Cosmin Catalin Ticu (2001492) - Covid-19 Analysis Project

Executive Summary

This study, linked here on Github, aims to uncover any patterns of association between global Covid-19 cases and deaths. Data was gathered, cleaned and merged using John Hopkins University Covid-19 reports (as of 04.11.2020) and World Bank data on population (as of 2019). This study worked with ratio data, namely coronavirus cases and deaths per 1000 people as the chosen metrics. By employing logarithmic transformations of the data to reduce skewness, the study found a population-weighted OLS linear regression model to estimate the pattern of association between global Covid-19 cases and deaths per 1000 capita. The final model moves away from a country-centric regression, instead tailoring to citizens. The bias of this model is given by the weight, thus prioritizing countries with large populations. In fact, the most interesting finding of the study was that model coverage (R-squared) increases dramatically once we discriminate against low-population countries or countries with close to no reported deaths.

Introduction

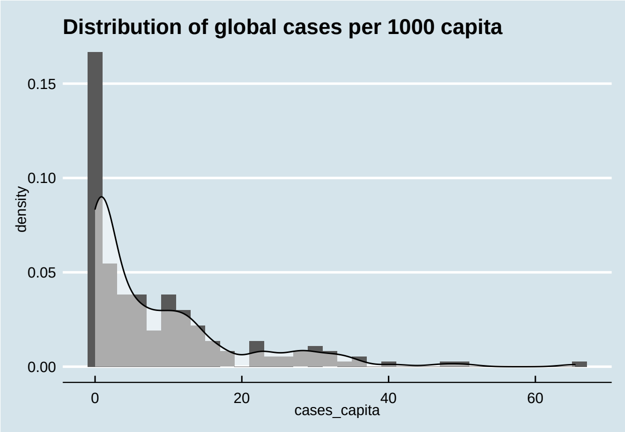
The aim of this study is to use Covid-19 data to explore a pattern of association between confirmed cases and confirmed deaths. Thus, the research question arises: "How do confirmed coronavirus cases per country affect confirmed coronavirus deaths per country?" The data gathered represents John Hopkins University Covid-19 global cases and deaths tracking as of the 4th of November 2020. The analysis takes the confirmed cases and the deaths and divides them by each country's population in order to work with ratio'd data. The population of this study is the entire world and the sample, containing close to all the internationally recognized territories of the world, can be considered as a representative batch. Many data quality issues are at play in this dataset, namely: questions of covid case tracking, improper death labels, confounding variables, improper cross-country comparisons and many more.

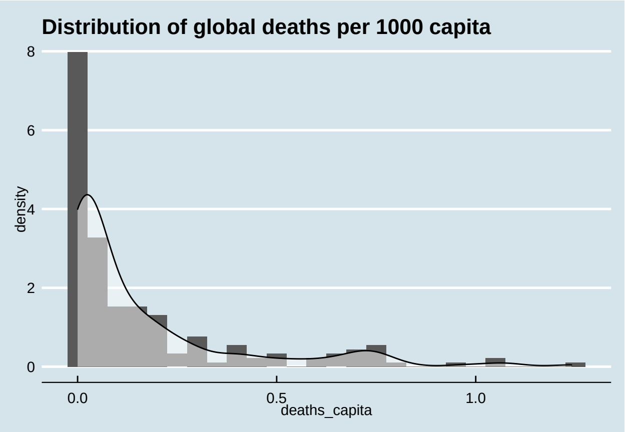
Selecting Observations

Only country instances not containing missing values for population, confirmed cases or confirmed deaths were kept. This meant that the sample size became 183 countries out of a possible almost 200 internationally recognized territories. As mentioned before, because the scope of this study is to get a per capita understanding of the pattern of association, each country's cases and deaths were scaled by population and then multipled by 1000 in order to signify coronavirus cases per 1000 capita and coronavirus deaths per 1000 capita respectively. This choice of scaling is done for ease of later interpretation. Example (in the case of a log-log regression): it is easier to say that with every 1% increase in coronavirus cases per 1000 people, we observe an increase of <...>% in coronavirus related deaths per 1000 people on average, rather than say this for a per capita scale (interpretation does not seem linguistically logical). Lastly, population was scaled to reflect millions (ex. Afghanistan has a value of ~38, which means ~38,000,000 people).

Histograms, Density Plots & Summary Statistics

From the histogram distributions of the x (cases per 1000 capita) and y (deaths per 1000 capita) variables as well as the population variable (see Appendix A.1 & A.2) which will later be used in model refinement, we can see a define right skew for all variables. This means that a few extreme cases such as China, India and US really skew the data distribution towards the right for both cases and deaths per 1000 capita. This skew happens when the mean is larger than the median.





