Covid Analysis (Deaths vs Cases)

Fasih Atif

Note: HTML version differs from html version in terms of formatting.

The COVID-19 pandemic is an unprecedented global crisis having affected every corner of the globe and the lives of every individual. Some countries have fared better in being able to control the Covid spread and keep the Covid death rate low while others have suffered huge causalities.

Using the John Hopkins database on Covid-19 data, i would like to observe if there is any pattern of association between deaths per capita and confirmed cases per capita. For research purposes, i would only be using the data for the day of 09-24-2020.

We will be working with 4 main variables during the analysis:

1. country: The name of each country
2. death: Total number of deaths in each country
3. confirmed: Total number of registered cases in each country
4. population: Number of people with permanent residence residing in each country

There were a lot of data quality issues that had to be dealt with before analysis. In many places, the country names were spelled differently which had to be corrected. There were some organizations, unions, and organizations listed in the country column which had to be removed so that we could keep our focus on countries. The numeric columns had missing values so the rows with missing values had to be entirely removed.

Our original table consists of population, no of deaths and no of confirmed cases variables. For easy interpretation of population variable, we first divide the population column by 100,000. In order to get per capita values for both no of deaths and confirmed cases, we will divide both columns by the population. We wont remove any extreme values as the count doesnt look like an error.

country	death	confirmed	cases_per_capita	deaths_per_capita	population
Afghanistan	1451	39170	1029.66	38.14	38.041754
Albania	370	12921	4527.03	129.63	2.854191
Algeria	1703	50579	1174.81	39.56	43.053054

Summary statistics for X and Y variables

Histograms for both ‘no of Covid related deaths’ and ‘no of confirmed Covid cases’ are right skewed (**Appendix: Figure 1**). The median is greater then the mean. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side). There are some extreme values in both histograms caused by the very high number of confirmed Covid cases and deaths in United States, India, and Brazil.

	mean	median	std	min	max
cases_per_capita	5372.69453	2625.77500	7300.6536	8.78000	43934.340
deaths_per_capita	132.94300	47.56000	211.1308	0.09000	1240.400
population	44.52144	10.27744	154.3139	0.03386	1397.715

Investigation of the transformation of our variables

We will be working with the following equation:

*deaths_per_capita = alpha + beta * cases_per_capita*

It is important to think beforehand what kind of transformations are feasible or not. We would like a transformation that explains the association between the dependent and independent variable clearly and interpretations that make the most sense. Hence, we will visualize the model with different log transformations.

Level-Level Model: deaths_per_capita = alpha + beta * cases_per_capita (**Figure 2**)

Level - Log Model: deaths_per_capita = alpha + beta * ln_cases_per_capita (**Figure 3**)

Log - Level Model: ln_deaths_per_capita = alpha + beta * cases_per_capita (**Figure 4**)

Log - Log Model: ln_deaths_per_capita = alpha + beta * ln_cases_per_capita (**Figure 5**)

Based on model comparison, our chosen model is ln_deaths_per_capita ~ ln_cases_per_capita

- Substantive Reasoning: Level changes are hard to interpret due to non linearity of curves. A percentage change gives a better summary of the relationship between deaths_per_capita and cases_per_capita.
- Statistical Reasoning: It is a linear model which is easy to interpret and captures the variation well

Log Transformation of variables for Regression

Based on our investigation, we found out that log transformations would work best on both independent and dependent variables. So, we created new log variables for deaths per capita and confirmed cases per capita.

Estimating different Regression models

Figure 5 shows a table with the comparisons of all the models listing the various average coefficients , R2, adjusted R2 and RMSE. The model with the highest R2 is the weighted OLS regression with 0.90 while all the models give the same R2 of 0.79. The weighted OLS also has the highest RMSE with 4.22 which means that the data points are far from the regression line.

The best choice for model according to our results is the Linear Regression model:

*ln_deaths_per_capita = alpha + beta * ln_cases_per_capita*

Interpretation of B1: This model and its results tells us that for a 10% increase in confirmed cases per capita per million, we observe 9.4% higher deaths per capita per million on average.

