Assignment: DA2 and Coding1

Xinqi Wang 2003316

Introduction

In this exercise, the main aim is to find the relationship between COVID-19 confirmed cases and its death cases across countries. Do they correlate with each other? If so, how? The variables I have in this analysis are list of country names where data was collected, counts of COVID19 confirmed cases include confirmed and probable (where reported), counts of COVID19 death cases include confirmed and probable (where reported), counts of estimated COVID19 recovered cases, counts of COVID19 active cases, and 2019 country population count in ten thousands. The population data is included becasue I am also interested to see if country's population would play a part in later analysis. One potential data quality issue would be the reliability of the data, if countries reported their true information.

The dependent variable here is death cases, and indepedent variable is confirmed cases. population data will be needed in later analysis for weighted regression. It is scaled in the data cleaning process with 10,000. Since the purpose of my analysis is to find the relationship between confirmed cases and death cases from countries, and due to my limited sample size, I have decided not to drop any data.

The combined histograms below shows the key variables used in this anlysis: my x variable: confimred, my y variable: death, and weighted variable: population. From the summary statistics table below, both confirmed and death variable's mean are higher than median, which means that both of them are skewed to the right, and have a long right tail. We can see that from the histograms as well.

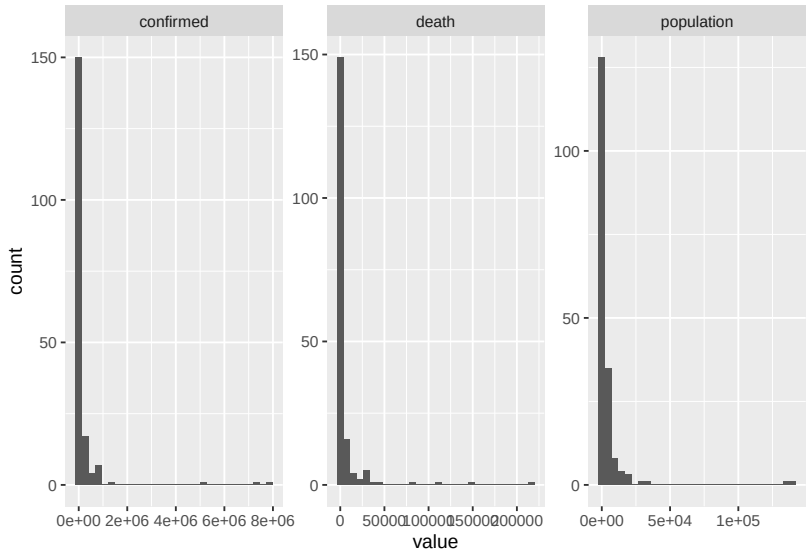

Figure 1: Histograms of Key Variables

```
##    country            confirmed           death            population
##  Length:182         Min.   :      2   Min.   :     0.0   Min.   :     3.39
##  Class :character   1st Qu.:   3583   1st Qu.:    59.5   1st Qu.:   253.53
##  Mode  :character   Median :  16274   Median :   276.5   Median :   975.80
##                     Mean   : 209431   Mean   :  5965.2   Mean   :  4175.21
##                     3rd Qu.:  90703   3rd Qu.:  1600.0   3rd Qu.:  3040.49
##                     Max.   :7860634   Max.   :216074.0   Max.   :139771.50
```

Next, I have to consider whether and where to use **log transformation**. I have 4 options here: **level-level**, **level-log**, **log-level** and **log-log**. I have tried each of them and plotted them with lo(w)ess for different scatterplots. The graphs are in the Appendix as figure 2 and 3. Based on the outputs, I decided to log both of my x and y. My **substantive reasons** are: Firstly, level changes is harder to interpret and our aim is not to get absolute based comparison. Secondly, log-log gives a better interpretation with percentage increases or decreases in confirmed and death cases. Finally, log transformation is a better approximation to make simplification. My **statistical reasons** are: It makes sense to take log as my variables have skewed distribution with long right tail. Since the distribustions of confirmed and death cases are skewed with a long right tail, taking natural log makes them closer to symmetric. The same results can be shown in figure 3, where the log-log transformation scatterplot gives the most linear relationship between 2 variables.

