Assignment 1 - Data analysis 2

Zsombor Hegedus

2020 november 29

Introduction

In this analysis my focus is uncovering a pattern of association between the number of confirmed COVID-19 cases and the number of deaths caused by the virus. In this analysis I will use admin data, where variables, such as the number of deceased and infected people are from the github repo of Johns Hopkins University, and population data is collected from World Development Indicators maintained by the World Bank. The population of the analysis covers every country where such data was available (each country is one observation in my dataset), while the sample used for the analysis is narrowed to one date - the 15th of Oct, 2020. Potential data quality issues could arose from the fact that:

- Joining the two dataframes required entity resolution which is a manual process that always leaves room for data quality errors
- There might be a bias in how countries collected or disclosed their information (e.g. some countries might not conduct enough tests hence the number of confirmed cases are low - or they might be using different criterion to indentify a COVID-19 death case)

After entity resolution, I ended up with 168 observations in my data, where I dropped out observations with zero reported death cases since I did ln transformation where records with zero deaths could not have been used. I used *per capita* variables so divided deaths and confirmed cases by the population - expressed in millions - of given country. My dependent variable y was the number of deaths per one million capita caused by COVID-19 and my explanatory variable x was the number of confirmed cases per one million capita in relation with the disease.

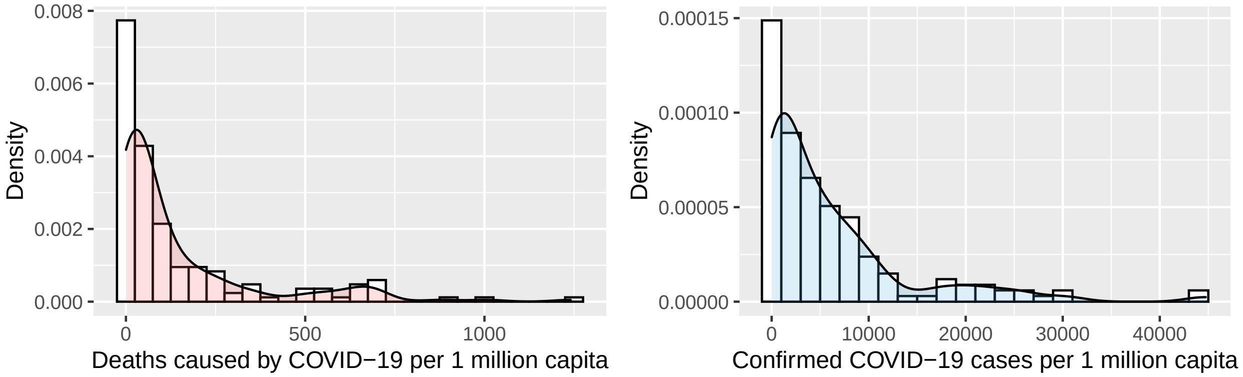


Figure 1: Distribution of deaths and confirmed cases in relation to COVID-19 for the examined countries on 15th Oct, 2020

variable	mean	median	std	iq_range	min	max	skew	numObs
Death per capita	142.92	51.38	218.83	148.00	0.09	1240.40	2.31	168
Confirmed cases per capita	6058.74	3209.04	8009.23	7371.35	8.78	44734.82	2.31	168

Table 1: Summary statistics of examined variables

EDA

Figure 1 shows the distribution of variables which is skewed with a fat right tail and a few extreme values for both. The number of confirmed cases (visible in Table 1) far exceed the deaths - in my sample, ca.2.4% of the people that contracted COVID-19 died from the disease. Since the density plots resemble a lognormal distribution, it indicates that ln transformations for both variables might strengthen the results of the analysis.

A substantive reason for choosing log-log model is that my variables are affected in multiplicative ways - we care about percentage change when talking about the number of infected and fallen people. Maybe we could argue that e.g. the absolute number of people died is an important measure in case a given country doesn't really have a lot of reported cases, but to have a hollistic view on this phenomenon, it makes more sense to look at percentages.

