DA2 - Assignment 1 - COVID-19 Analysis

Dominik Gulacsy

28/11/2020

General Introduction

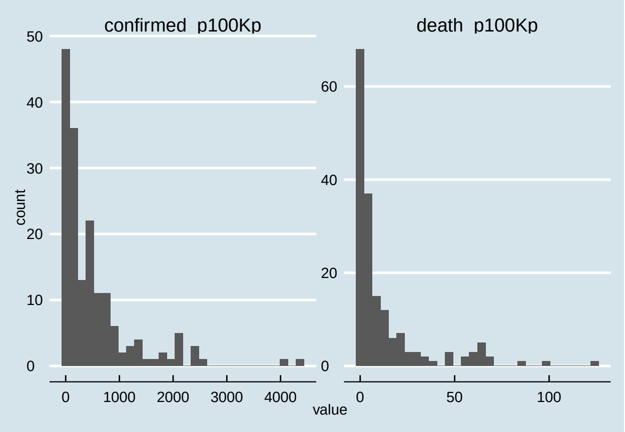
Aim of Analysis

The goal of this analysis is to look at what kind of patterns of association can be discovered between COVID-19 deaths and confirmed cases. One possible final objective of such analysis is to evaluate the healthcare system of different countries to a certain extent. However, this interpretation of model results should be considered with carefulness and limitations has to be made. This is partly because in the analysis only one particular aspect of the situation is taken into account, deaths compared to confirmed cases. Furthermore, results are not controlled for many other variables that may also affect a country's numbers (e.g.: demographic structure and sociological dynamics) AKA confounders.

Variable description and Data Quality Assessment

My variables of focus in this analysis are deaths per capita and confirmed cases per capita for each country it is available. All COVID-19 data is from administrative sources and reported by the countries themselves. The process of data gathering may differ from countries to countries and even within the country. For example, in larger countries like India the process may be different in different counties/municipalities. In case of this dataset, we may have multiple interpretations of how the population can be defined. One possibility is to consider the population as all COVID-19 infection cases that occurred until 21th Sept 2020. Therefore the gathered data represents only the part of the population which was effectively observed (mainly by testing). An other way to look at it is to say that the population is the infinite number of possible outcomes of COVID-19 infections and consequences and this data only shows one realization of such random variable. One of the issues regarding data quality is that there are some countries where the numbers reported may be influenced due to political reasons. So these countries' numbers may be significantly lower than actually they are. Secondly, reliability is also questionable. It is hard to believe that if cases were recounted they would be the same for the same observations. Most likely the data contains many errors due to duplication, mishandling and so on. ## Exploratory Data Analysis # Drop irrelevent observations In case of those countries that have very low confirmed cases, numbers do not really reflect their capability to combat the virus. To have countries where there is at least some information on their performance I dropped countries with lower than 100 confirmed cases to only include more exposed countries in the analysis. Death per capita and confirmed cases per capita are rather small numbers so I scaled the variables to show deaths and confirmed cases per 100K persons. This way they are more interpretable.

Summary Statistics



death_p100Kp	confirmed_p100Kp
Min. : 0.0000	Min. : 0.878
1st Qu.: 0.9725	1st Qu.: 62.749
Median : 4.5040	Median : 239.845
Mean : 12.8750	Mean : 515.034
3rd Qu.: 13.7129	3rd Qu.: 671.164
Max. :124.0402	Max. :4364.445

Both confirmed cases and deaths per 100K person have a distribution with a long right tail, similarly to a lognormal distribution. Most countries have a death rate between 0.97 and 13.71 deaths per 100K persons with a mean of 12.88 and maximum of 124.04 (San Marino) deaths per 100K persons. In case of confirmed cases most countries have a rate between 62.75 and 671.16 per 100K persons with a mean of 515.03 and a maximum of 4364 (Qatar) confirmed cases per 100K persons.

Variable Transformation (Taking logs)

Substantial: Level changes in both deaths and confirmed cases are hard to interpret, however it can be resolved by taking the logarithm of the variables. Percentage differences make more sense. Statistical: As it could be seen on the histograms of variables they have distribution similar to lognormal. Therefore taking the logarithm of these variables would result in distribution that are approximately normal. This is favorable characteristic in statistics. It provides much better approximation since it makes the association close to linear. In other cases the degree of non-linearity is very high. For these reasons I decided to take the ln of variables. In order to this I dropped those observations where the number of deaths is 0.

Model Selection

Running models

To find out which model specification is the best I ran the following 4 regression models: 1. Simple linear regression 2. Quadratic (linear) regression 3. Piecewise linear spline regression 4. Weighted linear regression, using population as weights

Results were the following:

	ln(Deaths/100K) - linear	ln(Deaths/100K) - quadratic	ln(Deaths/100K) -
(Intercept)	−3.65*	−3.62*	−3.82*
	[−4.14; −3.15]	[−4.57; −2.68]	[−4.37; −3.28
ln_confirmed_p100Kp	0.95*	0.93*	
	[0.85; 1.04]	[0.54; 1.33]	
ln_confirmed_p100Kp_sq		0.00	
		[−0.04; 0.04]	
lspline(ln_confirmed_p100Kp, cutoff_ln)1			0.99*
			[0.88; 1.09]
lspline(ln_confirmed_p100Kp, cutoff_ln)2			0.30
			[−0.56; 1.17]
R^2	0.78	0.78	0.79
Adj. R^2	0.78	0.78	0.78
Num. obs.	165	165	165
RMSE	0.85	0.85	0.84

* Null hypothesis value outside the confidence interval.

