Joint Assignment for DA2 and Coding 1

by Ozan Kaya (#2003859)

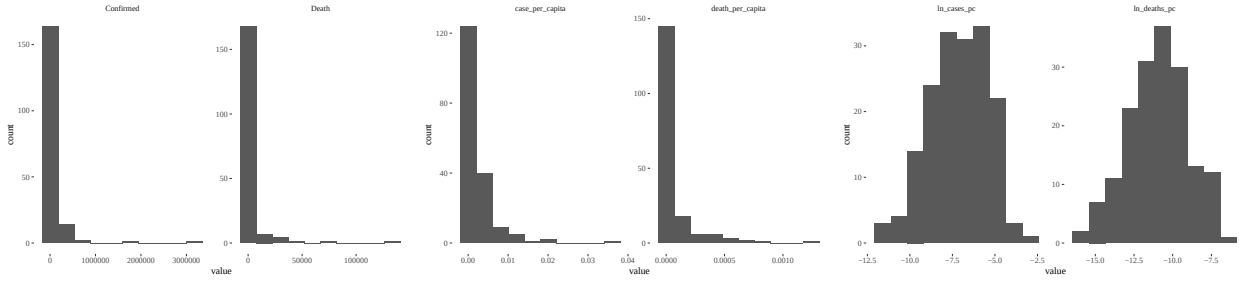
November 29, 2020

Introduction

This document is prepared as the first assignment for the Data Analysis 2 course in Central European University and it tries to document the relationship between confirmed covid cases and deaths in 07/10/2020 for various countries around the globe. To be more precise, I am interested to see if mortality rates differ accross countries and to see the pattern of association of mortality rates against confirmed covid cases in 167 distinct countries.

All of the input and output files can be seen at https://github.com/semihozankaya/Coding-1-Covid-Assignment.

First Controls



If we look at the first graph above, the confirmed covid cases shows a very skewed distribution. Considering the difference in country populations, this is expected. Accordingly, the covid deaths shows a very similar distribution with a certain percentage of confirmed cases resulted with the patients' deaths.

Per capita values also show a skewed distribution in the second graph, although a little less so. This could be due to the different testing procedures or recording methodologies between countries. Here the highest per capita case values consists of western countries as well as relatively small countries in terms of population. We can speculate that these countries can afford more testing for their citizens, thus record more cases.

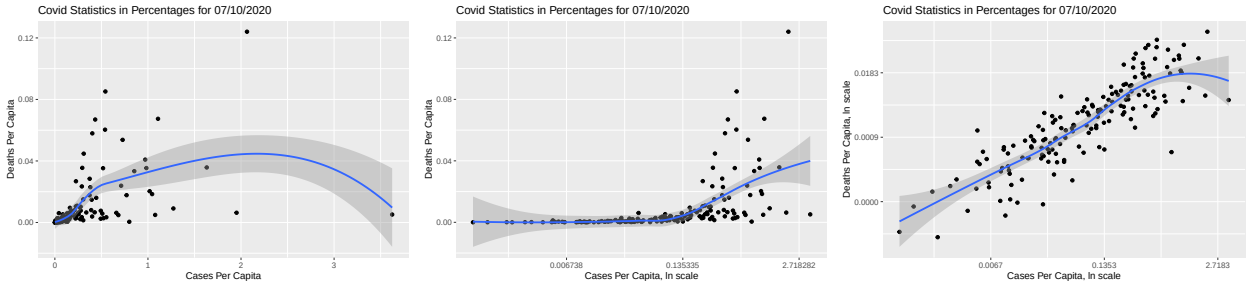
This distribution resembles a log-normal distribution with some outlying values in the right tail. We can think about a log transformation to correct for skewness and outliers for both of the variables. I am also going to transform the per capita measures to show as percentages so that it can be read easier.

It also seems that we have 0 deaths in 15 countries. Some of them are relatively small countries in terms of population but some of them are actually quite big. Among them, Uganda for instance had recorded 1000 covid cases but 0 deaths.

The mean death rate for our date seems to be 3%. The median is 2%. I am comfortable with dropping the beforementioned 15 countries from our dataset, either because they are small in size or the death counts seems unlikely.

The final distribution of our variables can be seen on the last graph above. The skew is not as problematic as before and the distribution resembles a normal distribution.

Checking scatter-plots



The first graph is deaths per capita against confirmed covid cases per capita in a linear sclae. As can be seen, there is clearly a non linear pattern of association. Most of the observations lies on the bottom left of the graph. This is not very informative as it is.