Residual Analysis

Countries with largest negative errors

This category includes countries like Maldives, Sri Lanka, Qatar, Burundi, and Singapore. The largest negative deviance from the predicted value is found in ‘Burundi‘ with predicted deaths per capita of ‘27.66‘, but the real value is only ‘25.1‘.These countries have predicted Y values greater than the log values of deaths_per_capita. This means that their death_per_capita per million is lower than what is predicted. It needs to be investigated further as to what initiatives have led to a lower deaths per capita per million value.

Countries with largest positive errors

This category includes countries such as Belgium, Italy, Mexico, United Kingdom, and Yemen. - The largest positive deviance from the predicted value is found in ‘Belgium‘ with predicted deaths per capita of ‘32.65’, but the real value is ‘34.29‘. These countries have predicted Y values lower than their actual ln_deaths_per_capita value. This means that their death_per_capita per million is higher than what is predicted. Again, further investigation would be required to see what policies and actions led to higher than predicted ln_death_per_capita values.

Testing hypothesis

We test the following hypothesis:

H0 - B= 0 : There is no relationship between ln_deaths_per_capita and ln_cases_per_capita

HA - B != 0 : There is a relationship between between ln_deaths_per_capita and ln_cases_per_capita

```
##
## Call:
## lm_robust(formula = ln_deaths_per_capita ~ ln_cases_per_capita,
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                     Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)          -3.4531    0.34222  -10.09 5.164e-19  -4.1287   -2.777 168
## ln_cases_per_capita   0.9379    0.04525   20.73 3.652e-48   0.8486    1.027 168
##
## Multiple R-squared:  0.7863 ,    Adjusted R-squared:  0.785
## F-statistic: 429.7 on 1 and 168 DF,  p-value: < 2.2e-16
```

We will use a significant value of 0.05 to check whether we can reject null hypothesis. Our results show that we received an estimated p value of 5.255e-48 which is very close to zero and t value of 20.66 which is greater than 2. Since the p value is less than 0.05, we reject null hypothesis and conclude that there is a relationship between between ln_deaths_per_capita and ln_cases_per_capita.

Conclusion

We used deaths_per_capita and cases_per_capita variables and transformed them into log variables which gave us a linear pattern of association. Then we ran several regression models and the simple linear regression model gave us the best fit in terms of good R2 and low RMSE. Hypothesis testing showed that there is a relation between deaths_per_capita and cases_per_capita. Our results would be weakened if it was revealed that the governments had lied about the Covid statistics in their countries. Ou results would be strengthened if we could account for variables such as underlying sicknesses that actually led to the death of the individuals.