Based on the summary statistics, we can see a definite right skew, with very high positive skewness values. With a high standard deviation for both variables as well as a large range of possible values, it is paramount to transform these variables to a more normal distribution.

Table 1: Summary statistics for cases and deaths per 1000 people on a country basis

variable	n	Mean	Median	Std	Min	Max	Skew
Cases per 1000 capita	183	8.656	4.255	11.15	0.003348	65.4	1.949
Deaths per 1000 capita	183	0.1653	0.05332	0.2459	0	1.24	2.059

Data Transformation

With data skewness observed, the variables of cases and deaths per 1000 capita need to be rescaled. See Appendix A.3 for graphs of logarithmic scale transformations with a non-parametric lowess regression model fitted. In conclusion, taking the natural logarithm of both variables is needed.

Substantive Reasoning

- Using only level models is not beneficial for interpretation, as our interest could also be to model percentage changes between cases and deaths per 1000 capita;
- Zero values need to be manually dealt with in the case of deaths per 1000 capita (logarithmic calculation in R cannot discern zero values and change them);
- It makes to either take level-level or log-log, not different scales between the variables as they are measured in the exact same way;

Statistical Reasoning

- Modeling non-linearity for extremely affected coronavirus countries and for isolated territories (such as Marshall Islands) would overfit any model due to complexity;
- Due to negative logarithmic values in the data, countries with 0 deaths cannot just be manually given a log value of 0;
- A workaround was found in the case of the zero values. The consensus online around data analysts is to add a small amount to each observation. The main suggested amount was half of the smallest non-zero value. See Appendix for detailed substantive and statistical explanation (see Appendix A.4);
- Both distributions have long right tails, and log transformation can make large differences smaller, but they also make small difference much larger;
- Taking log of both variables makes pattern closer to linear;
- Heteroskedasticity appears reduced as opposed to the level-level model and the interest is to reduce it even while using robust models (R-squared benefits);

Model of Choice & Hypothesis Test

See Appendix A.6 for model comparison discussion and reasoning for choice of weighted log-log OLS model by population (R-squared discussion, coefficient discussion, etc.). Also, see the standalone output html file (requires download to local machine) here for the model comparison statistics.

The final model (see Appendix A.6.2.3 for plot) takes the following form:

Linear.weighted.OLS.log.log.model.of.cases.and.deaths.per.1000.capita
ln_deaths_capita = -3.66 + 0.89 * ln_cases_capita

From the above model we can discern the alpha is -3.66. This theoretically means that the average deaths per 1000 people in logarithm is equal to -3.66 when the cases per 1000 people is equal to 1 (or that the cases per 1000 people in logarithm is equal to 0). As usual per log-log models, the alpha is meaningless in this case and its interpretation does not make sense. The measure of interest is the beta, which argues for all people across the globe (main difference to other models that are not weighted by population) that with every 10% increase in confirmed Covid-19 cases per 1000 people, we observe a 8.9% increase in confirmed Covid-19 deaths per 1000 people, on average. This can also be interpreted that on average global Covid-19 deaths do not increase as fast as global Covid-19 cases, both of which per 1000 capita.

The hypothesis test chosen in this case is:

H0: beta = 0 (i.e. there is no pattern of association between cases and deaths per 1000 capita)

Ha: beta neq 0 (i.e. there is a pattern of association between cases and deaths per 1000 capita)

The significance level of choice is that of 95% as it is the standard in beginner statistics. This value corresponds to a t-value of 1.96, which needs to be benchmarked to the model below.

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	-3.6570719	0.1587660	-23.03435	0	-3.9703422	-3.343802	181
ln_cases_capita	0.8857432	0.0770638	11.49364	0	0.7336842	1.037802	181

