COVID-19 Analysis Report

Mariam Bazzi

Introduction

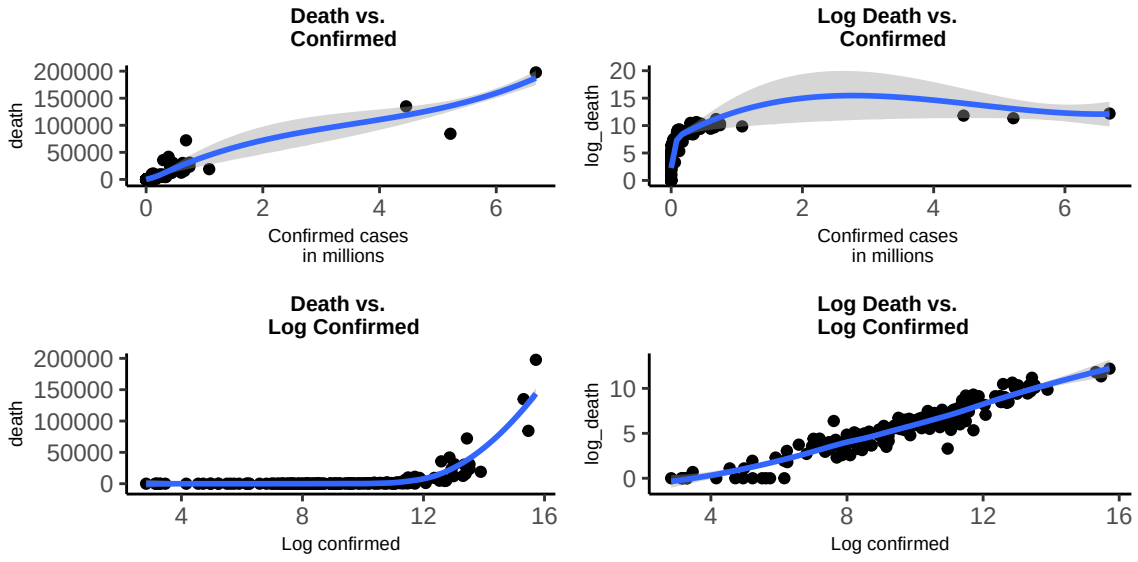
The main aim of this analysis is find out if we can predict the number of deaths from COVID-19 by knowing the number of registered COVID-19 cases. To find out, we analyze the pattern of association between registered COVID-19 cases and registered number of deaths due to COVID-19. There are two variables: the explanatory variable is the total number of registered COVID-19 cases in a country, and the dependent variable which is the total number of registered deaths from COVID-19 in the same country. Our sample contains a nearly complete population data, except 6 territories: Diamond Princess, MS Zaandam, Western Sahara, Taiwan, Ertirea, and Holy See (Vatican). The cleaned dataset consists of 182 observations, where each observation represents a country, and 6 variables which are: Country, Confirmed, Deaths, Recovered, Active, and Population. The dataset makes sense for our analysis; it is a good to model the research question.

Data Analysis

- Date to analyze: 17 September 2020.
- Explanatory variable: Number of registered cases.
- Dependent Variable: Number of registered deaths.

Note Although the variables are skewed and have outliers, these outliers will not dropped as it may bias our estimates and results. We do not scale our variables as they have the same unit of measurement, the total number of deaths or cases per country.

Different Scatter-plots with Lowess



When plotting the original variables, no clear correlation between the variables was observed. The log transformation can be used to disclose a relation between our variables. The best fitting line is for plotting the log_death vs the log_confirmed where most of data points do not deviate heavily from the fitted Lowess line. Also, the data points are equally distributed along the lowess line so this nearly linear relation can be modeled with our regression model.

