COVID-19 Analysis

Julianna Szabo

Importing data

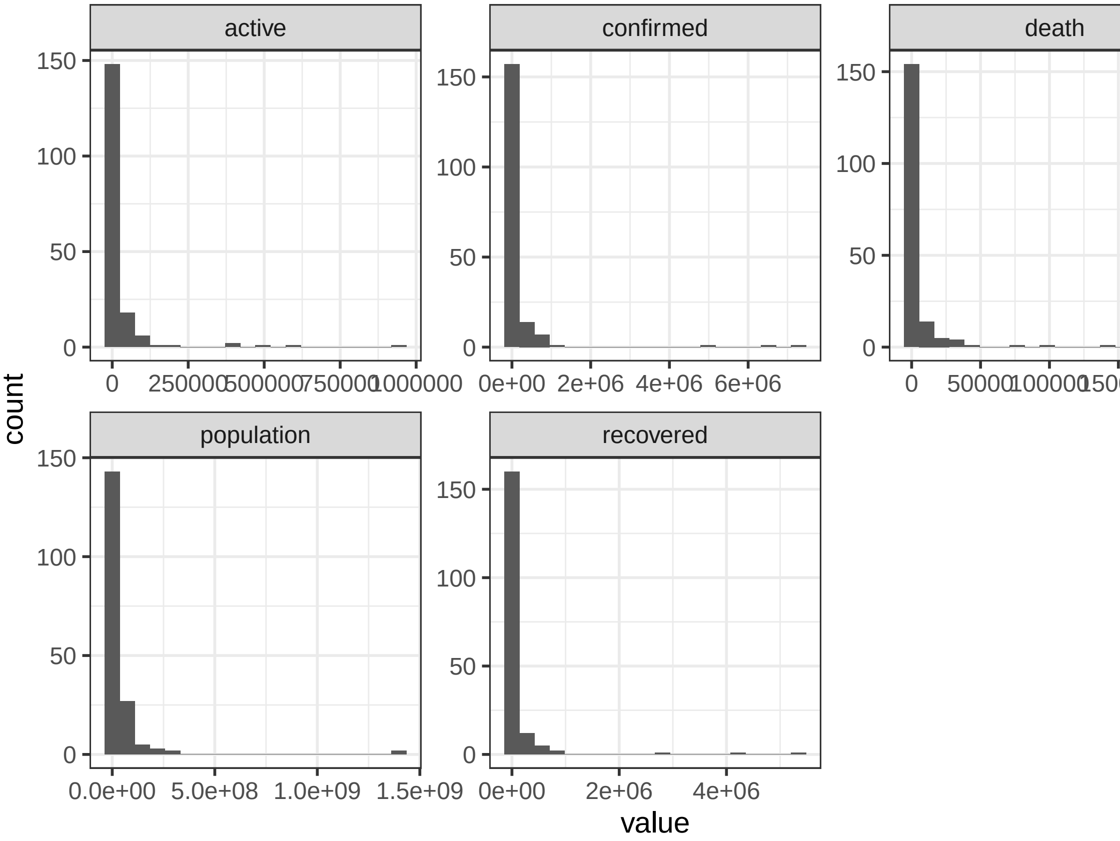
My question for this analysis is: What is the correlation between registered cases and deaths per capita?

My y variable will be Deaths per Capita and my x will be Registered Cases per Capita

The original data comes in absolute numbers instead of per capita so that transformation may cause some inaccuracies since population numbers are from last year. My population is all the cases and all the deaths by COVID throughout the world. My sample is therefore very relevant since it has that data, although some countries may not be as accurate when recording these, overall it is a relatively good representation of the population.