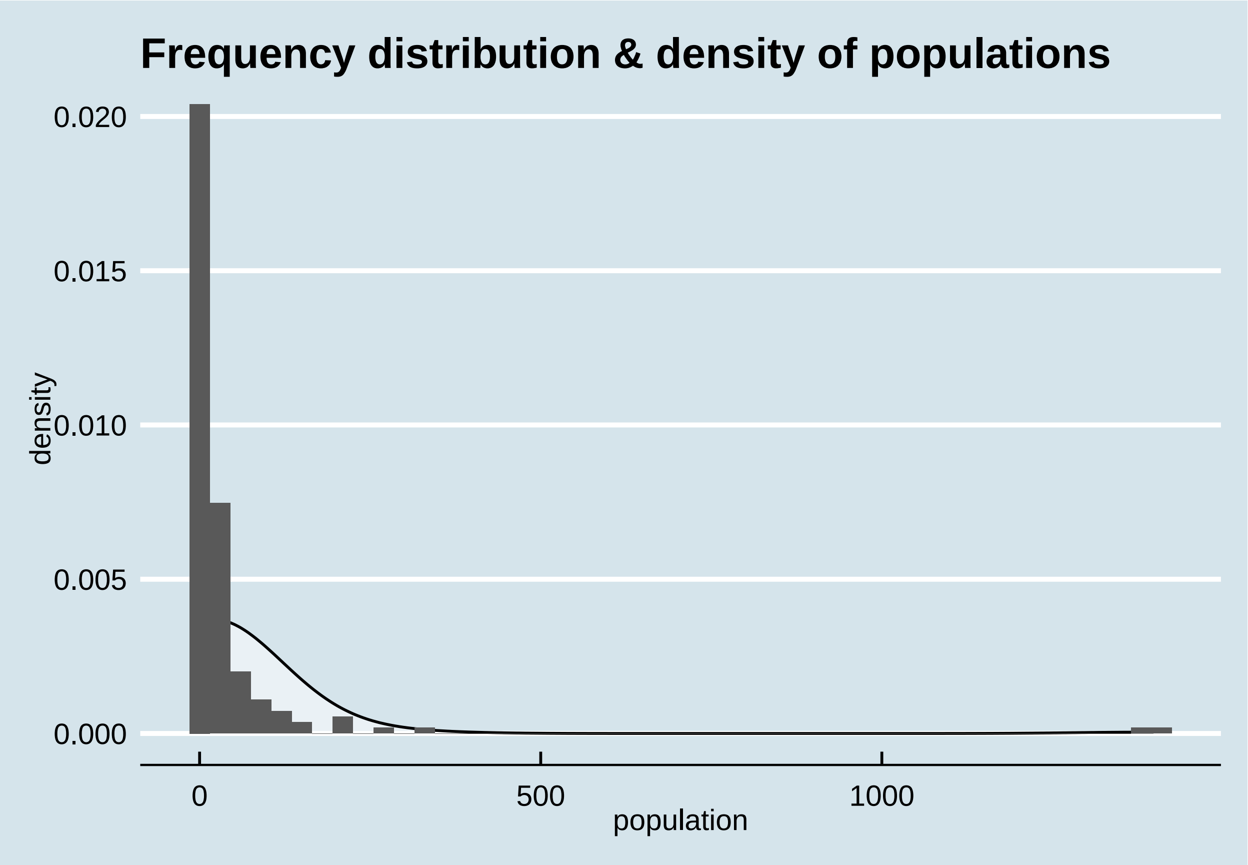
With a t-value for the beta coefficient of ~11, we are much above the threshold t-value of 1.96. The final model's (with 181 degrees of freedom, showing lower complexity than the rest) p-value is much below 0.05 (R has trouble displaying very small numbers) which means that we can reject the null hypothesis that there is no pattern of association between global Covid-19 cases per 1000 capita and global Covid-19 deaths per 1000 capita. The confidence interval states that we can be 95% confident that the true value of the slope parameter (interpreted above in tens of percentages as 8.9%) lies between 0.73 (7.3% change) and 1.03 (10.3% change).

Residual Analysis

The residual analysis (see Appendix A.5 for top 5 tables) uncovers top overestimated observations (by the top negative residuals) and the top underestimated observations (by top positive residuals). From the histogram of residuals (see Appendix) we can discern that some extreme negative values do exist. As suspected in the beginning, because this model discriminates against countries with low populations and subsequently (but this is applicable to all models) against extreme cases of isolated countries or countries with extremely low deaths, it estimates higher ln_deaths_capita for these special cases. Because this is a global model (pertaining to capita, not countries' capita) for instance it overestimates Bhutan's deaths (globally-praised for handling the pandemic) and underestimates Yemen's deaths (famished country in poverty with poor handling of pandemic).

