Covid Assigment

Yuri Almeida Cunha

2020-11-28

Executive Summary

On the following project, I checked the possible association between the number of confirmed Covid cases and their number of deaths associated with the disease by country on a given day. Using the Weighted - OLS model, considering the population size as the weight, I was able to reach a well-fitted model for the following dataset. The models states that the increase of the number of confirmed cases increases on average the number of deaths associated with the disease, and that the population numbers play an important role on that. It's important to mention that the data acquired may have some integrity and reliability issues due some government data collection standards.

Research Question and Variable Description

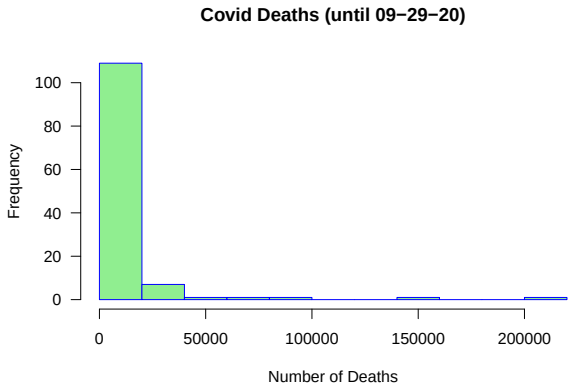
Research Question:
Is there any association between number of deaths and confirmed cases on this dataset? If yes, what is it?
Variables Description:
death (Explanatory): number of accumulated covid deaths until a given day by country
confirmed (Dependent): number of accumulated confirmed covid cases until a given day by country Sample: number of cases and deaths (until 29/09) / Population: Total number of cases and deaths whole Covid season Possible Quality Issues: Reliability (Data integrity - Governmental Interests) , Content (Caused by Covid or simply with Covid at death moment?

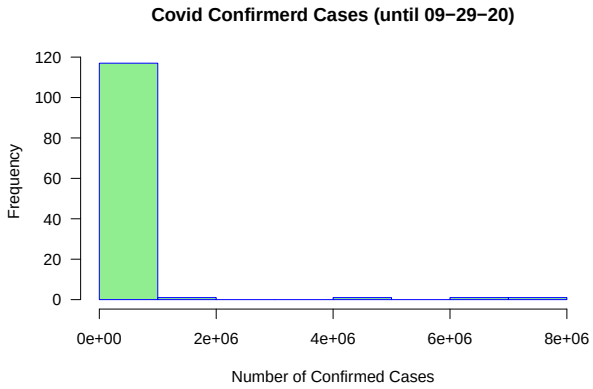
Scaling and Dropping Values

No scaling was used, used actual absolute numbers.
Few observations with really lower deaths and confirmed cases.
Removed all countries with low confirmed cases and really low population - not relevant for analysis.
Virus not sufficiently spread, and also helps to eliminate countries with specific geographic characteristics.
Ex: Brunei, Comoros, Dominica, Fiji and some other islands.

Variable Check

variable	mean	median	min	max	sd
confirmed	277081.364	49901	5008	7197687	962783.60
death	8301.256	813	27	206167	25779.07





Deaths and Confirmed Cases: Right tail Distributions: median < mean, outliers distributed on the right part of the graph

Log Transformations

Comparing the p-values (significance) and the R^2 adjusted (highness) for the four types of possible transformations (check the appendix),it´s possible to check that the lvl_lvl and log_log represented the best possible models to choose.
The models lvl_lvl and log_log showed good p and R^2-Ajusted values,checking the graphs its possible to determine that taking logs will be needed for both cases.Both deaths and confirmed cases are highly right-tailed skewed distributions with a few outliers.
Although, level is easier to interpret in general, the logs in this case produce a better fit, making the association pattern between them better to identify and analyze.
Absolute values are not that relevant in both cases.

