Assignment - Data Analysis 2 and Coding with R

David Utassy

11/27/2020

Introduction

My task is to analyze the pattern of association between registered covid-19 cases per capita and the registered number of death per capita in countries due to covid-19 on 05/11/2020.

My variables are: "Number of registered death per capita" and "Number of registered cases per capita" (Y and X). In every observation (row), I have both variables from a country on a given date. These are registered data from each country according to their official records. These data should be recorded by taking covid-19 tests. Each row is meant to represent the whole population of a country. Of course, there are a lot of possible data quality issues. For example, political, economical effects on a given country's data can be significant, as they try to hide their real data sometimes. On the other hand differences between infrastructure and health care system can also have a huge effect on our data.

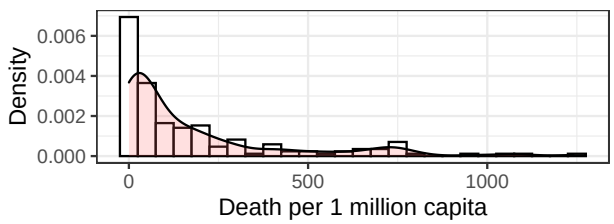
Selecting observations

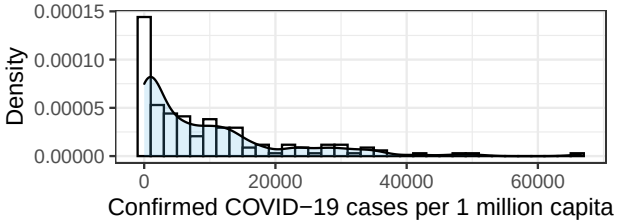
In the basic data cleaning I used the provided code by Ágoston Reguly, I used most of it to get a close to analyzable data set. In that data table, I kept only countries, that have all the needed variables (confirmed, death, recovered, active, population). At that point, we can see, that there are countries with 0 registered deaths. I have decided to drop these countries as taking the log of death cases would give us -Inf in the future that will ruin the analyses. For easier interpretation, I have applied a scaling, therefore from this point on I mean per 1 million capita, when I use the term "per capita".

(Another solution could have been to replace the zero with a very small positive number, but in that case, the question arises: -how small?-. A very small positive number we will get huge negative numbers after log transformation, that will "pull-down" our regression line.)

From the basic variables, I had to create confirmed_per_capita and death_per_capita variable by dividing with the population of the given country.

Histogram and summary statistics for x and y





variable	mean	median	std	iq_range	min	max	skew	numObs
Death per capita	179.75	66.12	252.20	199.13	0.09	1240.40	1.95	170
Confirmed cases per capita	9467.64	5500.87	11511.96	12284.90	8.78	66565.55	1.85	170

From the summary statistics, we can see, that both variables are skewed with a right tail. We can observe this on the distribution diagrams and also from the summary statistics. We can say that the distributions are mostly lognormal, hence it is likely that a log transformation on both axes will be beneficial. The mean is greater than the median with both variables which is a sign of skewness as well.

Investigate the transformation of variables

According to the x and y variable distributions, we should get the best result with the log-log option. In the appendix, I am showing the plots of all the possible transformations and offering substantive and statistical reasoning. From those graphs, we can also see the log-log option is the best as it makes the association close to linear! Statistically speaking, log transformations are needed also, as the original variables are skewed, and following a lognormal distribution. For a substantive reason, using log transformation is beneficial as after that, we can interpret results by using percentage changes, which is reasonable with these variables. For this reason from now on, I will use the ln_confirmed_per_capita and ln_death_per_capita variables as x and y variables.

Presentation of my model choice

According to the assignment description I experimented with 4 models, in the appendix, I am reasoning my choice of model.

My choice of model was Model 1:

$y^E = \alpha + \beta$x, where: α = -3.27, β = 0.90

It is a simple model, therefore it is easy to interpret, and it has comparatively high R2 and captures variation well. (see in appendix)

The interpretation of α is rarely meaningful as average ln(y) is difficult to interpret. In contrast, the β interpretation is possible. $\beta = 0.9$ means, that on average the number of death is 0.9 percent higher on average if the registered cases are higher by one percent.

