COVID-19 Analysis

Julianna Szabo

Executive summary

In this project I have analyzed the correlation between COVID-19 cases per 10'000 people and COVID-19 deaths per 10'000 people. I have found that they have a linear correlation best explained by a Weighted linear model using population as weights. The model shows that usually mode cases lead to more deaths, but population in the country play a role since countries with higher population doing a good or a bad job dealing with the pandemic have an affect on the curve. The strength of my results are the great fit of the model to the data, but the weakness is that there might be some misreposrting in the data, that could lead to skewed results.

Introduction

My question for this analysis is: What is the correlation between registered cases per capita and deaths per capita?

My y variable will be Deaths per capita and my x will be Registered Cases per capita

The original data comes in absolute numbers instead of per capita so that transformation may cause some inaccuracies since population numbers are from last year. My population is all the cases and all the deaths by COVID-19 throughout the world. My sample is therefore very relevant since it has that data, although some countries may not be as accurate when recording these, overall it is a relatively good representation of the population.

Both confirmed cases and deaths are skewed with a right tail (see Appendix 1). This may be lessened in the per capita numbers calculated later. Further, the analysis may benefit from a log transformation later. There seem to be some extreme values in both confirmed and deaths

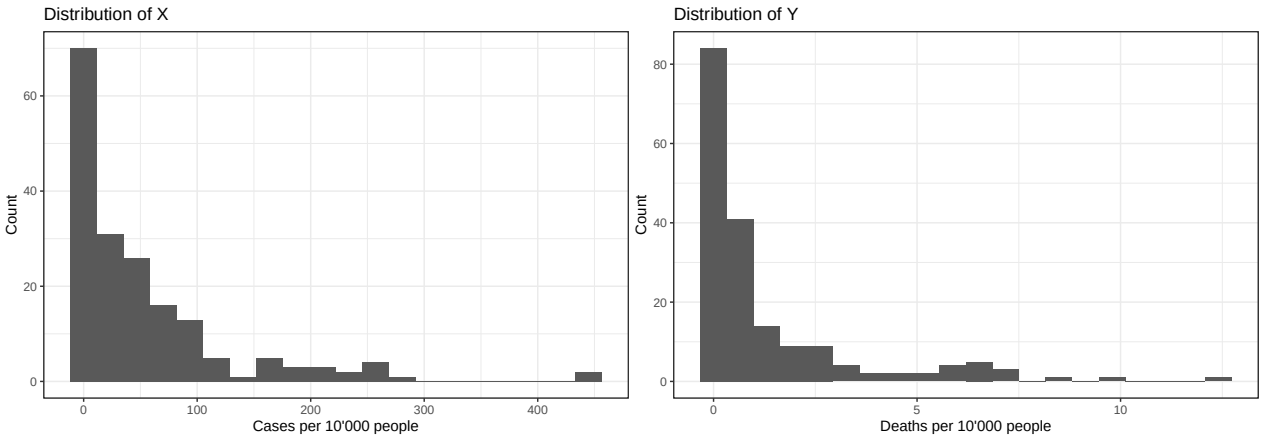
While with the summary one can see that the range is huge, I have decided to keep all the observations, since most of the extreme values will most likely be resolved once we do a per capita transformation.

Create new variables

Scaling

Looking at the graphs (see Appendix 1) from a scaling perspective, I have decided to transform the PPC numbers into per 10000 people instead of per one person that way they are easier to interpret because the numbers on the scale are easier to understand and compare.

Distribution of variables



variable	mean	median	min	max	sd
x	54.376880	26.2318309	0.0320805	444.75996	75.682308
y	1.312618	0.4027647	0.0000000	12.40402	2.115451

Overall both x and y have many observations around zero and have a long right tail. While x has higher number overall the distribution is very similar in shape.

Modeling

ln transformation

Looking at the graphs (see Appendix 2) the log - log transformation seems to be the best fit. Substantively it is easy to interpret and works with percentages which is great for comparison. It is also more pleasant to look at since the extreme values have been fixed by the log transformation. Statistically it seems to be the most collective and the CI seems the least spread out around most of it.

To be able to use this in the future I need to remove some values that give infinity when doing the transformation.

