Assignment 1 for Data Analysis 2 and Coding with R

Pauline Broussolle

2020-11-24

Introduction

My goal is to analyse the pattern of association between registered covid-19 cases and registered number of death due to covid-19 on **15/10/2020**. My dependent variable (y) is the **number of registered death due to covid** and my explanatory variable (x) is the **number of registered cases**. The aim of the analysis is to create a report on the pattern of association, choose and interpret a regression model and refer to robustness checks.

Introduction of the dataset and variables The aim of this assignment is to guide you through in creating a short report on the pattern of association. You will need to chose your final model, interpret the results and ## Executive summary

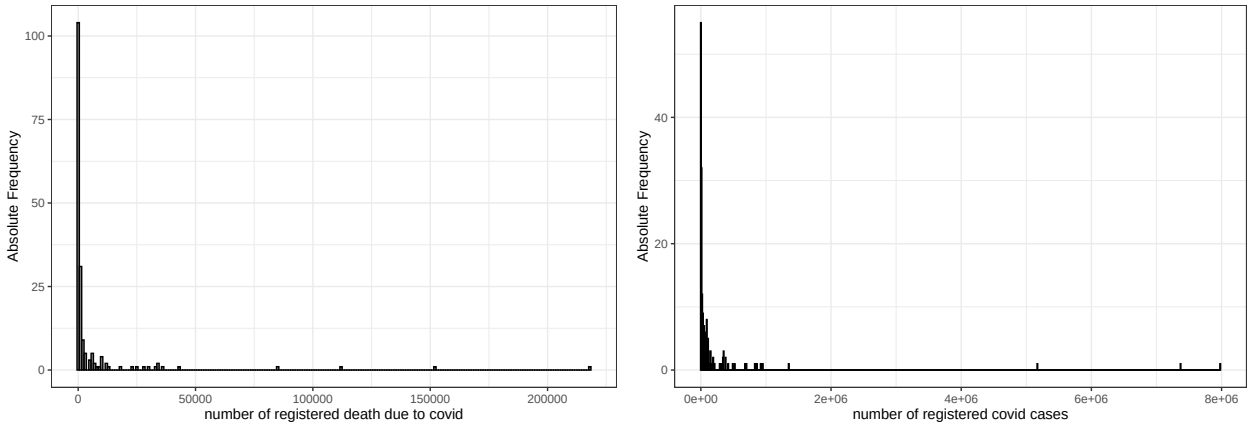
1. Preparatory Data Analysis

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

2-3 sentence, explain the main features and distribution - use histograms and summary statistics table (mean, median, min, max, standard deviation)

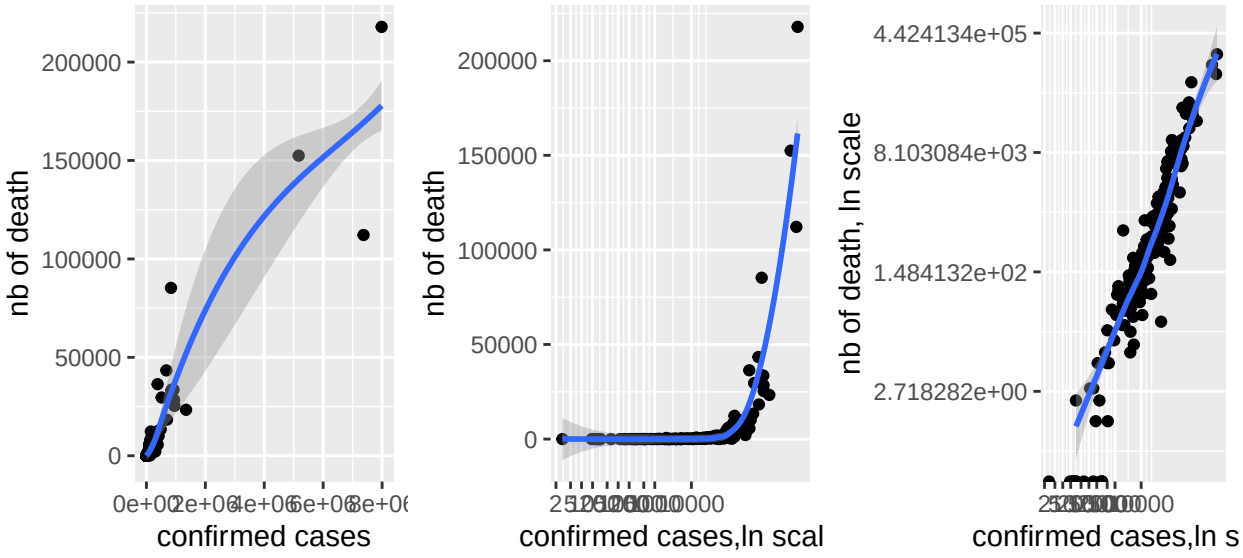


Select or drop observations, checking extreme values

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

2. Investigate the transformation of the variables

Scaling, Take Logs?



For the simple model without scaling the pattern is non-linear, most of observations are concentrated and there are some extreme observations, corresponding to Brazil, US, India.