In the next session, I examine 4 potential models to use in my analysis:

1. **The Simple Linear Regression Model (reg1)**:

$$ln(death) = \alpha + \beta * ln(confirmed)$$

   The slope of this regression is 1.03, means that death cases is higher, on average, by approximately 10.3% in countries with 10% higher in confirmed cases. The adj-R^2 is 0.89 which is pretty high.

The rest of the models are in the Appendix, there I compare the rest of three models with the Simple Regression Model (reg1). Based on model comparison, my choice is to go with **reg1 - Simple linear regression** (ln_death ~ ln_confirmed) My **substantive reasons** are: 1. log-log interpretation works properly for countries, and 2. magnitude of coefficients are meaningful. My **statistical reasons** are: 1. simple model, easy to interpret, and 2. Comparatively high R^2 and captures variation well (A more detailed explainations on model comparisons and selections are in the Appendix).

In this analysis, those more complicated specifications (reg2-reg4) didn't lead to very different conclusions becasue the pattern of association turned out to be fairly linear overall.

Next, a **hypothesis test** on β is conducted. I want to know if dependent variable and the explanatory variable are related at all? The significance level I choose is at 0.05. Baesd on the output table in the Appendix, the

p value of the hypothesis is < 2.2e-16 which is much lower than 0.05. Meaning that beta cannot be 0 and we should reject the null at significance level 0.05, and x and y are related. Below is the result table:

Summary Statistics for Hypithesis Testing

```
## Linear hypothesis test
##
## Hypothesis:
## ln_confirmed = 0
##
## Model 1: restricted model
## Model 2: ln_death ~ ln_confirmed
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    169
## 2    168  1 1285.8  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, I analyse the **residuals**. The top 5 countries who **lost** (relatively) the most people due to COVID are: **Burundi** (y = 0, $\hat{y}$ = 2.19), **Iceland** (y = 2.30, $\hat{y}$ = 4.18), **Qatar** (y = 5.39, $\hat{y}$ = 7.84), **Singapore** (y = 3.33, $\hat{y}$ = 7.02) and **Sri Lanka** (y = 2.56, $\hat{y}$ = 4.51). These countries are the ones with largest negative errors, which are also the ones that lies most far above the fitted line, that's why they are considered to be the ones who lost (relatively) the most people. On the other hand, the top 5 countries who **saved** (relatively) the most people due to COVID are: **Ecuador** (y = 9.41, $\hat{y}$ = 7.99), **Fiji** (y = 0.69, $\hat{y}$ = -0.70), **Italy** (y = 10.50, $\hat{y}$ = 8.92), **Mexico** (y = 11.34, $\hat{y}$ = 9.76), and **Yamen** (y = 6.39, $\hat{y}$ = 3.58). Similarly, these countries are the ones with largest positive errors, which are also the ones that lies most far below the fitted line, that's why they are considered to be the ones who saved (relatively) the most poeple.

To conclude, in this analysis, I use ln(confirmed) as my independent variable, and ln(death) as my dependent variable. The general pattern between these two variables is mostly linear. I choose to use the most simple linear regression model which gives the best fit and easy to interpret. The mean message from the results is death cases is higher, on average, by approximately 10.3% in countries with 10% higher in confirmed cases of COVID-19. The main advantage of this model is its simplicity and relatively high fits of my data. One weakness would be that linear regression only look at the mean of the dependent variable and the independent variables. However, sometimes we need to look at the extremes of the dependent variable, e.g., countries are at risk when they have extrame high numbers in death count, so we would want to look at the extremes in this example.

