DA2: Assignment 1 by Abduvosid Malikov

Introduction

This is the Assignment 1 for `Data Analysis 2` and `Coding` course. The aim of this project is to analyse the pattern of association between **registered COVID-19 cases** and **registered number of death** due to COVID-19 on 10-31-2020. As a source, covid-19 data from CSSE[1] was used. The *population* is all cases reported in CSSE dataset starting from 01-22-2020 until now (the last added report date was 11-28-2020). The *sample* is the dataset that contains confirmed and death cases only from one date: 10-31-2020. The variables that are used are:

- country: string value showing the name of the country
- confirmed: numeric value showing the number of confirmed cases of COVID
- death: numeric value showing the number of death cases from COVID
- population: numeric value showing the population of the country
- ln_confirmed: log of number of confirmed cases
- ln_death: log of number of death cases

Invalid data (measuring flu cases instead of COVID), unreliable data (showing one value today and showing enormously different value tomorrow) and missing values are the potential data quality issues.

The dependent (y) variable is *Number of registered death* and explanatory (x) variable is *Number of confirmed case*. Figure 1 below shows the histogram for both variables

Both of the histograms have long right tail because majority of observations have almost same values and only few observations have extreme values. Most of the observations for confirmed cases are between 0 and 2 million, only 2 observations have a value more than 4 million. Most of the death cases are between 0 and 50000 and only 3 observations have a value more than 50000.

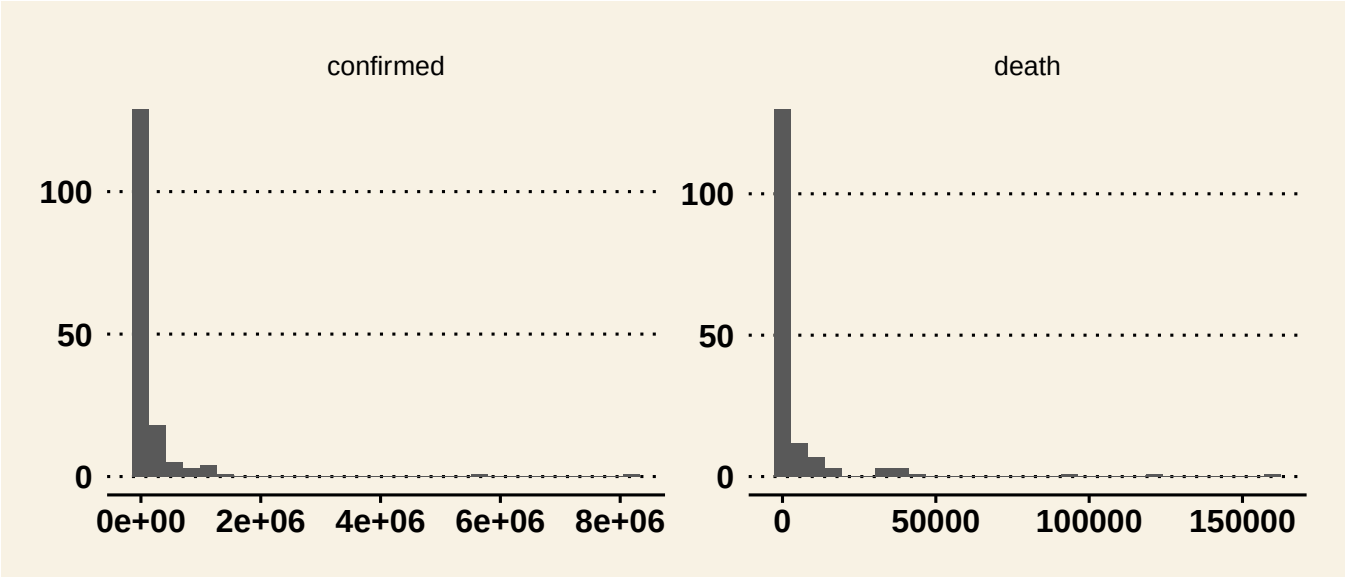

Figure 1: Histogram of confirmed and death cases

Tables below shows that on average there are 5472 death cases and 209452 registered cases. Mean is higher than median both for death numbers and registered cases indicating a skewed distribution with a long right tail.