Models

Based on model comparison our chosen model is reg4 - Weighted-OLS
Substantive:
- The model demonstrated a really great fit for the dataset, and can be used for prediction
- The beta parameter states that for the following dataset an increase of 10% on the number of confirmed cases will increase, on average, the number of deaths in 9.4%
- The alfa in that case represents the mean effect on number of deaths of variables not included in the model.
Statistical: - The models showed really similar statistical results (p-values and r-adjusted), capturing the variation really well.
- As logically expected, the weighted one by population size could demonstrate a even better fit.

Hypothesis Testing

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	-2.905429	1.1514831	-2.523206	0.0129479	-5.1854804	-0.6253773	119
ln_confirmed	0.943716	0.0911494	10.353509	0.0000000	0.7632311	1.1242009	119

Beta coefficient test is equal to 0? Test: H0: = 0, HA: ln_confirmed <> 0 Can reject null hypothesis: equal to 0

Residual Analysis

Top 5 Countries with lower death than predicted by regression

country	ln_death	reg4_y_pred	reg4_res
Bahrain	5.505	7.629	-2.123
Maldives	3.526	5.805	-2.278
Qatar	5.366	8.174	-2.808
Singapore	3.296	7.441	-4.145
Slovak Republic	3.807	5.745	-1.939

Top 5 Countries with higher death than predicted by regression

country	ln_death	reg4_y_pred	reg4_res
Belgium	9.21	8.109	1.102
Ecuador	9.334	8.248	1.086
Italy	10.49	9.036	1.451
Mexico	11.25	9.846	1.408
United Kingdom	10.65	9.376	1.273

Appendix

Log Transformations Decision

Check homoskedastic - Breusch-Pagan Test

```
lvl_lvl <- lm( death ~ confirmed , data = df)
log_lvl <- lm( death ~ ln_confirmed , data = df )
lvl_log <- lm( ln_death ~ confirmed , data = df )
log_log <- lm( ln_death ~ ln_confirmed , data = df )

bptest( lvl_lvl ) # small p-value can´t reject the null hypothesis (homoscedasticity)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lvl_lvl
## BP = 39.195, df = 1, p-value = 3.835e-10
```

```
bptest( log_lvl ) # small p-value can´t reject the null hypothesis (homoscedasticity)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  log_lvl
## BP = 18.553, df = 1, p-value = 1.653e-05
```

```
bptest( lvl_log ) # big p-value can t reject the null hypothesis (homoscedasticity)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lvl_log
## BP = 2.0841, df = 1, p-value = 0.1488
```

```
bptest( log_log ) # big p-value can reject the null hypothesis (homoscedasticity)
```

