Assignment 1 for Data Analysis 2 and Coding with R

Pauline Broussolle

2020-11-24

We analyse covid data from CSSE at Johns Hopkins University and population data for 2019 from World Bank (using WDI). The dataset contains 182 countries and registers the number of confirmed covid-19 cases and number of death due by covid on 15/10/2020.

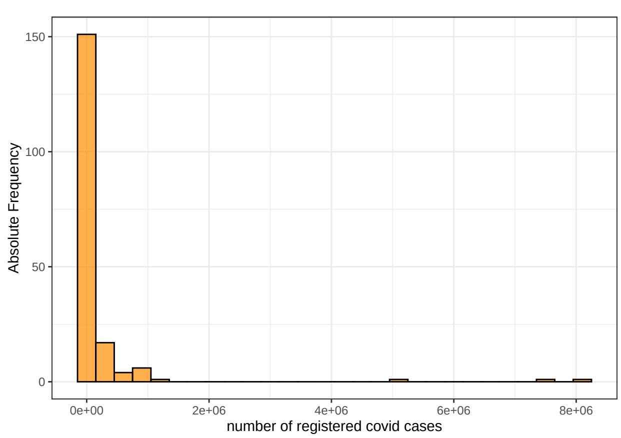
The aim of the analysis is to find the pattern of association between registered covid-19 cases and registered number of death due to covid-19 on **15/10/2020**. My outcome variable (y) is the **number of registered death due to covid** and my explanatory variable (x) is the **number of registered cases**.

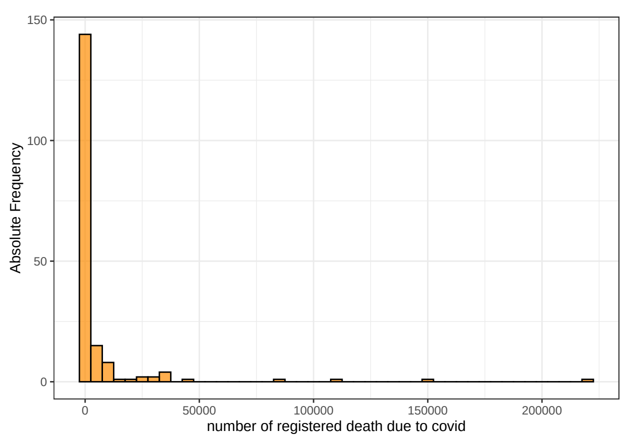
Executive summary: The **number of registered death due to covid** and the **number of registered covid cases** are positively correlated. According to our analysis countries with 10% more recorded covid cases have on average 9.5% more death due to covid.

1. Summary statistics and Distribution for x and y

Table 1: Summary for the number of registered death caused by covid and registered covid cases

mean	median	min	max	std	variable
213757	16483.0	3	7983919	899396	confirmed cases
6032	281.5	0	217883	22934	nb of death





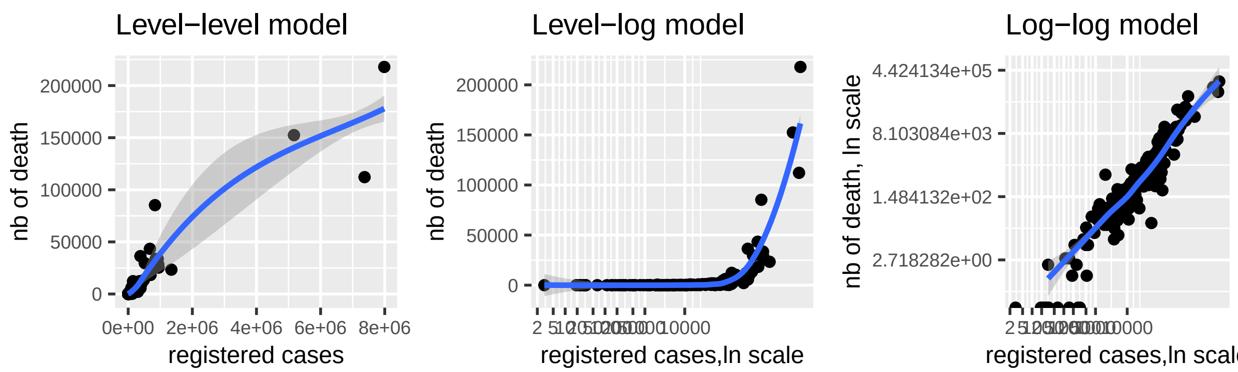
The "Confirmed" variable is the count of confirmed covid cases reported for one day and for each country. "Death" variable is the count of reported death due to covid for one day and for each country. Both distributions are skewed with long right tails. The 2 modes are located on low values, but we see that there are a few extreme values.

Table 2: Extreme values

country	confirmed	death	recovered	active	population
Brazil	5169386	152460	4526393	490533	211049527
India	7370468	112161	6453779	804528	1366417754
United States	7983919	217883	3177397	NA	328239523

We check countries which recorded confirmed covid cases above 2 milion and registered death above 50,000. These are India, Brazil and United States, which are not measurement errors. We keep these values.

2. Pattern of association



We check the possible different ln transformation for the variables with plotting different scatterplots with lowess. For the simple model without scaling and for the level-log model, the pattern are non-linear, most of observations are concentrated on the bottom and there are some extreme observations corresponding to Brazil, US and India. Taking log of confirmed cases and log of death creates a linear association. - Substantive reasoning: easier to interpret. - Statistical: it gives a better approximation to the average slope of the pattern. We should use the log-log transformation, which is the only to provide a linear pattern. We create new variables for log of the two variables: ln_confirmed and ln_death .