Statistical reasons for choosing ln-ln transformation is that:

- the distributions of both variables are skewed with a long right tail which resembles a log-normal distirbution
- comparing scatter plots with and without ln transformation resulted in a much better fit for the model (R^2 was about twice as good)
- Visual inspection of the scaterplots with lowess (available in the Appendix) also indicated that log-log transformation shows a clearer pattern.

Model Choice

For the analysis I chose a simple linear regression between the log transformed y and x variables:

$$ln(y) = \alpha + ln(x)$$

The estimated parameters can be found in the Appendix. α is not really meaningful as $ln(y)$ is difficult to interpret, but it shows how much is the ln of death per 1 million capita in case we have 1 confirmed case per 1 million capita. β is the slope parameter, and my model indicates that in this sample the number of death per 1m capital is 0.93% higher on average for observations having 1% higher confirmed cases per 1m capita.

I conducted hypothesis testing as I was interested to see if my estimated β parameter is significant at 5%:

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

The below table shows the result of the hypothesis test which resulted in a 95% CI of [0.84, 1.02], which means that I can reject the H_0 with 95% confidence. The overall conclusion is that β is significant in the level that I was interested in (even more, as we can see from the p value), therefore there is a positive pattern of association between the variables subject to my analysis.

variable	estimate	std.error	statistic	p.value	conf.low	conf.high
ln_conf_pc	0.93	0.04	20.76	0.00	0.84	1.02

Table 2: Hypothesis testing for the slope of the regression

Analysing residuals

Table 4 summarises the countries where my model projected $\hat{y}$ with the lowest negative errors. In these countries less people died than what the model implied. A number of reasons can explain the low number of deaths compared to the amount of confirmed cases. Some of these countries are very wealthy, so healthcare might be operating on a higher standard than in other countries. On the other hand it can also be that the density of population in given country is low, so that people don't interact and meet with others so much.

country	ln_death_pc	reg1_y_pred	reg1_res
Burundi	-2.45	0.12	-2.56
Iceland	3.32	4.98	-1.66
Qatar	4.33	6.56	-2.22
Singapore	1.55	5.17	-3.62
Sri Lanka	-0.52	1.31	-1.83

Table 3: 5 Countries with the biggest negative residuals

Table 5 on the other hand summarises the countries where my model projected $\hat{y}$ with the highest positive errors. In these countries more people died than what the model implied. This can also be the result of many reasons, for example Italy and the UK were one of the first to be hit with COVID-19 in Europe - they became an epicenter of the disease very soon, and their healthcare system operated on the limit of its capabilities. The relatively high deceased people might also be due to the lack of precautionary lock-down measures, or operational issues with local healthcare.

country	ln_death_pc	reg1_y_pred	reg1_res
Ecuador	6.51	4.97	1.54
Italy	6.39	4.59	1.80
Mexico	6.46	4.71	1.75
United Kingdom	6.45	4.92	1.53
Yemen	3.01	0.54	2.47