Appendix A (tables, charts, graphs & discussion)

A.1 Population Distribution

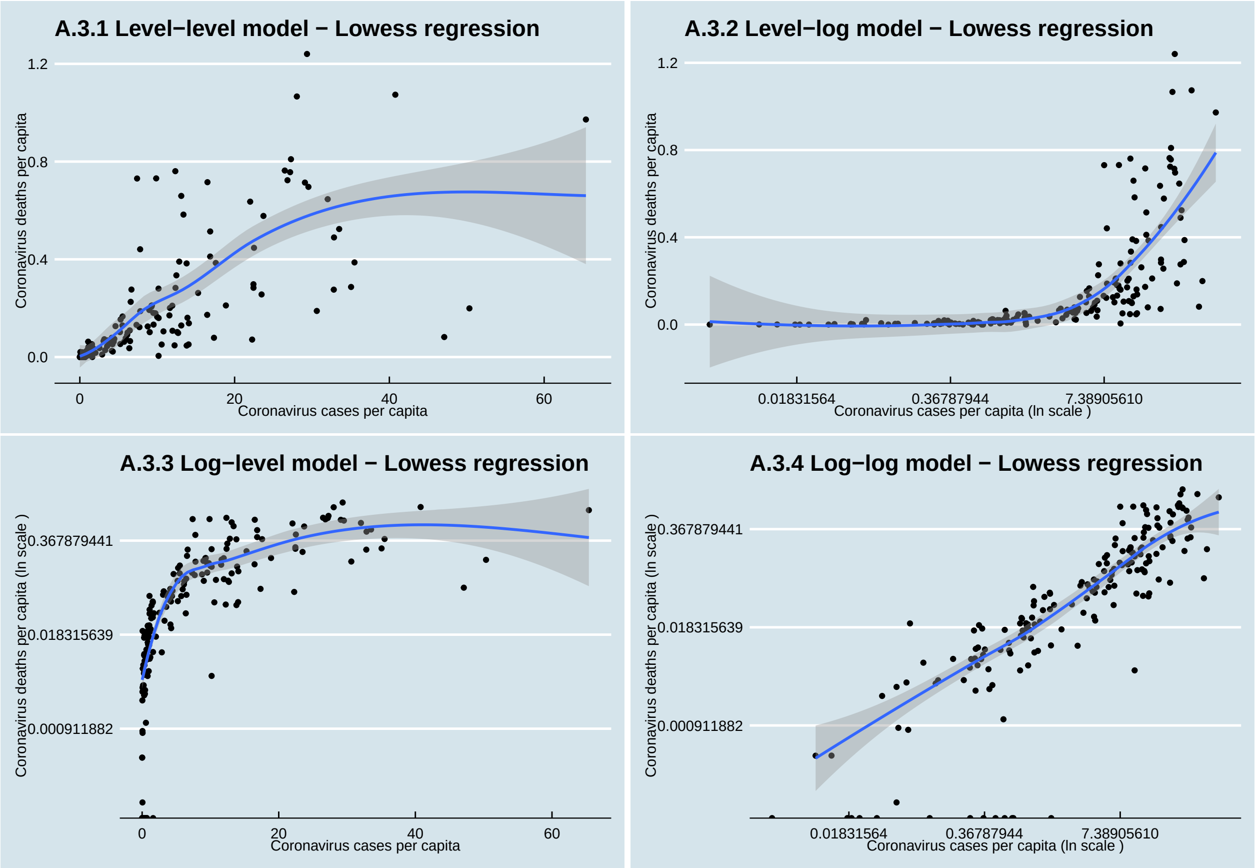


A.2 Population Summary Statistics

Table 4: Summary statistics for country populations

variable	n	Mean	Median	Std	Min	Max	Skew
Country populations	183	41.52	9.746	149.1	0.03386	1398	8.094

A.3 Transformation Investigation



A.3.1 Level−level model − Lowess regression

A.3.2 Level−log model − Lowess regression

A.3.3 Log−level model − Lowess regression

A.3.4 Log−log model − Lowess regression

A.4 Conclusion to dealing with zero values

It is worthwhile to manipulate the data using the half of the smallest non-zero values to keep what would otherwise be "-Inf" log values. This means that we are able to keep the 13 countries/territories that have reported no deaths from the coronavirus.

A.4.1 Substantive

This way we do not remove any more variables that contain (supposedly) correctly tracked data

A.4.2 Statistical

This is the mathematical equivalent of adding half of a death to each of the countries for every 11 million citizens (population of Burundi, country taken as benchmark for lowest deaths per 1000 people). One disavantage to this method is that it essentially adds about 63 deaths to a country like China, which in proportion to its population and coronavirus deaths is almost unnoticable. We are still talking about artificially inputting the death of 63 people. Some ethical considerations should be applied.