We can see that the simple linear regression provides a pretty good result. Both intercept and coefficient is significant at 5% significance level. R squared is high considering that it only contains one explanatory variable. The model explains 78% of variance in the data. The quadratic model does not really provides a better fit. The quadratic variable's coefficient is insignificant. R squared is nearly the same as in case of the linear model. The piecewise linear regression suffers from the same problems. The beta after the cutoff point is insignificant and R squared is only marginally better. The last model is the weighted linear regression with population as its weights. It has significant coefficients and R squared is considerably higher than in case of other models however the RMSE is quite high.

Finally, I decided to go forward with the simple linear regression. The main reasons behind it is that it is easier to interpret a log-log model for countries. Also it is a much simpler model so coefficients can be interpreted well while it gives a good approximation with a relatively high R squared.

Chosen model:

$$ln(deaths/100K) = \alpha + \beta * ln(confirmed/100K)$$

where the explanatory variable is the natural logarithm of confirmed cases per 100K persons and the dependent variable is the natural logarithm of death per 100K persons.

According to this model the average natural log of deaths per 100K persons is -3.597 when the natural log of confirmed cases per 100K persons is 0 (alpha cannot be intuitively interpreted). Furthermore, this model also tells that deaths per 100K persons is 0.937% higher for countries with one percent higher confirmed cases per 100K persons (beta).

Hypothesis Testing

Let's test the beta if it's zero. I pick a 5% significance level to test the following:

$$H_0 : \beta = 0; H_1 : \beta \neq 0$$

	x
Estimate	0.9452190
Std. Error	0.0473278
t value	19.9717695
Pr(>|t|)	0.0000000
CI Lower	0.8517645
CI Upper	1.0386736
DF	163.0000000

Based on the test results we can reject that the null hypothesis because the p value is close to zero. Therefore beta is statistically significant.

Residual Analysis

Now let's look at which countries did the best and worst relatively, based on model expectation. Those top 5 countries that had a lower death rate than we would have expected based on their number of confirmed cases, using the chosen regression model were the following:

country	death_p100Kp	reg1_y_pred	reg1_res
Bahrain	13.648783	66.20995	-52.56117
Israel	14.061171	36.09943	-22.03825
Kuwait	13.905121	40.40735	-26.50222
Maldives	6.403580	31.73352	-25.32994
Qatar	7.450389	71.78951	-64.33912

As we can see the best performing countries are mostly smaller countries in the Middle East. This might have to do with that in smaller countries measures can be more effectively executed. However, it can also be the case that there is lack of proper documentation procedure in these countries.

Those top 5 countries that had a higher death rate than we would have expected based on their number of confirmed cases, using the chosen regression model were the following:

country	death_p100Kp	reg1_y_pred	reg1_res
Belgium	86.64187	16.14649	70.49538
Ecuador	63.86103	13.23183	50.62920
Peru	96.48897	40.23022	56.25875
San Marino	124.04017	36.52511	87.51505
United Kingdom	62.65785	11.00552	51.65233