Choice of model

I have decided to pick the Weighted linear regression, based on Appendix 3.

ln(deaths per 10'000) = -3.3789 + 0.9 * ln(cases per 10'000), weighted by population

Alpha says there are on average -3.38 ln(deaths per 10'000 people) in a country because of COVID-19 when there are ln(cases per 10'000 people) is zero.

Beta means that there will be 0.9 ln(deaths per 10'000 people) in a country for every 1% increase in ln(cases per 10'000 people) of COVID-19.

The weights mean that countries with a higher population are given a larger impact on the slope.

Hypothesis testing

For this hypothesis test, I will use a 95% Confidence Interval (CI).

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	-3.3789366	0.2904825	-11.63215	0	-3.9523783	-2.805495	169
ln_cases_ppc	0.9007591	0.0812900	11.08080	0	0.7402844	1.061234	169

The t-value for the slope is aproximately 11, which is way above the value of 2 for CI 95%. Therefore we can reject the null. Also the p-value is very small making this prediction for the t-value correct with a high likeliness.

Residual analysis

Countries who saved the most people per 10'000 in COVID

country	ln_deaths_ppc	reg4_y_pred	reg4_res
Burundi	-7.05	-4.114	-2.936
Maldives	-0.4457	1.371	-1.817
Qatar	-0.2802	2.113	-2.394
Singapore	-3.05	0.781	-3.831
Sri Lanka	-5.122	-2.983	-2.139

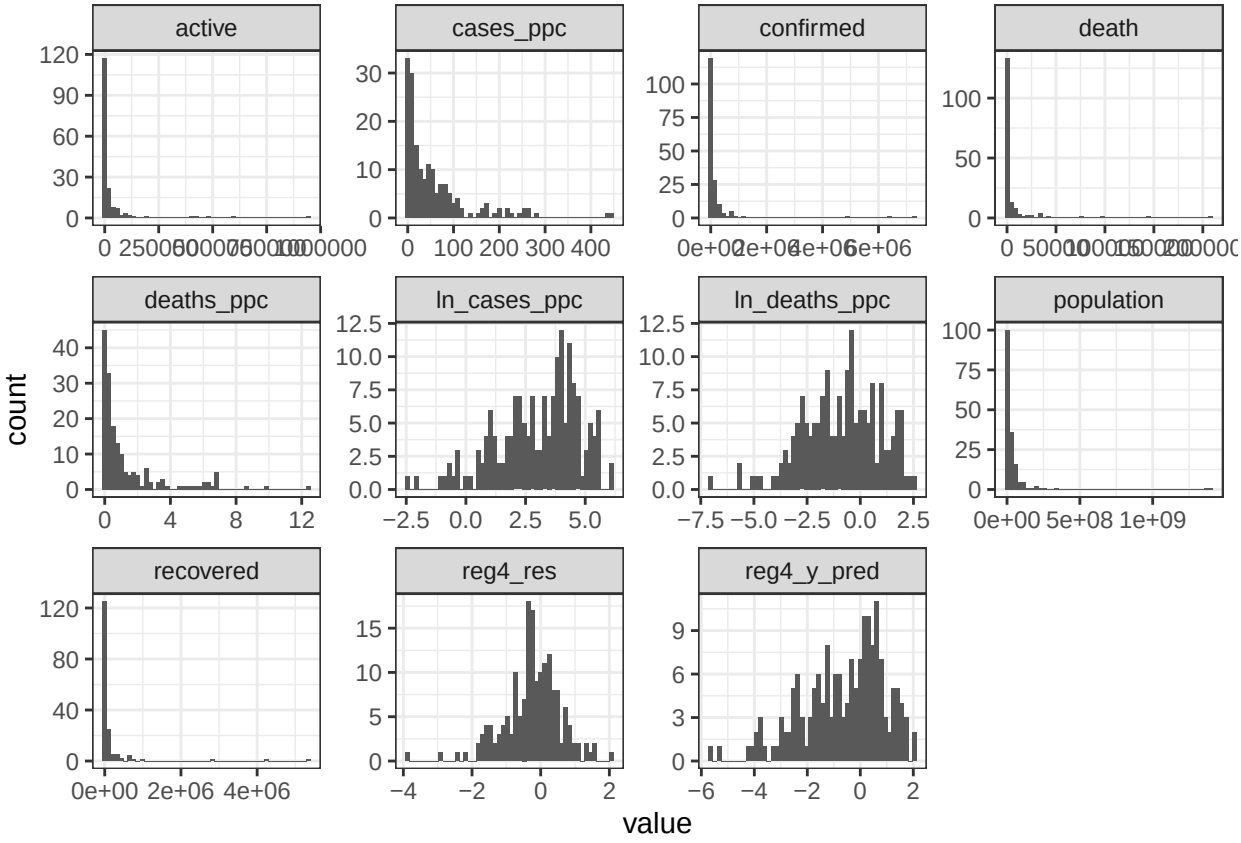
So these five countries had less deaths than expected. In the case of Singapore and the Maldives are known for their fast reaction, but Burundi and Sri Lanka are surprising, since they are not known for developed health care systems or strong leadership.

