Covid Analysis

Jia-li

11/29/2020

Introduction

This analysis aims to find the pattern of association between registered number of confirmed and death covid-19 cases on 23rd, Oct.2020.

Measured in numbers, variables include confirmed cases (as outcome variable) and death cases(as explanatory variable) of all countries, on 23rd, Oct.2020, with confirmed and death cases of all countries from day one till now, as population.

With data from CSSE and WDI R package, I selected columns of "country", "confirmed", "death" and "population", while deleting all else. These variables were also scaled for checking log-transformation. Also, I deleted countries with missing values of population, confirmed cases or death, which may raise potential data quality issue, as some countires are missing.

Pattern of Association

Main Feature of Histograms:

- For confirmed cases, most of the distribution is between 4 and 100,000 cases, with few observations above 7,500,000. For Death cases, most of the distribution falls within (0, 2000), with few observations above 5,000.
- Both distributions are skewed, with a long right tail.
- For confirmed cases, there is an extreme value above 800,000; for death cases, there is an extreme value above 220,000. We should not drop them.

Transformation of Variables

Variables are transformed by taking natural logarithms, and there are four models as listed below:

- level-level: death = alpha + beta * confirmed (Figure 1, in Appendix)
- log-level: ln_death = alpha + beta * confirmed (Figure 2, in Appendix)
- level-log: death = alpha + beta * ln_confirmed (Figure 3, in Appendix)
- log-log: ln_death = alpha + beta * ln_confirmed (Figure 4, in Appendix)

My final decisioin is the log-log model:

- Substantive reasoning:

  - Log transfomation of cases may remain valid in spite of pandemic waves and other seasonal fluctuations.
  - Also, choosing relative terms means being free from arbitrary units of measurement.

- Statistical reasoning:
  - Log-log model graph captures gives better approximation: the scatter plot suggests a good linear pattern.
  - Even though level-level model has a higher R-square, compared with level-log, its graph shows highly non-linear pattern. Compared with Log-level, log-log model has a relatively higher R-square.

Presentation of model choice

- My final decision is Simple Regression: ln_death = -4.319 + 1.029 * ln_confirmed
- When ln(confirmed cases) is zero, average ln(death cases) is -4.319, on average.
- For observations having one percent higher confirmed cases, death cases is 1.029% higher, on average.

Analysis of Hypothesis Testing

I am interested in $H_0 : \beta = 0$, $H_A : \beta \neq 0$ or not in the model, with 5% significance level.

```
hp_test
```

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    -4.319    0.31301  -13.80 1.934e-29  -4.9366   -3.701 168
## ln_confirmed    1.029    0.02926   35.16 2.250e-79   0.9712    1.087 168
##
## Multiple R-squared:  0.8877 ,    Adjusted R-squared:  0.887
## F-statistic:  1236 on 1 and 168 DF,  p-value: < 2.2e-16
```

- The estimated t-statistics is 35.16, with 95% confidence interval: [-4.94,-3.70] and p-value: 2.250267e-79
- The confidence interval does not contain zero, So true value of coefficient is unlikely to be zero.
- Also, p value is way below 5%. Thus we reject the H_0, which means the death cases is not uncorrelated with confirmed cases.

Analysis of the residuals

```
# Find countries with largest negative errors
neg5 <- df %>% top_n( -5 , reg1_res ) %>%
  select( country , ln_death,reg1_lny_pred, reg1_res)
print(neg5)
```

```
## # A tibble: 5 x 4
##   country   ln_death reg1_lny_pred reg1_res
##   <chr>        <dbl>         <dbl>    <dbl>
## 1 Burundi       0             2.18    -2.18
## 2 Iceland       2.40          4.29    -1.89
## 3 Qatar         5.43          7.80    -2.37
## 4 Singapore     3.33          6.97    -3.63
## 5 Sri Lanka     2.64          4.81    -2.17
```

```
# Find countries with largest positive errors
pos5 <- df %>% top_n( 5 , reg1_res ) %>%
  select( country , ln_death,reg1_lny_pred,reg1_res)
print(pos5)
```