[1] Center for Systems Science and Engineering

Table 1: Summary statistics on number of death

mean	median	std	min	max
5472.488	320.5	18583.42	0	159884

Table 2: Summary statistics on number of registered cases

mean	median	std	min	max
209452	21334	799982.2	2	8184082

Variable transformation

level - level

Substantive: distribution does not give meaningful interpretation

Statistical: there are some influential observations that governing the conditional mean. Graph is given in Appendix

level y - log x:

Substantive reasoning: level changes in death cases is harder to interpret

Statistical reasoning: log transformation shows the better approximation and capturing the non-linearity in data. Graph is given in Appendix

log y - level x:

Most of the observations are very close to each other at the bottom of the distribution of number of registered cases.

Substantive reasoning: level changes in registered cases is harder to interpret Statistical reasoning: log transformation of number of registered death and taking level of number of case does not give meaningful non-linear approximation. Graph is given in Appendix

log y - log x

Taking log of number of registered death and log of number of registered case is making the association close to linear

Substantive reasoning: it is easier to interpret the percentage change in number of death per percentage change in confirmed case

Statistical reasoning: it shows the better approximation: making the association close to linear

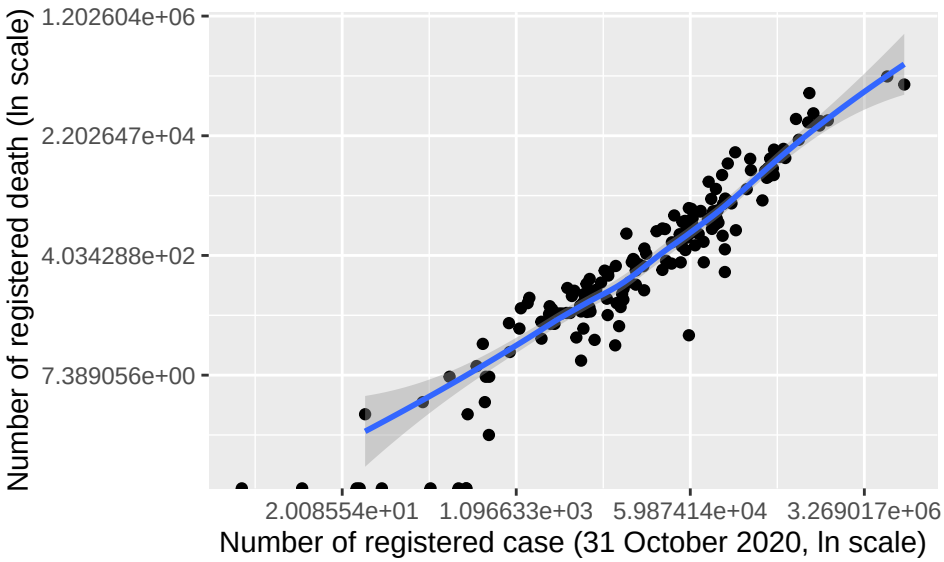

Figure 2: log y - log x transformation

Estimating different models

Four different models were created.

First model illustrates Simple Linear Regression for log of number of death regressed on log of number of registered cases:

Regression model 1: ln_death = alpha + beta * ln_confirmed

R squared: 0.8944

Second model shows Quadratic (linear) regression for og of number of death regressed on log of number of registered cases:

Regression model 2: ln_death = alpha + beta_1 * ln_confirmed + beta_2 * ln_confirmedˆ2

R squared: 0.90

Third model is for Piecewise linear spline regression. 3 cutoff points (knots) were chosen: 8,12,14

R-squared: 0.9037

Fourth model demonstrates Weighted linear regression, using population as weights.