First let's see what the data looks like

```
## Rows: 182
## Columns: 6
## $ country    <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angola"...
## $ confirmed  <dbl> 39285, 13806, 51690, 2050, 5114, 101, 765002, 50850, 271...
## $ death      <dbl> 1458, 388, 1741, 53, 185, 3, 20288, 963, 890, 802, 593, ...
## $ recovered  <dbl> 32842, 8077, 36282, 1432, 2082, 92, 603140, 44219, 24788...
## $ active     <dbl> 4985, 5341, 13667, 565, 2847, 6, 141574, 5668, 1431, 840...
## $ population <dbl> 38041754, 2854191, 43053054, 77142, 31825295, 97118, 449...
```



Let's take a look at the data

It looks like both confirmed cases and deaths are skewed with a right tail. This may be lessed in the per capita numbers calculated later. Further, the analysis may benefit from a log tranformation later. There seem to be some extreme values in both confirmed and deaths

```
##    country            confirmed          death             recovered
##  Length:182         Min.   :     19   Min.   :     0.0   Min.   :      0
##  Class :character   1st Qu.:   3141   1st Qu.:    55.0   1st Qu.:   1787
##  Mode  :character   Median :  14266   Median :   268.5   Median :   8088
##                     Mean   : 188374   Mean   :  5620.9   Mean   : 131059
##                     3rd Qu.:  79551   3rd Qu.:  1549.5   3rd Qu.:  56500
##                     Max.   :7280834   Max.   :207977.0   Max.   :5343837
##
##      active           population
##  Min.   :     0.0   Min.   :3.386e+04
##  1st Qu.:   410.5   1st Qu.:2.535e+06
##  Median :  2170.0   Median :9.758e+06
##  Mean   : 28867.6   Mean   :4.176e+07
##  3rd Qu.: 13172.5   3rd Qu.:3.040e+07
##  Max.   :943069.0   Max.   :1.398e+09
##  NA's   :3
```

While with the summary one can see that the range is huge, I have decided to keep all the observations, since most of the extreme values will most likely be resolved once we do a per Capita transformation

Create new variables

```
df <- df %>% mutate( deaths_ppc = death/population,
                     cases_ppc = confirmed/population)
```