```
## # A tibble: 5 x 4
##   country ln_death reg1_lny_pred reg1_res
##   <chr>      <dbl>         <dbl>    <dbl>
## 1 Chad       4.56          3.15      1.41
## 2 Ecuador    9.44          8.00      1.44
## 3 Fiji       0.693        -0.721     1.41
## 4 Mexico    11.4           9.77      1.62
## 5 Yemen      6.40          3.53      2.86
```

- Country with the largest negative errors is Singapore, with predicted logarithm death cases of 6.97, but the real value is only 3.33
- Country with the largest negative errors is Yemen, with predicted logarithm death cases of 3.53, but the real value is 6.40

Executive summary

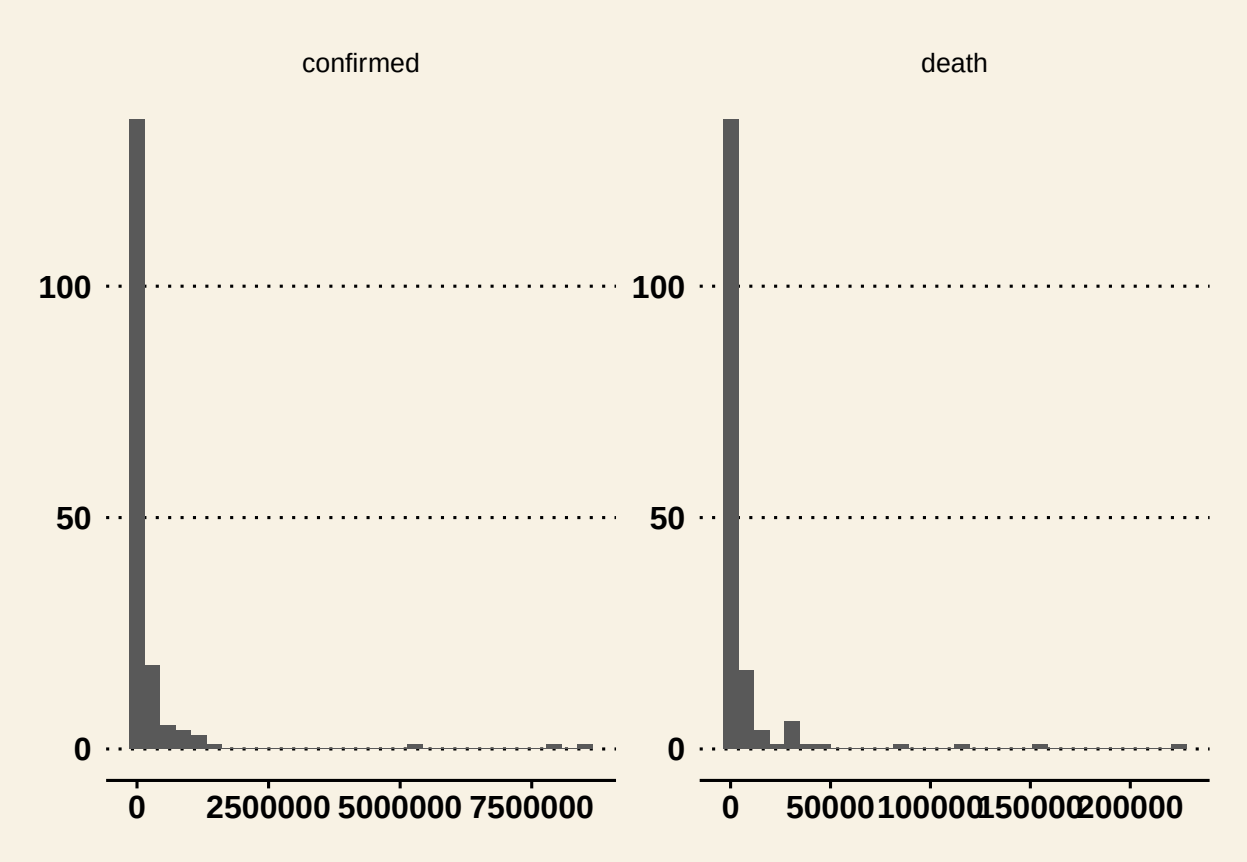
- Throught investigation, the report analyzed the association between global registered number of confirmed and death covid-19 cases on 23rd, Oct.2020. There is a positive correlation between registered number of confirmed and death covid-19 cases.
- The final model is ln(death)~ln(confirmed), which reveals death cases is 1.03 percent higher, on average, for observations having 1 percent higher confirmed cases.
- The analysis can be strengthened by invention and promotion of vaccines. Also, it might be weakened by some countries hiding true case numebers, or part of death cases caused by other pneumonia disease that are similar to Covid.

Appendix

Pattern of Association

```
# Quick check on all HISTOGRAMS
df %>%
  select(confirmed, death) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key,scales = "free") +
  geom_histogram()+
  theme_wsj() +
  scale_fill_wsj()
```

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




```
summary( df %>% select(confirmed, death) )
```

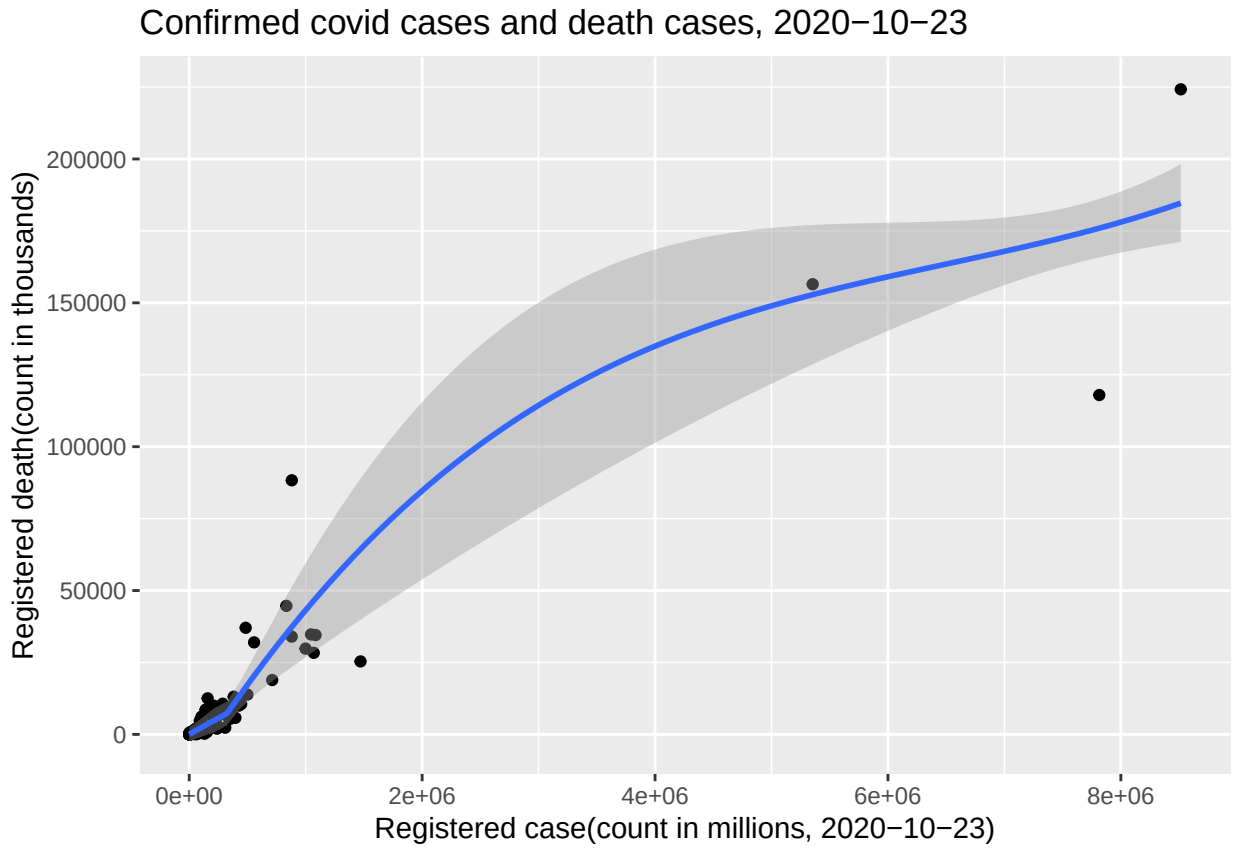
```
##    confirmed           death
##  Min.   :     33   Min.   :     1
##  1st Qu.:   5182   1st Qu.:    82
##  Median :  26743   Median :   419
```