A.5 Residual Analysis Resources

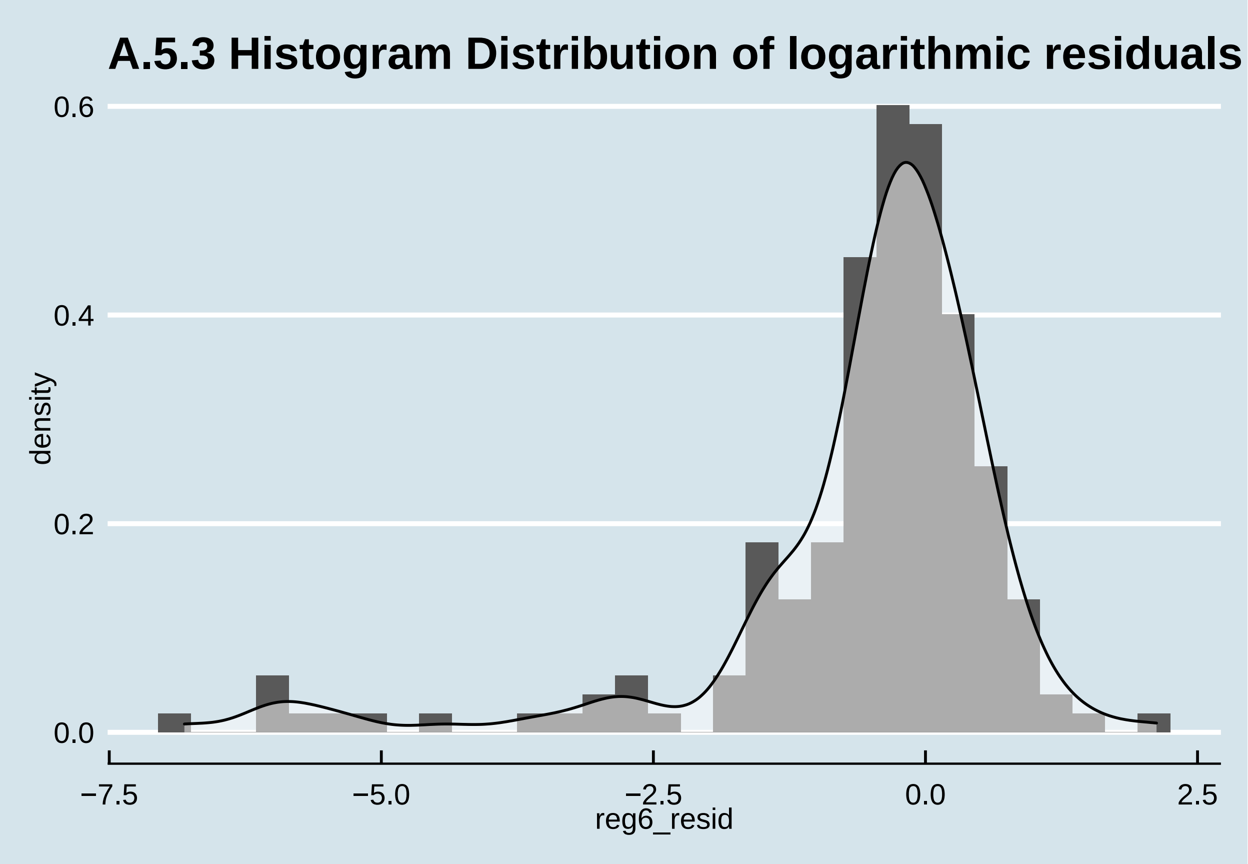
Table 5: A.5.1 Top 5 countries by negative residuals

country	ln_deaths_capita	reg6_y_pred	reg6_resid
Bhutan	-10.04591	-4.327443	-5.718462
Dominica	-10.04591	-3.977687	-6.068218
St. Lucia	-10.04591	-4.148108	-5.897797
St. Vincent and the Grenadines	-10.04591	-4.001035	-6.044870
Seychelles	-10.04591	-3.236245	-6.809661

Table 6: A.5.2 Top 5 countries by positive residuals

country	ln_deaths_capita	reg6_y_pred	reg6_resid
Bolivia	-0.2734609	-1.431396	1.157935
Ecuador	-0.3129792	-1.629396	1.316417
Iran	-0.8182302	-1.838419	1.020189
Mexico	-0.3136011	-1.884677	1.571076
Yemen	-3.8799223	-6.003142	2.123220

It is worthwhile to inspect the histogram distribution of residuals as well in order to see if any outliers or highly misclassified instances exist.



Because of our model's bias towards populous coutries and population masses, we have a few outliers in the distribution of residuals such as Bhutan and the Seychelles, which have been overestimated by the model. A last measure of interest for the residual analysis would be to inspect the residuals in level format. With the tables below, the measures of interest are now in their original scales of deaths per 1000 capita, rather than log, which can make individual country interpretations easier to discern in absolute numbers. However, since we are converting back to level scale from a log model, conclusions are subject to the log interpretation, not the level interpretation.

