COVID Analysis Report

Karola Takács, Student_id: 2003898

29/11/2020

Introduction

Outcome variable: **Number of registered deaths** Explanatory variable: **Number of registered case**

Research question: Which countries are under performing taken their number of registered cases in terms of number of deaths in our model?

Variables

- Population data is coming from The World Bank, which gathers cross-sectional internationally comparable statistical data, measured per country in each year.
- The COVID-related variables originate from the JHU CSSE COVID-19 Dataset, where daily case reports are collected. Variables included in the analysis after the cleaning can be found in the variables.xlsx file.
- The population is all the countries where there are reported COVID-cases. This covers most of the world by now. Potential issues: not that reliable data reporting in 3rd countries; delay in reporting times and date, so cases might get added to a different date when they actually were discovered, but this has a bigger relevance when the case numbers were still low. Deaths and confirmed cases are not 100% discovered and thus reported in general.
- Selecting observations: grouped my observations by country_region, then removed all the aggregate countries/regions since those were practically duplicates. Removed those observations which were lacking in population data and also dropped countries where either the y or the x variable was NA. Population could be scaled by dividing it by 1000 (the smallest value for that was San Marino with 33860 inhabitants), but this variable was only used for weighting. There are extreme values for China and India in terms of population, but those are not from error, so I kept them.Lastly I filters the rows where number of deaths was zero, since taking the ln of zero is not defined.

Summary statistics

```
## Rows: 1
## Columns: 5
## $ mean   <dbl> 279454.8
## $ median <dbl> 38919.5
## $ sd     <dbl> 1065018
## $ min    <dbl> 34
## $ max    <dbl> 9320266
```

```
## Rows: 1
## Columns: 5
```

```
## $ mean   <dbl> 7173.994
## $ median <dbl> 482.5
## $ sd     <dbl> 25464.22
## $ min    <dbl> 1
## $ max    <dbl> 231713
```

The mean of the left-hand side variable is 7174 and the median is 482. For the right-hand side variable these are 279454 and 38919.5 respectively. Both suggest that their distributions are skewed with a long right tail since mean is bigger than the median. The standard deviation is very big for both variables indicating a very wide spread around the mean.

Transformations

- Level-level: quasi linear (little curvy) but the problem is that the observations are grouped to the lower left corner. This skewness was already visible on the distribution histogram above. This is not giving us any useful patterns of association.
- Taking the natural logarithm of either X or Y variables indicates a non-linear pattern. In case of ln(confirmed cases) the distribution is less skewed though, but the shape is closer to exponential or maybe parabola (only a segment of the U shape)
- Log-log transformation: we could say that the the distribution of X and Y much more symmetric and not that many extreme values. It looks almost as a 45°line, but probably more difficult to interpret. Since I removed observations with zero values for the variable death, log can be applied to it as well.

Estimating different models

- **Linear model**: For countries with 1% more confirmed cases, we would expect the number of deaths to change by 1.02%, on average.
- **Quadratic model**: We can see that the parabola in convex (positive Beta2)
- **PLS**:
  - When comparing observations with number of confirmed cases lass than 5000, number of deaths is 0.81% higher, on average, for observations with one percent higher number of confirmed cases.
  - When comparing observations with number of confirmed cases over 5000, number of deaths is 1.08% higher, on average, for observations with one percent higher number of confirmed cases.
- **Weighted linear**: For countries with 10% more confirmed cases we would expect number of deaths to be higher by 9.5% on average.

R-squared: initially quite high, the best in case of the weighted model: this model explains 93% of the variation in the outcome variable and only 7% is left for residual variation.

Model formula: ln(death) = -3.01 + 0.95*ln(confirmed) Interpretation: - Alpha: not interpreted - Beta: for countries with 10% more confirmed cases we would expect number of deaths to be higher by 9.5% on average.

Hypothesis testing

Our null hypothesis says that beta = 0 Chosen significance level: 0.05 As the p-value is much less than 0.05, we can reject the null hypothesis that beta = 0. Hence there is a significant relationship between the variables.

Residual analysis

```
## # A tibble: 5 x 4
##   country   ln_death reg5_y_pred reg5_res
##   <chr>        <dbl>       <dbl>    <dbl>
## 1 Burundi       0           3.02    -3.02
## 2 Iceland       2.48        5.03    -2.55
## 3 Qatar         5.45        8.15    -2.70
## 4 Singapore     3.33        7.36    -4.03
## 5 Sri Lanka     3.04        5.82    -2.78
```

```
## # A tibble: 5 x 4
##   country ln_death reg5_y_pred reg5_res
##   <chr>      <dbl>       <dbl>    <dbl>
## 1 Bolivia     9.08        8.21    0.866
## 2 Ecuador     9.45        8.38    1.07
## 3 Iran       10.5         9.62    0.866
## 4 Mexico     11.4         9.99    1.44
## 5 Yemen       6.40        4.21    2.19
```

Countries with the largest negative errors are: Burundi, Iceland, Quatar, Singapore, Sri Lanka. These countries have a lower number of deaths than it was predicted by the regression model.Similarly countries with the largest positive errors are: Bolivia, Ecuador, Iran, Mexico, Yemen. These have higher actual number of deaths than it was predicted by the model. Since the residuals are not that big (though they are ln) probably outliers have not been overlooked.

