DA2 Assigment on covid-19 cases

AttilaSerfozo

2020-11-29

1. Executive Summary

In this project I have analyzed the association between confirmed covid-19 cases and number of deaths by covid-19. After checking the possible transformations I decided to go with a log-log transformation as we can have the best fit regression with this transformation. From the analyzed models I have selected the weighted linear regression using population weights as the best model, because it shows perfectly the correlation between confirmed cases and death. The results of the model showed that there is a strong connection between the two variables, it tells us that countries with 10% more confirmed covid cases can have on average 9.5% more deaths from covid-19. I think I could even strengthen further the results of my model by adding the number of chronicle diseases per country or proportion of elderly people as new variables and create a multiple regression model from it. On the other hand, I think that the accuracy of the model is hurt by the fact that the number of covid-19 deaths are measured variously across countries.

2. Introduction

The main goal of the analysis is to investigate the pattern of confirmed covid-19 cases and covid-19 death figures. The main question of the analysis is whether we can predict how many deaths can be expected, if we know the number of confirmed cases in a country. The population of the analysis is the number of people infected with covid-19 disease worldwide. The sample I am working with in this task is a proportion of that population. It includes only the confirmed covid-19 cases until 17.10.2020.

A potential quality issue with the data is about reliability, because the measurement of covid death figures is not unified across different countries. In some countries if someone had a chronicle disease and died due to covid-19, the reason of death counts toward death by chronicle disease. On the other hand, in other countries they count these as covid deaths, what's more if someone died due to a chronicle disease and later the virus is shown in the body it counts as covid-19 death case.

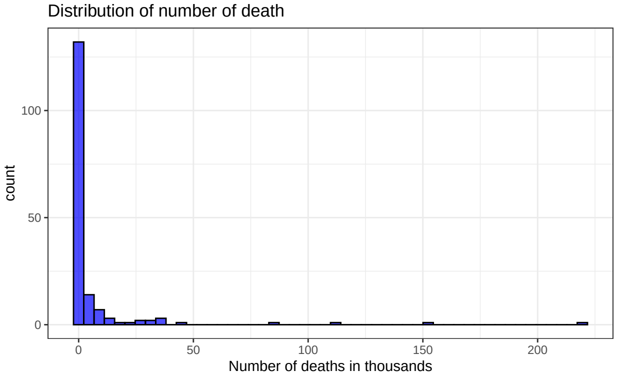
3. Selection and scaling of observations

The main variables of the dataset to be analyzed are the countries, their population in millions (downloaded from WDI), the number of confirmed and active covid cases in thousands and number of recoveries and deaths in thousands (collected by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University).

During the cleaning process I excluded all the grouping observations like EU, OECD and others. In addition all the countries missing population, confirmed cases or death (all together 39) are excluded. Furthermore I decided to narrow the scope to countries with non-zero death figures to make the logarithmic transformations interpretable. Thus I excluded 12 countries with zero registered death cases and ended up with 170 observations. The excluded 12 countries represented an insignificant ratio with only 4% from the population of the full sample and 0.003% of all confirmed cases.

4. Histograms and summary statistics of the X and Y variables

According to the histograms, both the number of deaths and confirmed cases are following an exponential distribution with most countries represented in the first bin. Both confirmed and death variables are right skewed with some extreme large values. These extreme values are not errors, the reason behind is the different handling of the pandemic by countries.



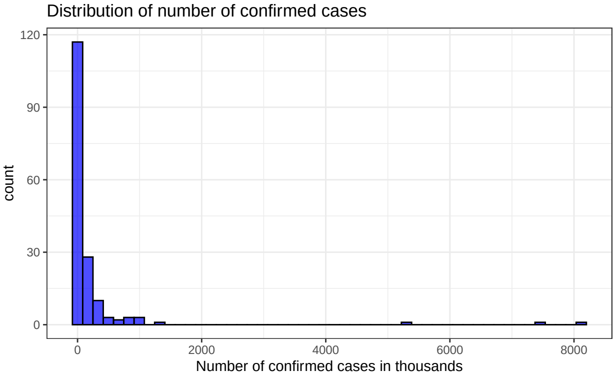


Table 1: Summary stat of death and confirmed cases (in thousands)