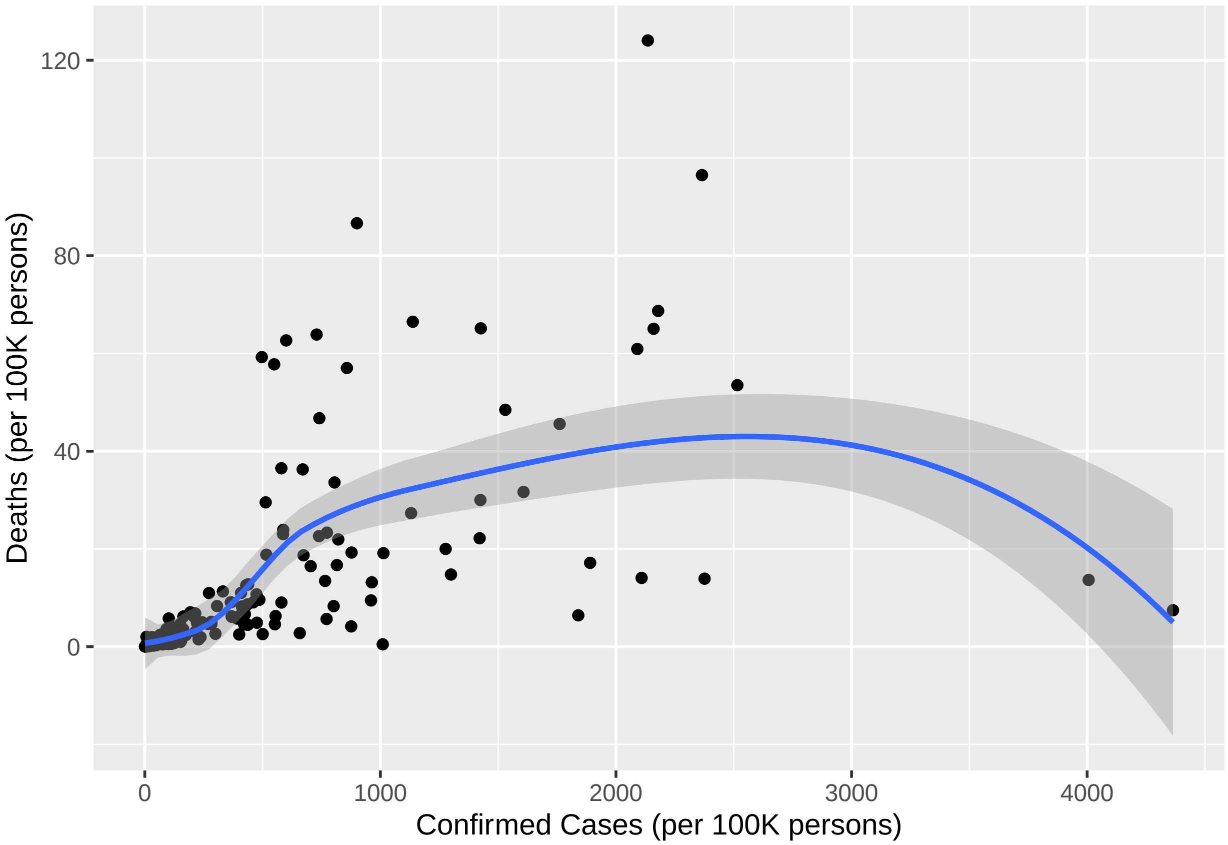
As we can see the worst performing countries are very different from each other. There are European and Latin American countries on the list. Some of them are larger like the UK or Peru and some of them is smaller like Ecuador and San Marino. We may drill down and consider other variables to understand this more deeply, but from this we cannot really say much more.

Executive Summar

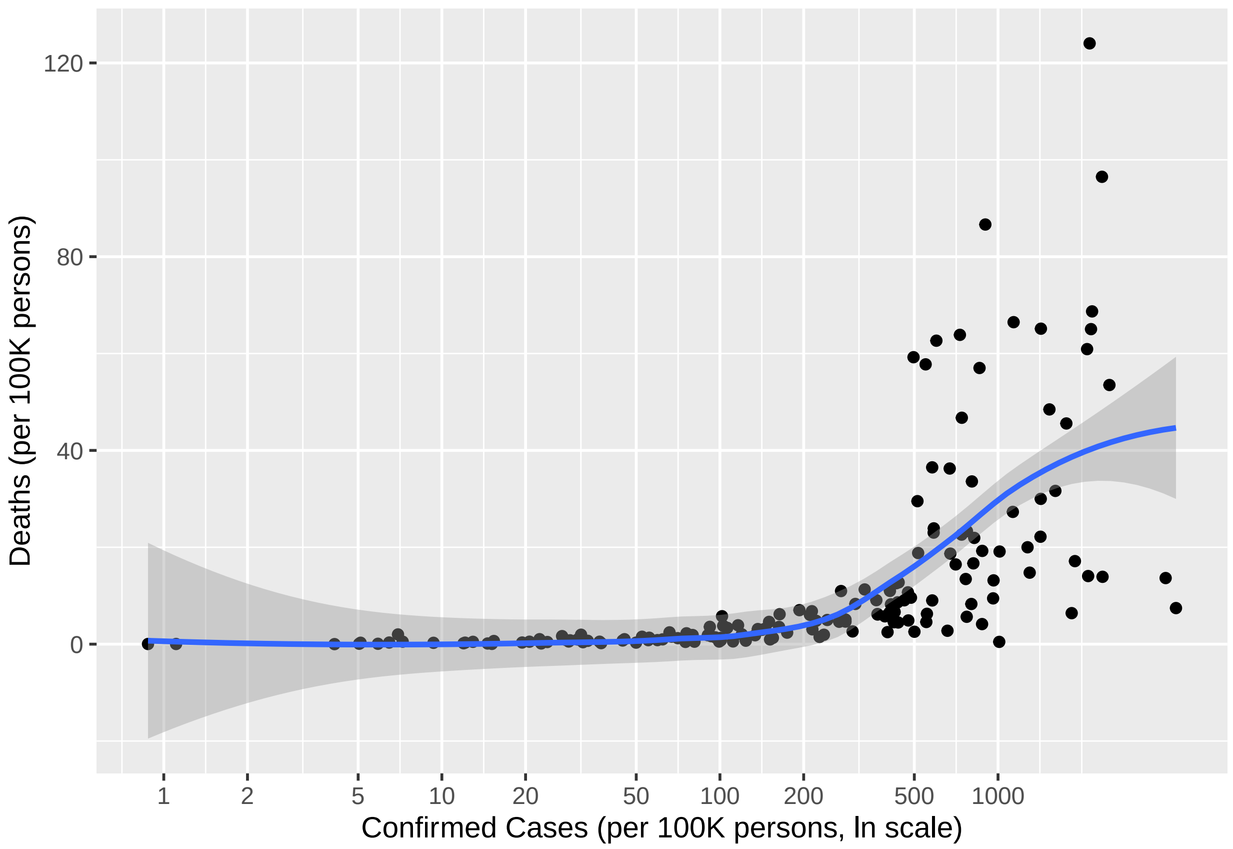
In this analysis I investigated the pattern of association between death rates and confirmed cases. I used deaths per 100K persons and confirmed cases per 100K persons metrics of each country. I took the natural logarithm of these variables to have better interpretation and the preferred characteristics of normality. I ran 4 different model specifications and picked the simple linear regression as my choice of model. The main message of this model is that deaths per 100K persons is 0.937% higher for countries with one percent higher confirmed cases per 100K persons. To have more faith in this model it would worth checking if leaving out some outliers would majorly change the model results. If results would be the same then we could be more confident in this model, however if it would differ dramatically then it would weaken the conclusion of the model.