```
##  Mean   : 248265   Mean   :  6728
##  3rd Qu.: 110508   3rd Qu.:  1952
##  Max.   :8514677   Max.   :224214
```

1) death - confirmed: level-level model without scaling

```
ggplot( df , aes(y = death, x = confirmed)) +
  geom_point() +
  geom_smooth(method="loess")+
  scale_x_continuous(
    breaks = pretty_breaks()) +
  scale_y_continuous(breaks = pretty_breaks()) +
  labs(y = "Registered death(count in thousands)",
       x = "Registered case(count in millions, 2020-10-23)",
       title = "Confirmed covid cases and death cases, 2020-10-23")
```

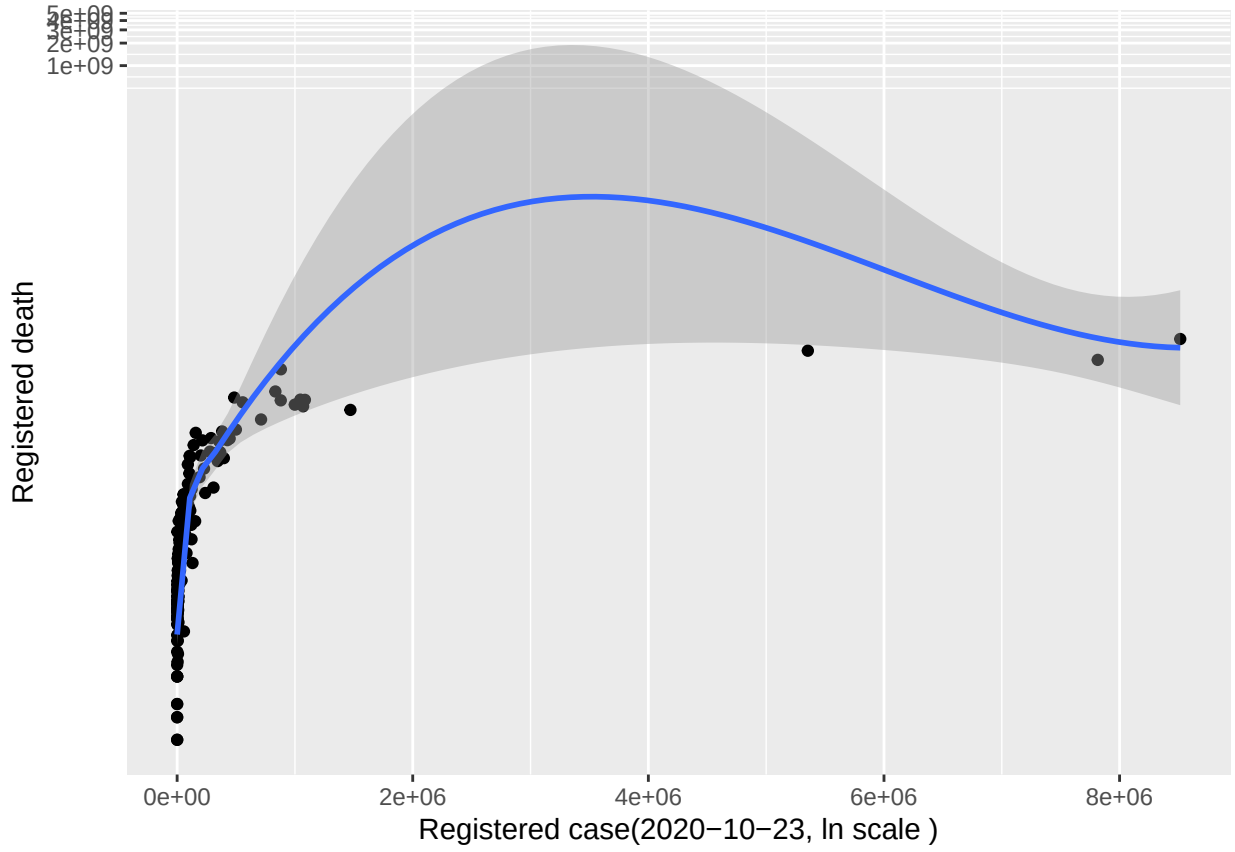
```
## `geom_smooth()` using formula 'y ~ x'
```



2) ln_death - confirmed: log-level model