Appendix

level - level

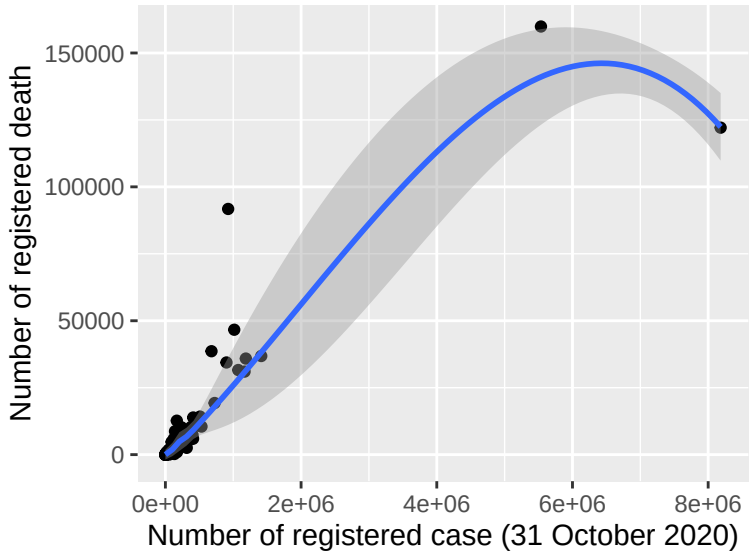

Figure 3: level - level distribution

level - log transformation

log y - level x transformation

```
##                Estimate Std. Error   t value     Pr(>|t|)   CI Lower  CI Upper
## (Intercept)  -4.406374 0.32435073 -13.58521 5.709025e-28 -5.0472256 -3.765522
## ln_confirmed  1.030951 0.03032022  34.00211 1.170499e-72  0.9710448  1.090858
##                DF
## (Intercept)   151
## ln_confirmed 151

##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
```