Appendix

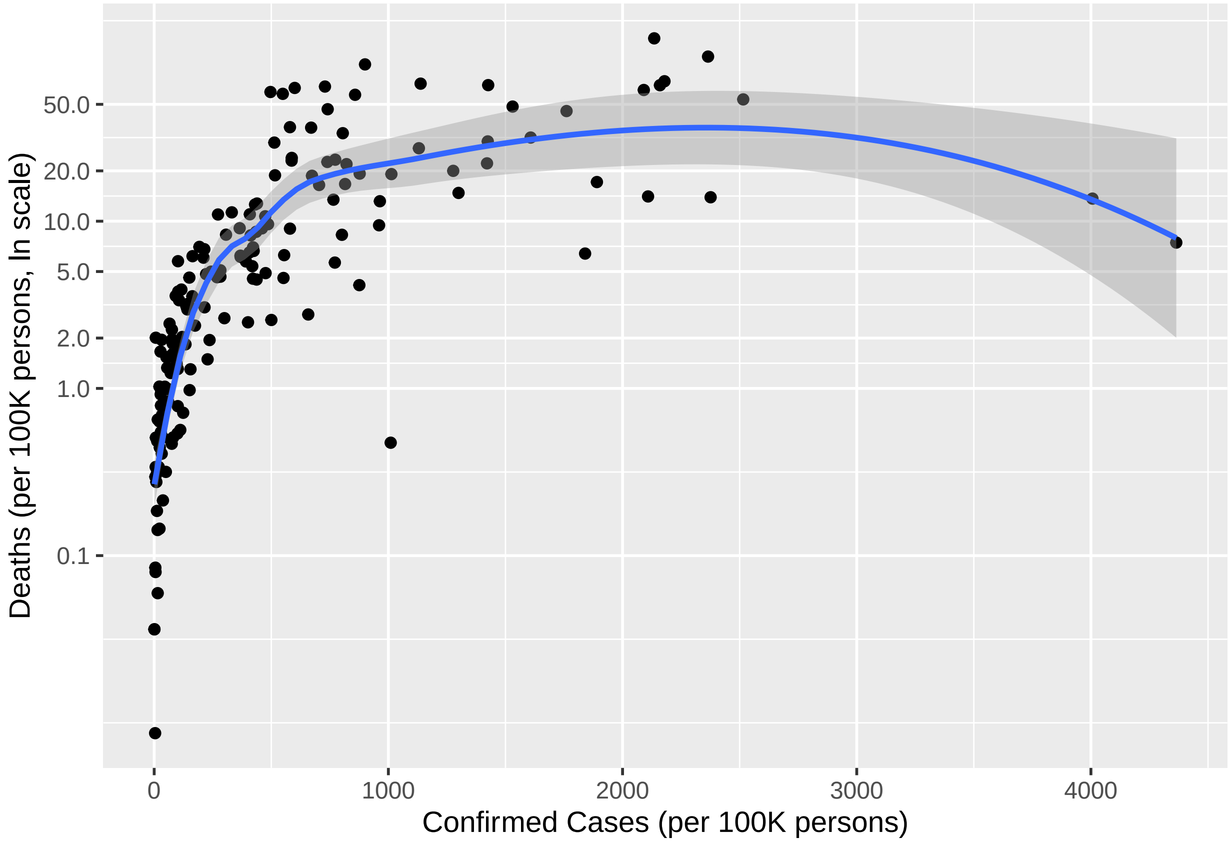
```
## `geom_smooth()` using formula 'y ~ x'
```



