster being able to implement reactionary measures to manage virus spread and especially deaths. 2nd, the difference between true virus related cases and deaths versus the currently available figures constitutes a larger proportion of true virus related figures, causing a stronger negative bias, if measured in percentage terms.

2) Filtering & Scaling Observations: Covid-19 related data was collected from Johns Hopkins University, aggregated to country-level, scaled by 1.000, and merged with national population data scaled by a 1.000.000. Countries whom did not report confirmed number of cases or population figures, were dropped (23), together with countries reporting 0 deaths (12) (treated as outliers), netting 170 observations.



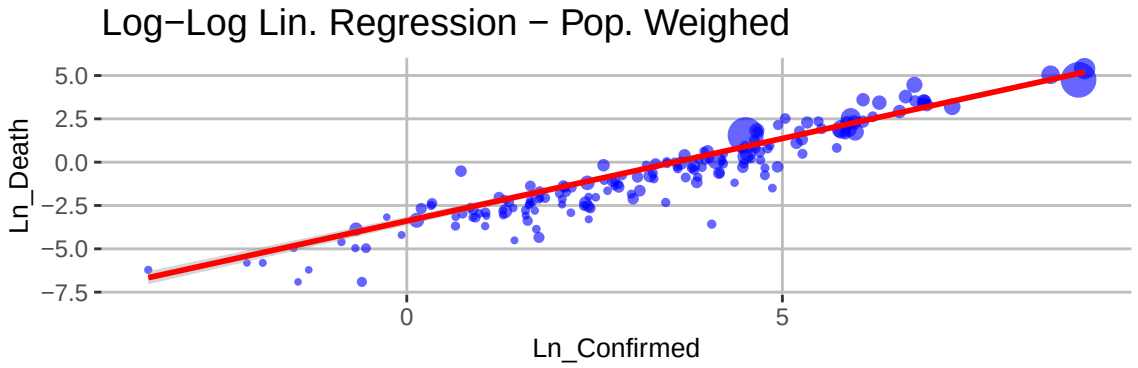
Variable	Min	1st IQR	Median	3rd IQR	Max	Mean	StDev	Skewness
Covid-19 Cases (1000s)	0	3.73	18.75	101.14	8295.14	224.10	932.02	7.30
Covid-19 Deaths (1000s)	0	0.06	0.30	1.66	221.27	6.18	23.36	6.63

3) Histograms & Summary Stats: Data sets' variables of interests' histograms resemble a power-law distribution, i.e. frequencies decreasing by a similar ratio interval-to-interval, with some positive outliers. Skewness can be observed in all cases (ranging 6.6-7.3), also by variable means being above 75th percentiles.

4) Ln Transformations: A theoretical power-law distribution approximates both observed variables well, thus logarithmic transformations created approximately normal distributions for both number of confirmed cases & deaths. Scatter-plots in turn show visibly strong log-log linear association, meaning percentage changes in confirmed cases numbers can be strongly associated with percentage changes in deaths. Substantively, population is power-law distributed, and pandemics are declared based on a given percentage of population becoming infected, while a viruses deadliness is measured in terms of the percentage of infected people it kills, thus both variables being linked to the population distribution. Another disease measure is infectiousness, based on the growth-rate of infections, as every sick person is able to infect approximately the same number of people. Given both virus growth-rate, and deadliness are relatively constant rates, being able to observe a linear association pattern between percentage increase in deaths and percentage increase in cases is expected.

5) Model Choice & Interpretation: The chosen regression model is the Log-Log Weighted Linear Regression, weighed by countries' populations. Please see the models' graph and formula below. For an argument on model choice, please visit the appendices.

Pop. Weighed Log-Log Linear Regression equation:
ln(Death) = -3.39 + 0.95 * ln(Confirmed Cases)



Though the intercept (alpha) parameter is not meaningfully interpretable with a log-log regression model, its slope (beta) parameter tells, that for people across all countries, a +10 percent change in confirmed Covid-19 cases is associated with +9.5 percent change in Covid-19 related deaths, on average. That is 0.8 percent less than the simple log-log linear regression model slope, highlighting that smaller countries experience disproportionately low number of deaths relative to their case numbers, which if not weighed, gives the perception that death figures' growth rate is larger than case numbers'.