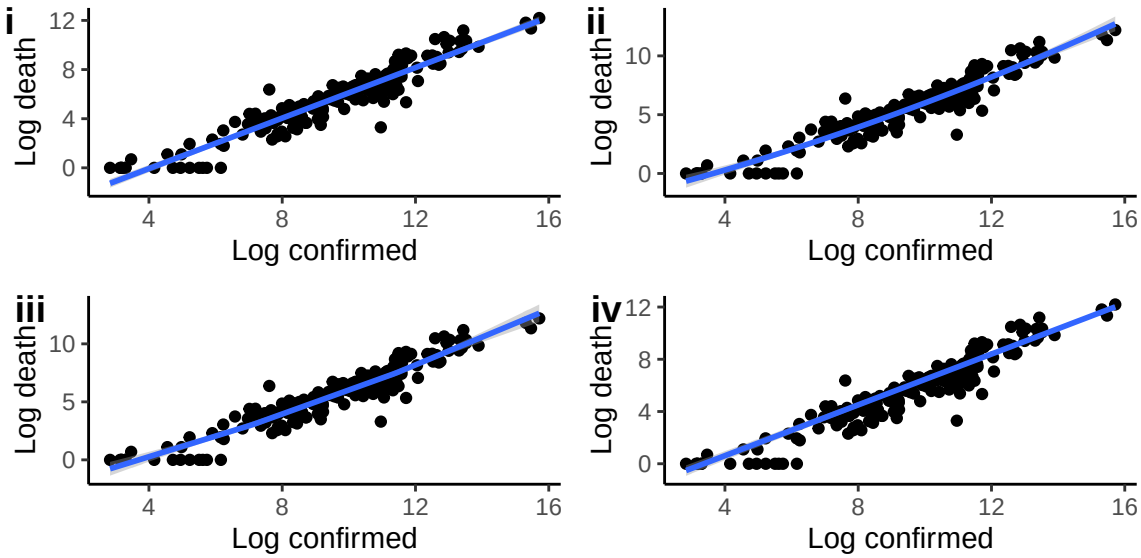
Regression Models

i. Simple linear regression

ii. Quadratic (linear) regression

iii. Piecewise linear spline regression

iv. Weighted linear regression, using population as weights.



Chosen Regression Model

Table 1: Comparing different regression models

model	R.squared	root_mean_squared_error
linear	0.900	0.870
quadratic	0.904	0.852
spline	0.904	0.855
weighted	0.898	0.961

The best model is quadratic model with smallest root mean squared error and highest R-squared. This is also evident when using the ANOVA test to compare the models, where the quadratic model has the lowest residual sum of squares (RSS)(See Appendix). This is also evident from the graphical smoothing of this data. To get the formula, we add the argument, raw = TRUE, to get the coefficients of the original variables.

Formula

$logdeath = -2.7 + 0.68(logconfirmed) + 0.02(logconfirmed)^2$

The interpretation of β_1 is for every 1% increase in the confirmed cases, the deaths will increase by 100*(1.01^0.68 -1)= 0.67%. *For every 10% increase in the confirmed cases, the deaths will increase by 100*(1.1^0.68 -1)= 6.7%. For every 100% increase in the confirmed cases (doubling of cases), the deaths will increase by 100*(2^0.68 -1)= 60%.

Hypothesis Testing

on β_1 (which interacts with x)

Test the following hypothesis: H0 : $\beta_1 = 0$; HA : $\beta_1 \neq 0$. Choose a significance level and make your conclusion. Our significance level is 0.05. We conclude that our sample indicates that β_1 is significantly larger than 0.

Table 2: summary of quadratic model

term	estimate	std.error	statistic	p.value
(Intercept)	-2.7352	0.5713	-4.7877	0.000
poly(log_confirmed, 2, raw = TRUE)1	0.6794	0.1277	5.3187	0.000
poly(log_confirmed, 2, raw = TRUE)2	0.0194	0.0070	2.7790	0.006

Table 3: summary of quadratic model fit

r.squared	adj.r.squared	sigma	statistic	p.value
0.9044	0.9034	0.8588	847.1487	0

The p-value of the total model is almost 0. This indicates that our model is significantly better than the null model. The r-squared is 0.90 or our models describes about 90% of the variability in the log deaths of COVID-19.