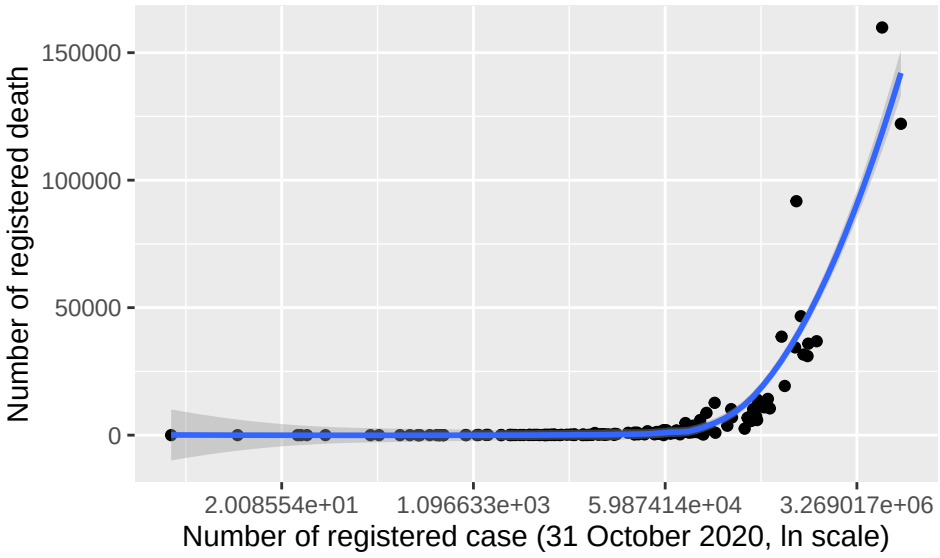

Figure 4: level -log transformation

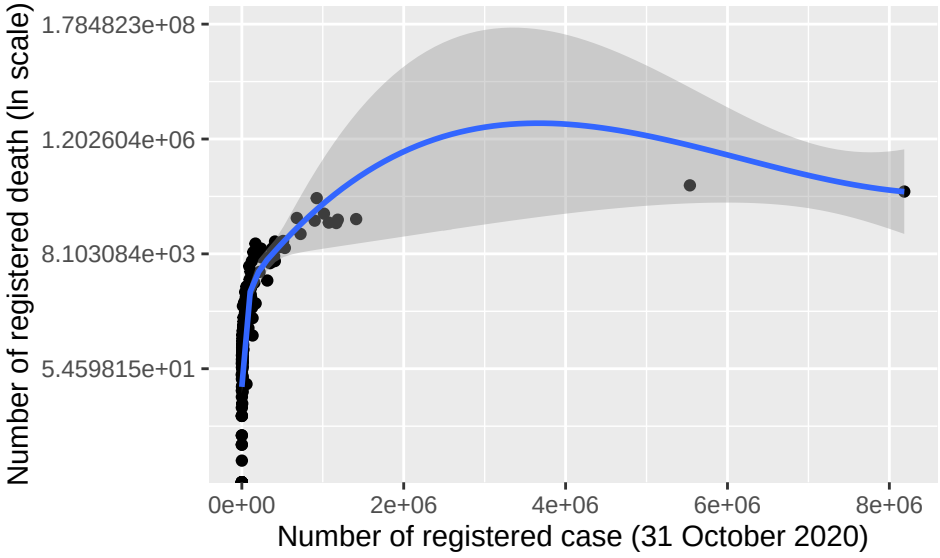

Figure 5: log y - level x transformation

```
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    -4.406    0.32435  -13.59 5.709e-28   -5.047   -3.766 151
## ln_confirmed    1.031    0.03032   34.00 1.170e-72    0.971    1.091 151
##
## Multiple R-squared:  0.8944 ,    Adjusted R-squared:  0.8937
## F-statistic:  1156 on 1 and 151 DF,  p-value: < 2.2e-16

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

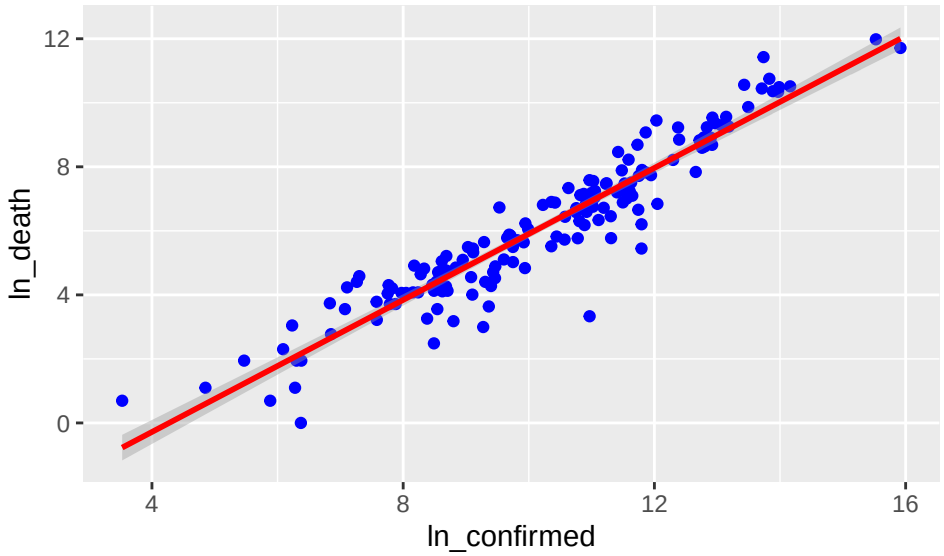

Figure 6: First model: Simple Linear Regression for ln_confirmed - ln_death

```
##
## 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)     -1.59313    0.95714  -1.664 0.098107 -3.48435  0.29809 150
## ln_confirmed     0.44012    0.19101   2.304 0.022583  0.06271  0.81753 150
## ln_confirmed_sq  0.02949    0.00934   3.157 0.001928  0.01103  0.04794 150
##
## Multiple R-squared:  0.9009 ,    Adjusted R-squared:  0.8996
## F-statistic:   683 on 2 and 150 DF,  p-value: < 2.2e-16


