718-Assignment2

Group2
01/03/2020

1. Understanding the Data

The purpose of the report is to examine the relationship between the treatments and death rate of HIV infected people having TB. This analysis is centered around the datasets *TB_burden_countries* and *TB_outcomes*.

TB_burden_countries dataset records the data from year 2000-2018 for all the WHO countries under six regions. The dataset describes the incidence and mortality for various TB patients, the number of HIV positive and patients having MDR/RR TB (i.e. drug resistant TB). From this table, the focus is on **e_inc_num** and **e_mort_tbhiv_num** variables.

TB_outcomes dataset records the resutls of various treatment outcomes (if the patients were cured or not), the focus is on the **tbhiv_succ** variable.

Firstly, we created 2 dataframes by importing csv files, and tidying the datasets by filtering, cleaning and removing the NA (Blank) values (using *is.na(Variable_Name)* function).

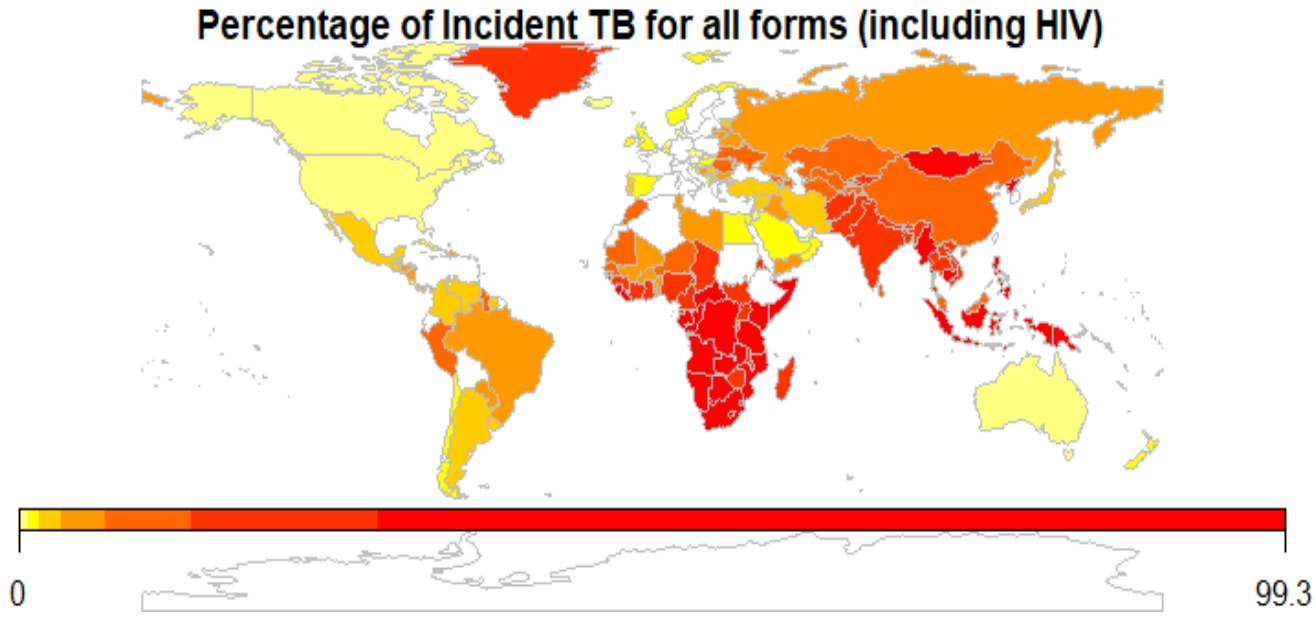
Structure of Dataset:

By joining and aggregating along the countries and year we made a merged dataframe for our analysis. Following is the structure of the new dataset.

```
## 'data.frame':    935 obs. of  5 variables:
##  $ country         : Factor w/ 216 levels "Afghanistan",..: 1 1 2 2 2 2 4 4 ...
##  $ year            : int  2016 2017 2012 2013 2014 2015 2016 2017 2016 2017 ...
##  $ g_whoregion.x   : Factor w/ 6 levels "AFR","AMR","EMR",..: 3 3 4 4 4 4 6 6 ...
##  $ e_mort_tbhiv_num: int  110 92 1 1 1 1 1 1 0 0 ...
##  $ tbhiv_succ      : int  0 2 6 1 2 3 4 3 0 0 ...

## [1] 935 120
```

The following World Heat Mapshows the percentage of incident cases of Tuberculosis.



2. Summarizing the Variable Selection

a. Identification of variables

From the merged dataset, we have chosen two variables, *e_mort_tbhiv_num* and *tbhiv_succ*. Our assumption is that more successfull the treatments are, lesser will be the mortality rate.