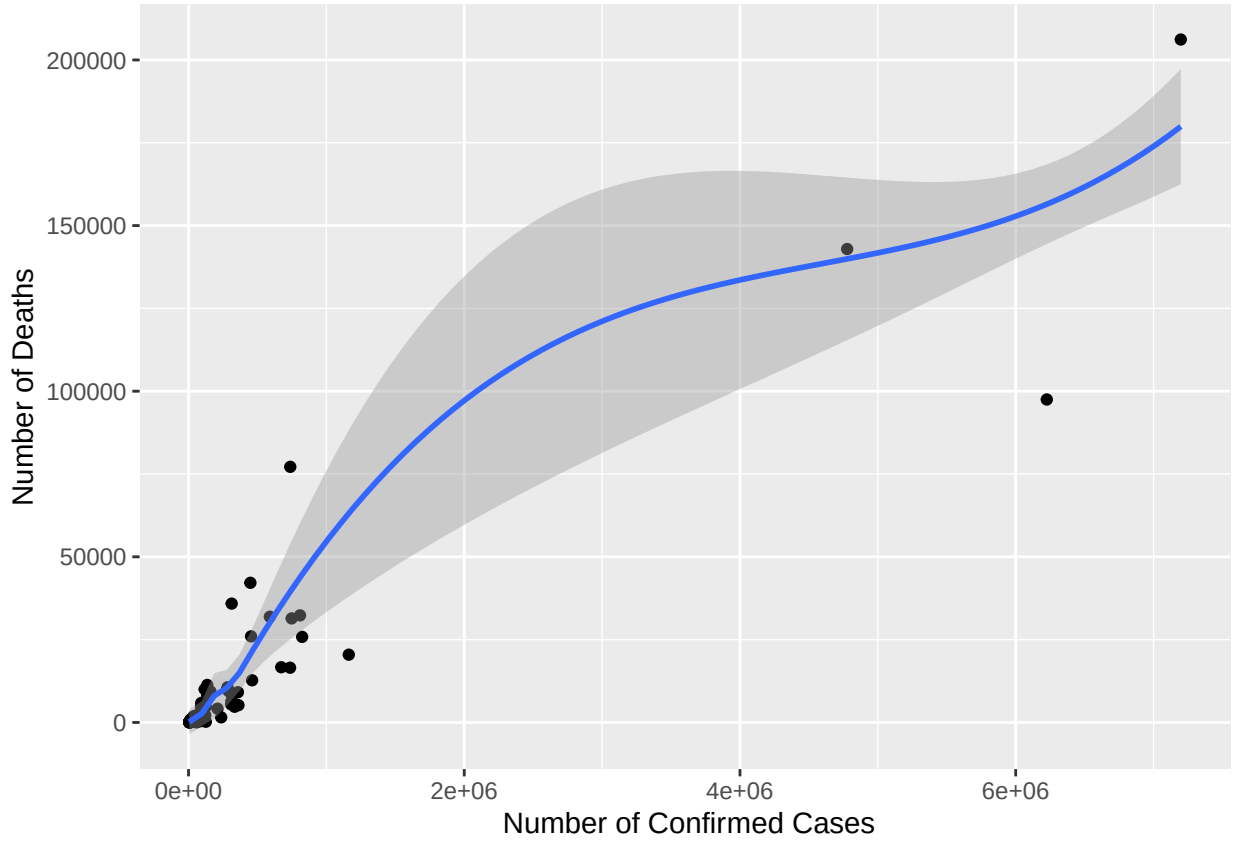
```
##
##  studentized Breusch-Pagan test
##
## data:  log_log
## BP = 0.29231, df = 1, p-value = 0.5887
```

```
# Use lm_robust lvl_log / log_log
lvl_log <- lm_robust( ln_death ~ confirmed , data = df , se_type = "HC2" )
log_log <- lm_robust( ln_death ~ ln_confirmed , data = df , se_type = "HC2" )
```