Summary

Dependent variable in the analysis is the number of registered death while the independent variable is the number of registered cases.The basic level-level plotting suggests a positive mean-dependence - as average y increases the values of x also seem to increase. After assessing which transformation to use and then plotting 5 different models, I chose the weighted linear model, because that produced an R-squared of 93%. This tells us that this linear approximation almost explains all the variations in the data, however the line in not a perfect fit: e.g. cubic model seems to have smaller heteroskedasticity. I would try to apply variable scaling (i.e ratios) or set more knots at the linear spline model, as it seems there could be another point at around ln 12. I also deleted observations with zero number of deaths, but if I would keep them and apply level-level transformation then that would also be a quite good model fit.

Appendix

Distributions

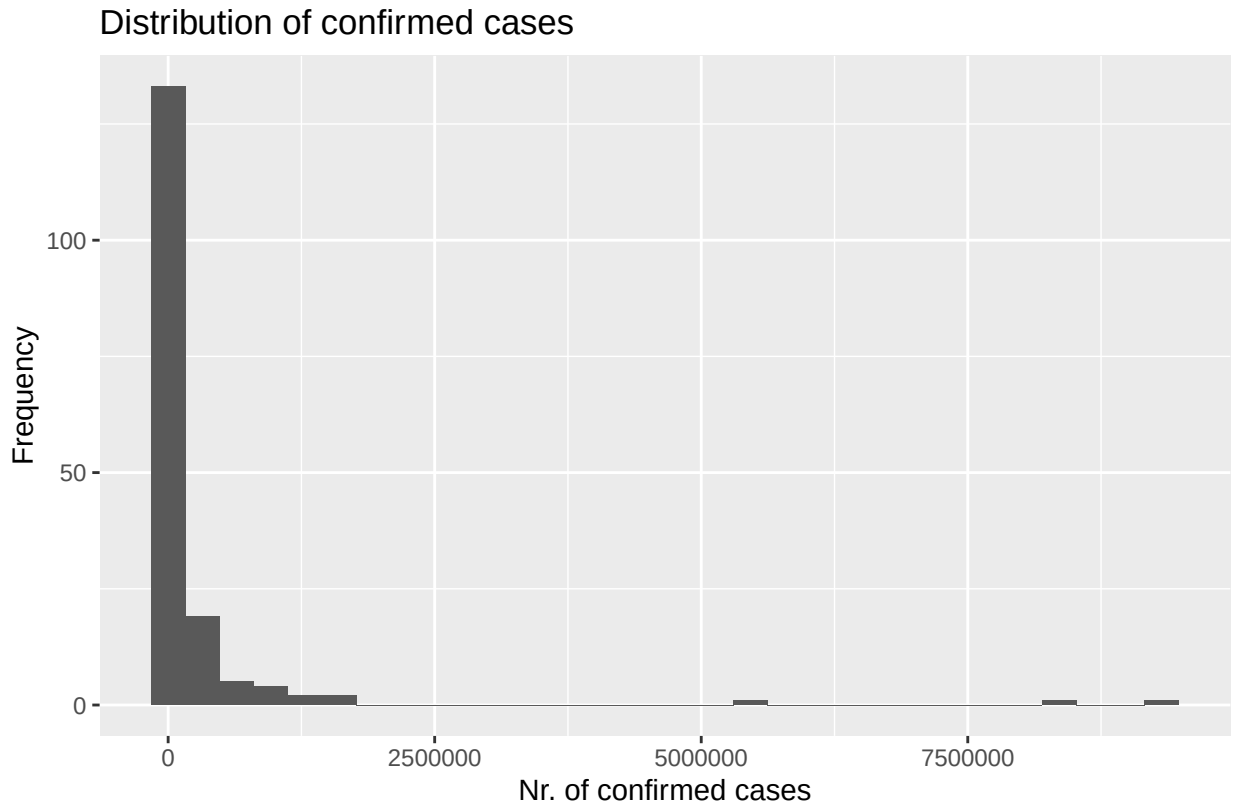

Figure 1.1

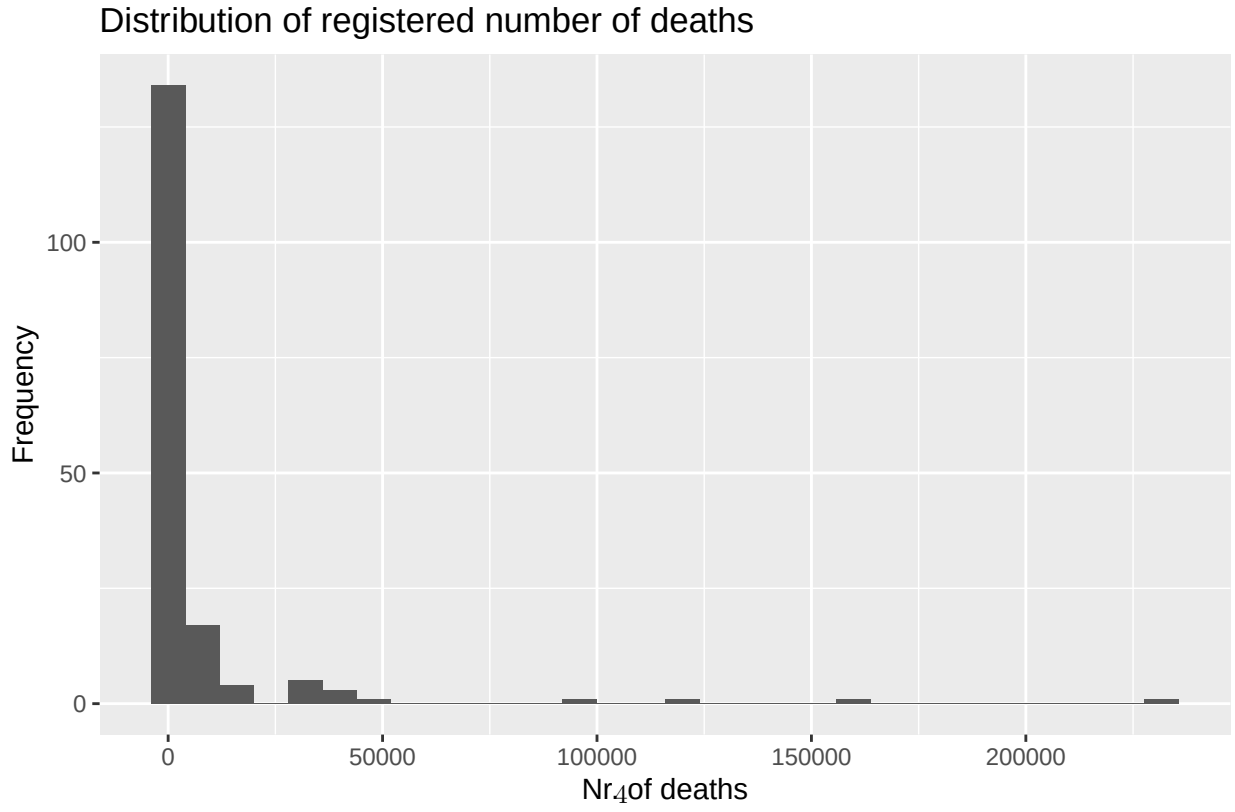

Figure 1.2

Investigating transformations

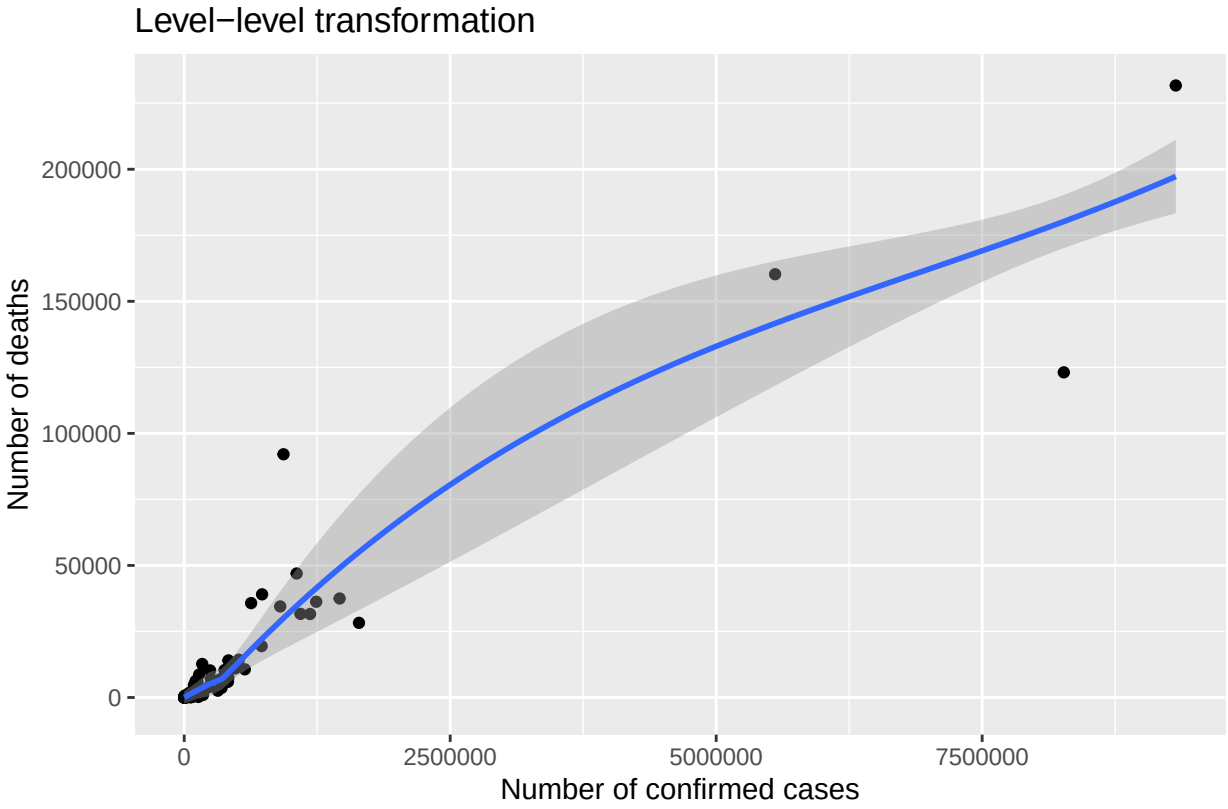

Figure 2.1

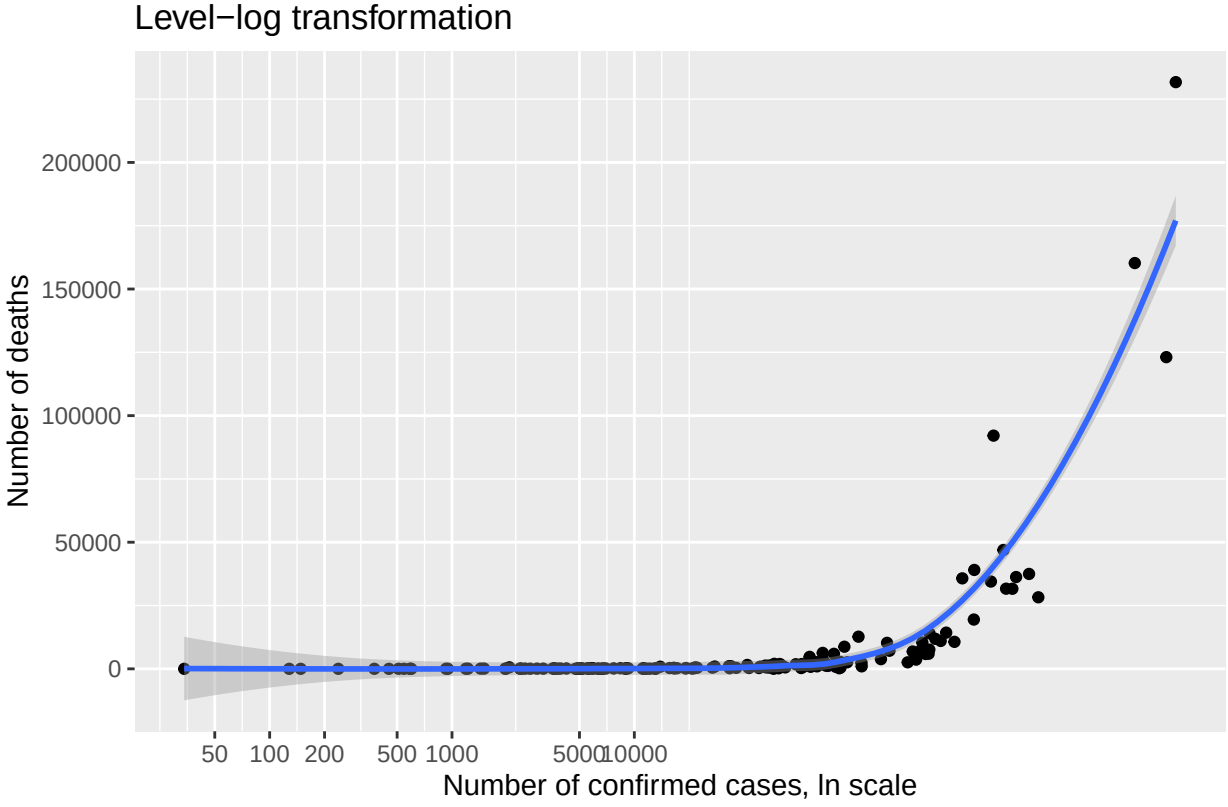

Figure 2.2