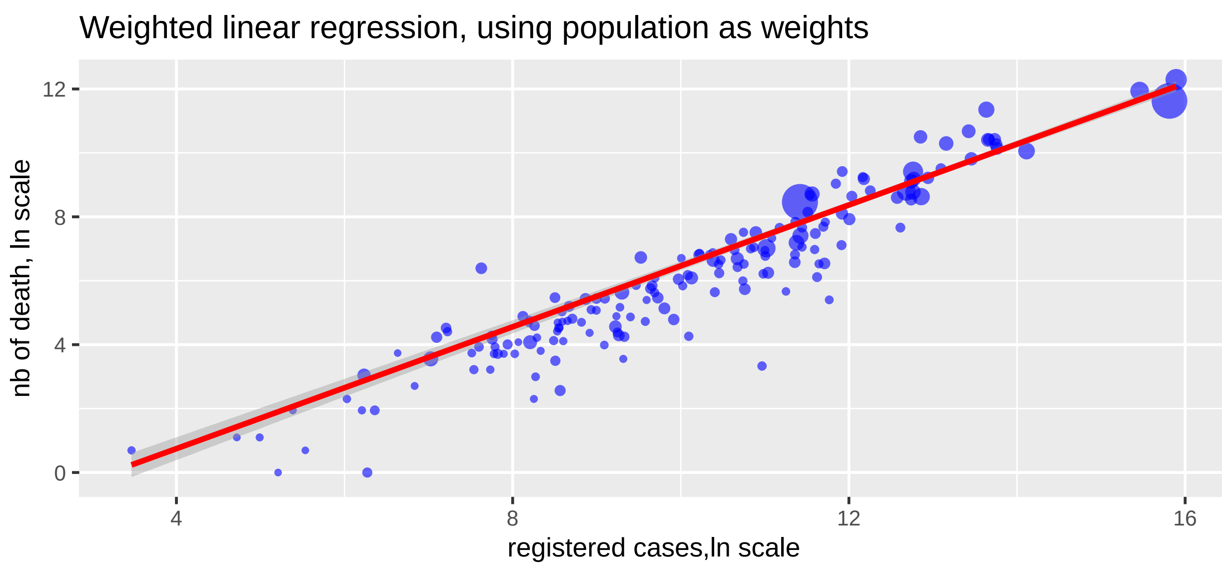
3. Estimating different models

We estimate four different regression models: Simple linear regression, Quadratic regression, Piecewise linear spline regression, Weighted linear regression using population as weights. The estimated model results are reported in the appendix of the report.

Based on model comparison we choose **Weighted linear regression model, using population as weights** (reg4). This regression has the higher R-squared out of the 4 models, which is 0.928. The coefficients can be well interpreted and an advantage is that the scatterplot for weighted regression shows the size of each country as the circles are proportionate to the population.

	Weighted linear regression, using population as weights
(Intercept)	-3.07^{***}
	(0.26)
ln_confirmed	0.95^{***}
	(0.02)
R^2	0.93
Adj. R^2	0.93
Num. obs.	170

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$



4. Hypothesis Test and Analysis of the residuals

We test our Beta parameter, carrying out the fellowing test: H0 :Beta=0, HA :Beta neq 0. The estimated t-statistics is 14.98, with p-value: 9.286e-33. Thus we reject the H0, which means that number of recorded death due to covid is not uncorrelated with number of confirmed covid cases.

Table 3: Countries with largest negative errors

country	ln_death	reg4_y_pred	reg4_res
Burundi	0.000000	2.910794	-2.910794
Iceland	2.302585	4.799315	-2.496730
Qatar	5.402677	8.148109	-2.745432
Singapore	3.332205	7.385913	-4.053708
Sri Lanka	2.564949	5.097056	-2.532107

Table 4: Countries with largest positive errors

country	ln_death	reg4_y_pred	reg4_res
Ecuador	9.417842	8.295599	1.1222430
Italy	10.501554	9.183262	1.3182927
Mexico	11.353754	9.929485	1.4242690
United Kingdom	10.677823	9.728898	0.9489249
Yemen	6.390241	4.203259	2.1869819

Conclusion

We investigated registered death due to covid and number of confirmed covid cases pattern of association. The slope for our chosen model is 0.953. This shows that countries with 10% more recorded covid cases have on average 9.5% more death due to covid. Our two variables X and Y are positively correlated. - Our analysis can be - strengthened by… - weakened by…

Appendix

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Jeu, nov 26, 2020 - 20:09:05

Table 5: Modelling number of registered death due to covid and number of confirmed covid cases

	Dependent variable:			
	ln_death			
	Linear model	Quadratic model	PLS model	Weighted linear model
	(1)	(2)	(3)	(4)
ln_confirmed	1.031^{***} (0.028)	0.583^{***} (0.178)		0.953^{***} (0.020)
ln_confirmed_sq		0.022^{**} (0.009)		
lspline(ln_confirmed, cutoff_ln)1			-2.309 (1.887)	
lspline(ln_confirmed, cutoff_ln)2			1.042^{***} (0.029)	
Constant	-4.290^{***} (0.291)	-2.176^{**} (0.880)	8.694 (7.343)	-3.066^{***} (0.260)
Observations	170	170	170	170
R^2	0.887	0.892	0.889	0.928
Adjusted R^2	0.887	0.890	0.888	0.928
Residual Std. Error	0.826 (df = 168)	0.813 (df = 167)	0.821 (df = 167)	4,289.901 (df = 168)
Note:				*p<0.1; **p<0.05; ***p<0.01