Appendix

Figure 1: Histogram for confirmed cases and deaths

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

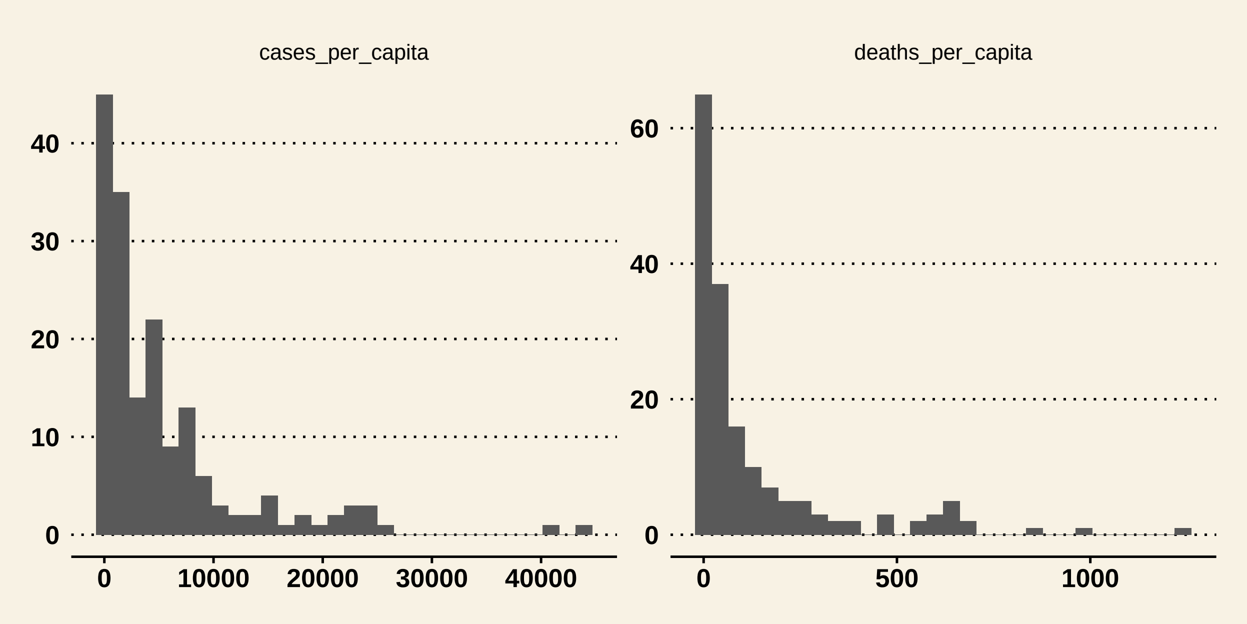


Figure 2: Level - Level Model

The graph shows a non-linear pattern with majority of values clustered towards the left while a few extreme values pull on the conditional mean towards the right hand side. We will run log models on the independent and dependent variables to check if we get better results.

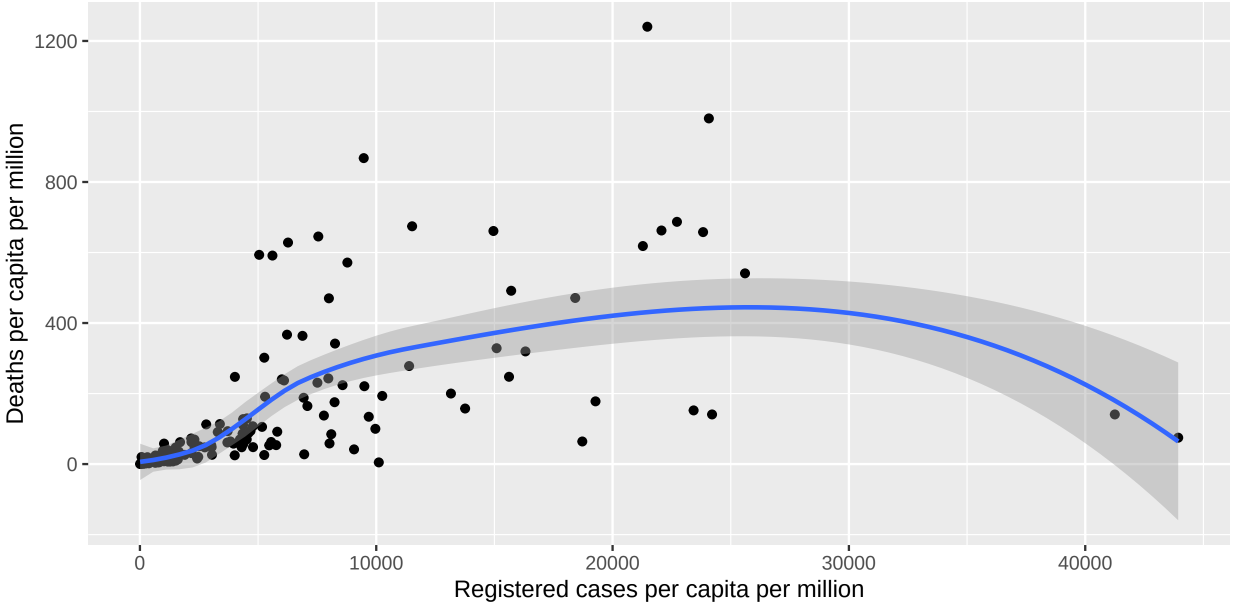


Figure 3: Level - Log Model

We transformed the independent variable (cases_per_capita) and visualized the model. Non-parametric regression on the model shows a non-linear pattern here as well. Non parametric regressions are very volatile at both ends of the curves as you can see by the wide shaded area (confidence intervals).