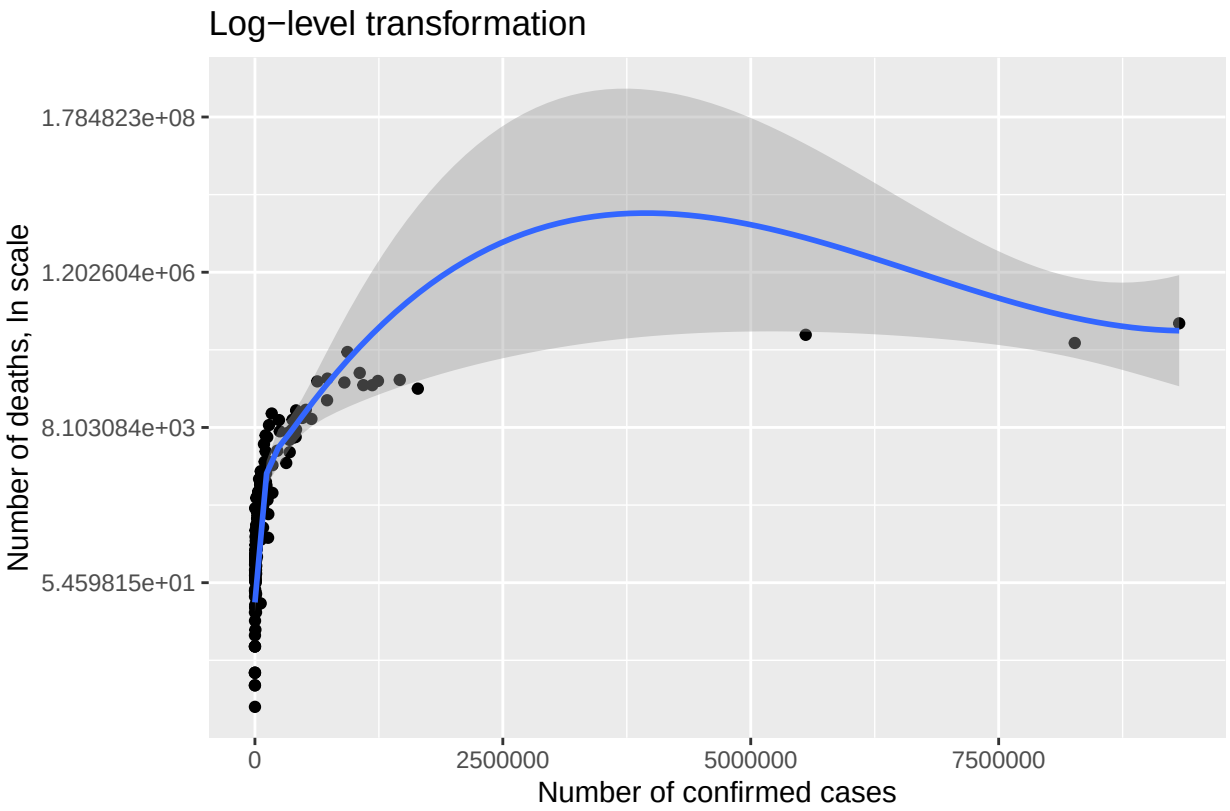
Log−level transformation
1.784823e+08
1.202604e+06
8.103084e+03
5.459815e+01
Number of deaths, ln scale
0
2500000
5000000
7500000
Number of confirmed cases

Figure 2.3

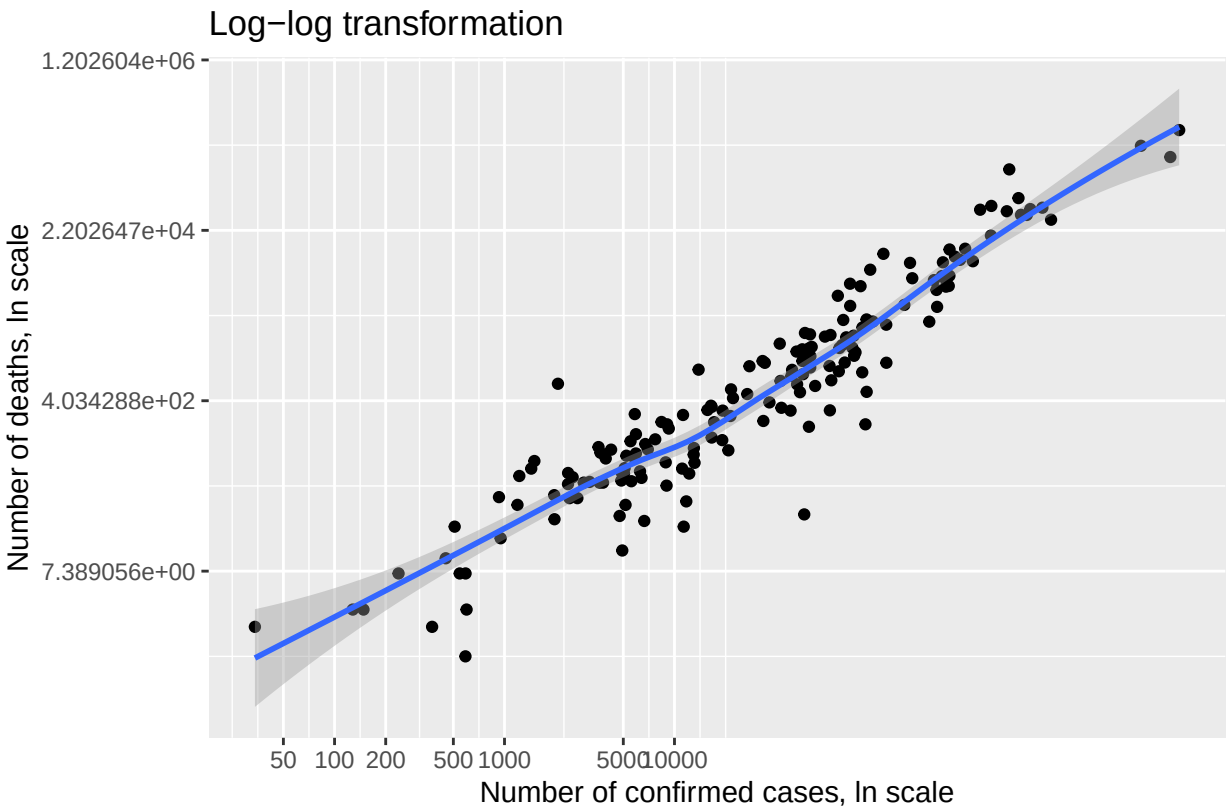
Log−log transformation
1.202604e+06
2.202647e+04
4.034288e+02
7.389056e+00
Number of deaths, ln scale
50
100
200
500
1000
5000
10000
Number of confirmed cases, ln scale