Graphical Check

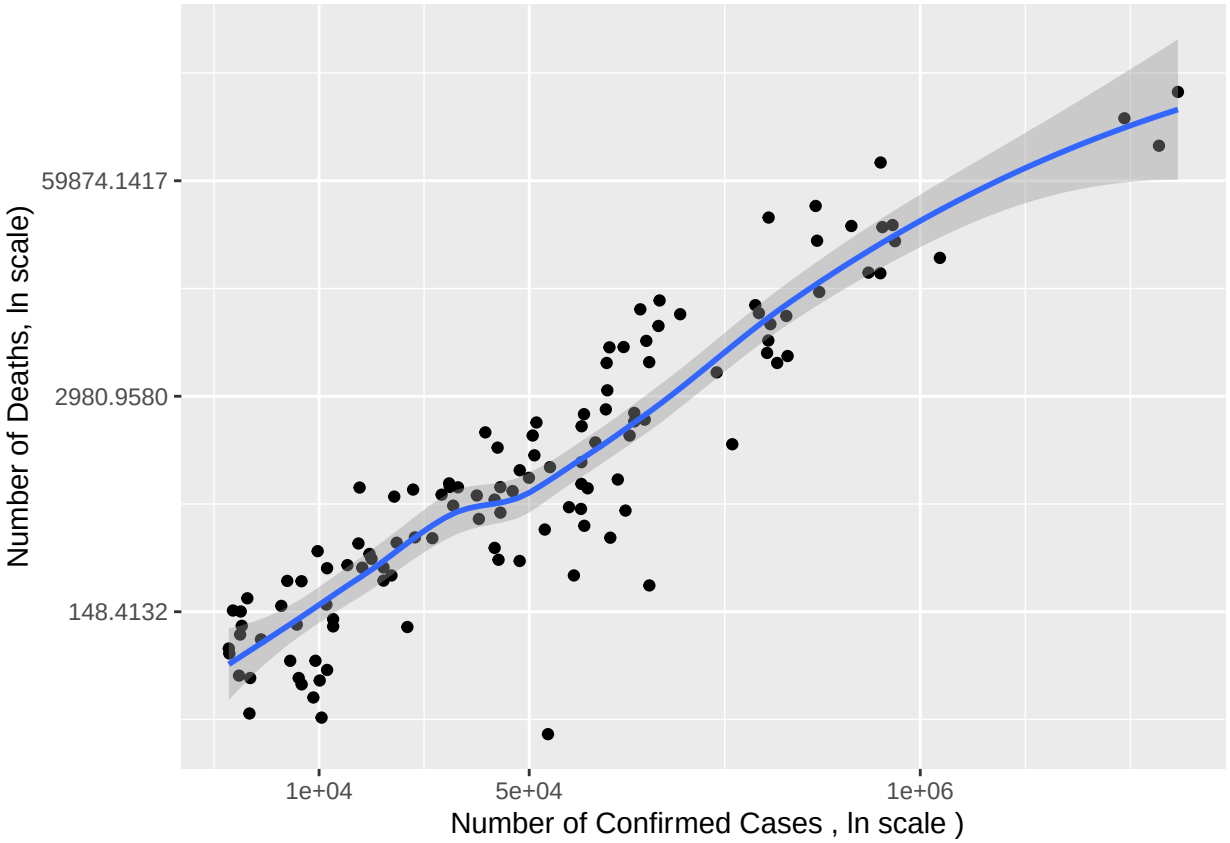
```
#Death - confirmed : level-level

ggplot( df , aes(x = confirmed , y = death)) +
  geom_point() +
  geom_smooth(method="loess")+
  labs(x = " Number of Confirmed Cases", y = "Number of Deaths")
```

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




```
# Death - confirmed : log-log
ggplot( df , aes(x = confirmed, y = death ))  +
  geom_point() +
  geom_smooth(method="loess")+
  labs(x = " Number of Confirmed Cases , ln scale )",y = "Number of Deaths, ln scale)") +
  scale_x_continuous( trans = log_trans(),  breaks = c(10000,50000,1000000))+
  scale_y_continuous( trans = log_trans())
```