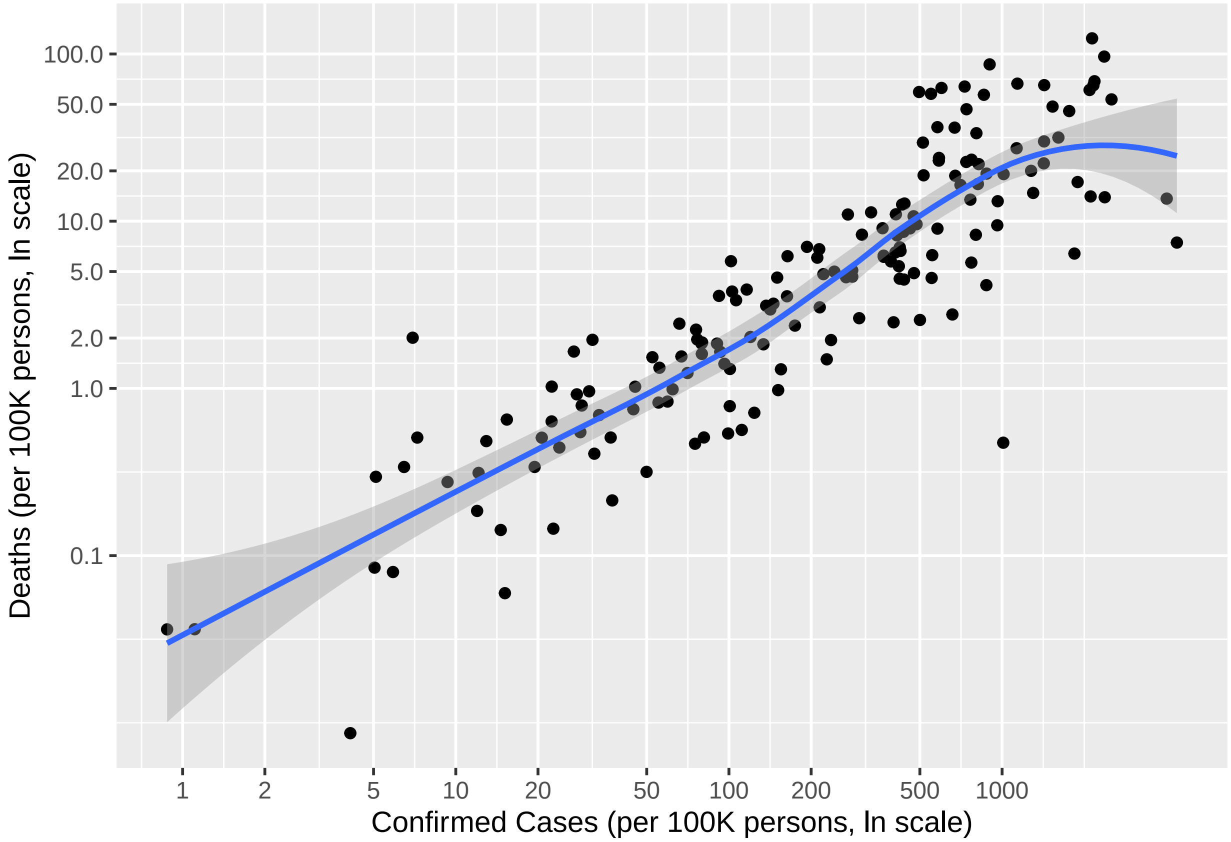
```
## `geom_smooth()` using formula 'y ~ x'
```



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



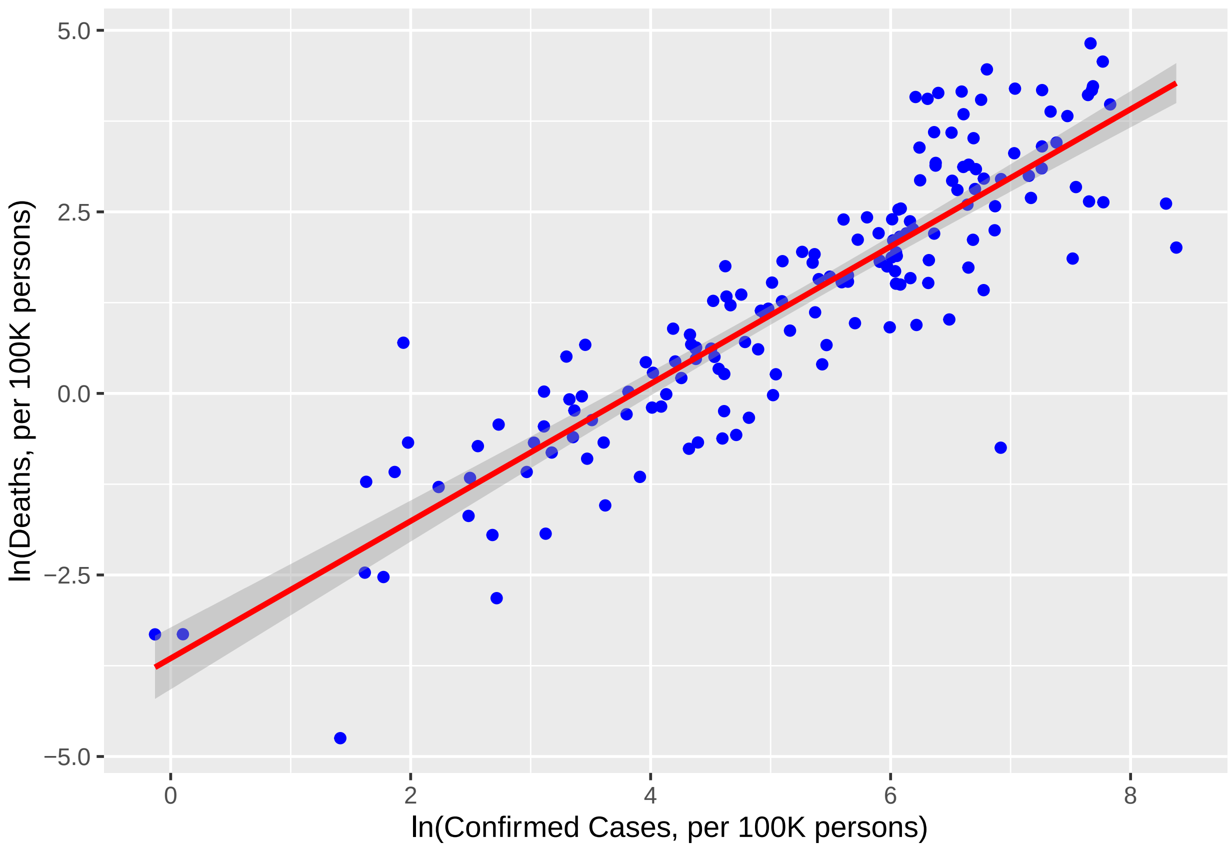
```
## `geom_smooth()` using formula 'y ~ x'
```