Figure 2.4

Model estimation

Figure 3.1

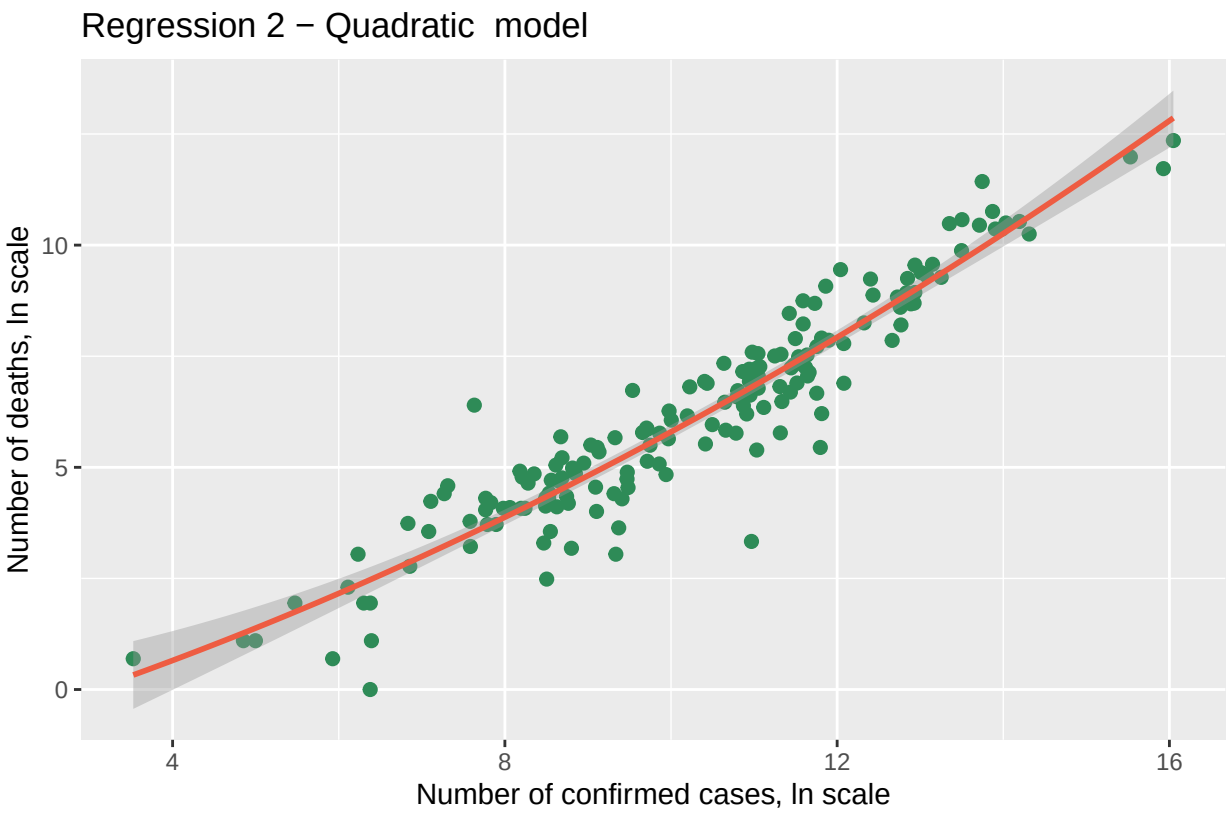
Regression 2 − Quadratic model
Number of deaths, ln scale
10
5
0
4
8
12
16
Number of confirmed cases, ln scale