##
## Call:
## lm_robust(formula = ln_death ~ lspline(ln_confirmed, cutoff),
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                Estimate Std. Error t value  Pr(>|t|) CI Lower
## (Intercept)                     -3.0891    0.93681  -3.297 1.222e-03  -4.9403
```

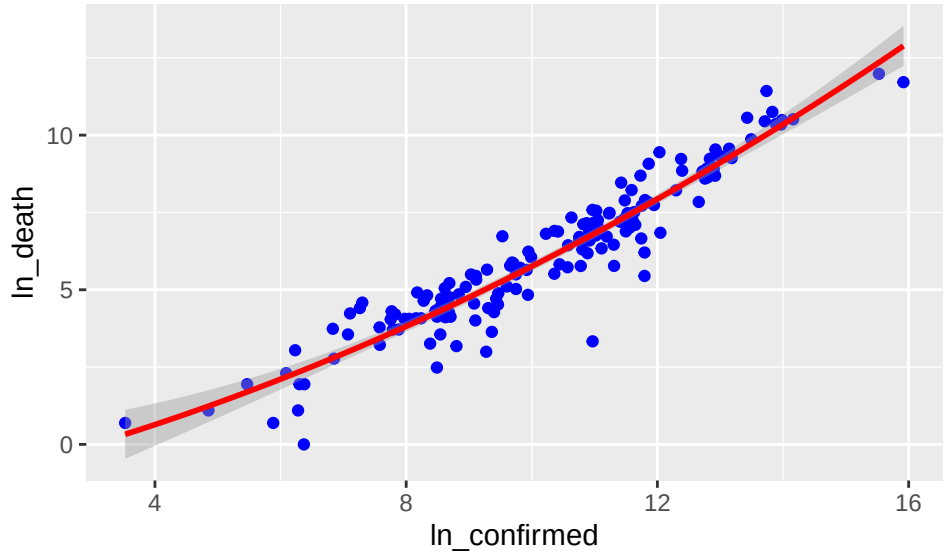

Figure 7: Second model: Quadratic (linear) regression for ln_confirmed - ln_death

```
## lspline(ln_confirmed, cutoff)1   0.8650    0.12196   7.093 5.028e-11   0.6240
## lspline(ln_confirmed, cutoff)2   0.9744    0.05436  17.925 6.448e-39   0.8670
## lspline(ln_confirmed, cutoff)3   1.5269    0.12380  12.333 1.745e-24   1.2822
## lspline(ln_confirmed, cutoff)4   0.5918    0.16813   3.520 5.733e-04   0.2596
##                                CI Upper  DF
## (Intercept)                     -1.2378 148
## lspline(ln_confirmed, cutoff)1   1.1060 148
## lspline(ln_confirmed, cutoff)2   1.0819 148
## lspline(ln_confirmed, cutoff)3   1.7715 148
## lspline(ln_confirmed, cutoff)4   0.9241 148
##
## Multiple R-squared:  0.9037 ,    Adjusted R-squared:  0.9011
## F-statistic: 586.3 on 4 and 148 DF,  p-value: < 2.2e-16
```

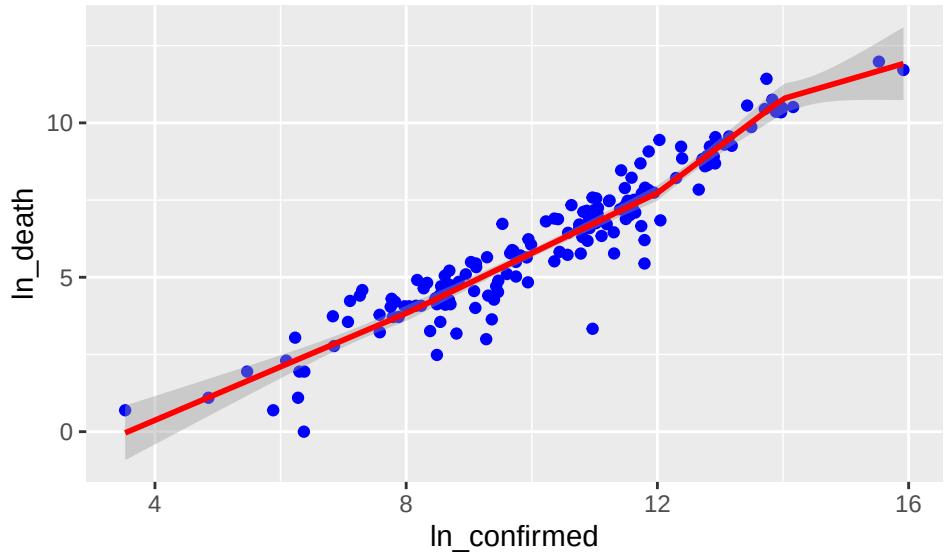

Figure 8: Third model: Piecewise linear spline regression