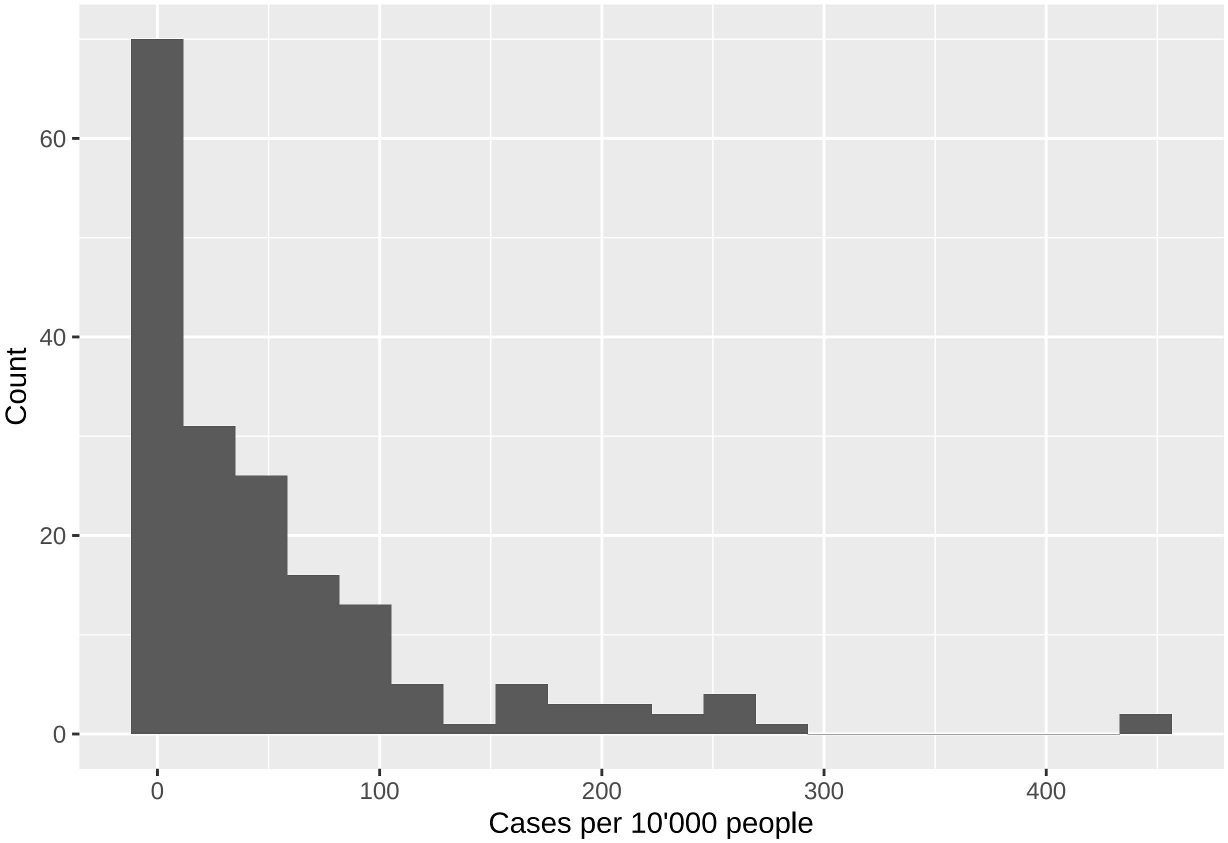
Scaling

Looking at the graphs from a scaling perspective, I have decided to transform the PPC numbers into per 10000 people instead of per one person that way they are easier to interpret because the numbers on the scale are easier to understand and compare.

```
df <- df %>% mutate( deaths_ppc = (death/population)*10000,
                     cases_ppc = (confirmed/population)*10000)
```