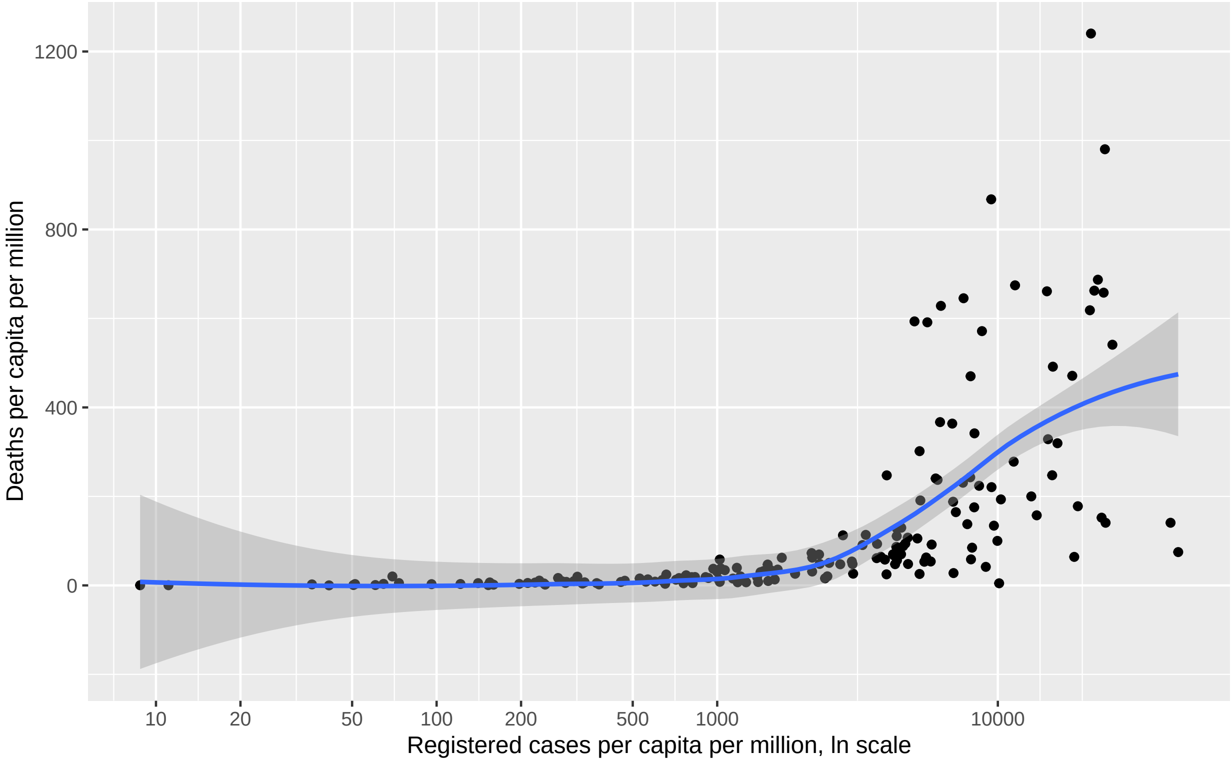


Figure 4: Log - Level Model

The graph shows a non linear pattern with majority of values clustered on the left side making the graph harder to interpret and analyze.