Table 4: 5 Countries with the biggest negative residuals

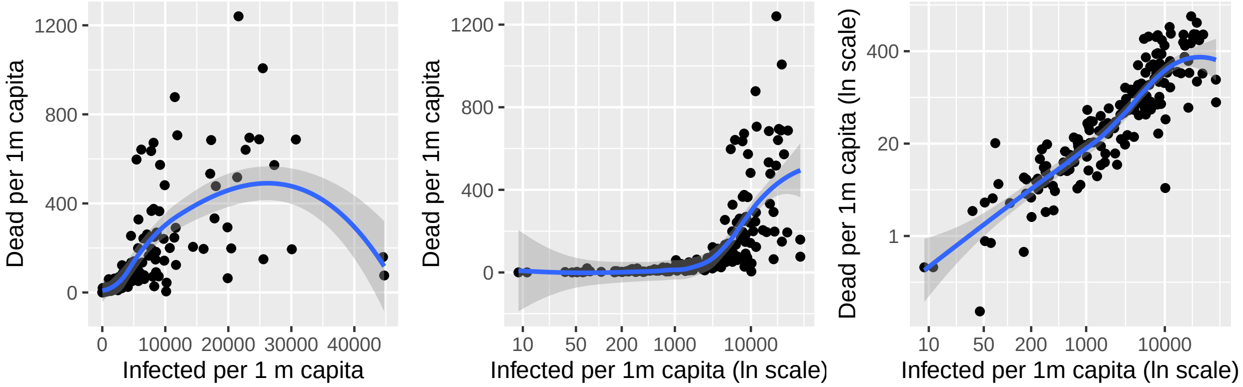
Executive summary

I set out to analyse the pattern of association between the number of confirmed COVID cases and the number of deaths caused by COVID-19 in all countries where this data is available. I used log-log transformation and scaled the number of confirmed and fallen people by the population of each country to arrive to 'per capita' variables. I used regression analysis to uncover these tendencies, and found in my sample for the 15th Oct, 2020 that there is a positive association between the two variables. The main message my model wishes to convey is that understanding the number of people contracted by the virus is crucially important for governments, so that they can have an expectation as to how many people might die due to the disease. On the one hand my results could have been strengthened by including more variables such as the age of people (or a categorical variable representing age groups) to see whether the virus is deadlier for the older population. My results, on the other hand could have been weakend by using level instead of log variables, as it was quite apparent in the EDA phase that without ln transformation my R^2 would have dropped approximately to the half of what I was able to achieve with my final model.

Appendix

Transforming variables

The below is to show why log-log model was the best choice when deciding on transformation of variables.



Model Comparison and model choice

The below table and scatter plots serve as summary for four regression models that I run in order to uncover the pattern of association between my log transformed x and y variables. The models were ther following:

- **Model 1** - simple linear regression
- **Model 2** - a quadratic regression
- **Model 3** - piecewise linear spline with two cutoffs that I determined by looking at the lowess function which is also visible in Figure 2
- **Model 4** - weighted regression where the weight was the population of a country

	Model 1	Model 2	Model 3	Model 4
(Intercept)	-3.42^*	-3.06^*	-2.67^*	-2.91^*
	$[-4.11; -2.73]$	$[-4.95; -1.16]$	$[-4.36; -0.97]$	$[-4.20; -1.62]$
ln_conf_pc	0.93^*	0.82^*		0.90^*
	$[0.84; 1.02]$	$[0.29; 1.35]$		$[0.74; 1.06]$
ln_conf_pc_sq		0.01		
		$[-0.03; 0.04]$		
lspline(ln_conf_pc, ln_cutoff)1			0.77^*	
			$[0.45; 1.09]$	
lspline(ln_conf_pc, ln_cutoff)2			0.97^*	
			$[0.81; 1.13]$	
lspline(ln_conf_pc, ln_cutoff)3			0.94^*	
			$[0.68; 1.20]$	
R^2	0.79	0.79	0.79	0.90
Adj. R^2	0.79	0.79	0.79	0.90
Num. obs.	168	168	168	168
RMSE	0.83	0.83	0.83	4176.82

* Null hypothesis value outside the confidence interval.

All the above models estimated a positive trend in our sample indicating that the log number of death per 1 million capita is higher for higher log number of confirmed cases per 1m capita. All models showed a very good fit, with R^2 being 0.79 in the first 3 models and 0.9 for the last one. In Model 4 we can see that countries with bigger population (indicated with bigger size for the dots) are very close to the trend line, which is probably the reason why the fit was much better. However weighting is not really sensible, since we already examine per capita variables. The estimated slope parameters were significant at 5% in every case, except the squared parameter for the quadratic case. We can also see from the scatter plots that the relationship is closer to linear, than non-linear, so probably not even higher order polynomials would provide a much better fit. From the looks of it, the dots are also spread around the estimated $\hat{y}$ variables in a quite symmetric, seemingly homoskedastic manner.