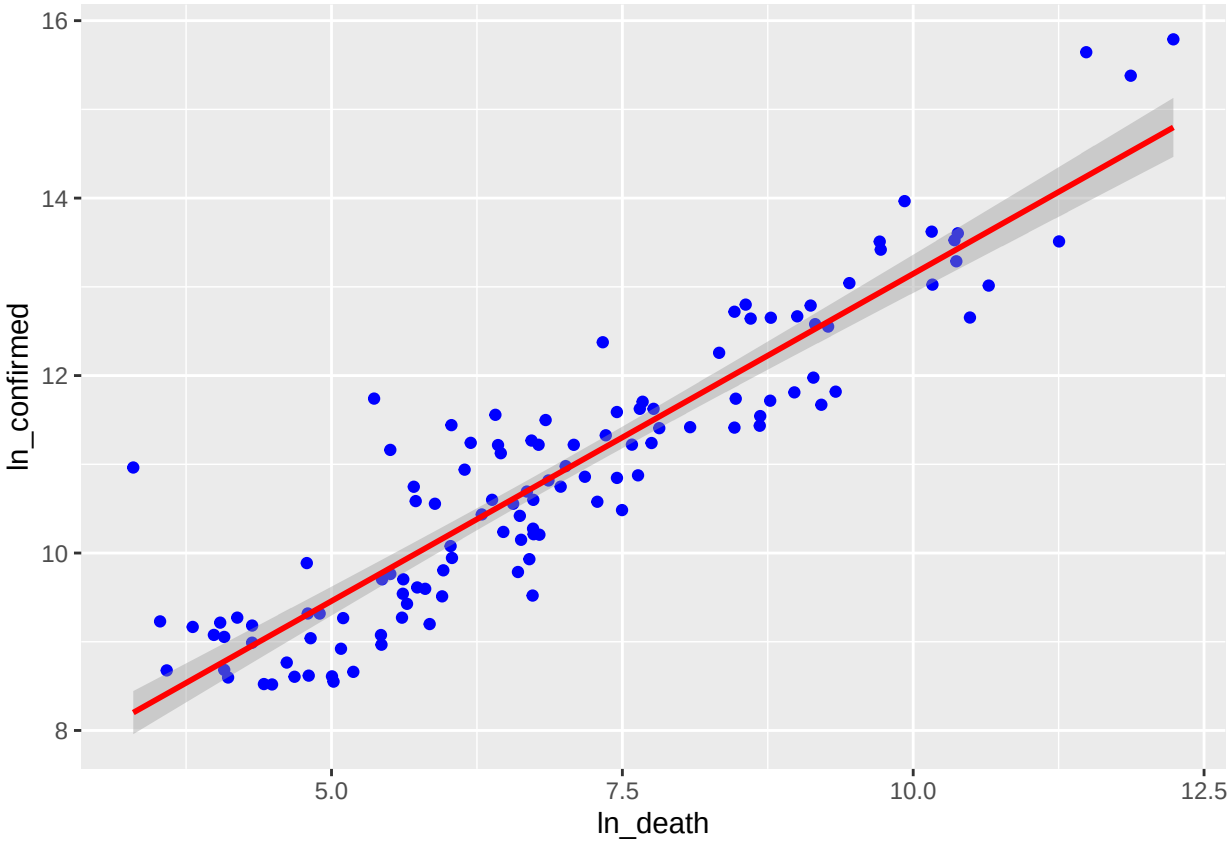
```
## `geom_smooth()` using formula 'y ~ x'
```