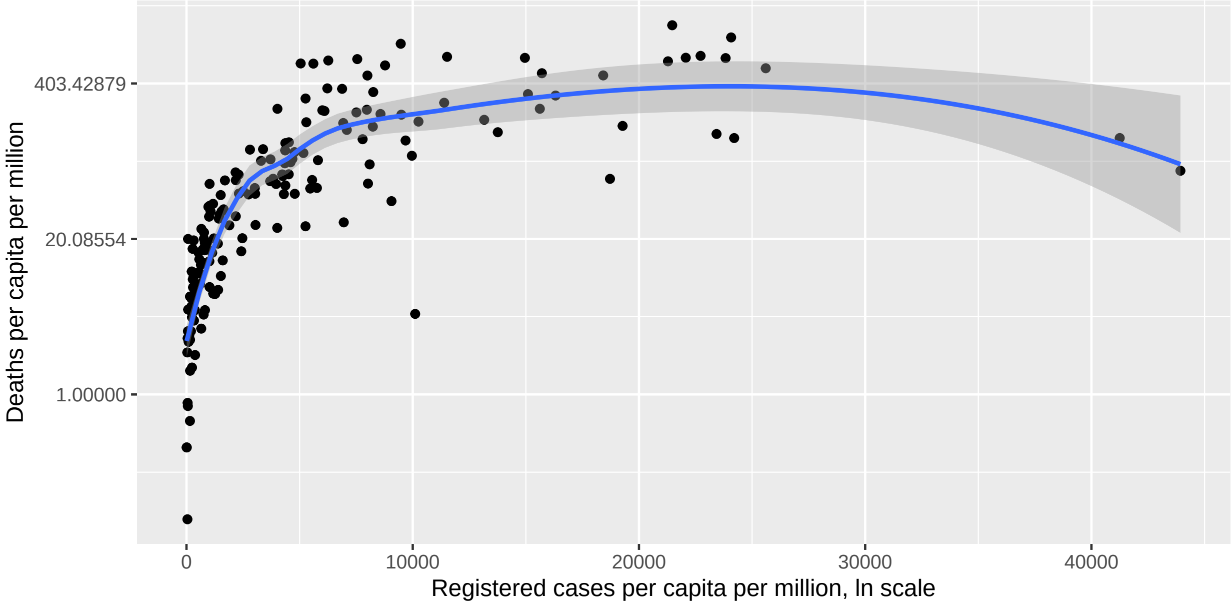


Figure 5: Log - Log Model

This transformation shows a linear pattern with the values distributed across the graph giving us a better approximation.