Table 7: A.5.4 Top 5 countries by positive level residuals

country	deaths_capita	reg6_y_pred_level	reg6_resid_level
Bahrain	0.1992479	0.8303216	-0.6310738
Israel	0.2868567	0.6025145	-0.3156578
Kuwait	0.1887293	0.5346122	-0.3458829
Maldives	0.0715694	0.4030995	-0.3315301
Qatar	0.0819190	0.7826400	-0.7007210

Table 8: A.5.5 Top 5 countries by positive level residuals

country	deaths_capita	reg6_y_pred_level	reg6_resid_level
Bolivia	0.7606987	0.2389751	0.5217236
Ecuador	0.7312218	0.1960480	0.5351738
Mexico	0.7307671	0.1518781	0.5788890
Peru	1.0664570	0.4945329	0.5719241
San Marino	1.2404017	0.5149424	0.7254593

A.6 Estimating Different Models

The following section provides argumentation of the models of estimation used within the regression analysis. The chosen base models were those of linear, quadratic, cubic, piecewise linear splines and weighted linear regression in order to tackle both linear and non-linear patterns observed in the non-parametric trial plots above (see Appendix A.3).

A.6.1 Linear quadratic and cubic regression discussion

The table on the next page compares the three variants of linear regression to each other. The first model of choice was the linear regression model showing a slope of 1.06 and an alpha of -4.4. With a statistically significant model, the simple interpretation of the linear regression states that with every 10% increase in Covid-19 cases per 1000 capita per country, we observe a 10.6% increase in Covid-19 deaths per 1000 capita per country on average. This model states that overall, deaths per 1000 capita per country tend to increase at a bit of a faster rate with an increase in their cases counterparts per 1000 people per country on average. This benchmark is used for the later models, having an adjusted R-squared of 72%, meaning that this model correctly encompasses more than 2/3 of the available refined global data. With the exact same coefficient (beta-1) for the quadratic regression, but a statistically insignificant second coefficient for the upper part of the parabola, as well as the same R-squared, we can omit the usage of the quadratic model as it does not add any benefit to our analysis and merely adds complexity. Lastly, the cubic regression appears to have a steeper slope than the other linear models, but at the expense of having the other 2 coefficients be statistically insignificant. As such, this model tends to overfit the overall data, tailoring to the lowest 1/3 of the plotted data.

A.6.2 PLS (1 knot), PLS (2 knots) and weighted OLS regression discussion

The table on the final page of this appendix compares the models that employ cutoff points and other arbitrary weights, unlike the first 3 models. The first piecewise linear splines regression model has statistically significant alpha and both beta coefficients, and the simple interpretation is that up to the threshold of 10 cases per

1000 capita per country, we observe a slope signifying just like the linear models argued that Covid-19 deaths per 1000 capita per country increase, on average, faster than cases per 1000 capita per country. After the threshold of cases per 1000 capita, the pattern of association of reversed, observing that deaths increase slower than cases. This pattern is very useful for understanding. However, it can be hypothesized that the slope around the middle of the scatterplot should be steeper but it is dragged downwards by isolated countries or special cases that handled the pandemic well. Investigating the noticeable heteroskedasticity of A.3.4 log-log model, we can see 3 different (more easily recognizable) point spreads that give the heteroskedasticity of the model. As such, the piecewise linear splines regression with 2 knots is introduced, with cutoff points at 0.2 cases per 1000 capita per country (or 2 cases in 10,000 people) and at 10 cases per 1000 capita. While this model has the same adjusted R-squared as the other models until now, we can observe a less than 1 slope for the first 1/3 of the plot, a steeper than 1 slope for the middle 1/3 of the plot and yet a less than 1 slope for the last 1/3 of the plot, which contains the countries that handled Covid-19 in the worst way in terms of infections. This model makes more intuitive sense because we allow the countries that handled Covid-19 well, that were isolated or that lied about their death rate to have their own grouping, skedastic spread and slope.

However, the adjusted R-squared is still the same, but the complexity of the model is too high for ease of interpretation. Furthermore, in the interest of global research, this problem should not be looked at from a country-centric approach, but rather from a citizen approach, taking a world-wide approach. This would discriminate against low population countries and subsequently against countries with much lower cases, but it will fit the global trend better, not requiring added complexity such as what the PLS models bring. Thus, the final model and the model of choice is that of an OLS weighted linear regression, using population as a weight. As the graph (A.6.2.3) shows, the countries with very few deaths and low populations are almost entirely exluded from the model, but this bias towards tailoring to more people overall means that the R-squared of this weighted OLS is around 90%, signifying a much greater fit when looking at global trends. The interpretation of this model can be found in the main body of the analysis.