Countries who lost the most people per 10'000 to COVID

country	ln_deaths_ppc	reg4_y_pred	reg4_res
Belgium	2.167	0.8167	1.35
Italy	1.785	0.1912	1.593
Mexico	1.812	0.2887	1.523
United Kingdom	1.845	0.4381	1.407
Yemen	-1.603	-3.701	2.098

These five counties have had more deaths per 10'000 people than expected. Countries like Italy and Mexico have been heard of as highly affected countries. Yemen also has more deaths than expected, probably because of their underdeveloped health care system.

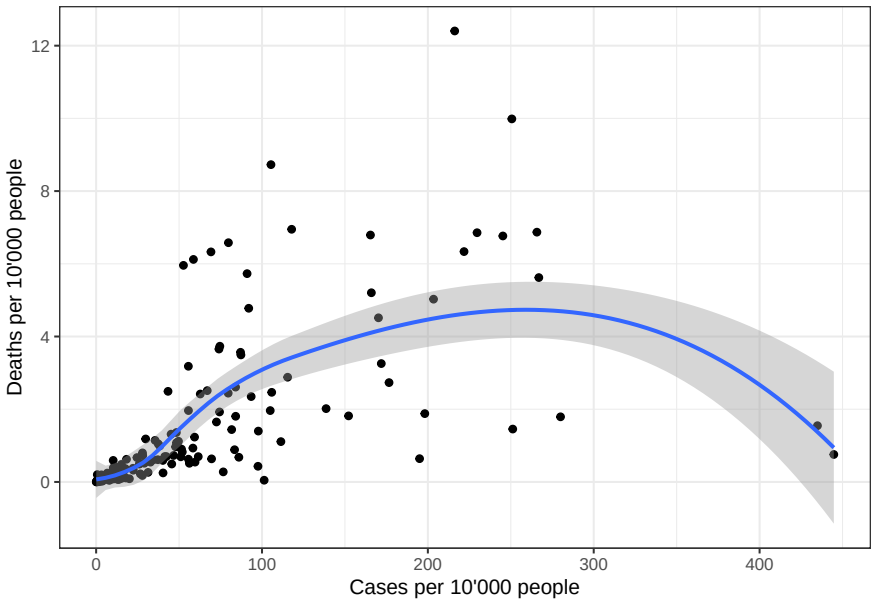
Appendix 1 - Looking at the data



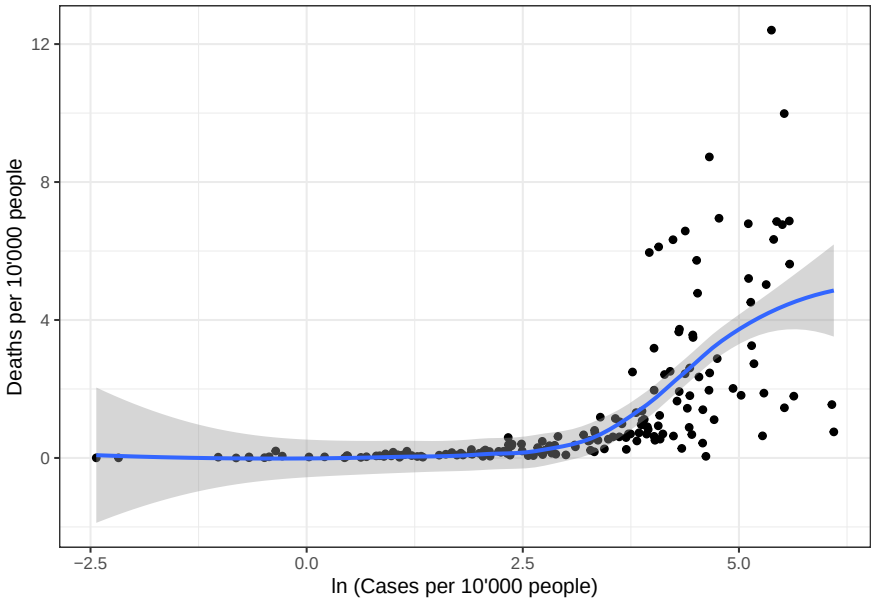
variable	mean	median	min	max	sd
confirmed	200484.678	16827.0	32	7280834	832563.39
death	5982.509	318.0	1	207977	22238.43
recovered	139483.287	10014.0	0	5343837	575463.89
active	30757.310	2813.5	1	943069	104517.70
population	44271570.865	10269417.0	33860	1397715000	153894099.59