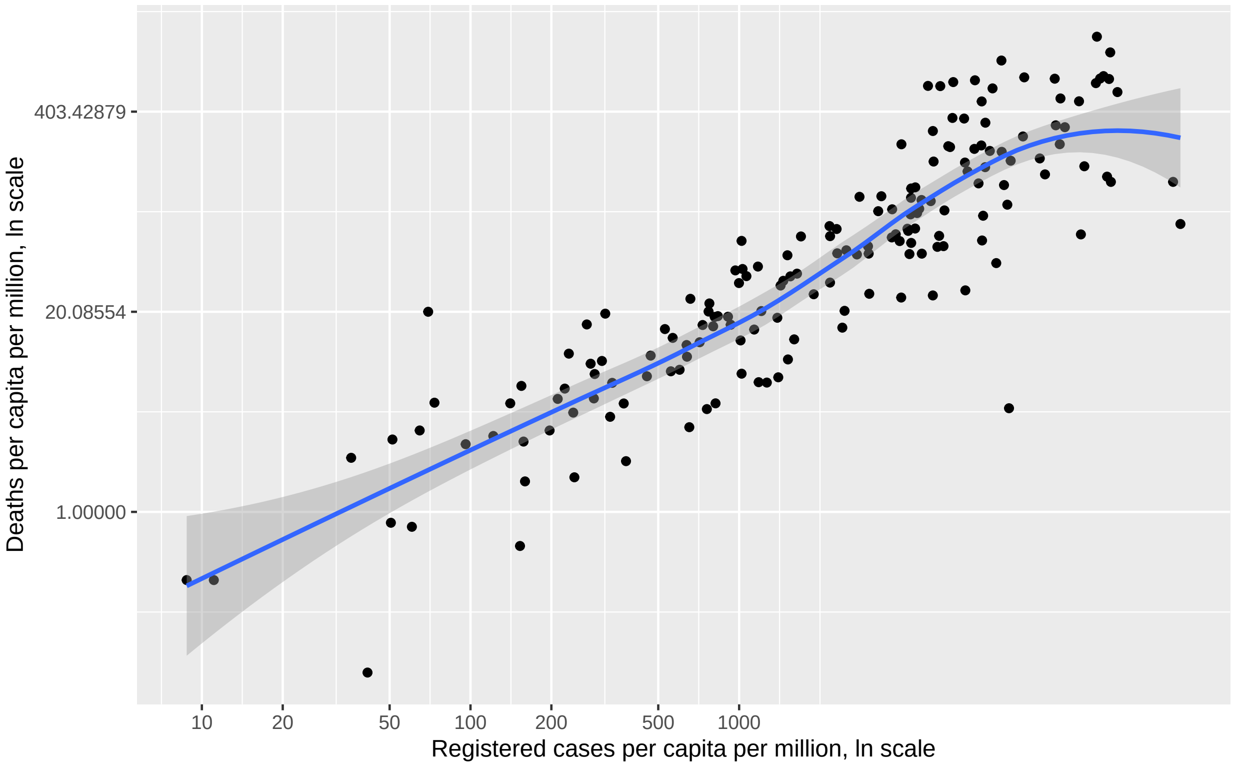


Figure 6: Estimating model comparison

	cases/capita - linear	cases/capita - quadratic	cases/capita - pls	cases/capita- weighted linear
(Intercept)	-3.45***	-3.18**	-3.26***	-2.94***
	(0.34)	(0.97)	(0.44)	(0.65)
ln_cases_per_capita	0.94***	0.86**		0.90***
	(0.05)	(0.28)		(0.08)
ln_cases_per_capita_sq		0.01		
		(0.02)		
lspline(ln_cases_per_capita, cutoff_ln)1			0.90***	
			(0.06)	
lspline(ln_cases_per_capita, cutoff_ln)2			1.62***	
			(0.41)	
lspline(ln_cases_per_capita, cutoff_ln)3			0.68*	
			(0.30)	
lspline(ln_cases_per_capita, cutoff_ln)4			-2.67***	
			(0.64)	
R^2	0.79	0.79	0.80	0.90
Adj. R^2	0.79	0.78	0.80	0.89
Num. obs.	170	170	170	170
RMSE	0.84	0.84	0.81	4.22

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Modelling deaths per capita and confirmed cases per capita

Figure 7: Linear Regression Visualization

This is the best model for estimating coefficients. R2 is 0.79 which tells us that 7.9% of the variation in the dependent variable (deaths per capita) is explained by the independent variable (cases per capita). It also has a small RMSE which means that the data points are concentrated around the regression line.

$ln\_deaths\_per\_capita = alpha + beta * ln\_cases\_per\_capita$

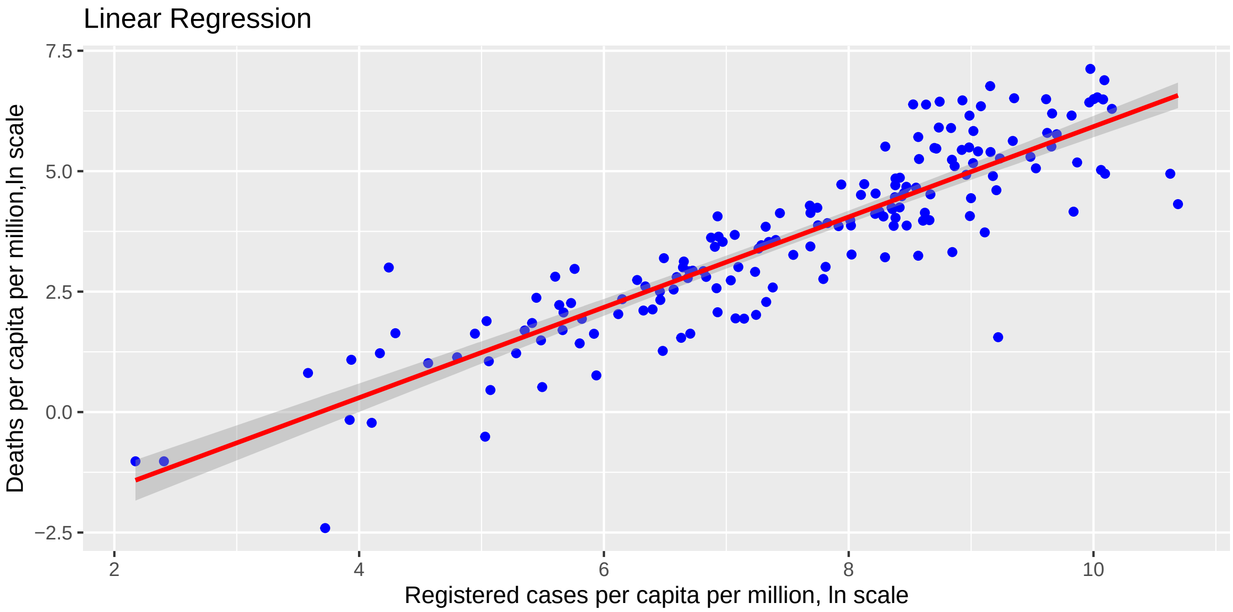


Figure 8: Quadratic Regression Visualization

$ln\_deaths\_per\_capita = alpha + beta\_1 * ln\_cases\_per\_capita + beta\_2 * ln\_cases\_per\_capita\hat{}2$