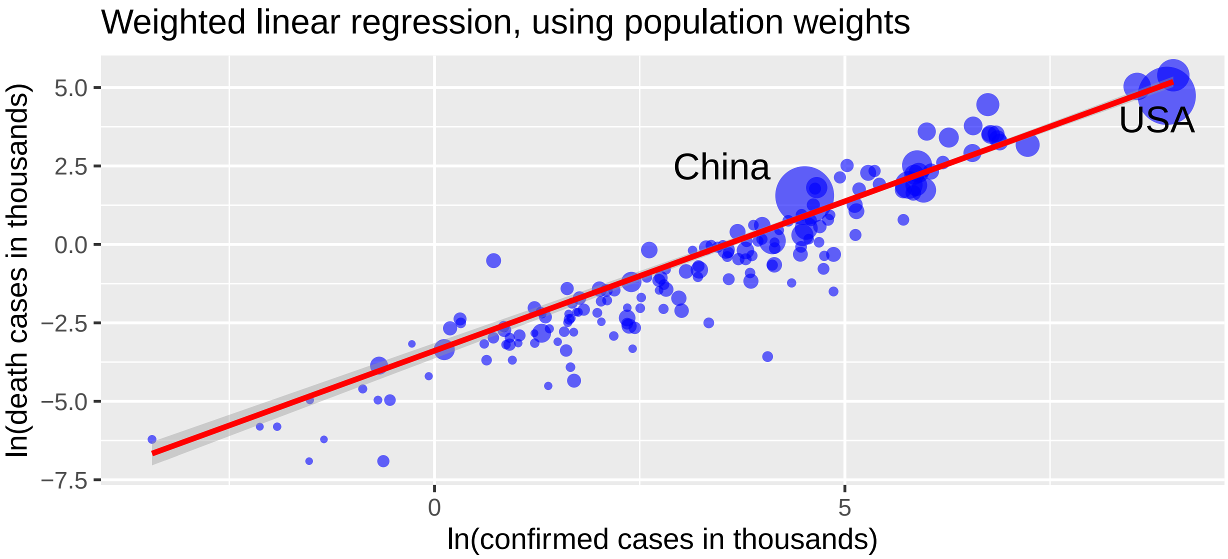
variable	mean	median	std	iq_range	min	max	skew	numObs
Death	6.53	0.35	23.89	1.77	0.00	219.49	6.42	170
Confirmed	233.30	24.85	943.56	99.69	0.03	8116.94	7.06	170

5. Transformation of variables and comparison

By checking the possible logarithmic transformations of confirmed and death variables, it seems that if we turn only one variable to logarithm and the other we keep level, the fit of the linear regression will be poor with an R-squared around 0.2. By transforming the variables to logarithm they take a normal distribution which is easier to interpret and analyze. The log-log transformed variables seem to have the best fit with a 0.886 R-squared. Taking the log-log transformation means that I will associate the percentage change of confirmed cases with the percentage change of death figures. The graphs of the transformed variables can be found in Appendix 1.

6. Choosen model

I selected the log-log linear regression model, weighted with population, as the best-fit regression. The model has the highest R-squared with 0.9283 and the magnitude of the coefficients are meaningful. The formula of the model is ln_death = -3.388 + 0.952 * ln_confirmed. Looking at the model, the alpha parameter can not be interpreted meaningfully, because in case of log the intercept is not useful. The beta parameter is more important, it tells us that countries with 10% more confirmed covid cases can see on average 9.5% more deaths from covid-19. Or in other words a 10% increase in confirmed covid cases can result on average 9.5% more deaths from covid-19 in the country. The graphs and comparison of the various models can be found in Appendix 2.



Hypothesis testing on the beta parameter

In the following I would like to the test the association between confirmed cases and number of deaths in a hypothesis. My null hypothesis is that beta equals to zero and the alternative hypothesis is that it is not equal to zero. I selected a 99% confidence interval for the test, which means we accept only 1% false positive. According to the results of the hypothesis testing, the t-value is 15.12 which is far away from 2.6 (99% CI), in addition the p-value is almost zero, which tells us, that the possibility of giving a false positive is almost 0%. As a result we can reject the H0 hypothesis that the beta coefficient on log confirmed cases is equal to zero. The correlation between the two variables seems to significant.