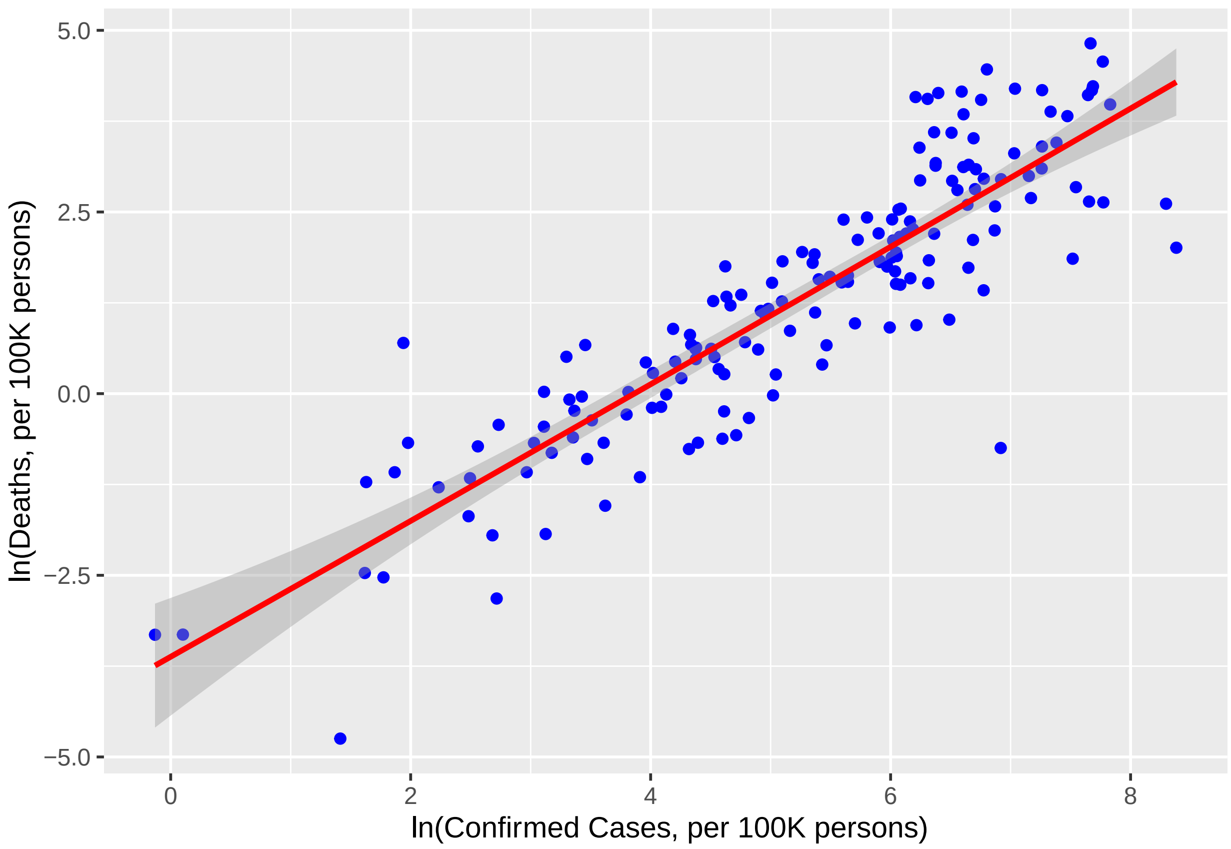
```
##                       Estimate Std. Error   t value     Pr(>|t|)   CI Lower
## (Intercept)          -3.648375 0.25049169 -14.56486 2.593993e-31 -4.1430022
## ln_confirmed_p100Kp   0.945219 0.04732776  19.97177 1.165151e-45  0.8517645
##                       CI Upper  DF
## (Intercept)          -3.153748 163
## ln_confirmed_p100Kp   1.038674 163
```

```
##
## Call:
## lm_robust(formula = ln_death_p100Kp ~ ln_confirmed_p100Kp, data = df,
##     se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                      Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)           -3.6484    0.25049  -14.56 2.594e-31  -4.1430   -3.154 163
## ln_confirmed_p100Kp    0.9452    0.04733   19.97 1.165e-45   0.8518    1.039 163
##
## Multiple R-squared:  0.7821 ,    Adjusted R-squared:  0.7807
## F-statistic: 398.9 on 1 and 163 DF,  p-value: < 2.2e-16
```