Hypothesis testing

H_0: $\beta = 0$

H_A: $\beta \neq 0$

I choose 5% as a significance level. The summary table of the hypothesis testing:

variable	estimate	std.error	statistic	p.value	conf.low	conf.high
ln_confirmed_per_capita	0.9	0.04	23.77	0	0.83	0.98

From the summary we can see that the p-value is almost zero which is definitely smaller than 0.05, therefore we can reject the null hypothesis. This means that we can state with 95% confidence, that death cases per capita and registered cases per capita correlate with each other.

Analysis of the residuals

Best countries These countries are under the prediction line. These are the countries who saved (relatively) the most people due to covid. They may be able to do it because of their health care system and infrastructure (Singapore), however, the quality of the data can affect these results in a positive direction as well in less developed countries.

country	ln_death_per_capita	reg1_y_pred	reg1_res
Botswana	2.4613216	4.0570749	-1.595753
Burundi	-2.4450026	0.2979289	-2.742932
Qatar	4.4057305	6.4278817	-2.022151
Singapore	1.5911124	5.0451046	-3.453992
Sri Lanka	0.2852483	2.4571069	-2.171859

Worst countries These countries are above the prediction line. These are the countries that lost (relatively) the most people due to covid. Possibly this is the result of bad decisions, bad health care system, and infrastructure, as a country would not want to publish the worst data than reality.

country	ln_death_per_capita	reg1_y_pred	reg1_res
Bolivia	6.635378	5.219689	1.415689
Ecuador	6.596761	5.019128	1.577633
Iran	6.100465	4.816708	1.283757
Mexico	6.599913	4.762775	1.837138
Yemen	3.025731	0.565827	2.459904