Lastly, the approach of a population-weighted OLS shows that because we take a global outlook, trying to tailor to the masses (more people), our intercept is higher than that of the simple OLS linear regression, but our coefficient is less steep than that of the linear model. This signifies that isolated cases such as Bhutan and some low population countries have dispropertionately lower deaths than what the rest of the country trends would lead us to believe and if their specific cases would not be weighted by population, they would bring the overall model to predict more deaths for high population countries. As such, if the linear model is employed, those isolated cases make the slope steeper, steering the country-trend towards having deaths per 1000 capita increase on average faster than cases per 1000 capita, while the global trend (factoring in more people) would dictate the opposite.

A.6.3 Summary of reasoning for model of choice

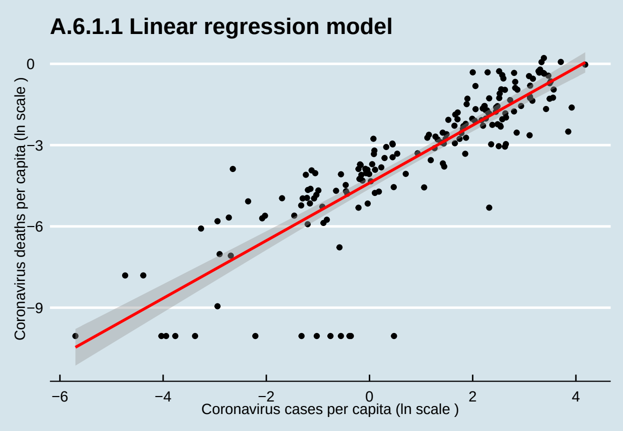
Substantive:

- Less complexity associated with model interpretation than potentially overestimating/underestimating PLS models;
- Gloabal outlook is more important than country-centric outlook for a global pandemic;
- The influence of the low population and isolated cases is less felt within the global picture;

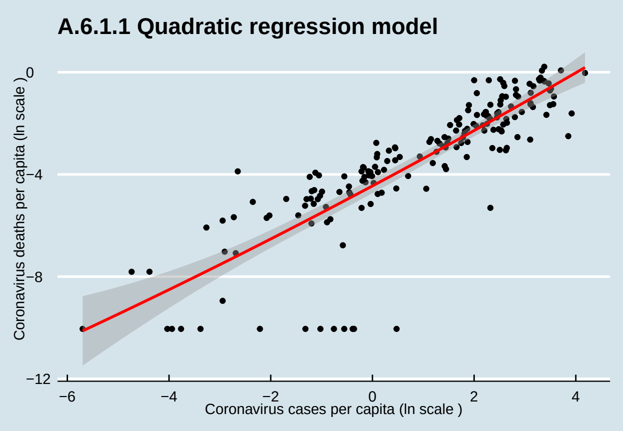
Statistical:

- R-squared is much higher than the other models;
- Standard errors and CI becomes smaller, which is of interest for more precise prediction;
- The percentage increase in deaths per 1000 global capita is smaller than the percentage increase associated with cases per 1000 global capita, which makes sense for larger populations, as people do not get sick in the same way in densly populated and modenr cities as opposed to sparsely populated more rural isolated areas;