Figure 3.2

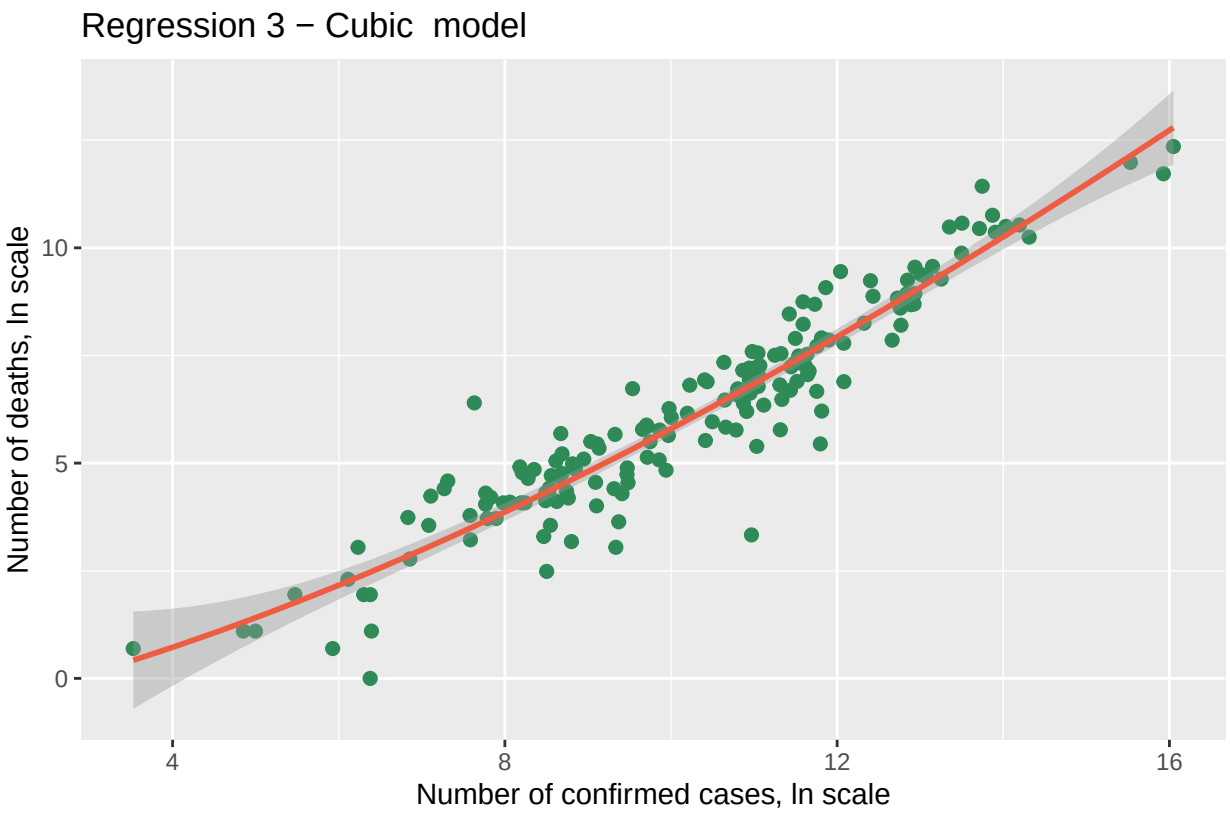
Regression 3 − Cubic model
Number of deaths, ln scale
10
5
0
4
8
12
16
Number of confirmed cases, ln scale

Figure 3.3

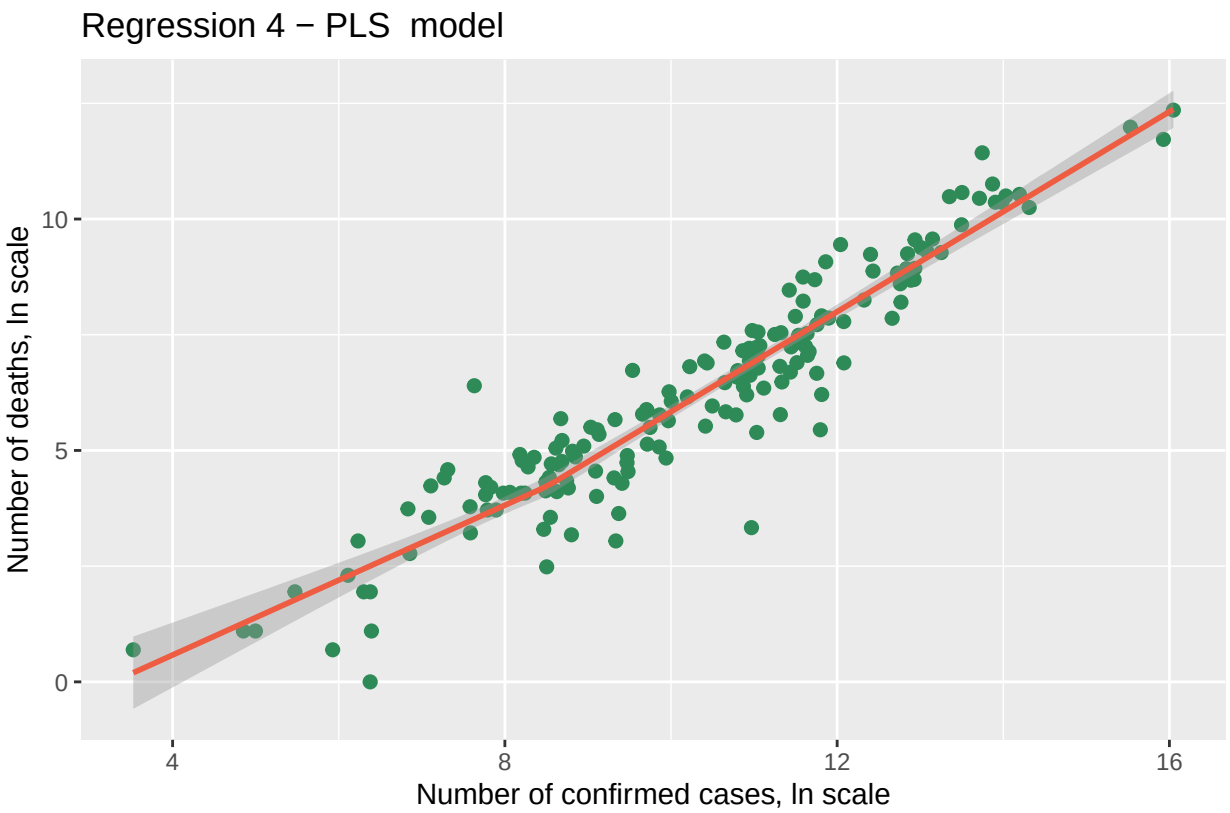
Regression 4 − PLS model
Number of deaths, ln scale
10
5
0
4
8
12
16
Number of confirmed cases, ln scale