6) Beta Parameter Hypothesis Test: The 1st necessity to establish any meaningful pattern association between the number of confirmed cases and deaths, is testing whether beta, the slope parameter is equal to zero. Thus, the null-hypothesis is Beta = 0, the alternative hypothesis being Beta not being equal to zero. The chosen significance level is 0.1%, due to the realistic need for leaders to be absolutely sure of such an association for decision-making, meaning the maximum accepted probability of a false-positive is 0.001. For the test to satisfy this requirement, the t-statistic must be at least 3.1 distance from 0.

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	-3.3906	0.3692	-9.1843	0	-4.1194	-2.6618	168
Ln_Confirmed	0.9516	0.0624	15.2531	0	0.8284	1.0748	168

As can be seen from the table above, the beta parameters' (Ln_Confirmed) t value is 15.25, providing ample evidence to reject the null hypothesis, and accept the alternative hypothesis, that the beta parameter is not equal to zero, with 99.9% confidence. In fact, there is less than 1 in a million chance that accepting the alternative hypothesis would result in a false positive error.

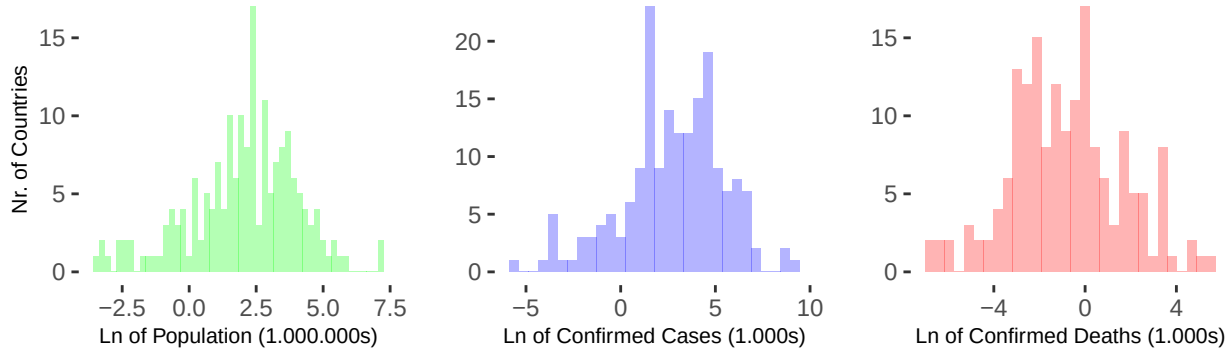
7) Residual Analysis: Best countries were defined as those with largest negative residuals versus respective model-predicted values, while on the flip side, Worst countries as those with largest positive residuals versus model predicted values. If one accepts countries' deviations from residuals as a measure of job quality nations' governments have performed in pandemic management thus far, the Singaporean leadership qualifies 1st worldwide, despite being the most densely populated nation, which is supposed to influence its' pandemic management very negatively.

Best	Ln_Deaths	Pred	Error	Worst	Ln_Deaths	Pred	Error
Singapore	-3.58	0.47	-4.05	Yemen	-0.52	-2.70	2.19
Burundi	-6.91	-3.96	-2.95	Mexico	4.46	3.04	1.42
Qatar	-1.50	1.24	-2.74	Italy	3.60	2.39	1.21
Sri Lanka	-4.34	-1.72	-2.63	Ecuador	2.52	1.40	1.11
Iceland	-4.51	-2.03	-2.48	United Kingdom	3.79	2.93	0.86