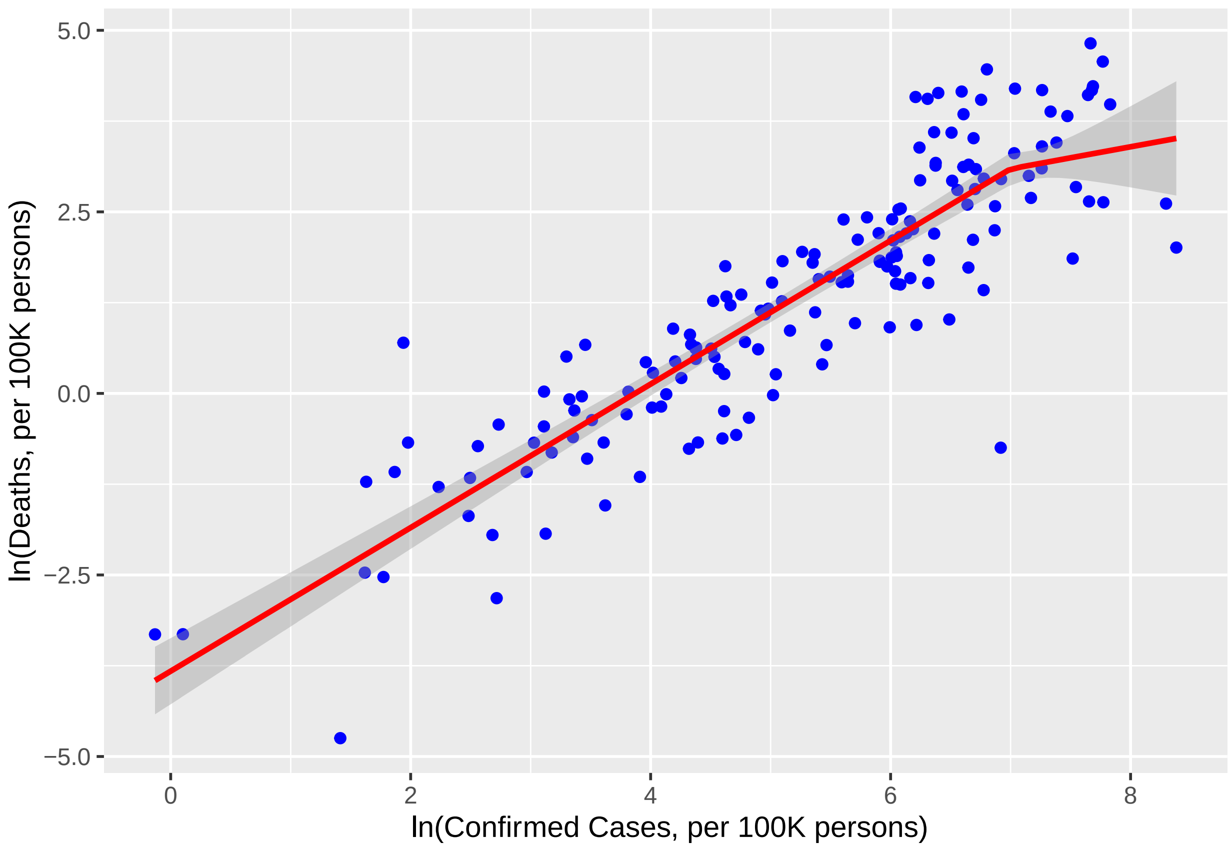
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
##
## Call:
## lm_robust(formula = ln_death_p100Kp ~ ln_confirmed_p100Kp + ln_confirmed_p100Kp_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                         Estimate Std. Error  t value  Pr(>|t|) CI Lower
## (Intercept)            -3.621433    0.47905 -7.55965 2.832e-12 -4.56742
## ln_confirmed_p100Kp     0.931901    0.20098  4.63669 7.252e-06  0.53501
## ln_confirmed_p100Kp_sq  0.001413    0.02115  0.06678 9.468e-01 -0.04036
##                        CI Upper  DF
## (Intercept)            -2.67545 162
## ln_confirmed_p100Kp     1.32879 162
## ln_confirmed_p100Kp_sq  0.04318 162
##
## Multiple R-squared:  0.7821 ,    Adjusted R-squared:  0.7794
## F-statistic: 196.8 on 2 and 162 DF,  p-value: < 2.2e-16
```

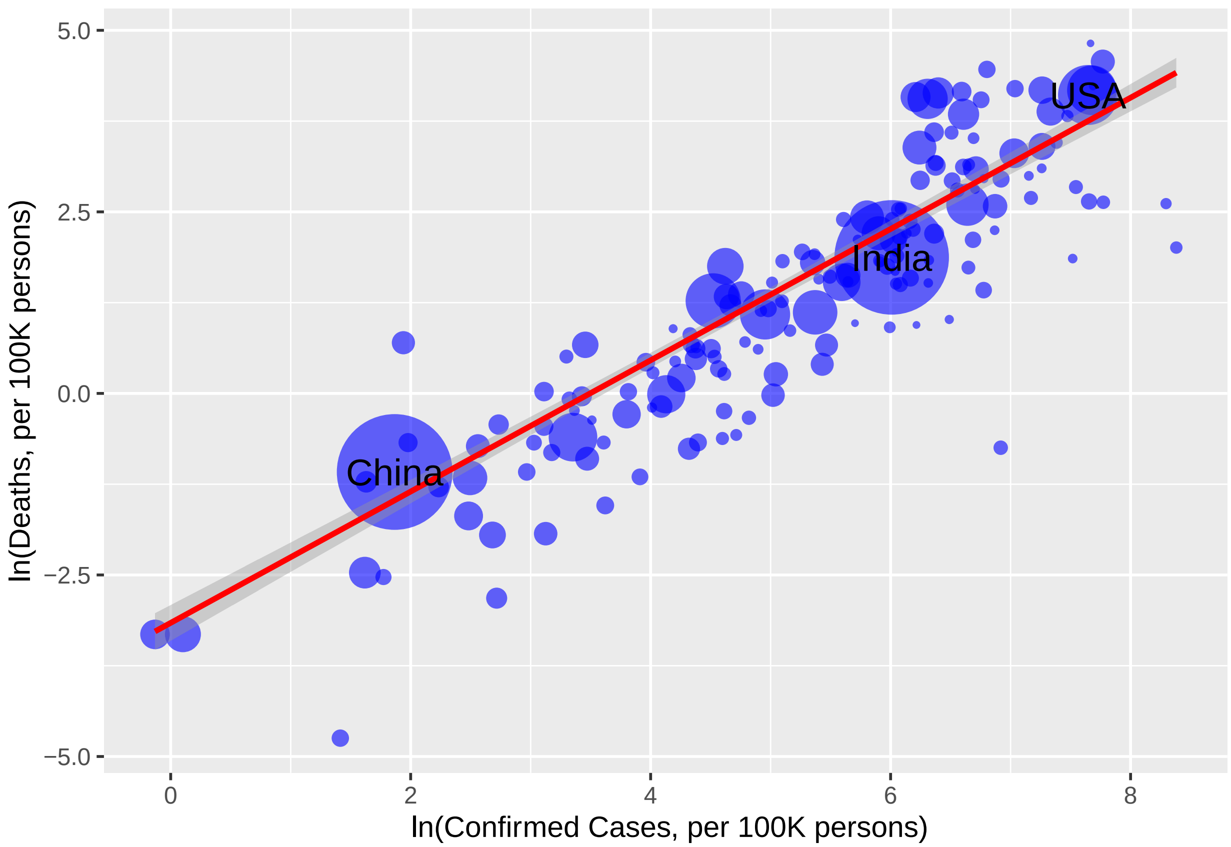