```
ggplot( df , aes(y = death, x = confirmed ))  +
  geom_point() +
  geom_smooth(method="loess")+
  labs(y = "Registered death",x = "Registered case(2020-10-23, ln scale )") +
  scale_y_continuous( trans = log_trans(), breaks = pretty_breaks())
```

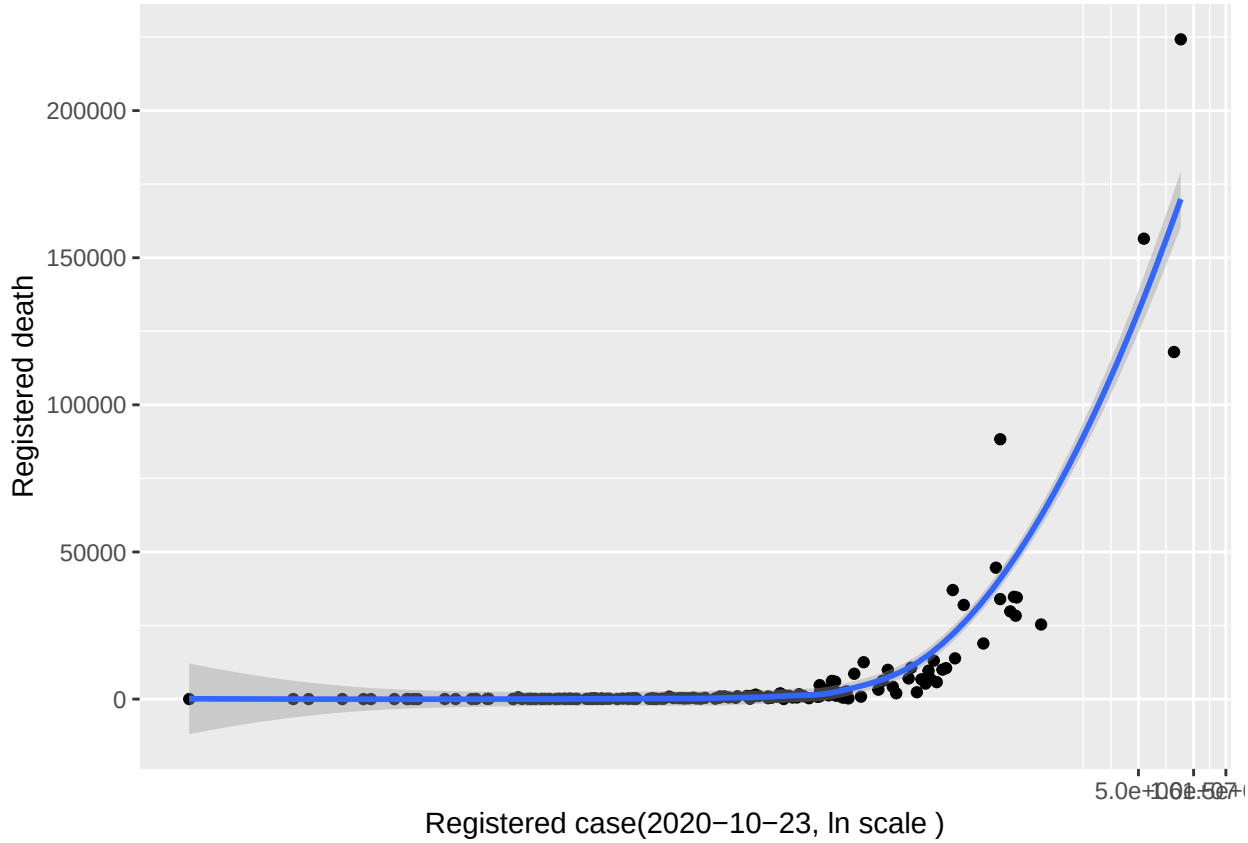
```
## `geom_smooth()` using formula 'y ~ x'
```



3) death - ln_confirmed: level-log model

```
ggplot( df , aes(y = death, x = confirmed)) +
  geom_point() +
  geom_smooth(method="loess")+
  labs(y = "Registered death",x = "Registered case(2020-10-23, ln scale )") +
  scale_x_continuous( trans = log_trans(), waiver(),
                      breaks = pretty_breaks()) +
  scale_y_continuous(breaks = pretty_breaks())
```

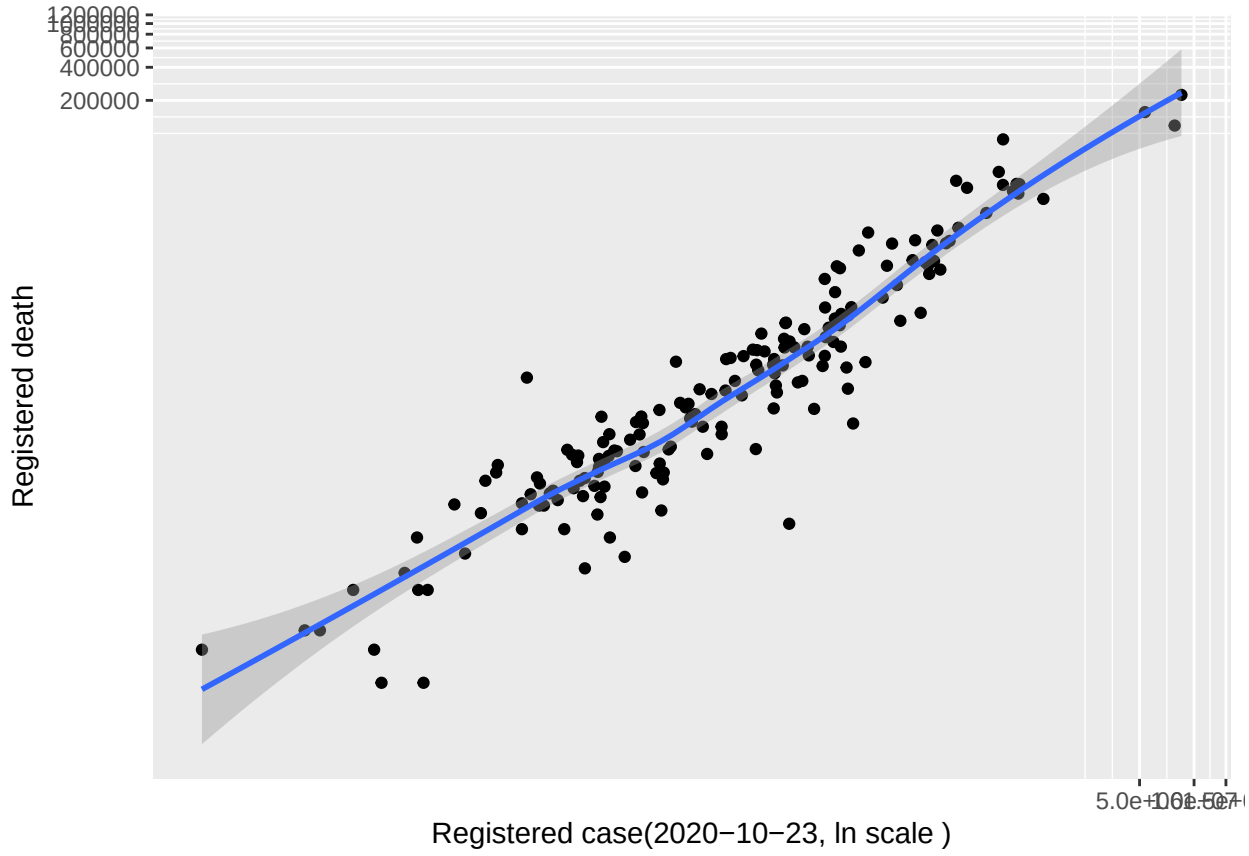
```
## `geom_smooth()` using formula 'y ~ x'
```



4) ln_death - ln_confirmed: log-log model