```
##
## 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)   -3.0667     0.9382  -3.269 1.337e-03  -4.9203   -1.213 151
## ln_confirmed   0.9487     0.0744  12.752 9.682e-26   0.8017    1.096 151
##
## Multiple R-squared:  0.9281 ,    Adjusted R-squared:  0.9276
## F-statistic: 162.6 on 1 and 151 DF,  p-value: < 2.2e-16

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

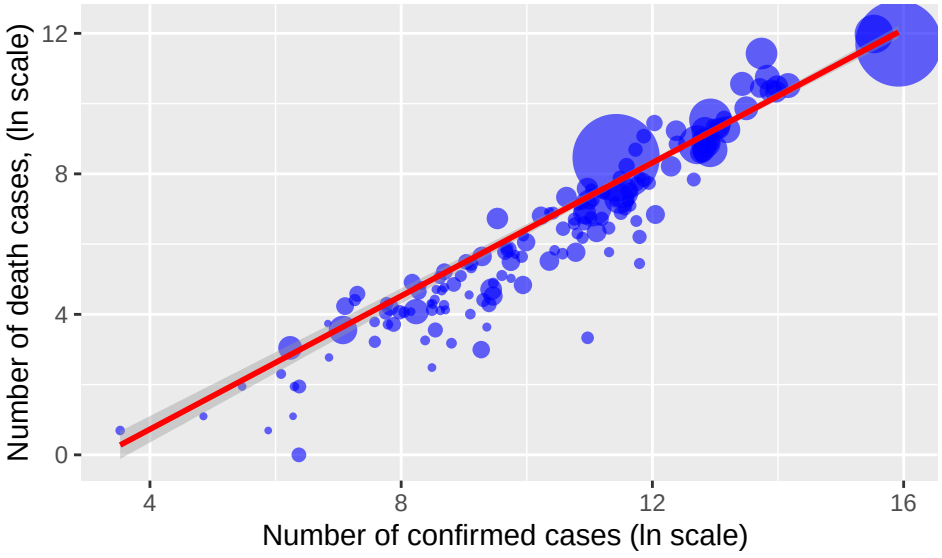

Figure 9: Fourth model: Weighted linear regression, using population as weights

	Confirmed (ln)- linear	Confirmed (ln) - quadratic	Confirmed (ln) - PLS	Confirmed (ln) - weighted linear
(Intercept)	-4.42***	-1.03	-3.11**	-4.62***
	(0.34)	(0.88)	(0.94)	(0.46)
ln_confirmed	1.03***	0.31		1.07***
	(0.03)	(0.17)		(0.04)
ln_confirmed_sq		0.04***		
		(0.01)		
lspline(ln_confirmed, cutoff)1			0.87***	
			(0.12)	
lspline(ln_confirmed, cutoff)2			0.97***	
			(0.05)	
lspline(ln_confirmed, cutoff)3			1.54***	
			(0.12)	
lspline(ln_confirmed, cutoff)4			0.75*	
			(0.29)	
R^2	0.89	0.90	0.90	0.93
Adj. R^2	0.89	0.90	0.90	0.93
Num. obs.	151	151	151	151
RMSE	0.79	0.76	0.76	3147.57

***p < 0.001; **p < 0.01; *p < 0.05

Modelling log of death numbers and log of confirmed numbers of COVID in all countries

Figure 10: Model comparison