Analysis of the residuals

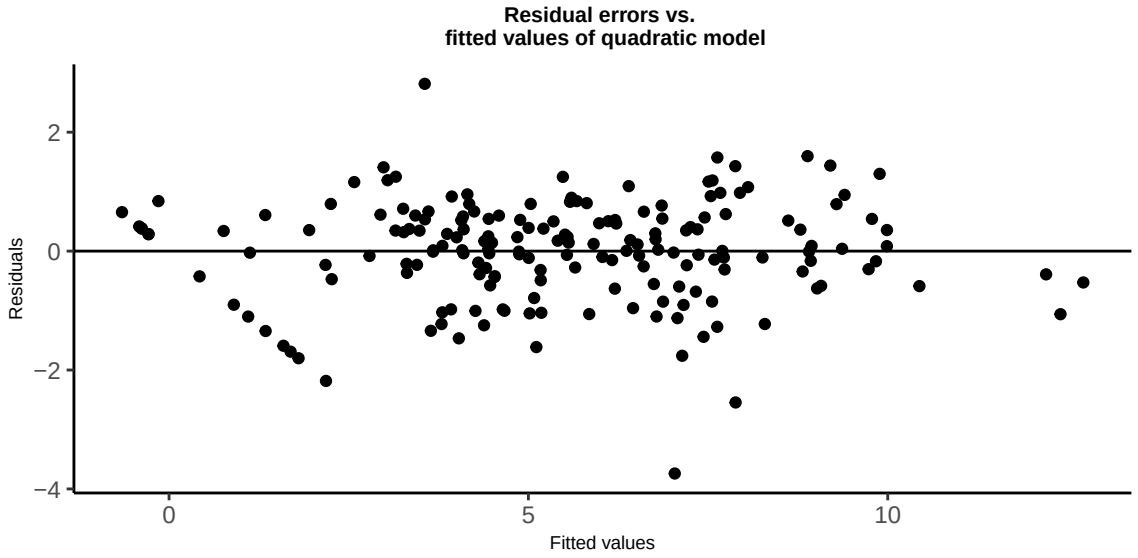


Table 4: Largest countries with negative errors

country	residual	fitted	actual
Singapore	-3.740553	7.036390	3.295837
Qatar	-2.545849	7.883387	5.337538
Burundi	-2.183622	2.183622	0.000000
Mongolia	-1.802169	1.802169	0.000000
Bahrain	-1.760180	7.140077	5.379897

Table 5: Largest countries with positive errors

country	residual	fitted	actual
Yemen	2.813737	3.557875	6.371612
Italy	1.597279	8.884450	10.481729
Belgium	1.574445	7.629475	9.203920
United Kingdom	1.438907	9.201601	10.640508
Ecuador	1.428930	7.879354	9.308283

The residual errors are almost equally dispersed around the horizontal line at zero with no patterns, which is good. We create a data frame that contains the fitted or predicted log_death and the residual of each country along with the country name and the actual log_death. residual = actual value - fitted value The largest countries with negative errors, meaning that our model overestimate their log_death value, Singapore, Qatar, Burundi, Mongolia, and Bahrain. The largest countries with positive errors, meaning that our model underestimate their log_death value, Yemen, Italy, Belgium, United Kingdom and Ecuador.

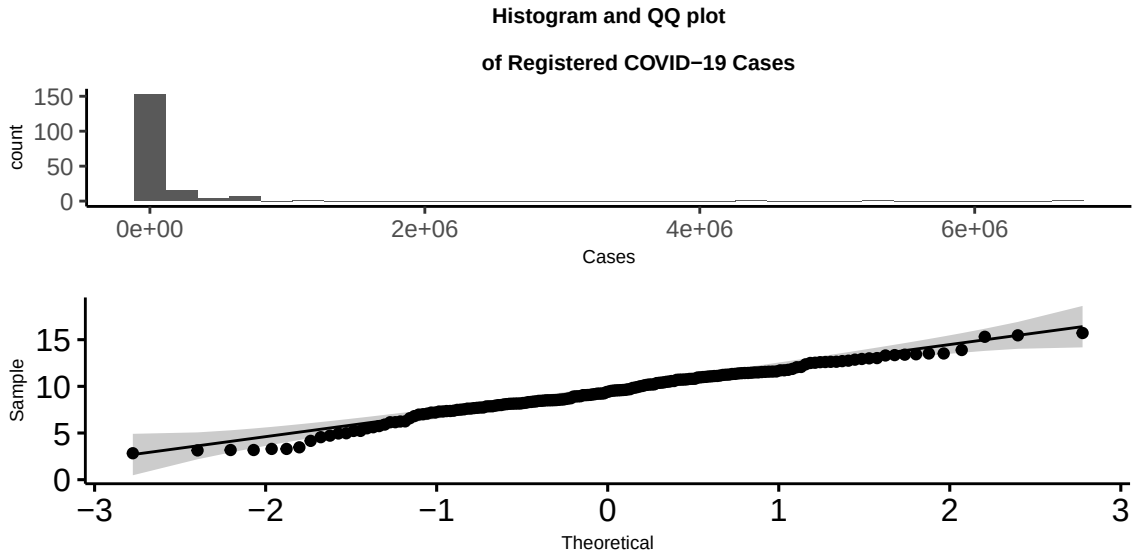
Results of the analysis

The aim of this analysis was to predict the deaths from COVID-19 using confirmed cases of COVID-19. As both variables were heavily skewed, the variables were log transformed, and used the quadratic regression to predict the log deaths. Our model may get better if we included other variables in our model, for example, total tests carried, co-morbidity, hospital care received for cases, the age of patients. The inclusion of only 1 predictor certainly will weaken our model as it assumes that the relation between cases and deaths is the same in all countries regardless of underlying health and co-morbidity issues, national income effect, etc.