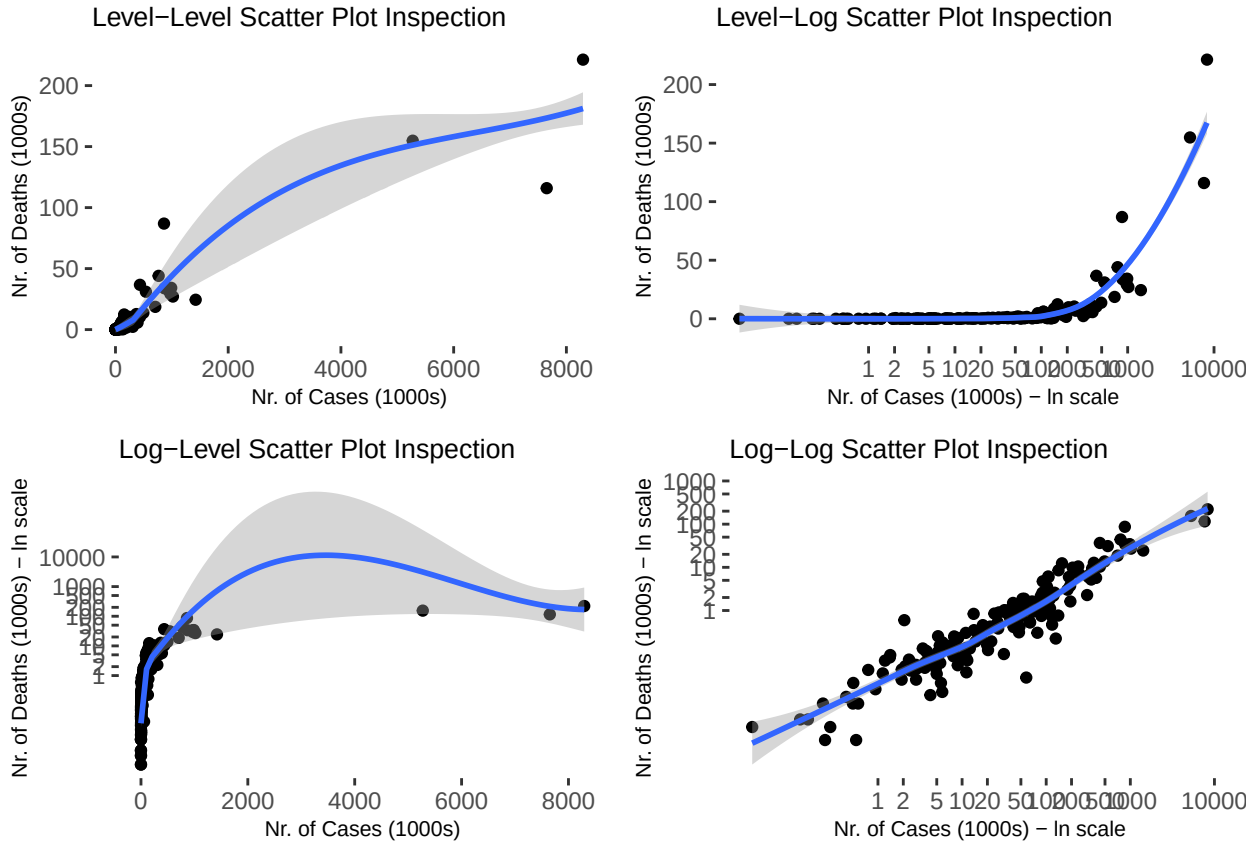
Iceland is the best-performing European country, while Yemen's pandemic deaths in light of their covid-19 cases are worst worldwide, and Italy was the worst performing country in Europe, after the UK. Also, best performing countries' absolute residual values are approximately twice to that of worst performing countries, possibly due to large-population countries, e.g. China boasting positive errors (ranking 159th), negatively influencing the models' absolute-value beta parameter. On a final note, Hungary ranks 114th worldwide.