Make a substantive and statistical reasoning, where and when to use ln transformation. You do not need to fit any model here, only use statistical reasoning based on the graphs. i. Take care when it is possible to make ln transformation: you may need to drop or change some variables.

2) using only gdppc is possible, but need to model the non-linearity in data

- Substantive: Level changes is harder to interpret and our aim is not to get $ based comparison

- Statistical: log transformation is way better approximation make simplification!

3) taking log of confirmed cases and log of death is making the association close to linear!

4) taking log for life-expectancy does not matter -> use levels!

- Substantive: it does not give better interpretation

- Statistical: you can compare models with the same y, no better fit

- Remember: simplest the better!

We should use the log-log transformation, which is the only to provide a linear pattern.

We create new variables which are ln_confirmed and ln_death .

3. Estimating different models

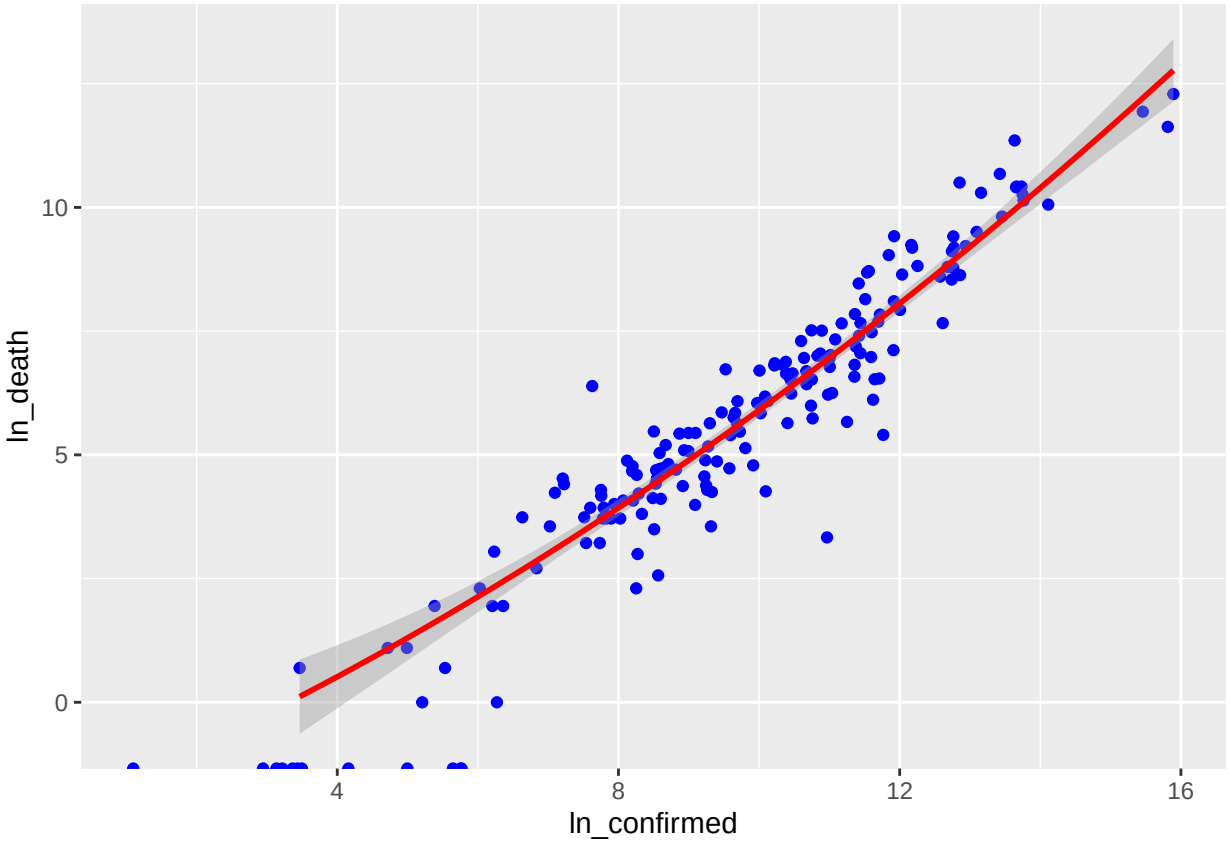
I chose the log-log transformation : ln_death = alpha + beta*ln_confirmed

Simple linear regression

```
## 2 coefficients  not defined because the design matrix is rank deficient
```

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df)
##
## Standard error type:  HC2
##
## Coefficients: (2 not defined because the design matrix is rank deficient)
##              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)       NaN         NA     NaN       NA      NaN      NaN NA
## ln_confirmed      NaN         NA     NaN       NA      NaN      NaN NA
##
## Multiple R-squared:    NaN , Adjusted R-squared:    NaN
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```



Quadratic regression

Piecewise linear spline regression

Weighted linear regression, using population as weights.

Presentation of model choice

(i) Compare the models and choose your preferred one

Hypothesis testing on beta (which interacts with x)

Analysis of the residuals

Conclusion