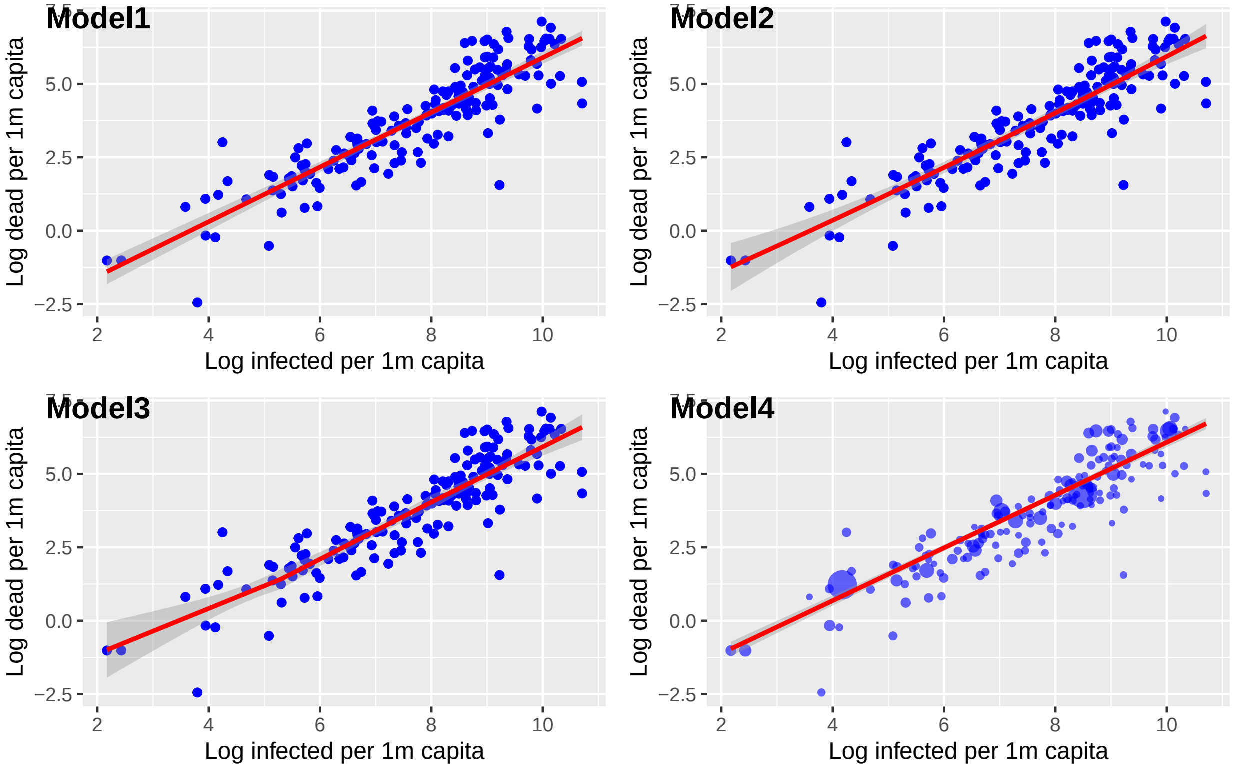


Figure 2: Scatter plots with regression models visualised for confirmed and died COVID-19 cases for 15th Oct, 2020

For my analysis I chose **Model 1**, the simple linear regression. The pattern of association in the data clearly indicates a linear relationship between the log number of confirmed cases per 1m capita and log deaths per 1m capita. This linear pattern is also understood between virologist and other experts - if someone contracts this disease, there is a possibility that it kills them. In addition to that, regression estimates in the simple case are the easiest to translate.