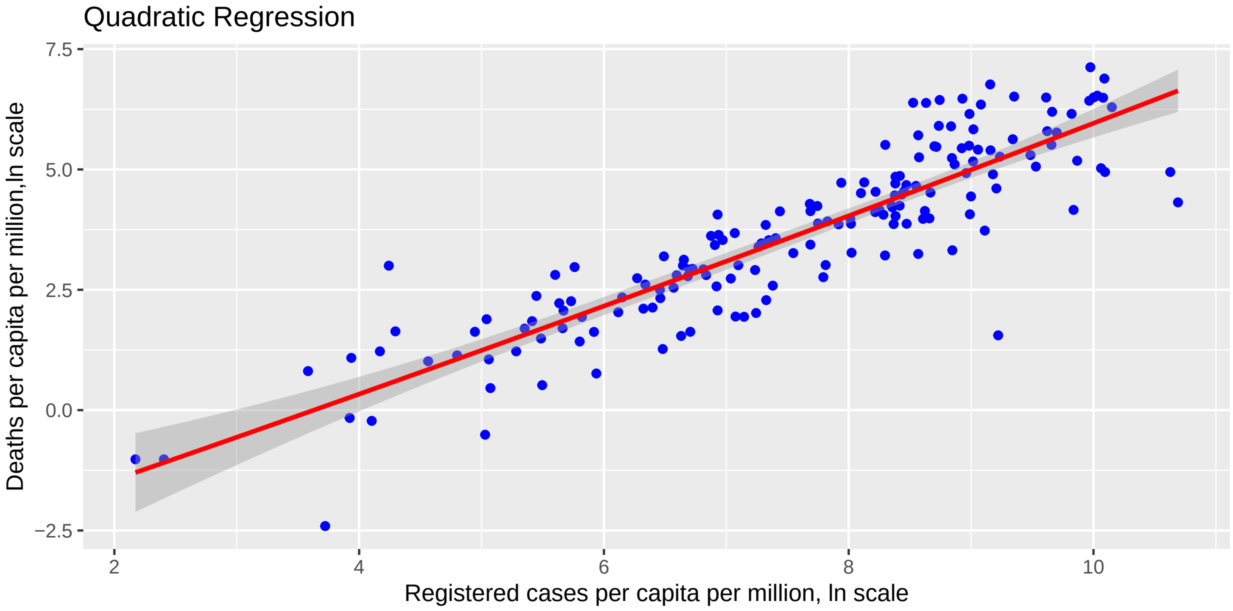


Figure 9: Piecewise Linear Spline Regression Visualization

*ln_deaths_per_cap = alpha + beta_1 * ln_cases_per_capita * 1(confirmed < 50) + beta_2 * ln_cases_per_capita * 1(confirmed >= 50)*