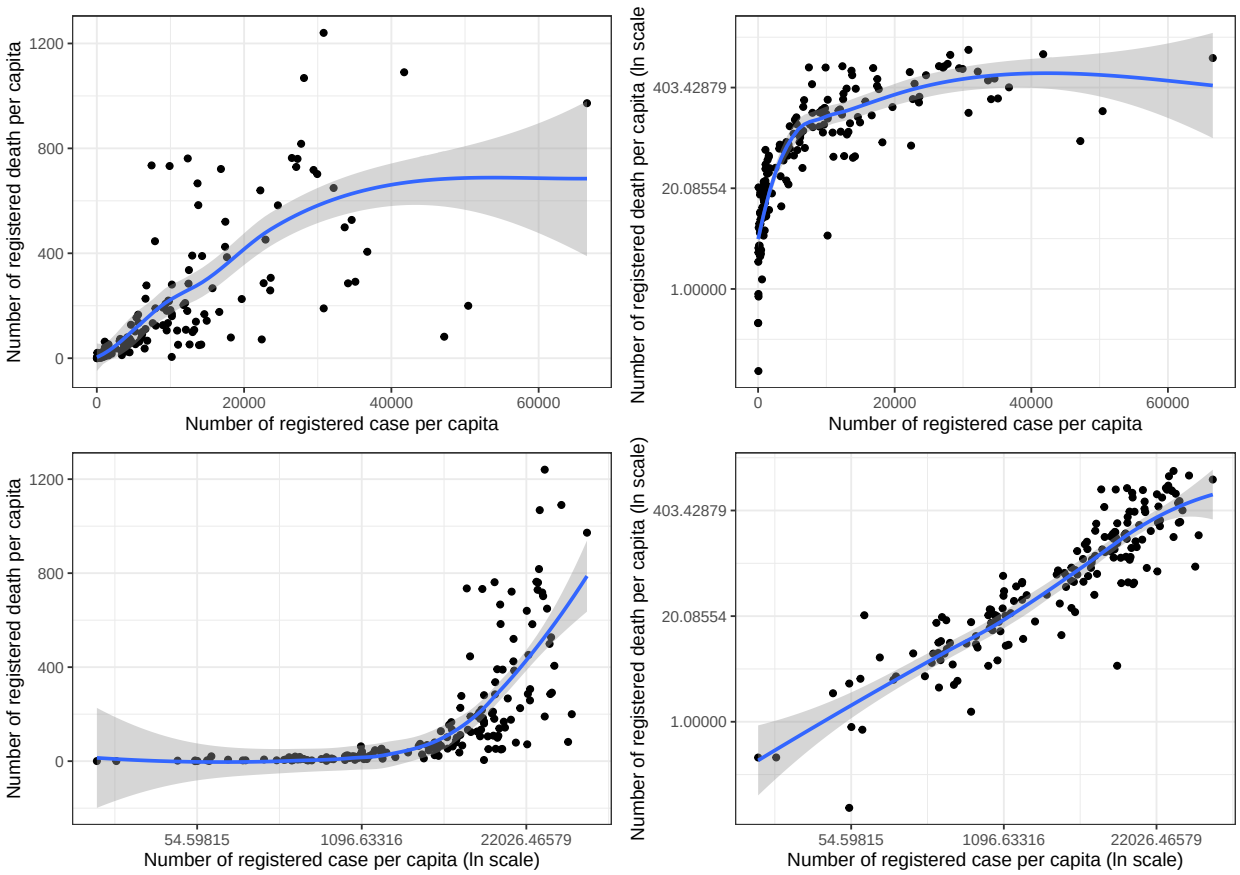
Executive summary

The main result of my analysis is that there is a positive correlation between the number of registered covid-19 cases per capita and the number of registered covid death per capita. It means that the number of death per capita is higher as the number of registered cases per capita is higher. The model I used on my variables (Y: number of registered covid-19 cases per capita, X: number of registered covid death per capita) was a simple linear regression.

My model's main message is, that on average the number of death is 0.9 percent higher on average if the registered cases are higher by one percent. Better data quality would strengthen my model to have better external validity. The problem is that each and every country's data is originated from its country's government, which is an unreliable factor as they might want to hide the truth.

Appendix

Investigate the transformation of variables



Estimating different models

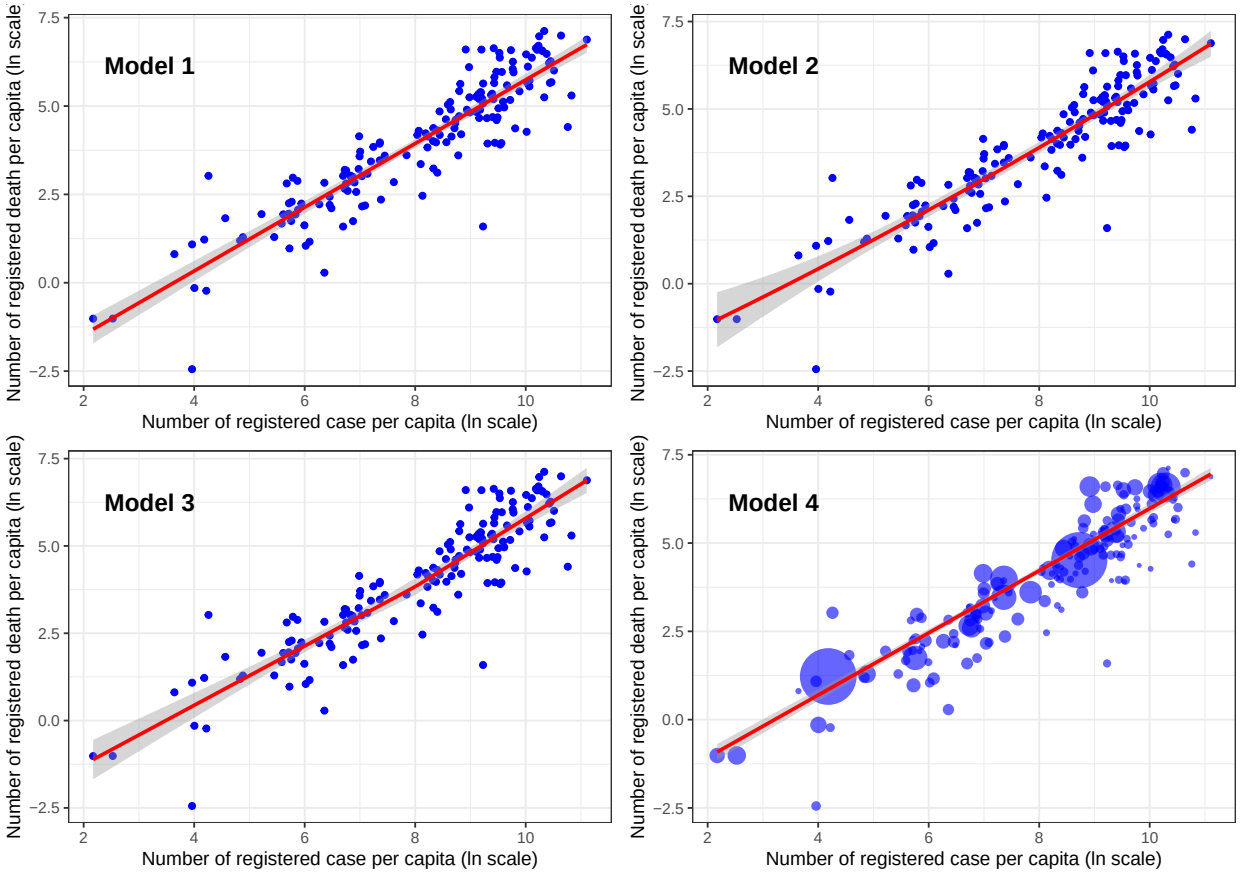
According to the assignment description, I experimented with the following models:

Model 1: $y^E = \alpha + \beta \text{x}$

Model 2: $y^E = \alpha + \beta_1 \text{x} + \beta_2 x^2$

Model 3: $y^E = \alpha_1 + \beta_1 \text{x}[\text{if x} < 3000] + (\alpha_1 + \beta_2 \text{x})[\text{if x} \geq 3000]$

Model 4: Simple linear regression weighted with population



	Model 1	Model 2	Model 3	Model 4
(Intercept)	−3.27*	−2.65*	−2.96*	−2.83*
	[−3.90; −2.65]	[−4.42; −0.89]	[−3.99; −1.93]	[−4.05; −1.60]
ln_confirmed_per_capita	0.90*	0.72*		0.88*
	[0.83; 0.98]	[0.26; 1.18]		[0.73; 1.03]
ln_confirmed_per_capita_sq		0.01		
		[−0.02; 0.04]		
lspline(ln_confirmed_per_capita, cutoff_ln)1			0.85*	
			[0.70; 1.00]	
lspline(ln_confirmed_per_capita, cutoff_ln)2			0.98*	
			[0.81; 1.15]	
R^2	0.82	0.82	0.82	0.92
Adj. R^2	0.82	0.82	0.82	0.91
Num. obs.	170	170	170	170
RMSE	0.79	0.79	0.79	3919.44

* Null hypothesis value outside the confidence interval.

Substantive reasoning I have chosen Model 1 from these four models, however, all of them catch the pattern of the data pretty well. In this case, it is beneficial to choose the simplest model, which is Model 1 in my case. The log-log interpretation works well, and the magnitude of coefficients are meaningful. The next model, which is fairly good also, is Model 4 with population weight. With that model, we are taking greater countries into account with more weight, however, as we are already using per capita measures, we do not need to take this into account again.

Statistical reasoning As we can see from the summary table, the R2 of the models are almost the same, except for Model 4 which is higher. However, I have not chosen Model 4 because we already took the population into account in Model 1 as well, as we are using per capita measures. Additionally, the first model is simpler also, therefore I will go on with model 1. (The quadratic and a PLR is overkill in this case as they are also resulting in a fairly straight line.)