#Measures of central tendency

```
##    Measure | tbhiv_succ | e_mort_tbhiv_num
## 1:    Mean |   2066.141 |         2042.003
## 2:  Median |         17 |               13
## 3:    Mode |          0 |                0
```

#Measure of Dispersion:

```
##                       Measure |    tbhiv_succ | e_mort_tbhiv_num
## 1:                   Variance | 124300291.573 |     70296490.575
## 2:         Standard Deviation |     11149.004 |           8384.3
## 3: Inter-Quartile Range (IQR) |           268 |              320
## 4:                      Range |      0 147605 |          0 89000
```

b. Assumptions for Normality Test and log Transformation

e_mort_tbhiv_num - ***normality check***

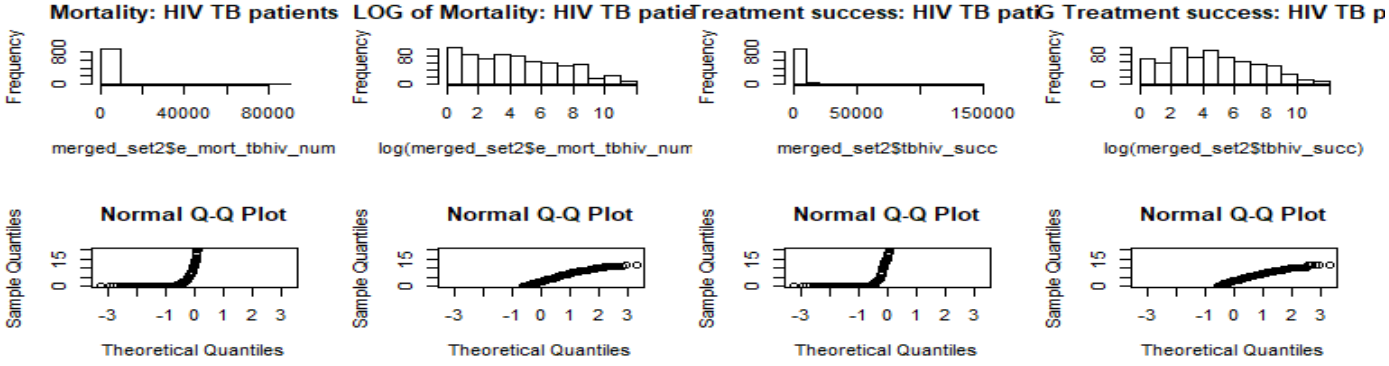
```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
## 1.300000e+01 2.042003e+03 2.741959e+02 5.381115e+02 7.029649e+07 8.384300e+03
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
## 4.105919e+00 6.407067e+00 4.005480e+01 4.699415e+01 1.470515e+02 2.556335e-01
##   normtest.p
## 1.982412e-51
```

The skew.2SE(4.005480e+01) and kurt.2SE (1.470515e+02) values are >1, indicating skewness. Further confirming the results visually using histogram & Q-Q plot, we see that the data is right skewed.

tbhiv_succ - ***normality check***

```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
## 1.700000e+01 2.066141e+03 3.646114e+02 7.155525e+02 1.243003e+08 1.114900e+04
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
## 5.396051e+00 1.029775e+01 6.437803e+01 1.182106e+02 3.698983e+02 1.675072e-01
##   normtest.p
## 2.936082e-53
```

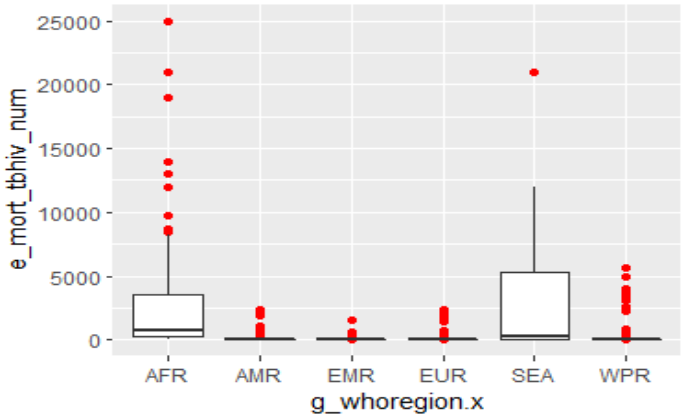
The skew.2SE(6.437803e+01) & kurt.2SE (3.698983e+02) values are >1, indicating skewness. Further confirming the results visually using histogram and Q-Q plot, we see that the data is right skewed.