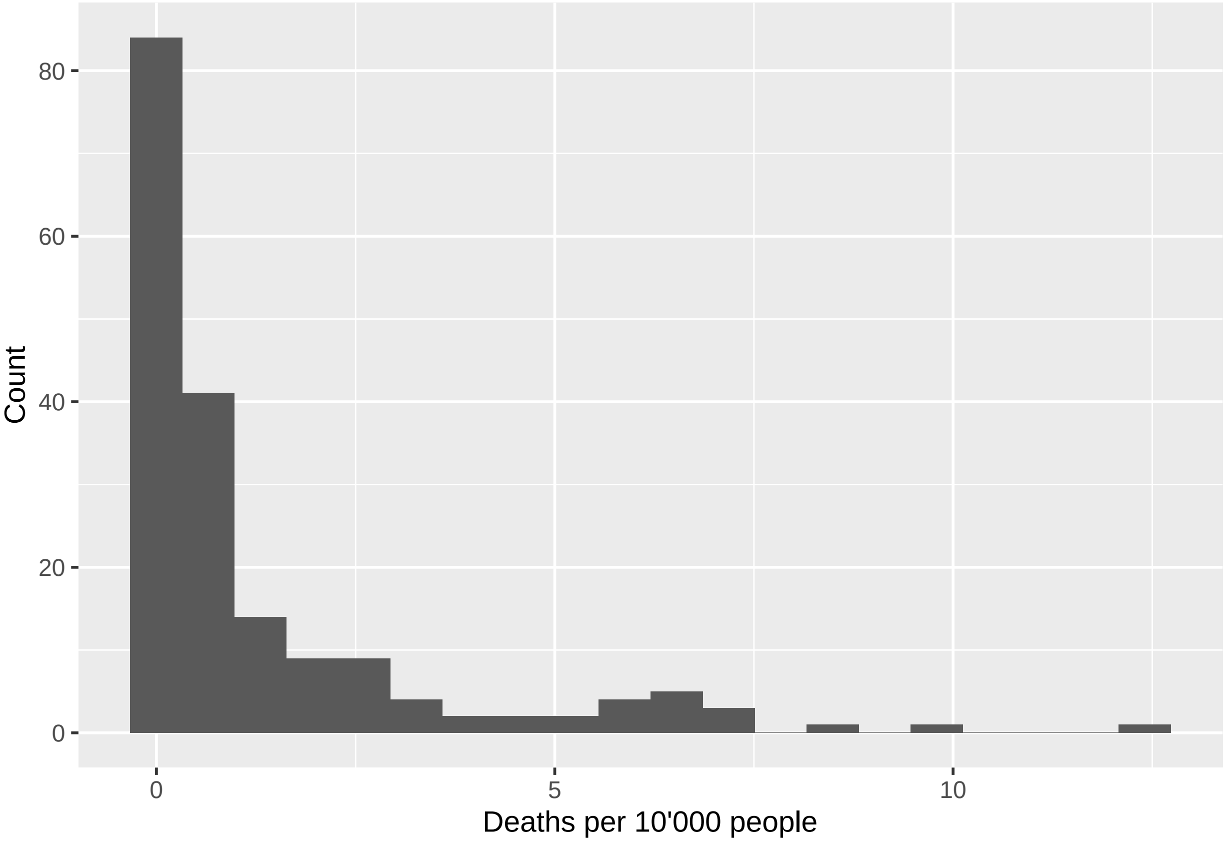
Distribution of variables

Distribution of x



```
## # A tibble: 1 x 5
##    mean median    min   max    sd
##   <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1  54.4   26.2 0.0321  445.  75.7
```

Distribution of y



```
## # A tibble: 1 x 5
##    mean median   min   max    sd
##   <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1  1.31  0.403     0  12.4  2.12
```

Overall both x and y have many observations around zero and have a long right tail. While x has higher number overall the distribution is very similar in shape.

Modeling

##ln transformation

```
df <- df %>% mutate( ln_cases_ppc = log( cases_ppc ),
                     ln_deaths_ppc= log( deaths_ppc) )
```

The log - log transformation seems to be the best fit. Substantively it is easy to interpret and works with percentages which is great for comparison. It is also more pleasant to look at since the extreme values have been fixed by the log transformation. Statistically it seems to be the most collective and the CI seems the least spread out around most of it.

To be able to use this in the future I need to remove some values that give infinity when doing the transformation.

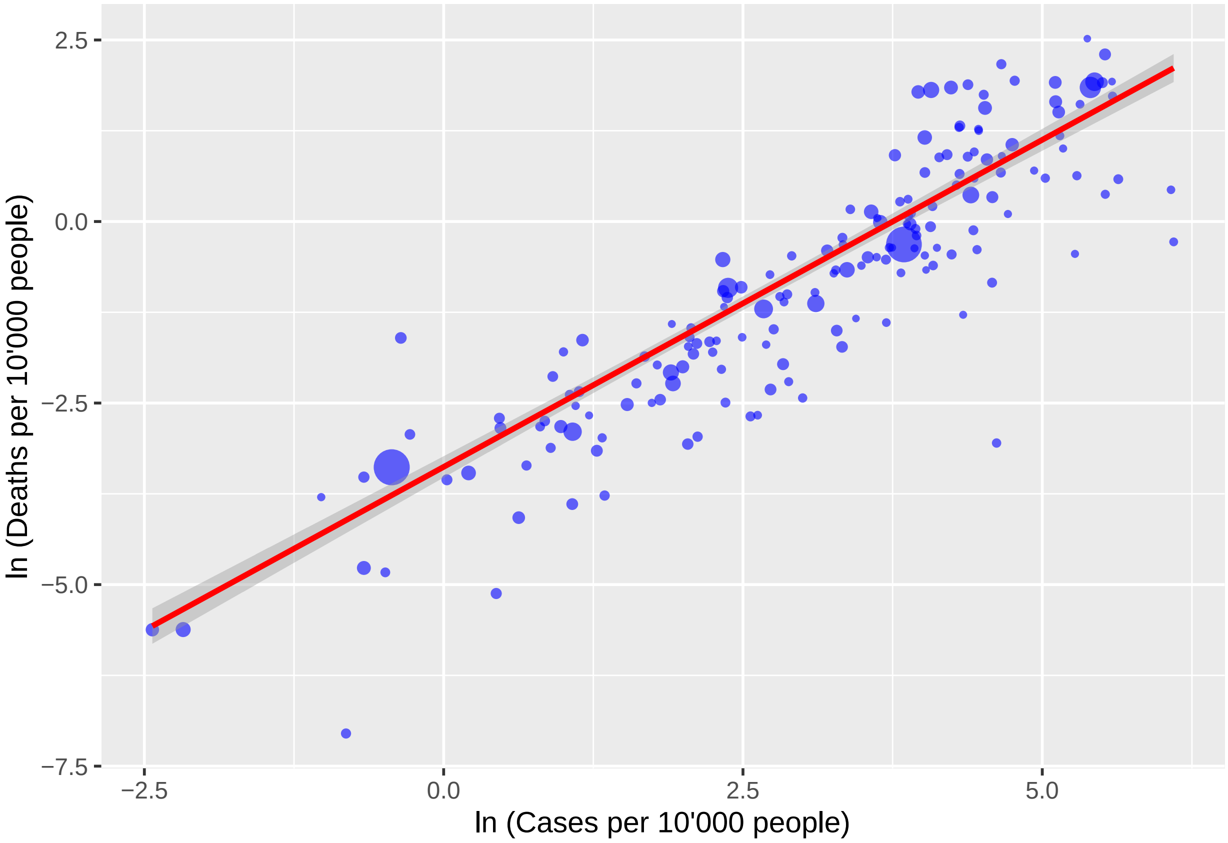
```r
df <- df[!is.infinite(df$ln_deaths_ppc),]
```