Appendix 2 - Modeling

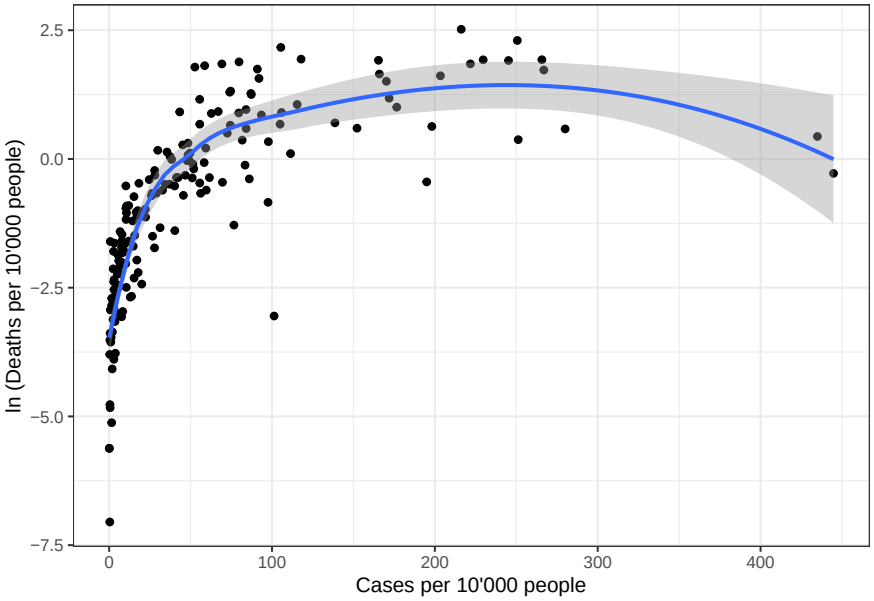
1, Level - level regression



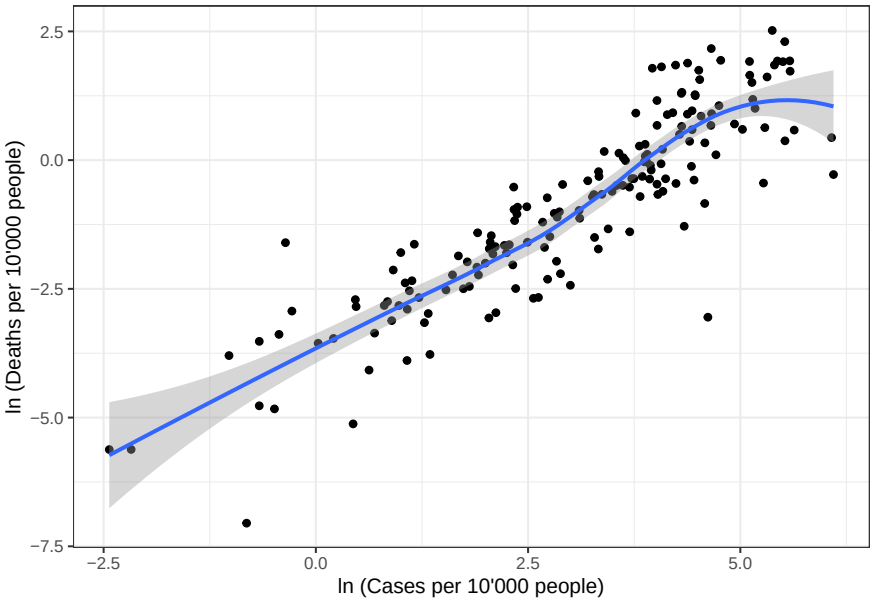
2, Log - level regression



3, Level - log regression



4, Log - log regression



I have decided to go with a log-log transformation. For more details, please see the body of the report.

Apendix 3 - Regression Models

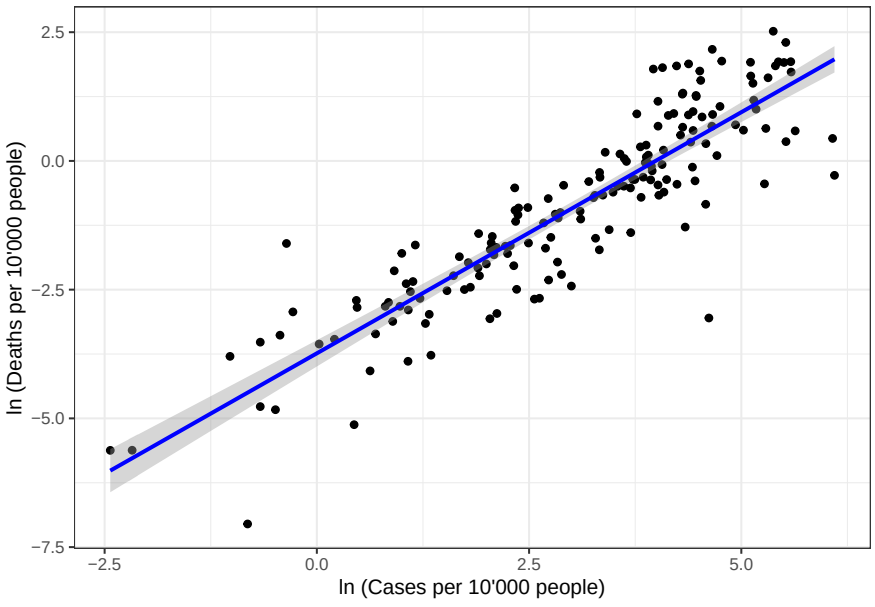
Different models: reg1: ln_deaths_ppc = alpha + beta * ln_cases_ppc reg2: ln_deaths_ppc = alpha + beta_1 * ln_cases_ppc + beta_2 * ln_cases_ppc^2 reg3: ln_deaths_ppc = alpha + beta_1 * ln_cases_ppc * 1(ln_cases_ppc < 50) + beta_2 * ln_cases_ppc * 1(ln_cases_ppc >= 50) reg4: ln_deaths_ppc = alpha + beta * ln_cases_ppc, weights: population

First I will add the square and cube of the x variable to df

```
df <- df %>% mutate( ln_cases_ppc_sq = ln_cases_ppc^2)
```