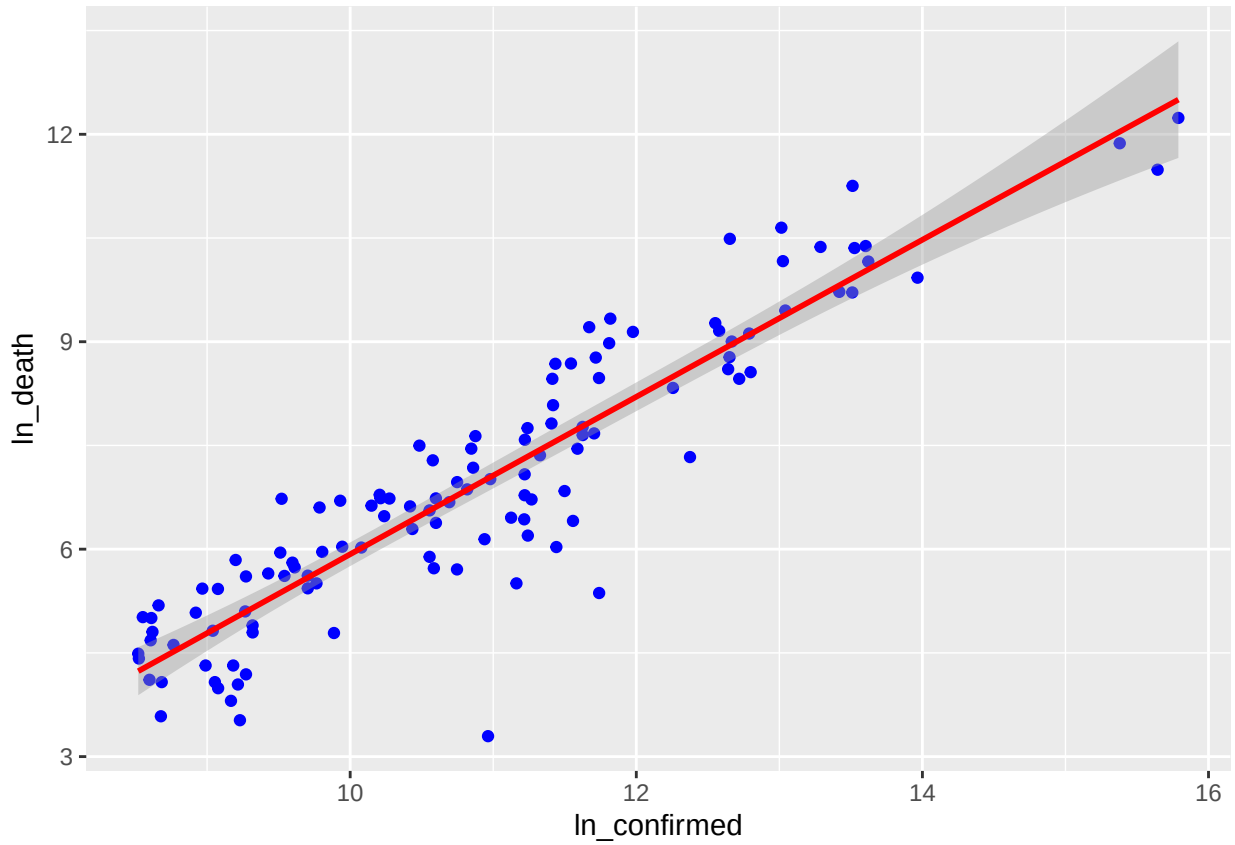
Considering the graphs you can definetily see that you reach a better fit using the log_log, instead of the lvl_lvl. The distribution of deaths is a really right-tailed skewed one.