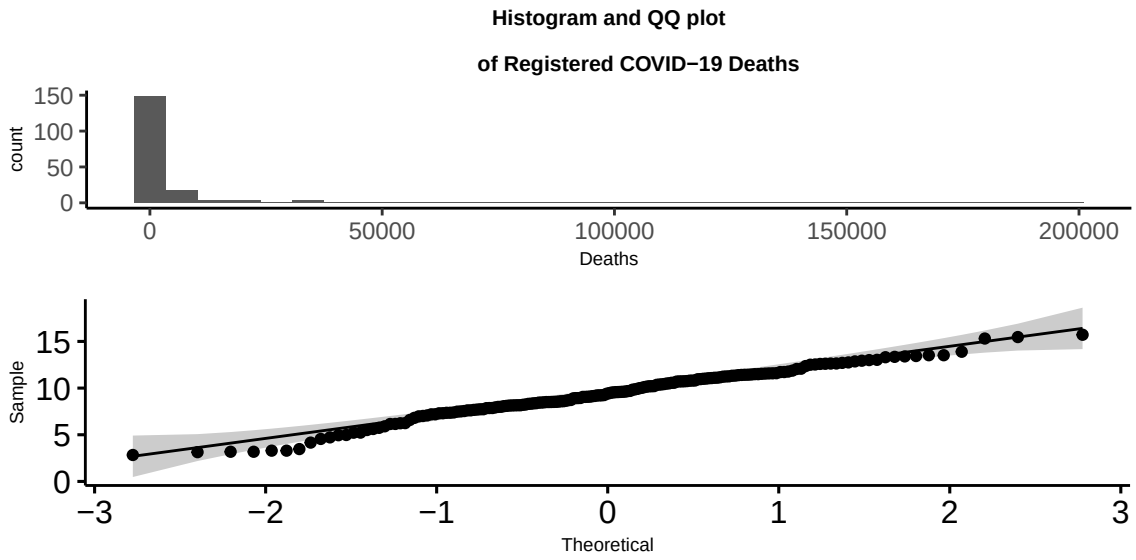
Appendix

1.Number of registered cases

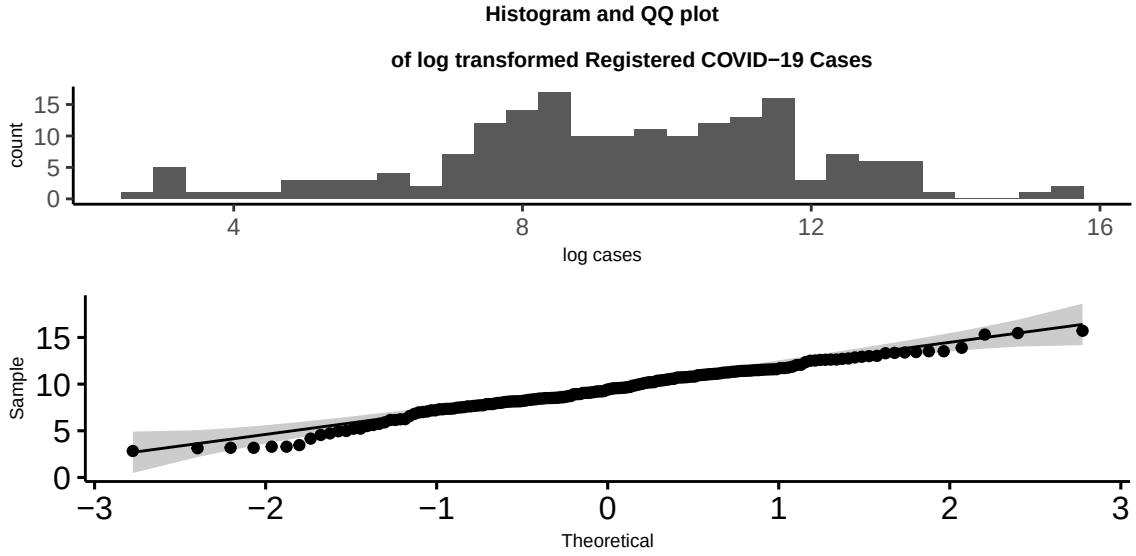
The histogram below shows the number of registered cases without transforming the variables. The histogram for the number of registered COVID-19 cases is skewed with a long right tail. The quantile-quantile plot also shows the skewness.