Appendices

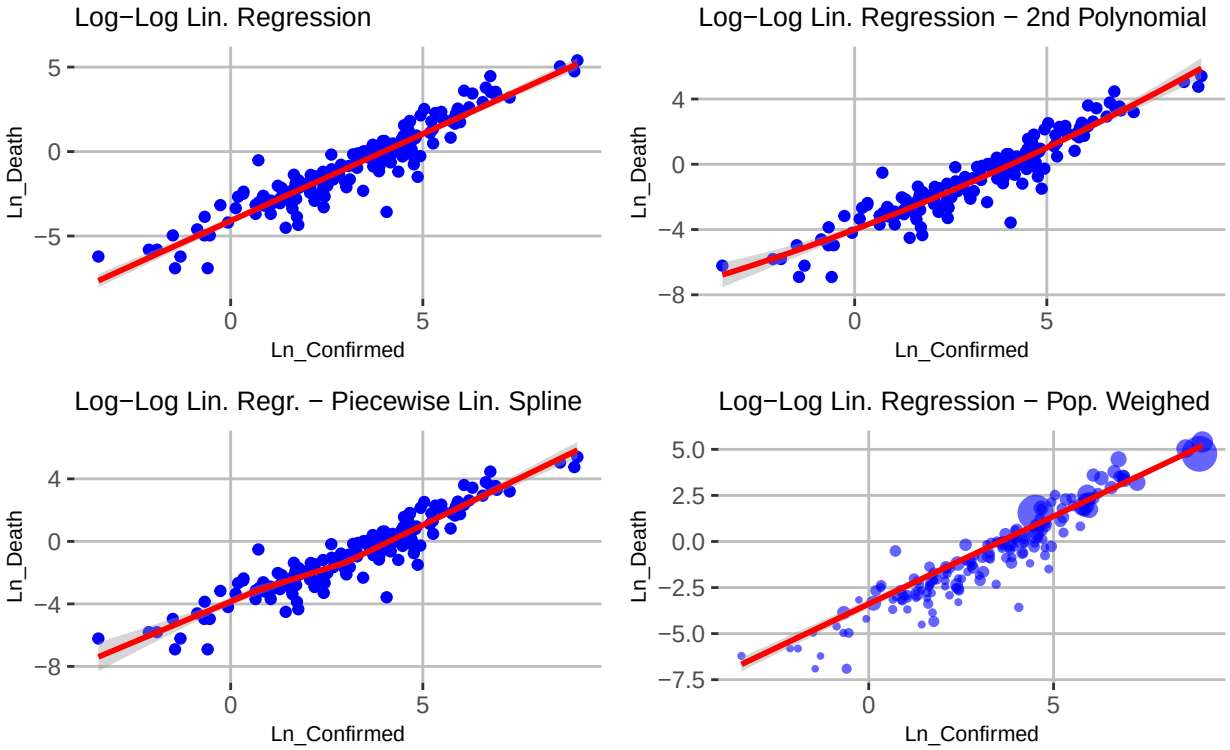
1) Variable Ln-Transformations



2) Variable Scatter-Plots: Level-Level, Level-Log, Log-Level, Log-Log



3) Model Visuals



4) Model Selection & Statistical Summaries

Based on scatter-plot visuals, both variables of interest were log-transformed, from which 4 regression models were estimated: 1st a Log-Log Linear Regression, 2nd a Log-Log 2nd degree Polynomial regression, 3rd a Log-Log Piecewise Linear Spline Regression, and 4th a Log-Log Weighted linear Regression, weighed by countries' respective populations. Using log-log models, their interpretations are "a certain percentage change in Confirmed Cases being associated with a certain percentage change in Deaths". Overall, the 4 models perform well, each explaining at least 89% of variation in percentage change in Deaths, by the percentage change in Confirmed cases, while producing statistically significant parameters. The population weighed log-log regression stands out however, explaining 93% of variation in the outcome variable with variation in the explanatory variable. This model however, produced wider standard errors, causing its' parameters' sampling distributions to overlap with model 1, the log-log linear regression, both in case of the intercept and slope parameter. Hypothesis tests therefore, would not be able to reject with 95% confidence the hypothesis that these models are statistically the same.

The 2 models however, differ in their interpretation significantly. The simple log-log regression outputs show how much percentage more deaths can a country expect on average, give how much percentage more confirmed cases it has country to country, as opposed to the population weighted regression outputs showing how much percentage more deaths can people expect on average, given how much percentage more confirmed cases there are.

Substantively, this makes a significant difference due to the data quality issues outline in the introduction. Firstly, the complexity associated with managing the pandemic at a national level is different depending on country size, thus smaller countries' leaders arguably have an easier job managing infection rates once the virus is identified, as they are faster able to implement directives versus significantly larger countries. Thus it can be expected smaller countries are better able to minimize the number of deaths relative to the number of cases observed, skewing the true association pattern.

Secondly, the percentage difference between measured- and true virus-related figures is greater in smaller countries, and weighing smaller-country observations equally to large ones (e.g. Liechtenstein versus China) exposes the model to measurement errors in smaller countries. For these reasons, despite their parameters' sampling distributions overlapping, my chosen model is the population weighted log-log regression model.

	Ln_Deaths	Ln_Deaths	Ln_Deaths	Ln_Deaths - Weighted
(Intercept)	-4.10 ***	-4.00 ***	-3.85 ***	-3.39 ***
	(0.12)	(0.13)	(0.15)	(0.37)
Ln_Confirmed	1.03 ***	0.89 ***		0.95 ***
	(0.03)	(0.06)		(0.06)
Ln_Confirmed_sq		0.02 **		
		(0.01)		
lspline(Ln_Confirmed, cutoff_ln)1			1.03 ***	
			(0.17)	
lspline(Ln_Confirmed, cutoff_ln)2			0.78 ***	
			(0.11)	
lspline(Ln_Confirmed, cutoff_ln)3			1.19 ***	
			(0.05)	
nobs	170	170	170	170
r.squared	0.89	0.89	0.89	0.93
adj.r.squared	0.89	0.89	0.89	0.93
statistic	1257.39	718.98	523.14	232.66
p.value	0.00	0.00	0.00	0.00
df.residual	168.00	167.00	166.00	168.00
nobs.1	170.00	170.00	170.00	170.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.