```
ggplot( df , aes(y = death, x = confirmed ))  +
  geom_point() +
  geom_smooth(method="loess")+
  labs(y = "Registered death",x = "Registered case(2020-10-23, ln scale )") +
  scale_x_continuous( trans = log_trans(),breaks= pretty_breaks())+
  scale_y_continuous( trans = log_trans(),breaks = pretty_breaks())
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Model Estimation

Here we have four models:

i: Simple regression

ln_death = alpha + beta * ln_confirmed

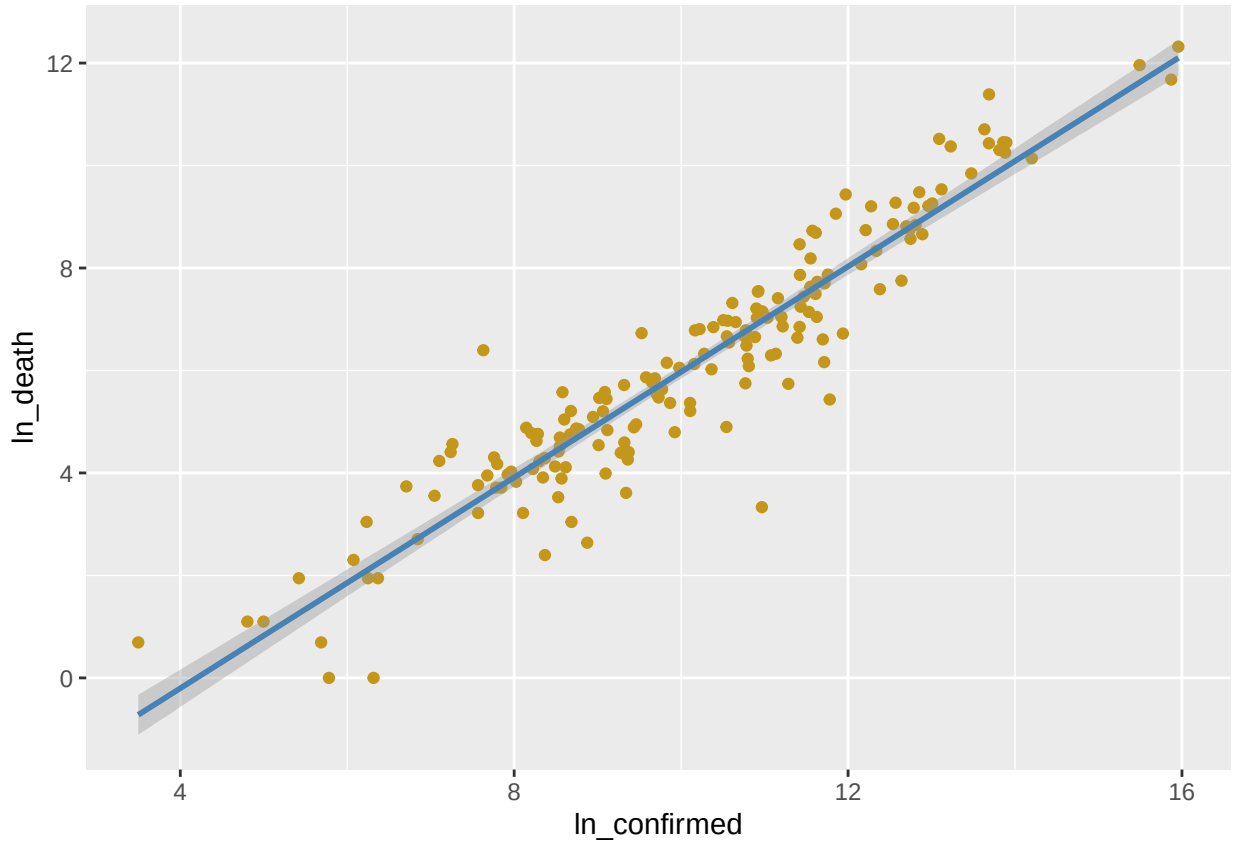
```
reg1 <- lm_robust(ln_death ~ ln_confirmed, data = df , se_type = "HC2" )
# Summary statistics
summary( reg1 )
```

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    -4.319    0.31301  -13.80 1.934e-29  -4.9366   -3.701 168
## ln_confirmed    1.029    0.02926   35.16 2.250e-79   0.9712    1.087 168
##
```

```
## Multiple R-squared:  0.8877 ,    Adjusted R-squared:  0.887
## F-statistic:  1236 on 1 and 168 DF,  p-value: < 2.2e-16
```