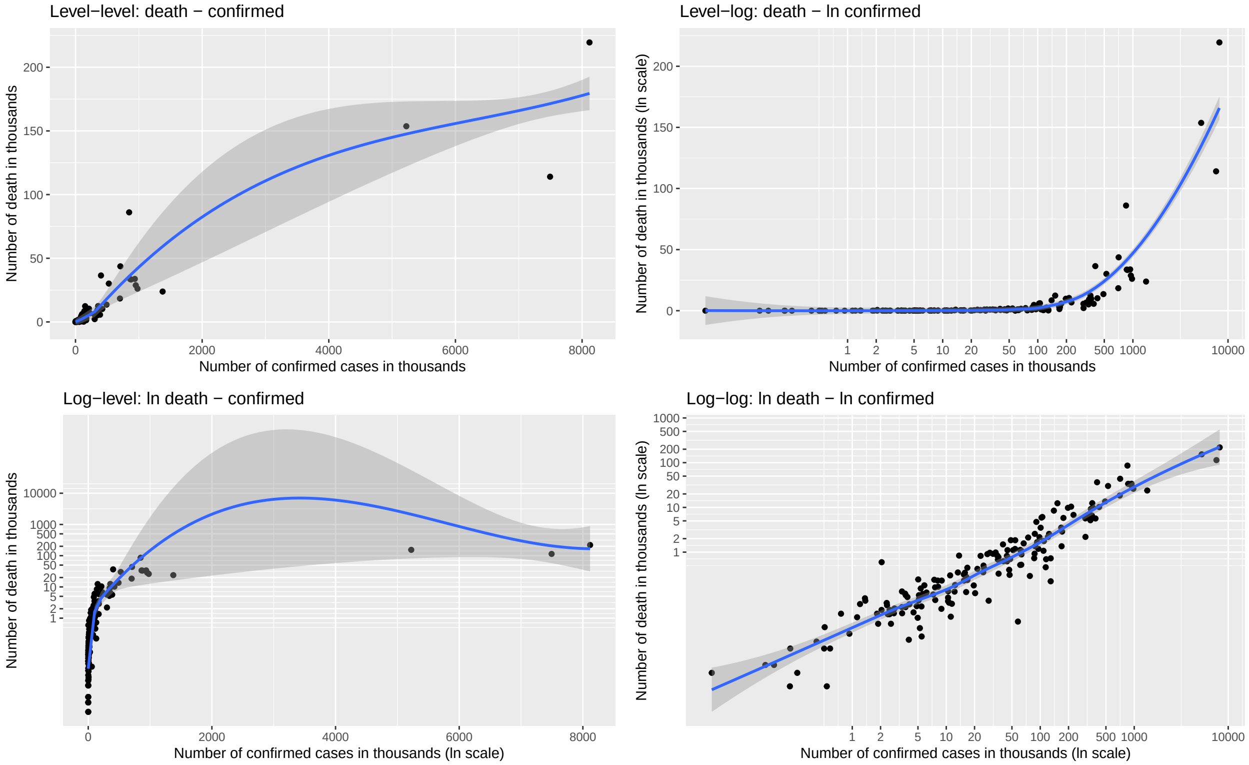
Analysis of the residuals

The table below contains the best 5 and worst 5 countries regarding actual log number of deaths versus the predicted numbers by the model. In the best 5 countries we can see the countries like Singapore, where the number of deaths are far better than the predicted number, probably because the country has one of the highest quality healthcare system around the world. I means that it is the best place to be in these troubling times. On the other hand we can find countries like Italy, Mexico or the United Kingdom between the worst performers, where the number of death is well above the model predictions. These bad results can be outcome of the bad society habits about social distancing or because of the underestimatement of the power of covid-19.

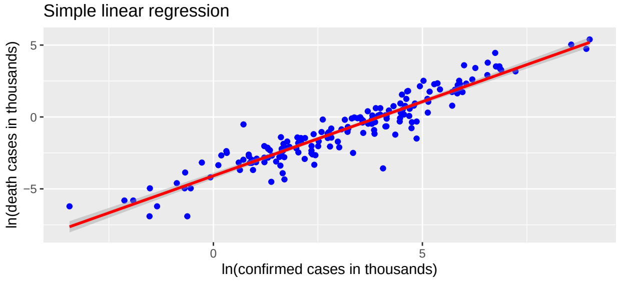
Table 2: Countries with largest negative and positive errors

Best_country	Ln_death	Prediction	Residual	Worst_country	Ln_death	Prediction	Residual
Burundi	-6.91	-3.98	-2.93	Ecuador	2.52	1.40	1.12
Iceland	-4.51	-2.07	-2.44	Italy	3.60	2.32	1.27
Qatar	-1.50	1.24	-2.74	Mexico	4.46	3.03	1.42
Singapore	-3.58	0.48	-4.05	United Kingdom	3.78	2.86	0.91
Sri Lanka	-4.34	-1.77	-2.57	Yemen	-0.52	-2.70	2.18