Model Comparison - Graphs

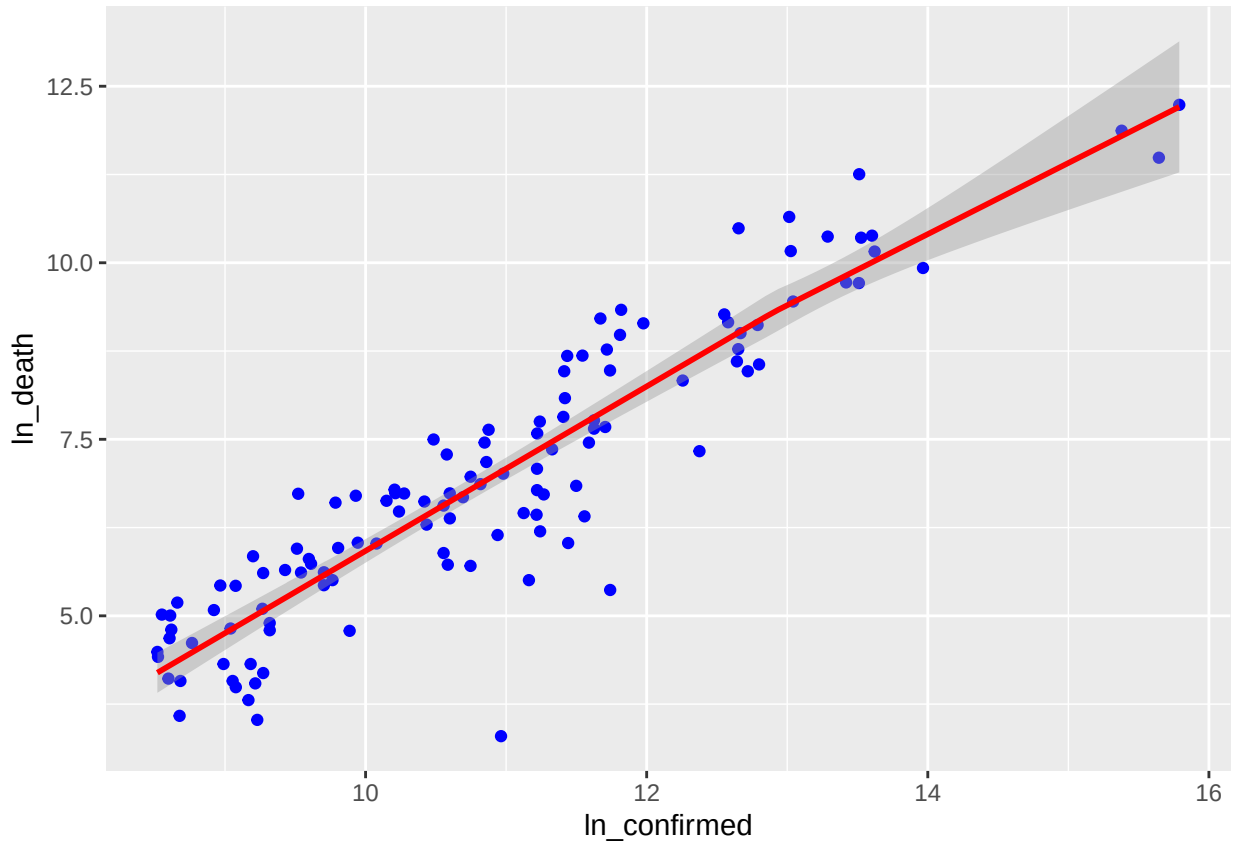
```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    -5.454    0.39650  -13.76 2.360e-26   -6.240   -4.669 119
## ln_confirmed    1.138    0.03677   30.95 9.413e-59    1.065    1.211 119
##
## Multiple R-squared:  0.8395 ,    Adjusted R-squared:  0.8382
## F-statistic: 957.7 on 1 and 119 DF,  p-value: < 2.2e-16

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



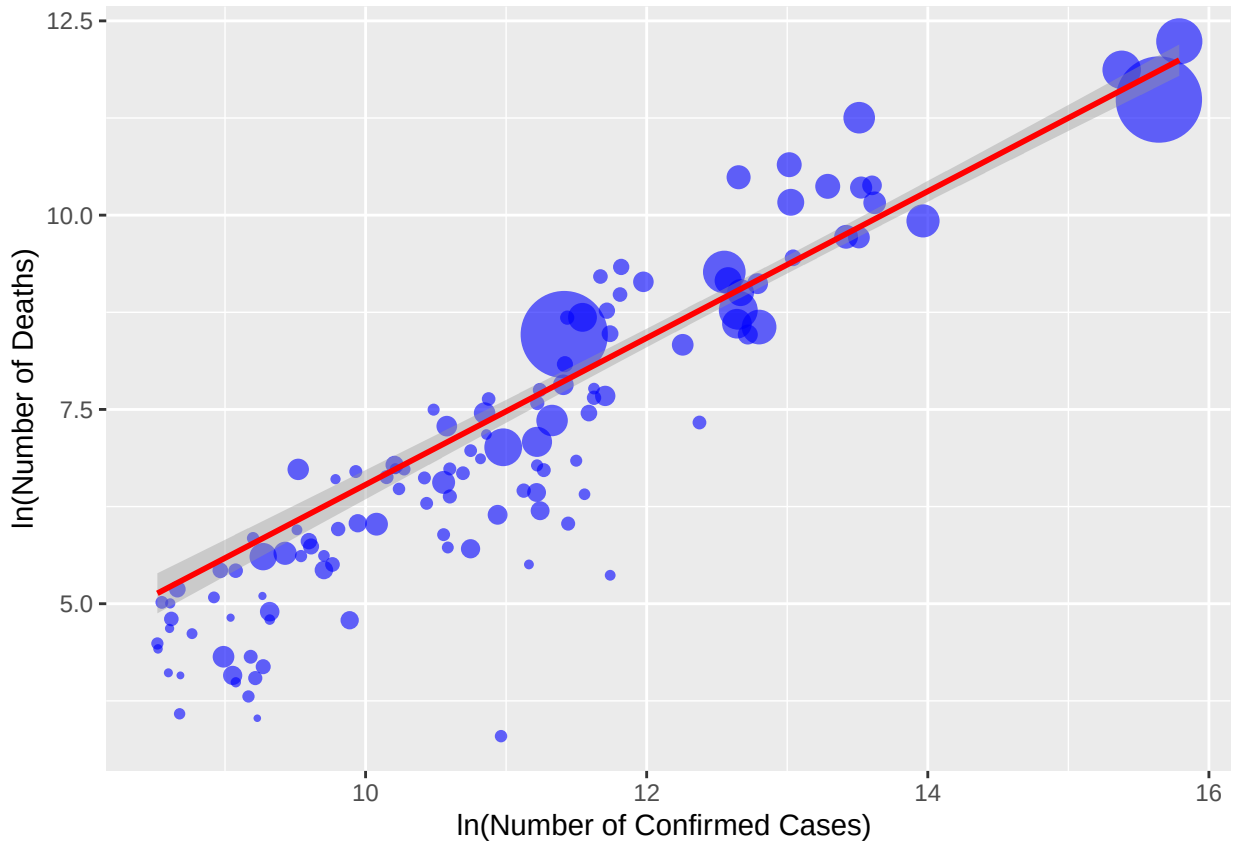

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed + ln_confirmed_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept)     -5.5522636    2.50606  -2.216  0.02864 -10.51495 -0.58958 118
## ln_confirmed     1.1555397    0.44522   2.595  0.01065   0.27388  2.03720 118
## ln_confirmed_sq -0.0007698    0.01924  -0.040  0.96816  -0.03888  0.03734 118
##
## Multiple R-squared:  0.8395 ,    Adjusted R-squared:  0.8368
## F-statistic: 472.7 on 2 and 118 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)                        -5.728    0.53375 -10.732 3.654e-19
## lspline(ln_confirmed, cutoff_ln)1   1.165    0.05161  22.571 1.239e-44
## lspline(ln_confirmed, cutoff_ln)2   1.007    0.11384   8.848 1.049e-14
##                                  CI Lower CI Upper  DF
