Assignment 1 for Data Analysis 2 and Coding with R

Pauline Broussolle

2020-11-24

Introduction

- Introduction of the data
  - Outcome variable: **number of registered death due to covid**
  - Explanatory variable: **number of registered cases**
  - Potential data cleaning (1-2 bullet points)

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**. 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.

Executive summary

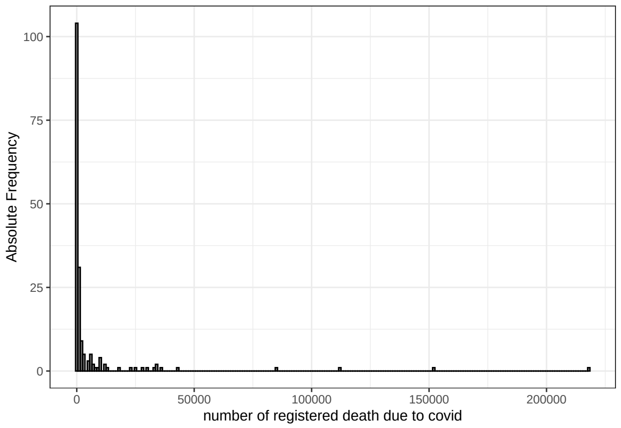
1. Preparatory Data Analysis

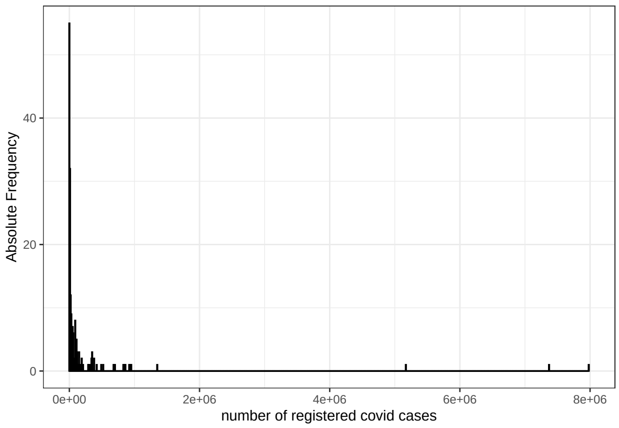
Summary statistics and Distribution for x and y

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

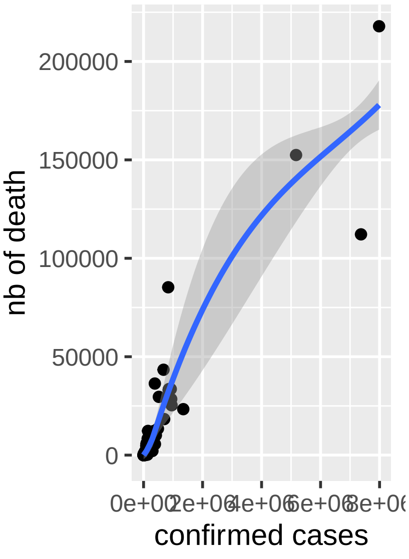
```
df %>% filter( confirmed > 2000000, death > 50000 )
```

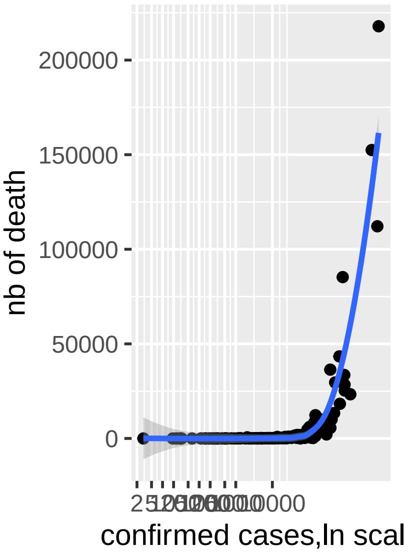
```
## # A tibble: 3 x 6
##   country       confirmed  death recovered active population
##   <chr>             <dbl>  <dbl>     <dbl>  <dbl>      <dbl>
## 1 Brazil          5169386 152460   4526393 490533  211049527
## 2 India           7370468 112161   6453779 804528 1366417754
## 3 United States   7983919 217883   3177397     NA  328239523
```