Figure 3.4

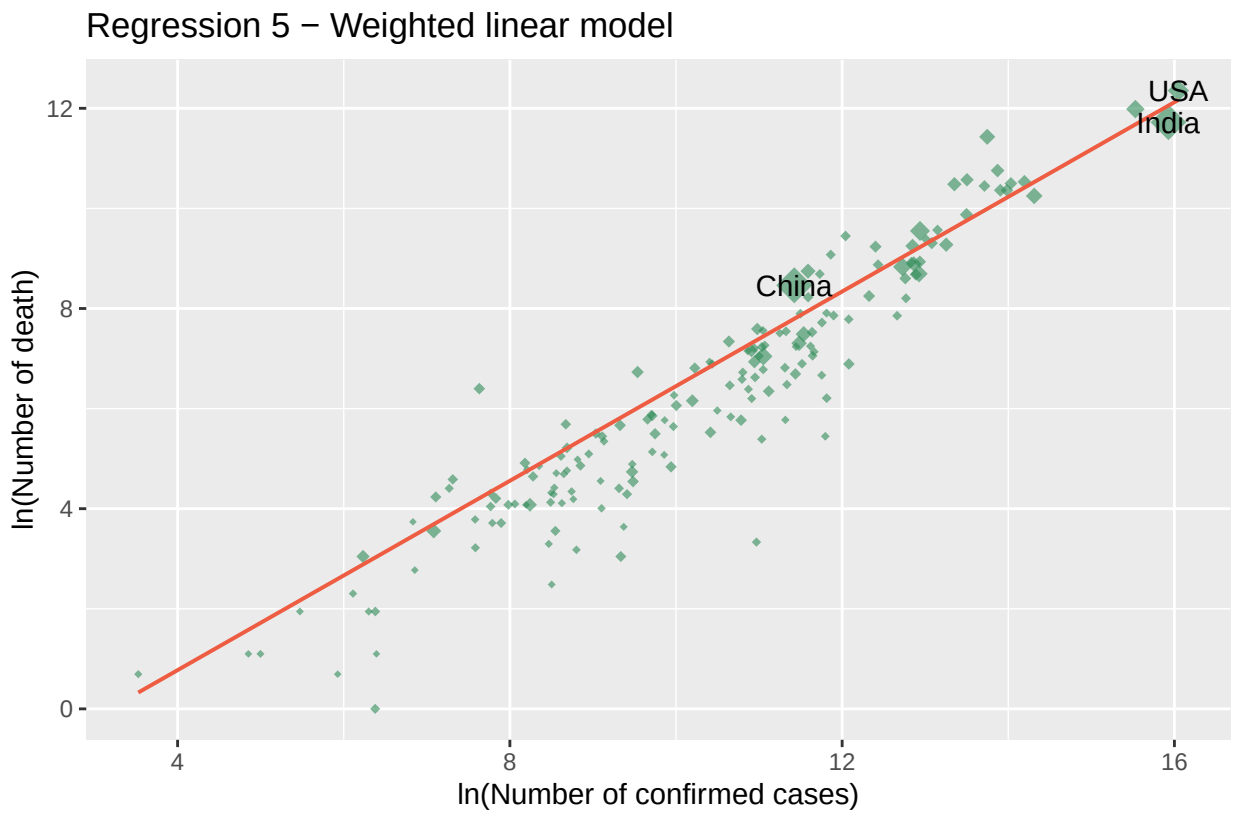

Figure 3.5

Based on model comparison my chosen model is reg5 - weighted linear

Substantive: - log-log interpretation works properly with these type of variables - magnitude of coefficients are meaningful Statistical: - linear is a good approximation - very high R-squared and captures variation well

```
##
## 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)   -3.0086    0.78467  -3.834 1.786e-04   -4.558   -1.459 166
## ln_confirmed   0.9457    0.06013  15.728 1.030e-34    0.827    1.064 166
##
## Multiple R-squared:  0.9288 ,    Adjusted R-squared:  0.9284
## F-statistic: 247.4 on 1 and 166 DF,  p-value: < 2.2e-16
```