```
# Visual inspection:
ggplot( data = df, aes( x = ln_confirmed, y = ln_death) ) +
  geom_point( color="#C4961A") +
  geom_smooth( method = lm , color = "steelblue")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



ii: Quadratic regression

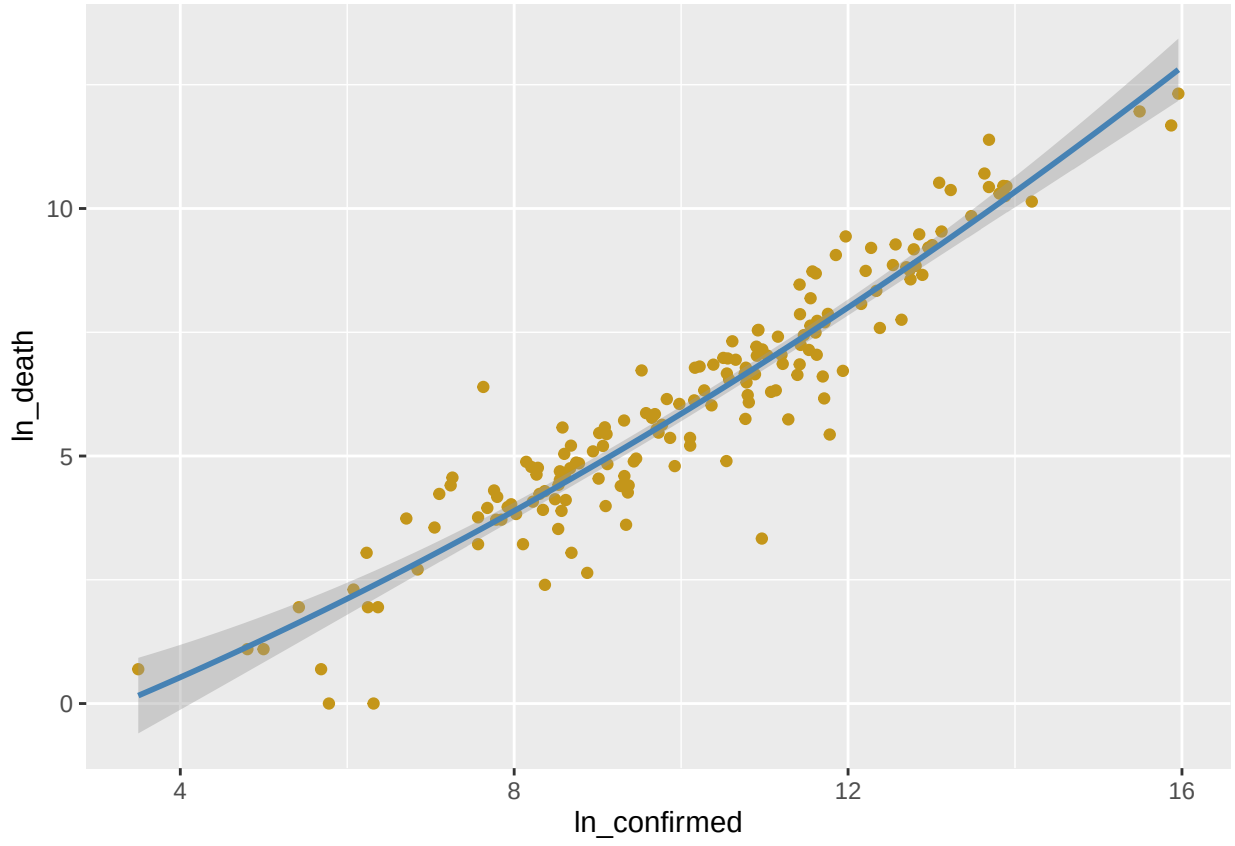
ln_death = alpha + beta_1 * ln_confirmed + beta_2 * ln_confirmed^2

```
reg2 <- lm_robust( ln_death ~ ln_confirmed + ln_confirmed_sq , data = df )
# Summary statistics
summary( reg2 )
```

```
##
## Call:
## lm_robust(formula = ln_death ~ ln_confirmed + ln_confirmed_sq,
##     data = df)
##
```

```
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)     -2.08400   0.893825  -2.332 0.020918 -3.84865 -0.31935 167
## ln_confirmed     0.55996   0.172295   3.250 0.001396  0.21980  0.90012 167
## ln_confirmed_sq  0.02338   0.008178   2.858 0.004801  0.00723  0.03952 167
##
## Multiple R-squared:  0.8921 ,    Adjusted R-squared:  0.8908
## F-statistic: 719.7 on 2 and 167 DF,  p-value: < 2.2e-16
```

```
# Visual inspection:
ggplot( data = df, aes( x = ln_confirmed, y = ln_death ) ) +
  geom_point( color="#C4961A") +
  geom_smooth( formula = y ~ poly(x,2) , method = lm , color = "steelblue")
```