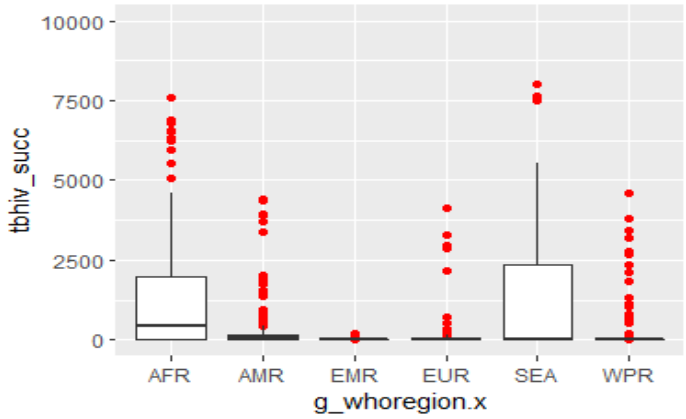


Shapiro test

```
## Shapiro-Wilk normality test
## data:  merged_set2$e_mort_tbhiv_num
## W = 0.25563, p-value < 2.2e-16
```

Since the p-value is too low we can conclude that the data is not normal. Upon further analysis of the skewness, there were many outliers identified as shown by the below box plots.





We did not find it appropriate to remove these outliers as the sample population of each region varies distinctly. Some regions have high mortality and HIV-TB incidence than others. This is the region for skewness and is necessary for our analysis.

c. Appropriate Correlation Test

Correlation is way of measuring the extent to which **e_mort_tbhiv_num** & **tbhiv_succ** are related, and the pattern of responses across variables. We have found the correlation on aggregated data by years.

Spearman Correlation test Since our data is non-parametric we cannot perform Pearson's correlation test and therefore testing our assumption using Spearman test.

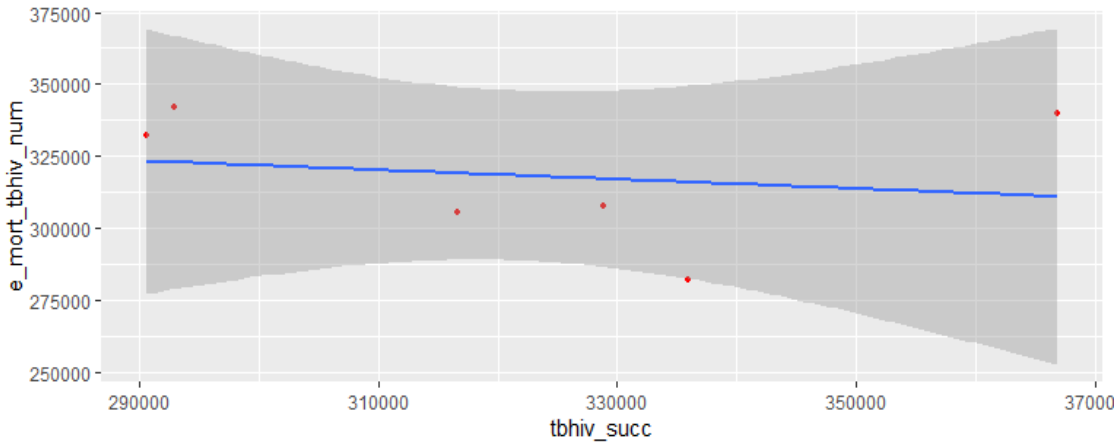
```
## [1] -0.2571429
```

From Spearman's test the Rho is -0.2571 which indicates negative correlation.

Kendall Correlation test

```
## [1] -0.2
```

From Kendall's test p value is very small and true tau is not equal to zero. Thus, test statitics prove that x(***e_mort_tbhiv_num***) & y(***tbhiv_succ***) are slightly negatively correlated which confirms our ***hypothesis***.



3. Analysis

To corroborate the negative correlation between the number of mortality in HIV patients with Tuberculosis, and the Number of Successful treatments in HIV patients with Tuberculosis, we determine the based on the 6 main sub-regions as follows:

4. Conclusion and Future Scope

1. Using Spearman's Correlation test we get co-relation of -0.2571 i.e Estimated number of deaths from TB in people who are HiV positive (aggregated e_mort_tbhiv_num) is negatively co-related with Treatment success for HIV positive TB cases (aggregated tbhiv_succ), which confirms ours asumption.
2. For region AFR (Africa), the mortality rate decreases with increasing treatment success.
3. For region AMR (America), the mortality rate is almost constant over time.
4. For EMR (East Mediterranean) region, mortality increases intially from 2011 to 2013 and decreases steeply over the years while treatment success slightly increased.
5. For EUR (European) regions, the mortality has been increasing due to increase in HIV and also, the treatment success rate has increased drastically. This, is the only region which doesn't have negative correlation as shown in graphs.
6. For WPR (West Pacific), it follows our assumption completely, Mortality rate has decreased drastically with high increase in Treatment success rate.
7. For SEA (South East Asian), mortality rate has decreased sharply also the treatment success rate has been decreasing gradually.
8. AS per our assumption, mortality rate decreases with increase in the treatment success rates, aggregated over the years as evident from our overall plot. This is because, as the technology has improved, TB is more likely to be cured than older times.