Regression Analysis

Eszter Diamant

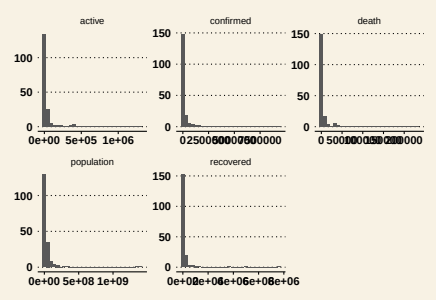
2020 11 29

Introduction

The aim of this analysis to identify linear regression between number of confirmed COVID-19 cases and death due to COVID-19. The main research question whether can be any regression model interpreted that determines which countries managed to save their people efficiently or not. The dataset contain data on countries, their population, the number of confirmed COVID-19 cases in the country, the number of death resulted from the infection, the number of recovered cases, and the number of currently active COVID-19 cases. Considering the high differences among the popuation, scaling was implemented during the analysis. Also, some extreme values were found such as Qatar and Singapore where high number of confirmed infection resulted in significantly lower number of death due to the quality of the healthcare system. However, because of the aim of the research question, these extreme values were kept. On the other hand, ue to errors in the regression analysis, those countries were droped where there were not deaths which might influence potential results.

Summary statistics

The mean of the distribution is 25,6647 while the median is 21,523. It is already indicates a high skeweness that is more than 7 for confirmed covid cases and more than 6 for deaths, meaning a long right tail in both distribution.



	min	max	mean	median	sd	skew	iq_range
Confirmed	2	9320266	256647.89	21523	1023053.09	7.27	107784.00
Death	0	231713	6588.15	322	24471.42	6.56	1809.00

Table 1: Summary statistics for the confimred COVID-19 cases and death resulted due to the infection

Investigate the transfomration of variables:

Considering the fact that the analysis used number of confirmed cases and number of deaths, significant transformation among he values did not occur.

Investigate different models:

Make the three models:

1) level-level:

```
(chart 1): reg1: death = alpha + beta * confirmed
(chart 5): reg2: death = alpha + beta_1 * confirmed + beta_2 * confirmed^2
(chart 6): reg3: death = alpha + beta_1 * confirmed + beta_2 * confirmed^2 + beta_3 * confirmed
```

2) log -log:

```
(chart 4): reg4: log_death = alpha + beta * log_conf
(chart 7): reg5: log_death = alpha + beta_1 * log_conf + beta_2 * log_conf^2
(chart 8): reg6: log_death = alpha + beta_1 * log_conf + beta_2 * log_conf^2 + beta_3 * log_conf
(chart 9): reg7: log_death = alpha + beta_1 * log_conf * 1(log_conf < 7) + beta_2 * log_conf * 1
```

3) extra: weighted-ols:

```
chart 10): reg8: log_death = alpha + beta* log_conf, weights: population
```

Model choice

log_death = 0.30 + 0.03 * log_conf

The model states that death 0.03% higher on average in a country, for observation with one percent higher confirmed cases. Based on model comparison our chosen model is reg4 - log_death ~ log_conf

Substantive:

- the log-log transformation gives the most fitting linear model
- the coefficients are meaningful

Statistical:

- the R^2 of reg4-6 exceeds the R^2 of the reg1-3. This means that those regrssion models capture the variables better
- it has one of the lowest p value compared to the other log regressions that indicates the lowest chance of false positive
- log-log transformation works better considering the high skewness

Hypothesis testing:

The hypothesis test whether the ß of chosen model can equal to zero:

H0: ß = 0 HA: ß != 0

Because the p-value is lower than 0.05 the null hypothesis, that ß=0, can be rejected.

```
##
## Call:
## lm_robust(formula = log_death ~ log_conf, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -4.241    0.30343  -13.98 6.074e-30  -4.8396   -3.642 168
## log_conf       1.018    0.02786   36.53 7.869e-82   0.9626    1.073 168
##
## Multiple R-squared:  0.8901 ,    Adjusted R-squared:  0.8895
## F-statistic:  1334 on 1 and 168 DF,  p-value: < 2.2e-16
```

Analysis of residuals:

The residuals consider the distance of the actual value from the predicted value. The lower the residual values indicates which countries lost relatively the most infected people while the highest residuals tell which ones saved the most.