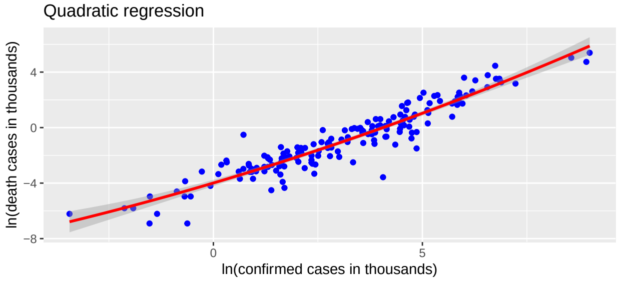
Appendix 1 - Logarithmic transformations



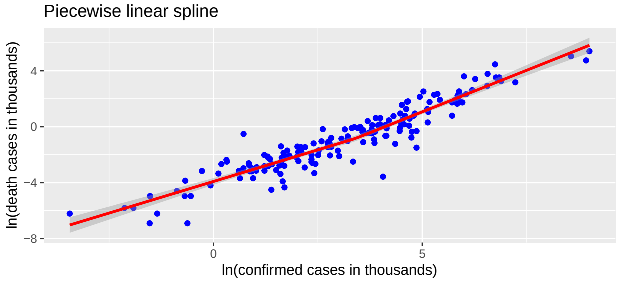
Appendix 2 - Regression models



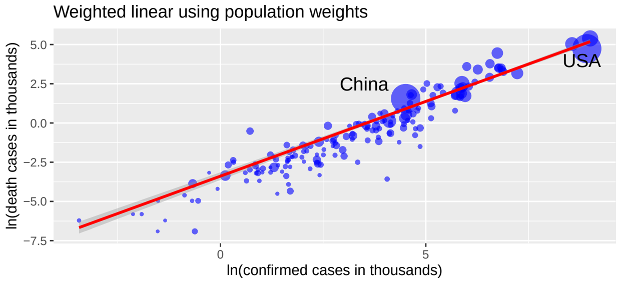
```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    -4.091    0.11672  -35.05 3.615e-79  -4.3216   -3.861 168
## ln_confirmed    1.031    0.02908   35.44 7.081e-80   0.9733    1.088 168
##
## Multiple R-squared:  0.8868 ,    Adjusted R-squared:  0.8861
## F-statistic:  1256 on 1 and 168 DF,  p-value: < 2.2e-16
```



```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed + ln_confirmed_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept)     -3.99104   0.123800 -32.238 1.313e-73 -4.235455 -3.74662 167
## ln_confirmed     0.89054   0.062509  14.247 1.184e-30  0.767135  1.01395 167
## ln_confirmed_sq  0.02287   0.008289   2.759 6.438e-03  0.006508  0.03924 167
##
## Multiple R-squared:  0.8911 ,    Adjusted R-squared:  0.8898
## F-statistic: 708.9 on 2 and 167 DF,  p-value: < 2.2e-16
```



```
##
## Call:
## lm_robust(formula = ln_death ~ lspline(ln_confirmed, cutoff_ln),
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                         -3.9170    0.13789  -28.41 7.680e-66
## lspline(ln_confirmed, cutoff_ln)1   0.9045    0.05581   16.21 4.256e-36
## lspline(ln_confirmed, cutoff_ln)2   1.1890    0.05031   23.64 3.746e-55
##                                    CI Lower CI Upper  DF
## (Intercept)                         -4.1892   -3.645 167
## lspline(ln_confirmed, cutoff_ln)1   0.7943    1.015 167
## lspline(ln_confirmed, cutoff_ln)2   1.0897    1.288 167
##
## Multiple R-squared:  0.8925 ,    Adjusted R-squared:  0.8912
## F-statistic: 780.6 on 2 and 167 DF,  p-value: < 2.2e-16
```



```
##
## 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)   -3.3883    0.36663  -9.242 1.059e-16   -4.112   -2.664 168
## ln_confirmed   0.9523    0.06297  15.124 3.624e-33    0.828    1.077 168
##
## Multiple R-squared:  0.9287 ,    Adjusted R-squared:  0.9283
## F-statistic: 228.7 on 1 and 168 DF,  p-value: < 2.2e-16
```

Model comparison

As we can see from the results all four models are providing a good fit with an R-squared above 0.886. All of the models have a very small p-value and large t-values. On the graphs it is easy to see that the lines barely changes. But the reason behind choosing the weighted linear regression was it's highest R-squared value of 0.9283 and had one of the smallest p-values. In addition the weighted model works well with the population weights considering larger countries as more important and smaller countries with less priority. I think it is a usable method as smaller countries can treat the pandemic better and can focuse easier on stopping th virus than larger countries where it needs more resources and effort.