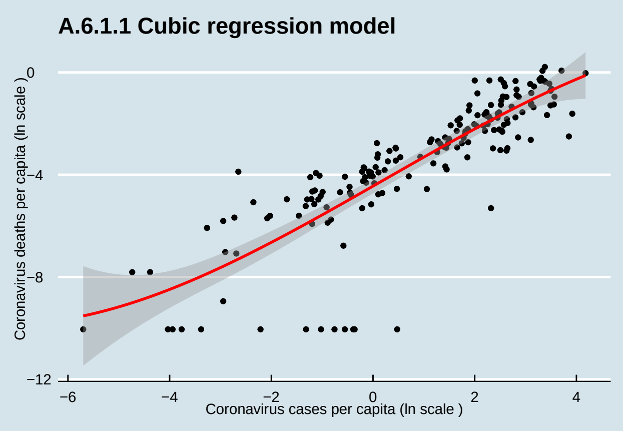
Plots of the different models



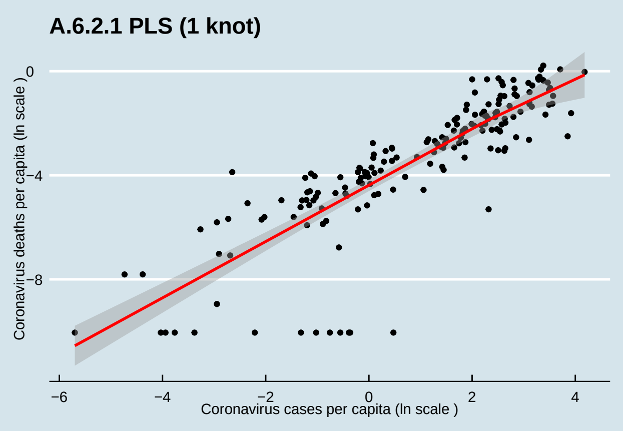
A.6.1.1 Linear regression model



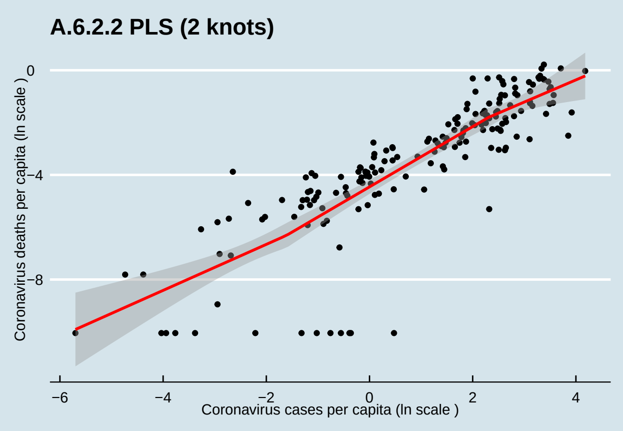
A.6.1.1 Quadratic regression model



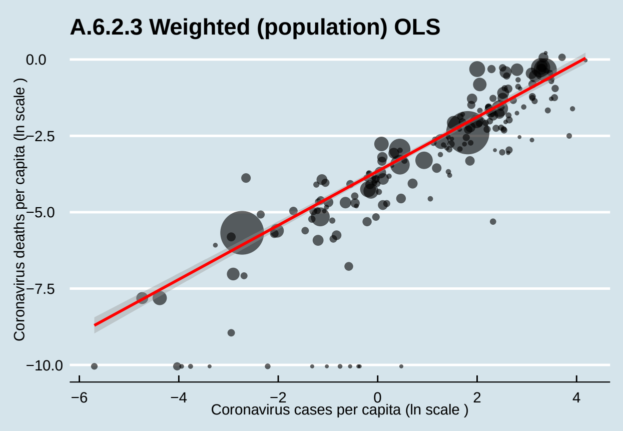
A.6.1.1 Cubic regression model



A.6.2.1 PLS (1 knot)



A.6.2.2 PLS (2 knots)



A.6.2.3 Weighted (population) OLS

	Linear	Quadratic	Cubic
(Intercept)	-4.40 ***	-4.45 ***	-4.44 ***
	(0.13)	(0.18)	(0.18)
ln__cases__capita	1.06 ***	1.06 ***	1.14 ***
	(0.05)	(0.05)	(0.11)
ln__cases__capita__sq		0.01	0.00
		(0.02)	(0.02)
ln__cases__capita__cb			-0.01
			(0.01)
nobs	183	183	183
r.squared	0.72	0.72	0.72
adj.r.squared	0.72	0.72	0.72
statistic	413.74	205.81	159.98
p.value	0.00	0.00	0.00
df.residual	181.00	180.00	179.00
nobs.1	183.00	183.00	183.00
se__type	HC2.00	HC2.00	HC2.00

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

	PLS (1 knot)	PLS (2 knots)	Weighted (population) OLS
(Intercept)	-4.38 ***	-4.90 ***	-3.66 ***
	(0.13)	(0.59)	(0.16)
lspline(ln_cases_capita, cutoff_ln)1	1.08 ***		
	(0.06)		
lspline(ln_cases_capita, cutoff_ln)2	0.93 ***		
	(0.20)		
lspline(ln_cases_capita, c(cutoff_ln, cutoff_ln2))1		0.88 ***	
		(0.21)	
lspline(ln_cases_capita, c(cutoff_ln, cutoff_ln2))2		1.15 ***	
		(0.10)	
lspline(ln_cases_capita, c(cutoff_ln, cutoff_ln2))3		0.85 ***	
		(0.20)	
ln_cases_capita			0.89 ***
			(0.08)
nobs	183	183	183
r.squared	0.72	0.72	0.91
adj.r.squared	0.72	0.72	0.91
statistic	213.51	147.04	132.10
p.value	0.00	0.00	0.00
df.residual	180.00	179.00	181.00
nobs.1	183.00	183.00	183.00
se_type	HC2.00	HC2.00	HC2.00

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