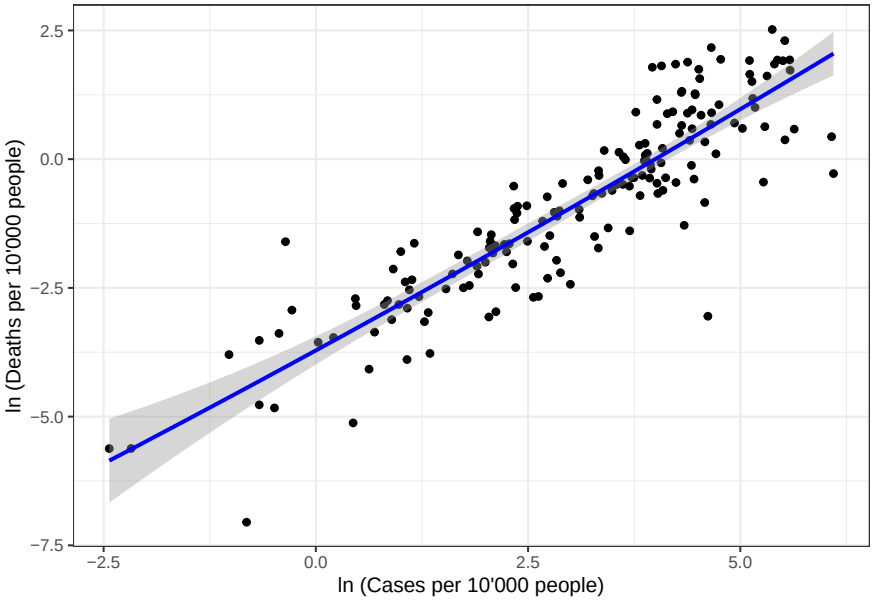
Regression 1 - Simple linear regression

```
##
## Call:
## lm_robust(formula = ln_deaths_ppc ~ ln_cases_ppc, data = df,
##     se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -3.7364    0.14610  -25.57 5.506e-60  -4.0249   -3.448 169
## ln_cases_ppc   0.9364    0.04472   20.94 7.931e-49   0.8482    1.025 169
##
## Multiple R-squared:  0.7911 ,    Adjusted R-squared:  0.7898
## F-statistic: 438.6 on 1 and 169 DF,  p-value: < 2.2e-16
```



Regression 2 - Quadratic (linear) regression

```
##
## Call:
## lm_robust(formula = ln_deaths_ppc ~ ln_cases_ppc + ln_cases_ppc_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                  Estimate Std. Error  t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)     -3.714940    0.16147 -23.0075 8.394e-54 -4.03370 -3.39618 168
## ln_cases_ppc     0.898380    0.10253   8.7620 2.018e-15  0.69596  1.10080 168
## ln_cases_ppc_sq  0.007721    0.01923   0.4016 6.885e-01 -0.03024  0.04568 168
##
## Multiple R-squared:  0.7913 ,    Adjusted R-squared:  0.7888
## F-statistic: 215.9 on 2 and 168 DF,  p-value: < 2.2e-16
```



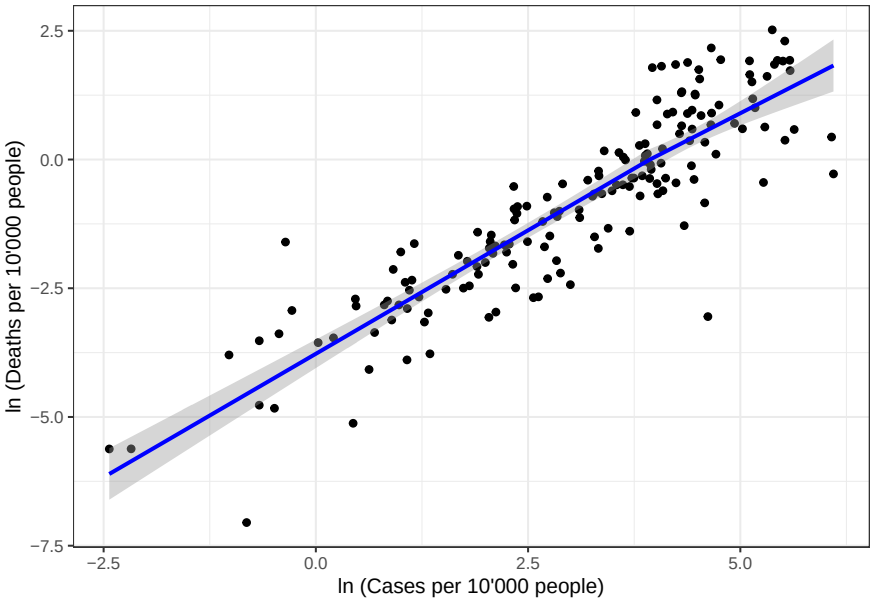
Regressipn 3 - Piecewise linear spline regression

First we create a cut off point and trasnform it into a log

```
cutoff <- 50
cutoff_ln<- log( cutoff )
```