Appendix

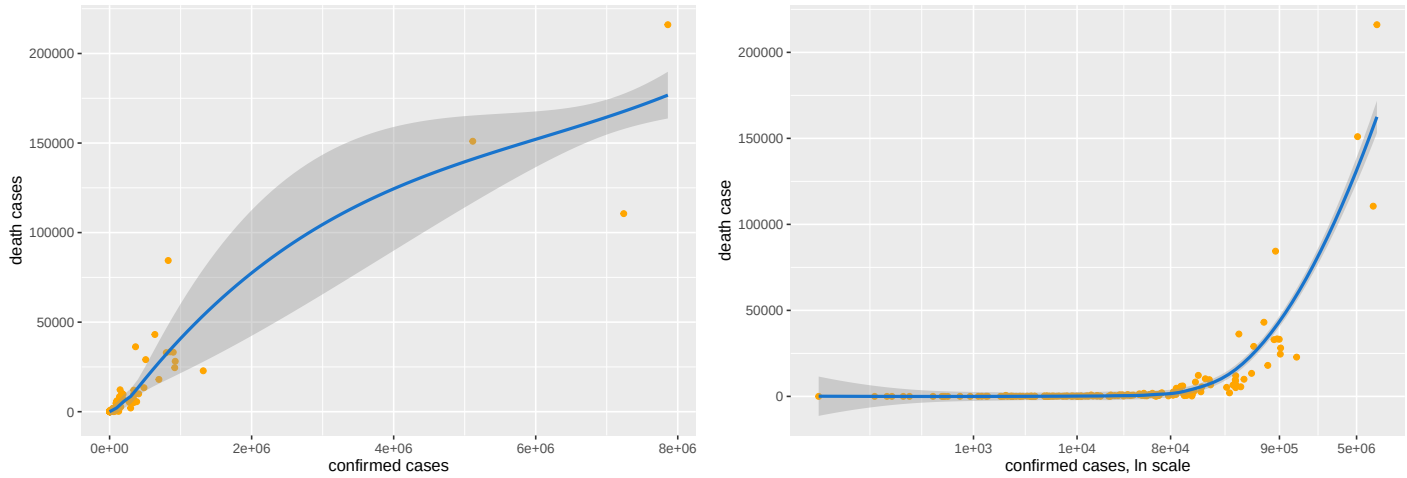

Figure 2: Level-level model without Scaling and Level-log model with Scaling for Confirmed

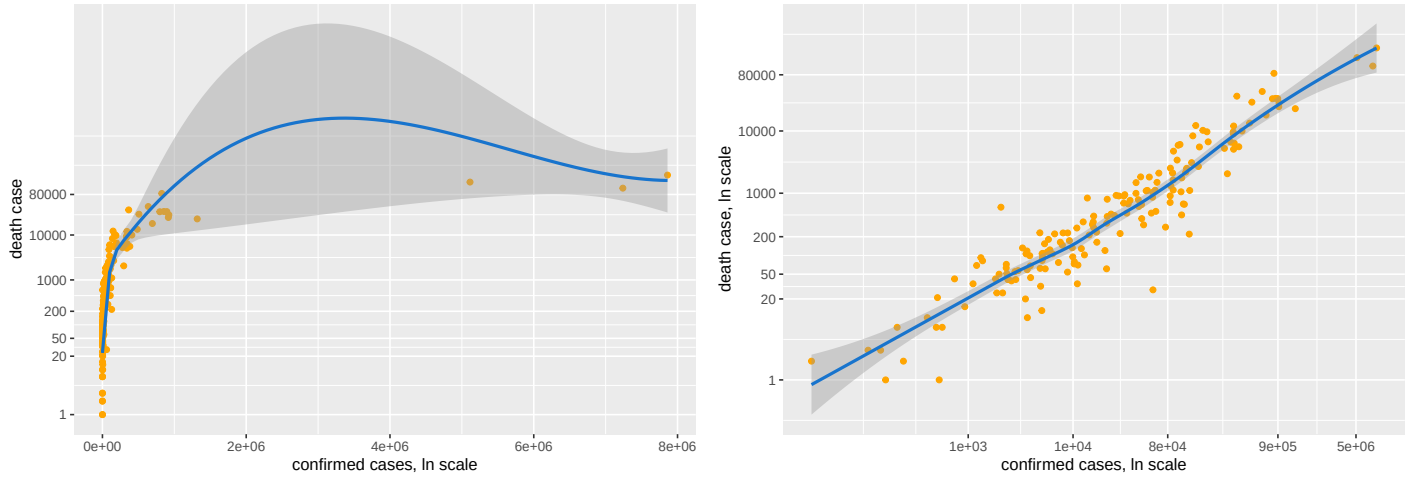

Figure 3: Log-level model with Scaling for Death and Log-log model with Scaling for both Confirmed and Death

2. **The Quadratic (Linear) Regression Model (reg2)**:

$$ln(death) = \alpha + \beta_1 * ln(confirmed) + \beta_2 * ln(confirmed)^2$$

The graph shows the quadratic fit settles for slight nonlinearity. The pattern is a positive association through the entire range of observed ln_confirmed. Based on the summary statistics in figure 6, the two slope parameters are 0.59 and 0.02, with no clear interpretation except the second, positive, number showing that the parabola is convex. It didn't lead to very different conclusions with the simple linear model because the pattern of association turned out to be fairly linear overall.

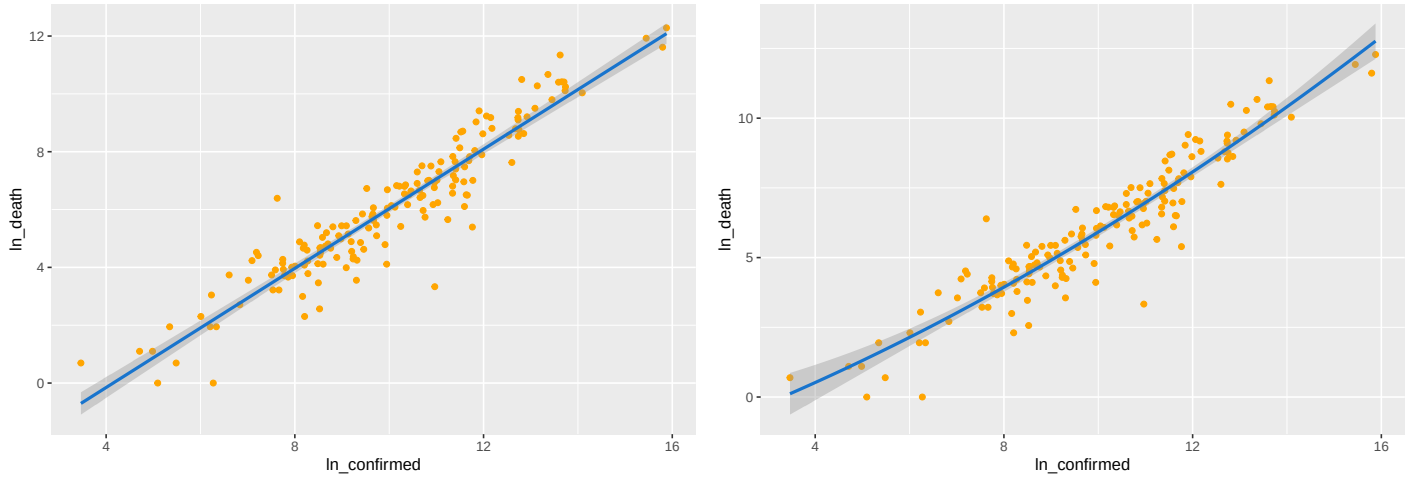

Figure 4: Simple Linear Regression (reg1) and Quadratic (Linear) Regression (reg2)

3. **The Piecewise Linear Spline Regression Model (reg3)**:

$$ln(death) = \alpha + \beta_1 * ln(confirmed) * 1(confirmed < 15000) + \beta_2 * ln(confirmed) * 1(confirmed >= 15000)$$

The slope of first line segment is 0.87, just a little flatter than reg1- the simple linear regression, where the slope is 1.03. The slope of the other line segment is 1.15. Comparing countries with confirmed cases below 15000, death cases is higher, on average, by approximately 8.7% in countries with 10% higher in confirmed cases. With confirmed cases above 15000, death cases is also higher, on average, by approximately 11.5% in countries with 10% higher in confirmed cases. The adj-R^2 is 0.89 here, same with the one from the simple linear regression. The improvement on the fit is very small in this case and only provides a better fit for only a few observations.