## (Intercept)                       -6.7850   -4.671 118
## lspline(ln_confirmed, cutoff_ln)1  1.0626    1.267 118
## lspline(ln_confirmed, cutoff_ln)2  0.7818    1.233 118
##
## Multiple R-squared:  0.8403 ,    Adjusted R-squared:  0.8376
## F-statistic: 617.4 on 2 and 118 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)   -2.9054    1.15148  -2.523 1.295e-02  -5.1855  -0.6254 119
## ln_confirmed   0.9437    0.09115  10.354 2.659e-18   0.7632   1.1242 119
##
## Multiple R-squared:  0.9068 ,    Adjusted R-squared:  0.906
## F-statistic: 107.2 on 1 and 119 DF,  p-value: < 2.2e-16

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



Model Comparison - HTML

```
data_out <- 'D:/CEU/Data_Analysis_2/Covid_Assigment_Yuri_Cunha/out/'
htmlreg( list(reg1 , reg2 , reg3 , reg4),
         type = 'html',
         custom.model.names = c("Linear Regression", "Quadratic (Linear) Regression","Piecewise Linear S
                                "Weighted-OLS"),
         caption = "Modelling number of deaths by covid based on number of confirmed cases",
         file = paste0(data_out,'model_comparison.html'), include.ci = FALSE)
```

```
## The table was written to the file 'D:/CEU/Data_Analysis_2/Covid_Assigment_Yuri_Cunha/out/model_compar
```

For the analysis, check the main body