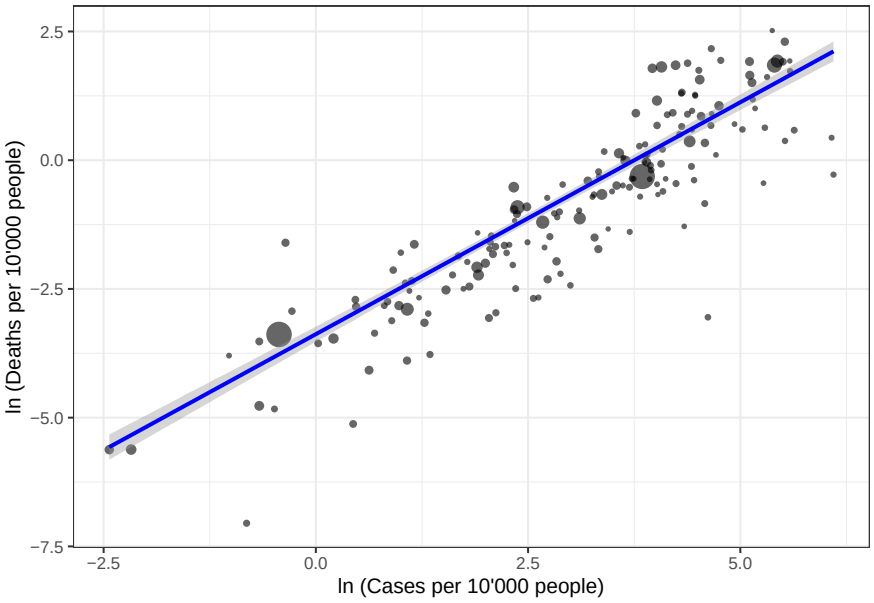
```
##
## Call:
## lm_robust(formula = ln_deaths_ppc ~ lspline(ln_cases_ppc, cutoff_ln),
##     data = df)
##
## Standard error type:  HC2
##
```

```
## Coefficients:
##                                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                        -3.7730    0.16458 -22.925 1.328e-53
## lspline(ln_cases_ppc, cutoff_ln)1   0.9593    0.05819  16.486 6.086e-37
## lspline(ln_cases_ppc, cutoff_ln)2   0.8445    0.17648   4.785 3.719e-06
##                                   CI Lower CI Upper  DF
## (Intercept)                        -4.0979   -3.448 168
## lspline(ln_cases_ppc, cutoff_ln)1   0.8444    1.074 168
## lspline(ln_cases_ppc, cutoff_ln)2   0.4961    1.193 168
##
## Multiple R-squared:  0.7916 ,    Adjusted R-squared:  0.7892
## F-statistic: 220.1 on 2 and 168 DF,  p-value: < 2.2e-16
```



Regression 4 - Weighted linear regression, using population as weights

```
##
## Call:
## lm_robust(formula = ln_deaths_ppc ~ ln_cases_ppc, data = df,
##     weights = population)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -3.3789    0.29048  -11.63 2.381e-23  -3.9524   -2.805 169
## ln_cases_ppc   0.9008    0.08129   11.08 8.523e-22   0.7403    1.061 169
##
## Multiple R-squared:  0.8981 ,    Adjusted R-squared:  0.8975
## F-statistic: 122.8 on 1 and 169 DF,  p-value: < 2.2e-16
```



Comparing models

On the table, it is clear to see that all four of these models fit well to the data. In the linear model we have an alpha of -3.74 and a beta of 0.94. Both of these values have a very low p-value of less than 0.1. The model also has a good R°2 value at .79. These numbers do not change significantly for the quadratic or the PLS model either, which have an apha of -3.71 and -3.77 respectively. In the quardatic, one can see that the sqared term has almost no impact on the line since it's coefficient is only 0.01. The PLS also does not change the line-of-best-fit much since the slopes of the two lines are relatively close togehter. While the alpha and beta also do not change significantly for the Weighted linear regression the R^2 increases from 0.79 for all the other models to 0.9.

Reasons for picking Weighted linear regression

Substansive:

Since more people usually means more cases, it makes sense to add * population as a weight, even with per capita numbers. Especially in his case, some large counties who are doing great could be gives more weight than some small country that is doing very bad.

Statistical:

The weighted linear regression has the best R^2 of the four models. Further its coefficients are similar to the other regressions and have very small p-values (under 0.01).