4. **The Weighted Linear Regression, using Population as Weights (reg4)**:

$$ln(death) = \alpha + \beta * ln(confirmed), weights : population$$

The scatterplot for the weighted regression shows the size of each country: the area of the circle is proportionate to their population. The same linear regression using population as weight gives a slope

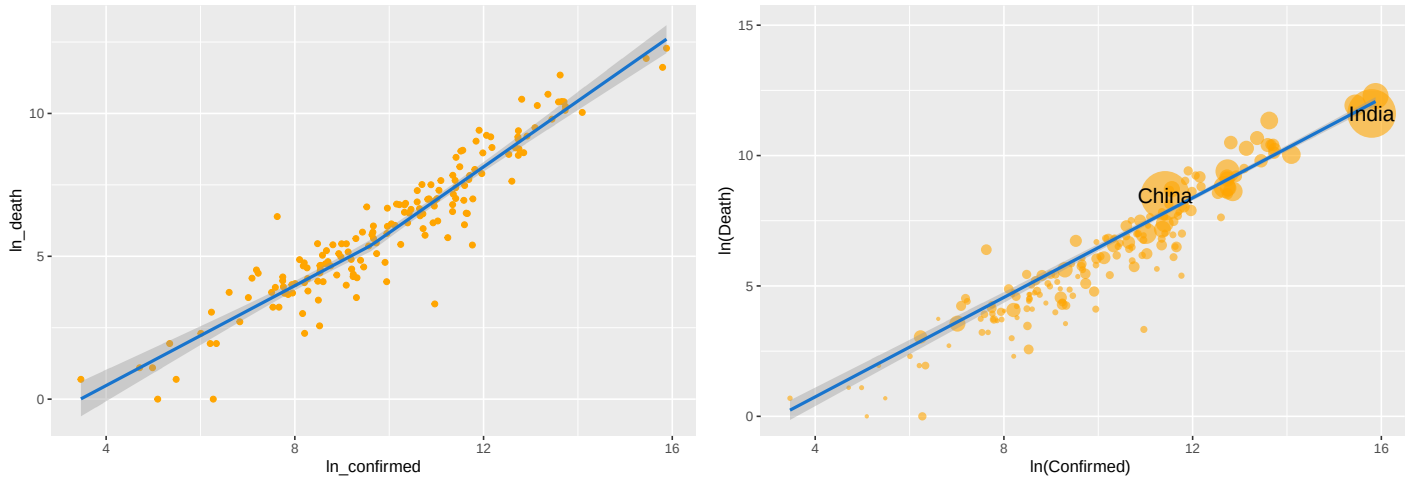

Figure 5: Piecewise Linear Spline Regression (reg3) and Weighted Linear Regression (using Population as Weights) (reg4)

of 0.95, which turns out to be similar. This shows that countries with 10% more confimred cases have, on average, 9.5% more death cases. The adj-R^2 is improved by 0.04 compare with reg1, but is not a big difference. And RMSE is very high at 42.97 reflects the poor ability of the model to accurately predict the data even though my goal here is not model prediction, it is worthwhile to point out. Overall, the two regressions (reg1 and reg4) show similar results because larger countries do not tilt the regression line much. As the weighted regression produces results that are similar to the unweighted regression.

Figure 6 gives the summary statistics for the 4 model regressions results.

	ln_confirmed - linear	ln_confirmed - quadratic	ln_confirmed - PLS	ln_confirmed - weighted linear
(Intercept)	-4.28***	-2.18*	-3.02***	-3.07***
	(0.30)	(0.86)	(0.55)	(0.78)
ln_confirmed	1.03***	0.59***		0.95***
	(0.03)	(0.17)		(0.06)
ln_confirmed_sq		0.02**		
		(0.01)		
lspline(ln_confirmed, cutoff_ln)1			0.87***	
			(0.06)	
lspline(ln_confirmed, cutoff_ln)2			1.15***	
			(0.04)	
R^2	0.89	0.89	0.89	0.93
Adj. R^2	0.89	0.89	0.89	0.93
Num. obs.	170	170	170	170
RMSE	0.83	0.81	0.81	42.97

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

Modelling COVID-19 Death Cases and Confirmed Cases of Countries

Figure 6: Modelling COVID-19 Death Cases and Confirmed Cases of Countries