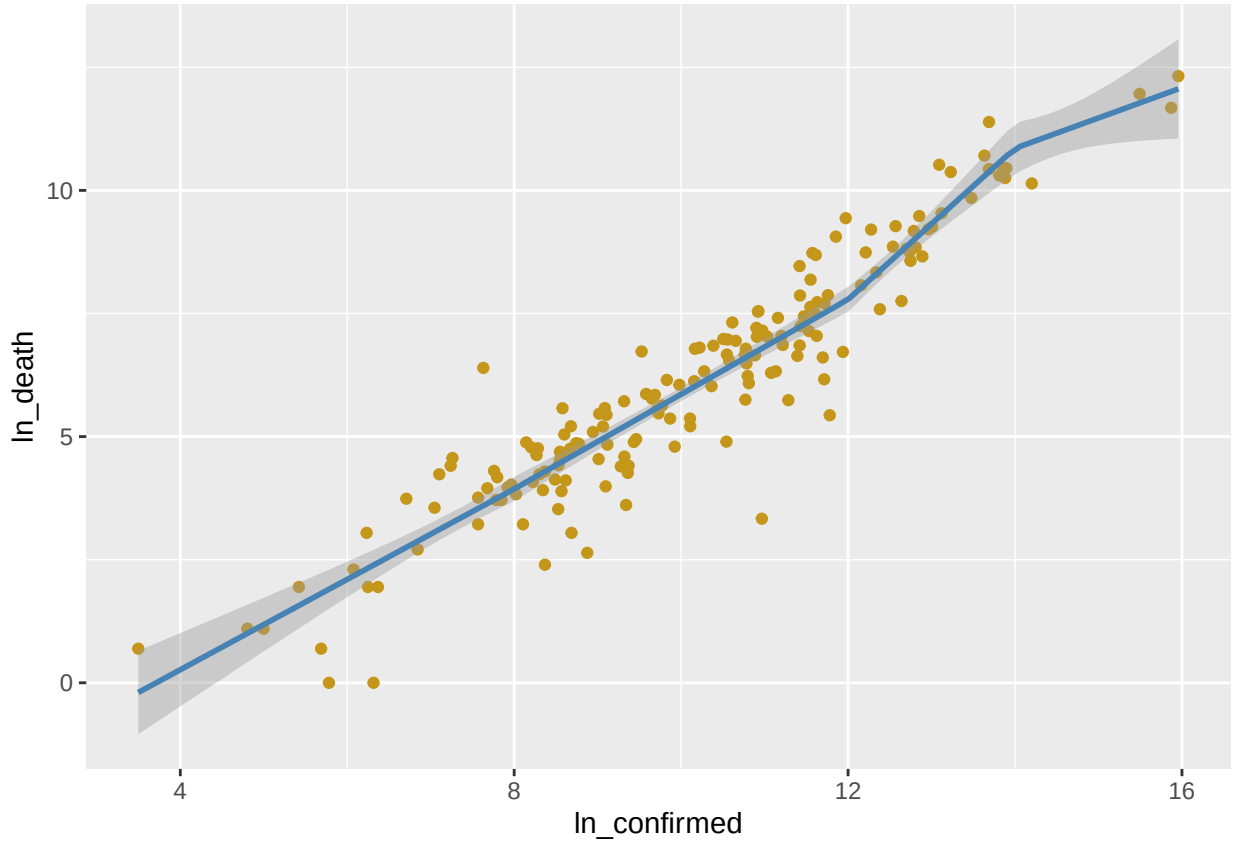
iii: Piecewise linear spline regression

ln_death = alpha + beta_1 * ln_confirmed * 1(ln_confirmed <= 8) + beta_2 * ln_confirmed * 1(8 < ln_confirmed <= 12) + beta_3 * ln_confirmed * 1(12 < ln_confirmed <= 16)

```
reg3 <- lm_robust(ln_death ~ lspline( ln_confirmed , c(8,12,14)), data = df )
# Summary statistics
summary(reg3)
```

```
##
## Call:
## lm_robust(formula = ln_death ~ lspline(ln_confirmed, c(8, 12,
##     14)), data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                          -3.4058    0.91774  -3.711 2.818e-04
## lspline(ln_confirmed, c(8, 12, 14))1   0.9178    0.11997   7.651 1.579e-12
## lspline(ln_confirmed, c(8, 12, 14))2   0.9618    0.05711  16.842 1.126e-37
## lspline(ln_confirmed, c(8, 12, 14))3   1.5356    0.13668  11.235 4.123e-22
## lspline(ln_confirmed, c(8, 12, 14))4   0.6159    0.14956   4.118 6.021e-05
##                                     CI Lower CI Upper  DF
## (Intercept)                          -5.2179  -1.5938 165
## lspline(ln_confirmed, c(8, 12, 14))1   0.6809   1.1547 165
## lspline(ln_confirmed, c(8, 12, 14))2   0.8491   1.0746 165
## lspline(ln_confirmed, c(8, 12, 14))3   1.2657   1.8055 165
## lspline(ln_confirmed, c(8, 12, 14))4   0.3206   0.9112 165
##
## Multiple R-squared:  0.8953 ,    Adjusted R-squared:  0.8927
## F-statistic: 549.1 on 4 and 165 DF,  p-value: < 2.2e-16
```

```
# Visual inspection:
ggplot( data = df, aes( x = ln_confirmed, y = ln_death ) ) +
  geom_point(color = "#C4961A") +
  geom_smooth( formula = y ~ lspline(x, c(8,12,14)) , method = lm , color = "steelblue" )
```



iv: Weighted linear regression, using population as weights.