The second graph is in a log scale in the x axis. The data shows very small values of death per capita for a relatively large range of cases per capita. Within a log scale for the x axis, we can see that a large subset of death per capita values have clustered around 0.00 to 0.01. So perhaps we can benefit from a change in scale for Y axis as well.

After changing both of the axes' scales to log scaling, as we have suspected, lots of deaths per capita values have been clustered between 0 and 0.001. Thus, changing the scale seems to be resulting with a more informative visualization for our case. The fit is much better now. It is almost linear.

The log-log model would help us overcome the skewness in our data and ease out some of the outlying values as well. We will continue with the ln transformations of our values.

Testing models:

1) First Model: $ln\_deaths\_pc = \alpha + \beta * ln\_cases\_pc$

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-8.465	0.1662	-50.95	0
ln_cases_pc	0.9743	0.05232	18.62	0

First model shows a very robust β estimate of 0.9743 (with a t statistic of 18.62). Here, note that β is $\frac{dY}{dX}$, which can be rewritten as $\frac{dY}{Y}$ over $\frac{dX}{X}$ since both x and y are ln transformations of our original variables, which is basically the elasticity coefficient of deaths per capita with respect to cases per capita.

Here, we can interpret the regression coefficient as a relative change in the covid cases per capita between countries are associated with a 97% relative change in covid deaths per capita. The elasticity is constant between countries in this model. The fit seems relatively strong (with adjusted R square to be 0.74 and F-statistic being 346, the model seems statistically significant in overall.)

Our findings are intiutive as well. We shouldn't expect death rate of the Covid-19 to change between countries. Covid doesn't seem to differentiate between cultures or geography.

2) Second Model: $ln\_deaths\_pc = \alpha + \beta_1 * ln\_cases\_pc + \beta_2 * ln\_cases\_pc^2$

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-8.537	0.2462	-34.67	0
ln_cases_pc	0.8924	0.1695	5.266	0
ln_cases_pc_sq	-0.01456	0.02536	-0.574	0.5667

In our second model, β_2 is not statistically significant. Adjusted R square is actually smaller than our first model. F-statistic is also considerably lower than our first model. The overall model is still statistically significant though. But individually β_2 is insignificant.

β_1 shows a less stronger pattern of association between x and y than our first model.

Since the visual inspection showed no non-linear association in our first model and since we don't expect higher confirmed cases per capita to change the mortality rate of the virus, the squared variable being statistically insignificant is also in line with intiution and is not surprising.

3) Third Model: $ln\_deaths\_pc = \alpha + \beta_1 * ln\_cases\_pc + \beta_2 * ln\_cases\_pc^2 + \beta_3 * ln\_cases\_pc^3$

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-8.644	0.2452	-35.25	0
ln_cases_pc	0.4947	0.2936	1.685	0.09384
ln_cases_pc_sq	-0.1973	0.1061	-1.86	0.06466
ln_cases_pc_cub	-0.02016	0.01087	-1.856	0.06531

In our third model, we also included the cubic transformation of our independent variable. The variables are all individually statistically significant in 90% significance level. Even though, the model is alltogether statistically significant, the F-statistic has declined to a third of our first model. Adjusted R square on the other hand improved very slightly.

The third model seems to have better represent the dispersion in both ends of the ln_cases_pc. However, its interpretation is a lot harder than our first model. Here, we are increasing complexity for a very little extra information.

4) Fourth Model: $ln\_deaths\_pc = \alpha + \beta_1 * ln\_cases\_pc * (ln\_cases\_pc < 0.05) + \beta_2 * ln\_cases\_pc * (ln\_cases\_pc > 1)$

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-8.928	1.418	-6.297	0
ln_cases_pc_cutoff1	0.9067	0.2508	3.615	0.0004
ln_cases_pc_cutoff2	1.023	0.06068	16.87	0
ln_cases_pc_cutoff3	-0.624	0.8344	-0.7478	0.4557

In our fourth model, we can see that before the first cutoff point, the slope coefficient is 0.90, suggesting a 90% association of deaths per capita in relative changes of confirmed cases per capita below countries with 0.005 percent infection rate. β_1 is statistically significant but its standard error is considerably large. This could be due to the small sample size as well.

β_2 also shows a similar pattern of association. It is very robust with a t statistic of almost 17 and it implies an association of 102 percent relative change between relative changes in cases per capita in countries between 0.005 percent and 1 percent infection rates.