```
##
## Call:
## lm_robust(formula = ln_death_p100Kp ~ lspline(ln_confirmed_p100Kp,
##     cutoff_ln), data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                           Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                                -3.8248    0.27546 -13.8851 2.235e-29
## lspline(ln_confirmed_p100Kp, cutoff_ln)1   0.9882    0.05393  18.3227 2.486e-41
## lspline(ln_confirmed_p100Kp, cutoff_ln)2   0.3035    0.43841   0.6922 4.898e-01
##                                           CI Lower CI Upper  DF
## (Intercept)                                -4.3687   -3.281 162
## lspline(ln_confirmed_p100Kp, cutoff_ln)1   0.8817    1.095 162
## lspline(ln_confirmed_p100Kp, cutoff_ln)2  -0.5623    1.169 162
##
## Multiple R-squared:  0.7875 ,    Adjusted R-squared:  0.7849
## F-statistic:   208 on 2 and 162 DF,  p-value: < 2.2e-16
```



```
##
## Call:
## lm_robust(formula = ln_death_p100Kp ~ ln_confirmed_p100Kp, data = df,
##     weights = population)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                       Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)             -3.160    0.46836  -6.746 2.513e-10  -4.0846   -2.235 163
## ln_confirmed_p100Kp      0.904    0.08281  10.917 3.587e-21   0.7405    1.068 163
##
## Multiple R-squared:  0.8938 ,    Adjusted R-squared:  0.8932
## F-statistic: 119.2 on 1 and 163 DF,  p-value: < 2.2e-16
```

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


5.0
2.5
0.0
−2.5
−5.0
ln(Deaths, per 100K persons)
0
2
4
6
8
ln(Confirmed Cases, per 100K persons)
China
India
USA