Countries that relatively saved the most people due to COVID-19:

country	log_death	reg4_y_pred	reg4_res
Burundi	0.000000	2.249810	-2.249810
Iceland	2.484907	4.411990	-1.927083
Qatar	5.446737	7.764013	-2.317276
Singapore	3.332205	6.920522	-3.588317
Sri Lanka	3.044522	5.258958	-2.214436

Countries that relatively lost the most people due to COVID-19:

country	log_death	reg4_y_pred	reg4_res
Chad	4.5849675	3.200339	1.384628
Ecuador	9.4487272	8.011783	1.436945
Fiji	0.6931472	-0.652333	1.345480
Mexico	11.4306302	9.747078	1.683552
Yemen	6.3985949	3.525310	2.873285

Summary:

Based on the above findings, the log number of deaths and log number of confirmed cases presents the most meaningful linear regression for estimation. It was determined that 0.03% higher on average in a country, for observation with one percent higher confirmed cases. Considering tha fact that the analysis excluded those countries that do not have any deaths it might affect the the regression model potentially. On the other hand, those countries, that were considered as extreme values for their significantly low deaths rates, would illustrate that better health care system do reduce the number of death even in pandemic.

Appendix:

Chart 1: level - level linear regression

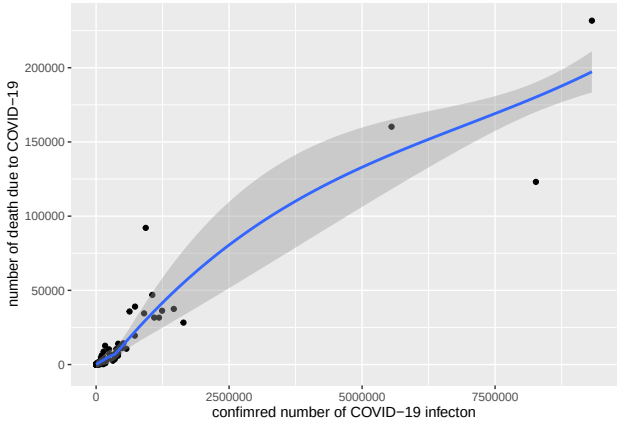


Chart 2: level - log linear regression

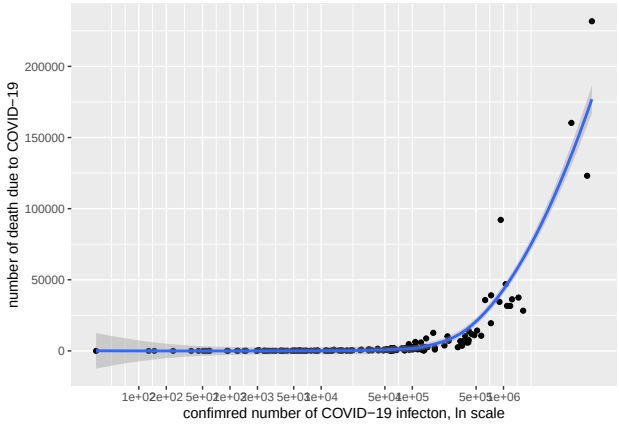


Chart 3: log - level linear regression

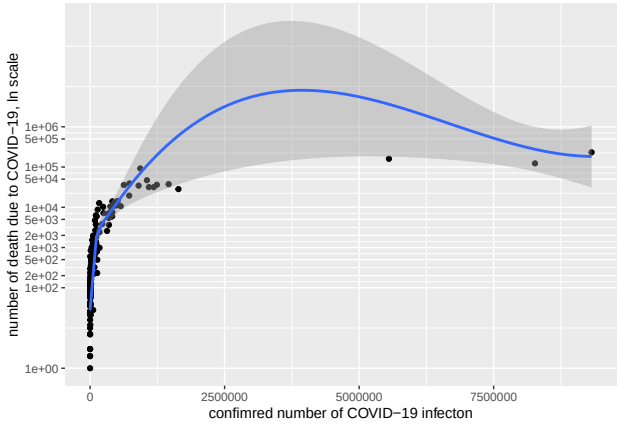


Chart 4: log - log regression linear regression

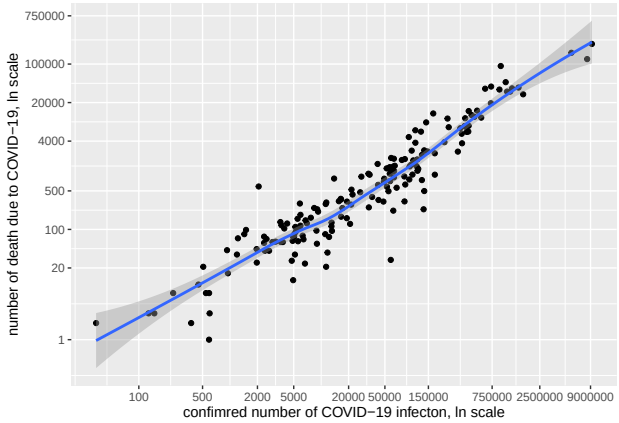


Chart 5: level-level quadratic regression

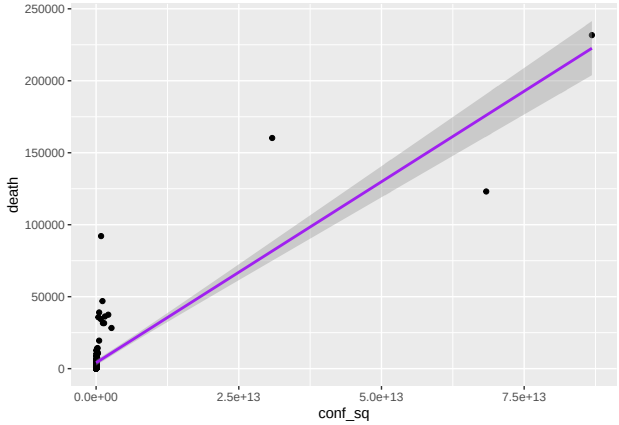


Chart 6: level - level cubic regression

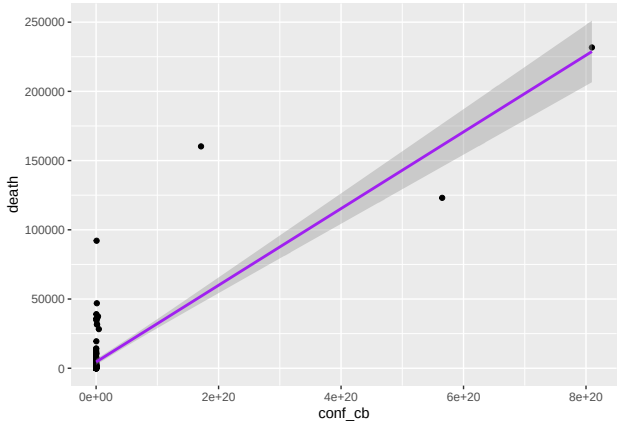


chart 7: log(death) - log(confirmed) quadratic regression

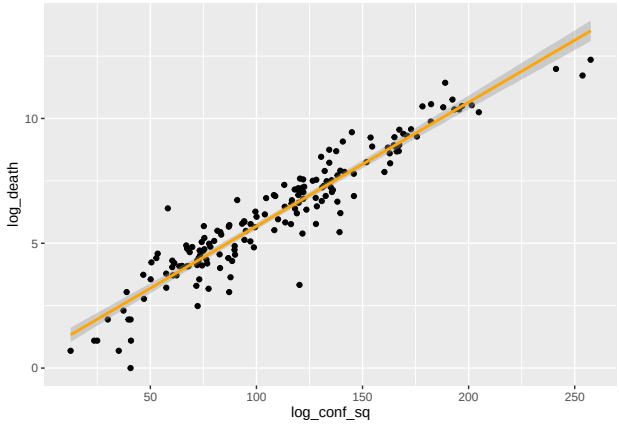


chart 8: log(death) - log(confirmed) cubic regression

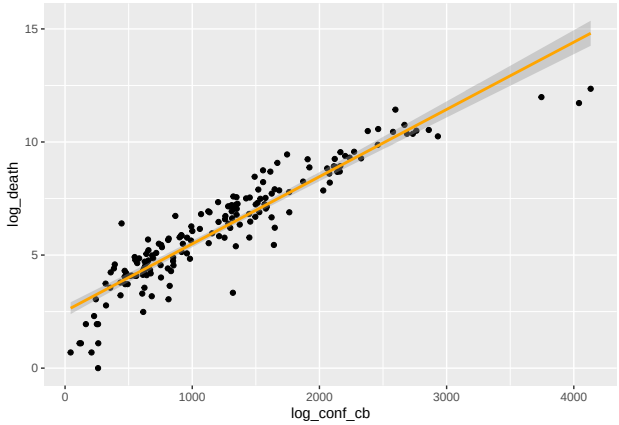


chart 9: Piecewise linear spline regression

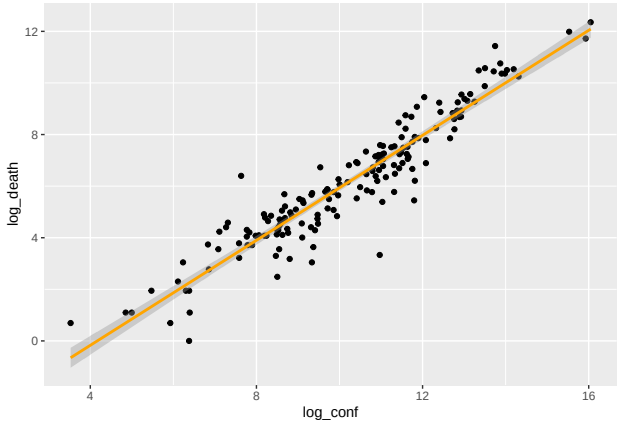


chart 10: Weighted linear regression, using population as weights

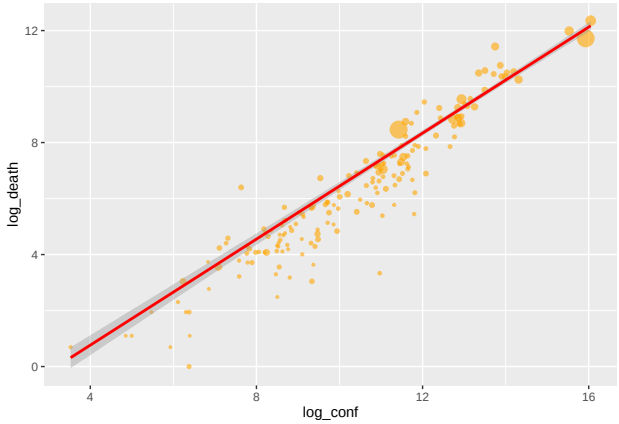


chart 11: Model comparisons