2.Number of registered deaths The summary statistics for the number of registered deaths are indicating right tailed distribution. This is shown in the histogram and the quantile-quantile plot of this variable.

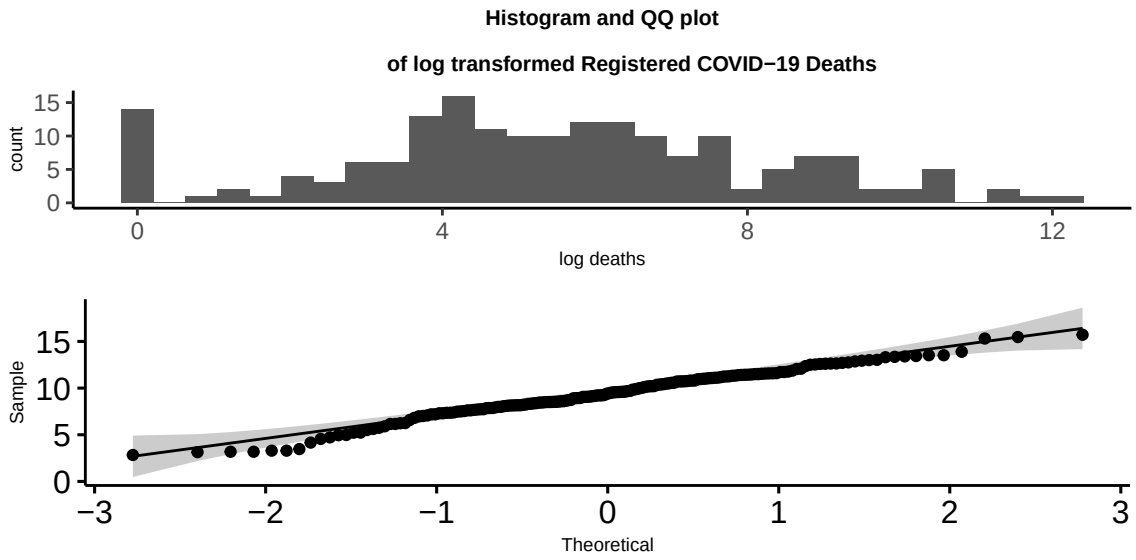


3.Log Number of registered cases



The log transformed variable of confirmed cases is more normally distributed as shown by the histogram and quantile-quantile plot. Using the transformed variable is a better representation as the original variable were skewed with a long right tail. (See Appendix)

4.Log Number of registered deaths



To make death variable more normally distributed, death was transformed log_death. The log transformed variable is more normally distributed as shown in the histogram and quantile-quantile plot.

Summary statistics of the variables

Both the confirmed and death variables are skewed and not normally distributed. However, the log transformed variables are more normally distributed than the original variables.

Table 6: Summary statistics of COVID cases and deaths and their transformed variables

summary	confirmed	death	log_confirmed	log_death
Min.	17.00	0.00	2.83	0.00
1st Qu.	2643.75	49.00	7.88	3.89
Median	12296.00	222.00	9.42	5.40
Mean	165828.33	5198.75	9.37	5.46
3rd Qu.	73770.25	1369.50	11.21	7.22
Max.	6679265.00	197648.00	15.71	12.19
Standard deviation	674664.34	19545.43	2.54	2.77

ANOVA Test

```
## Analysis of Variance Table
##
## Model 1: log_death ~ log_confirmed
## Model 2: log_death ~ log_confirmed
## Model 3: log_death ~ elspline(log_confirmed, 3)
## Model 4: log_death ~ poly(log_confirmed, 2, raw = TRUE)
##   Res.Df        RSS Df   Sum of Sq          F    Pr(>F)
## 1    180        138
## 2    180 3388154424  0 -3388154286
## 3    178        133  2  3388154291 2267791406 < 2.2e-16 ***
## 4    179        132 -1           1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```