ln_death = alpha + beta * ln_confirmed, weights: population

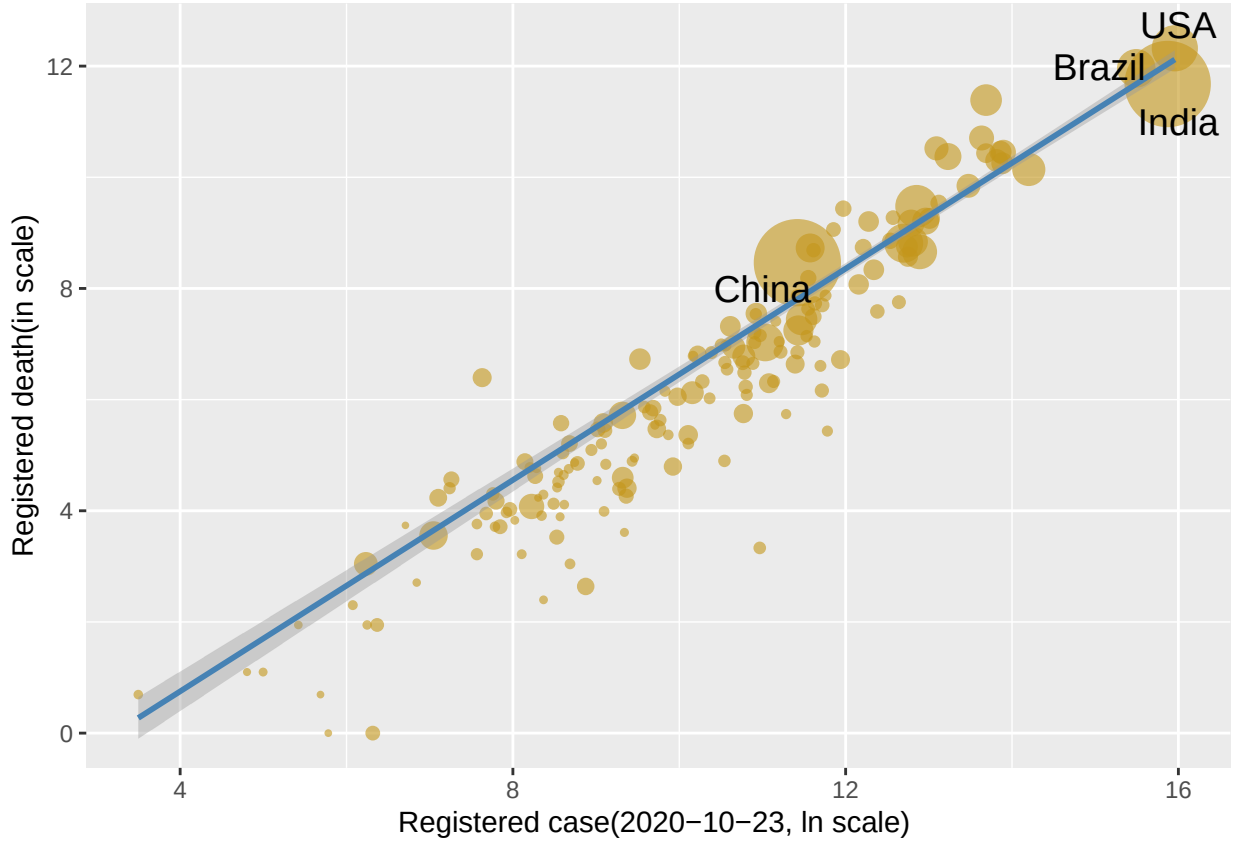
```
reg4 <- lm_robust(ln_death ~ ln_confirmed, data = df , weights = population)
# Summary statistics
summary( reg4 )
```

```
##
## 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)   -3.0510    0.77349  -3.944 1.173e-04  -4.5780   -1.524 168
## ln_confirmed   0.9506    0.06153  15.449 4.477e-34   0.8291    1.072 168
##
## Multiple R-squared:  0.9294 ,    Adjusted R-squared:  0.929
## F-statistic: 238.7 on 1 and 168 DF,  p-value: < 2.2e-16
```

```
# Visual inspection:
ggplot(data = df, aes(x = ln_confirmed, y = ln_death)) +
```

```
  geom_point(data = df, aes(size=population),  color = "#C4961A", shape = 16, alpha = 0.6,  show.legend=
  geom_smooth(aes(weight = population), method = "lm", color= "steelblue")+
  scale_size(range = c(1, 15)) +
  coord_cartesian(ylim = c(0, 12.5)) +
  labs(x = "Registered case(2020-10-23, ln scale) ", y = "Registered death(in scale)") +
  annotate("text", x = 16, y = 11, label = "India", size=5)+
  annotate("text", x = 11, y = 8, label = "China", size=5)+
  annotate("text", x = 16,  y = 12.75, label = "USA", size=5) +
  annotate("text", x = 15.05,  y = 12, label = "Brazil", size=5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Summary of All models

Model Summaries

With model comparision table and plot visualization, we can conclude:

- For linear regression:
  - For observations having one percent higher confirmed cases, death cases is 1.029% higher, on average.
  - r-square = 0.89 means 89% of the variation in ln(death cases) is captured by the regression, and 11% is left for residual variation.

	Linear	Quadratic	P.L.S	WOLS
(Intercept)	-4.32***	-2.08*	-3.41***	-3.05***
	(0.31)	(0.89)	(0.92)	(0.77)
ln_confirmed	1.03***	0.56**		0.95***
	(0.03)	(0.17)		(0.06)
ln_confirmed_sq		0.02**		
		(0.01)		
lspline(ln_confirmed, cutoff)1			0.92***	
			(0.12)	
lspline(ln_confirmed, cutoff)2			0.96***	
			(0.06)	
lspline(ln_confirmed, cutoff)3			1.54***	
			(0.14)	
lspline(ln_confirmed, cutoff)4			0.62***	
			(0.15)	
R^2	0.89	0.89	0.90	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	4260.74

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

Modelling death cases and confirmed cases of countries(2020-10-23)

Figure 1: model comparision

- For quadratic regressioin:
  - 0.02 is positive; so the relationship is convex
  - r-square = 0.89 means 89% of the variation in ln(death cases) is captured by the regression, and 11% is left for residual variation.
- For Piecewise linear spline regression:
  - Among obervations with ln(confirmed_case) values less than 8, ln(death_cases) is 0.92 units higher, on average, for observations with one unit higher ln(confirmed_case) value.
  - Among obervations with ln(confirmed_case) values between 8 and 12, ln(death_cases) is 0.96 units higher, on average, for observations with one unit higher ln(confirmed_case) value.
  - Among obervations with ln(confirmed_case) values between 12 and 14, ln(death_cases) is 1.54 units higher, on average, for observations with one unit higher ln(confirmed_case) value.
  - Among obervations with ln(confirmed_case) values above 14, ln(death_cases) is 0.62 units higher, on average, for observations with one unit higher ln(confirmed_case) value.
  - For Weighted linear regression: USA, Brazil and India have most servere covid pandemic situation.

Final Choice

Based on model comparison, my chosen model is reg1(ln_death ~ ln_confirmed):

- Substantive:
  - The other three make little change to what is basically a linear association: they are over-complicated, to some degree.
- Statistical:
  - Conpared with other three, simple regression(log-log) model is easy to interpret.
  - Log-log model has comparatively high R-square and captures variation well.