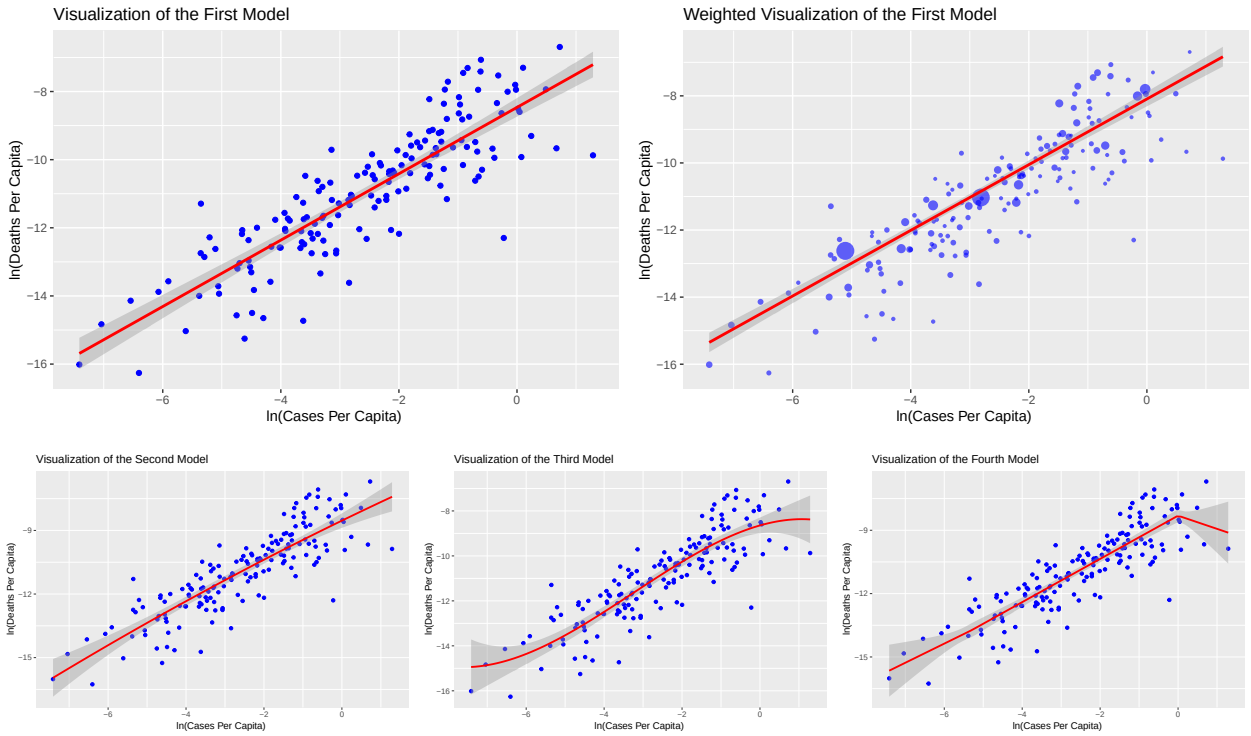
β_3 on the other hand, is not statistically significant and shows a negative pattern between our variables.

This model has the highest adjusted R square among the alternative models. Its F statistic is one of the lowest but the model is still very significant alltogether.

However, this approach is again overly complex and in fact it is identical with dropping the outlying values from our analysis. Even though these outlying values might show some faulty record keeping, fraud or differences in methodologies, our first model implies a very similar result with a much simpler approach.

I believe that we can comfortably choose the first model for our analysis.

5) Visual Inspections of the Models



Residual analysis.

1) Find countries with largest negative errors

```
## # A tibble: 5 x 5
##   Country   Confirmed Death ln_deaths_pc reg1_res
##   <chr>         <dbl> <dbl>        <dbl>    <dbl>
## 1 Singapore     45614    26        -12.3    -3.62
## 2 Namibia         668     1        -14.7    -2.74
## 3 Qatar        102630   146         -9.87   -2.66
## 4 Nepal         16649    35        -13.6    -2.38
## 5 Rwanda         1252     3        -15.3    -2.29
```

2) Find countries with largest positive errors

```
## # A tibble: 5 x 5
##   Country        Confirmed Death ln_deaths_pc reg1_res
##   <chr>              <dbl> <dbl>        <dbl>    <dbl>
## 1 Yemen               1380   364       -11.3      2.39
## 2 Belgium            62357  9781        -7.07     1.99
## 3 United Kingdom    289678 44735        -7.31     1.97
## 4 Italy             242639 34938        -7.45     1.90
## 5 France            208015 30007        -7.71     1.89
```

Testing hypothesis

The most common test is already tested where H0 is $\beta = 0$, HA: β is not zero and the t statistic can be find in the regression summary of our first model and it is 18.62. β is statistically different than 0 in 99.9 significance leve, even higher. We can reject the null.