	Confirmed cases - linear	Confirmed cases - quadratic	Confirmed cases - cubic	log confirmed cases - linear	log confirmed cases - quadratic	log confirmed cases- cubic	log confirmed cases - PLS	log confirmed cases - weighted linear
(Intercept)	892.20	-420.02		-4.24***	-1.77*	-1.37		-3.02***
	(660.12)	(619.48)		(0.30)	(0.82)	(1.56)		(0.76)
confirmed	0.02***	0.03***	0.04***					
	(0.00)	(0.01)	(0.01)					
conf_sq		-0.00	-0.00					
		(0.00)	(0.00)					
conf_cb			0.00					
			(0.00)					
log_conf				1.02***	0.50**	0.37		0.95***
				(0.03)	(0.16)	(0.53)		(0.06)
log_conf_sq					0.03***	0.04		
					(0.01)	(0.06)		
log_conf_cb						-0.00		
						(0.00)		
lspline(log_conf, cutoff_ln)1							-1.35***	
							(0.14)	
lspline(log_conf, cutoff_ln)2							1.02***	
							(0.03)	
R^2	0.88	0.90	0.90	0.89	0.90	0.90	0.89	0.93
Adj. R^2	0.88	0.89	0.90	0.89	0.89	0.89	0.89	0.93
Num. obs.	170	170	170	170	170	170	170	170
RMSE	8767.93	8217.81	8027.17	0.82	0.80	0.80	0.82	4223.30

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

Modelling confirmed COVID-19 cases and death resulted among them

Figure 1: Regression Models

Github repository:

Github