Pick Weighted OLS MORE TEST TO BE ADDED

```r
reg4 <- lm_robust(ln_deaths_ppc ~ ln_cases_ppc, data = df , weights = population)
summary( reg4 )
```

```
##
## Call:
## lm_robust(formula = ln_deaths_ppc ~ ln_cases_ppc, data = df,
##     weights = population)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -3.3789    0.29048  -11.63 2.381e-23  -3.9524   -2.805 169
## ln_cases_ppc   0.9008    0.08129   11.08 8.523e-22   0.7403    1.061 169
##
## Multiple R-squared:  0.8981 ,    Adjusted R-squared:  0.8975
## F-statistic: 122.8 on 1 and 169 DF,  p-value: < 2.2e-16
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Hypothesis testing

```
## Linear hypothesis test
##
## Hypothesis:
## ln_cases_ppc = 0
##
## Model 1: restricted model
## Model 2: ln_deaths_ppc ~ ln_cases_ppc
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    170
## 2    169  1 122.78  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual analysis

First I will create my variables that I will need for the analysis

```
df$reg4_y_pred <- reg4$fitted.values
df$reg4_res <- df$ln_deaths_ppc - df$reg4_y_pred
```

Countries who lost the most people to COVID

```
## # A tibble: 5 x 4
##   country   ln_deaths_ppc reg4_y_pred reg4_res
##   <chr>             <dbl>       <dbl>    <dbl>
## 1 Burundi          -7.05       -4.11     -2.94
## 2 Maldives         -0.446       1.37     -1.82
## 3 Qatar            -0.280       2.11     -2.39
## 4 Singapore        -3.05        0.781    -3.83
## 5 Sri Lanka        -5.12       -2.98     -2.14
```

Countries who saved the most people in COVID

```
## # A tibble: 5 x 4
##   country        ln_deaths_ppc reg4_y_pred reg4_res
##   <chr>                  <dbl>       <dbl>    <dbl>
## 1 Belgium                 2.17       0.817     1.35
## 2 Italy                   1.78       0.191     1.59
## 3 Mexico                  1.81       0.289     1.52
## 4 United Kingdom          1.84       0.438     1.41
## 5 Yemen                  -1.60      -3.70      2.10
```