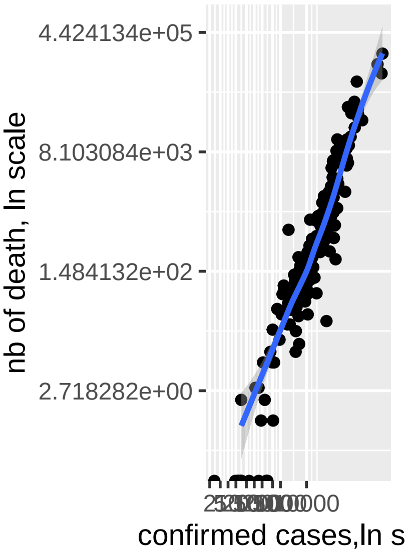
We check countries which have confirmed cases aboves 2 milion and registered number of death above 50,000. These are India, Brazil and United States, which are not measurement errors. We keep these values.

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 transformation : ln_death = alpha + beta*ln_confirmed

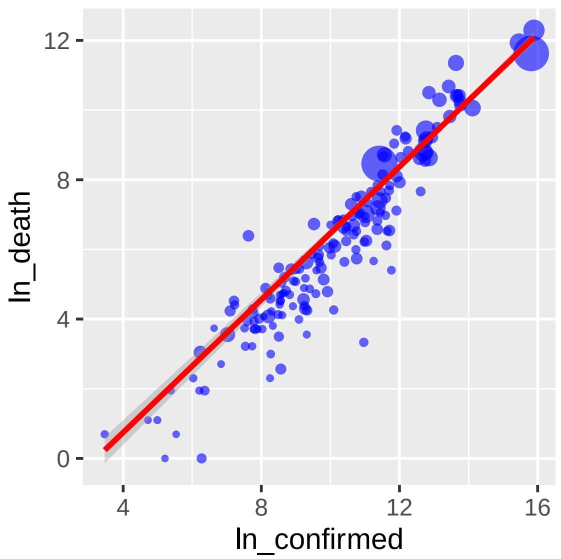
Presentation of model choice

The model comparison (all the estimated model results) is reported in the appendix of the report.

Weighted linear regression, using population as weights.

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, weights = population)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -3.0661    0.78208   -3.92 1.285e-04  -4.6100   -1.522 168
## ln_confirmed   0.9531    0.06364   14.98 9.286e-33   0.8275    1.079 168
##
## Multiple R-squared:  0.9285 ,    Adjusted R-squared:  0.9281
## F-statistic: 224.3 on 1 and 168 DF,  p-value: < 2.2e-16

## ‘geom_smooth()‘ using formula ’y ~ x’
```



Compare the models and choose your preferred one Use substantive and statistical reasoning for your chosen model. ii. Show the model results in the report along with the graph.

4. Hypothesis Test and Analysis of the residuals

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, weights = population)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -3.0661    0.78208   -3.92 1.285e-04  -4.6100   -1.522 168
## ln_confirmed   0.9531    0.06364   14.98 9.286e-33   0.8275    1.079 168
##
## Multiple R-squared:  0.9285 ,    Adjusted R-squared:  0.9281
## F-statistic: 224.3 on 1 and 168 DF,  p-value: < 2.2e-16
```

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 2: 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 3: 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 …
- and we have found
  - X and Y are … correlated
- Our analysis can be
  - strengthened by…
  - weakened by…

Appendix

	Linear model
(Intercept)	-4.29^*
	$[-4.88; -3.69]$
ln_confirmed	1.03^*
	$[0.97; 1.09]$
R^2	0.89
Adj. R^2	0.89
Num. obs.	170
RMSE	0.83

* 0 outside the confidence interval.

“‘