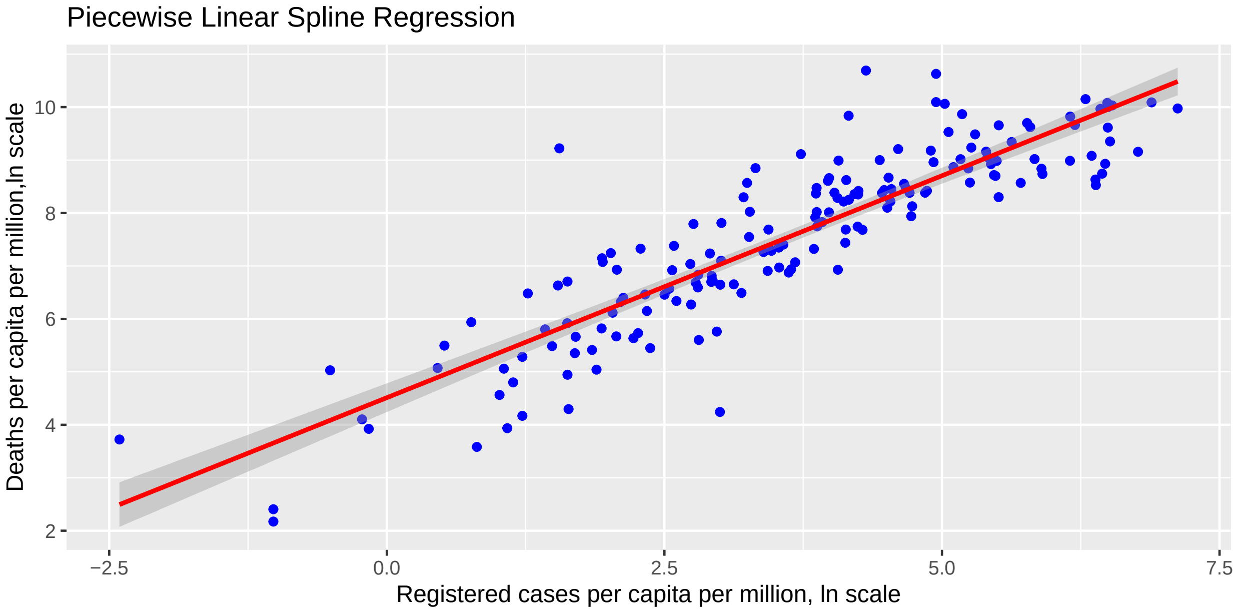
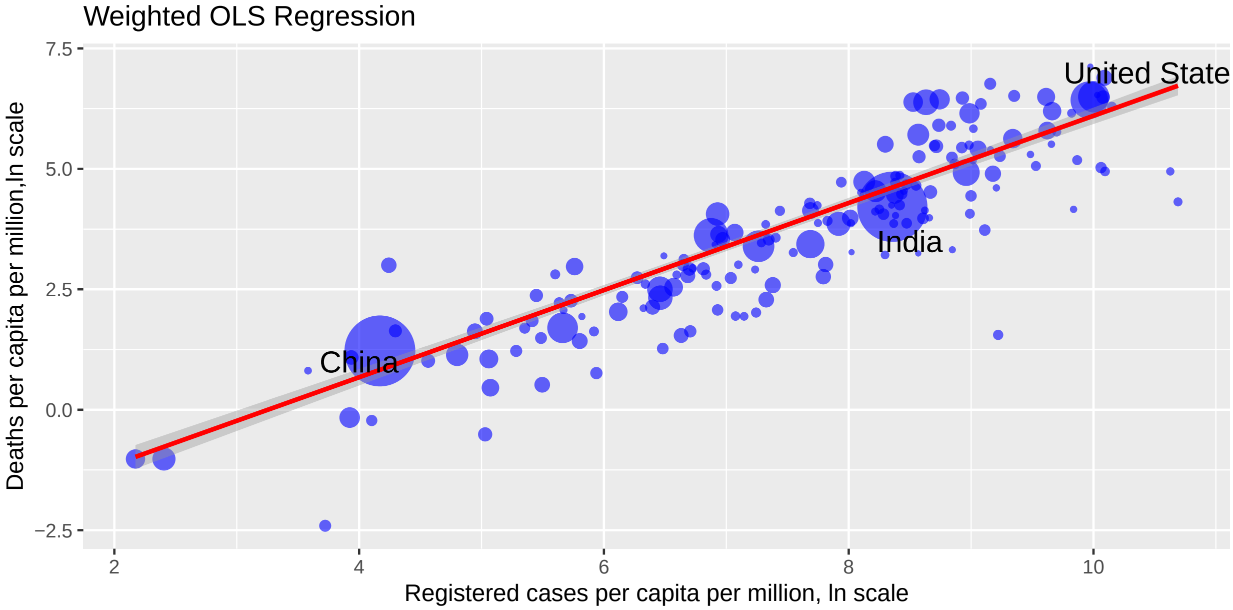


Figure 10: Weighted OLS Regression Visualization

*ln_deaths_per_capita = alpha + beta * ln_cases_per_capita, weights: population*


Weighted OLS Regression
Deaths per capita per million,ln scale
Registered cases per capita per million, ln scale
United State
India
China